Essential Coding Theory

Venkatesan Guruswami Atri Rudra[1] Madhu Sudan

March 15, 2019

[1]Department of Computer Science and Engineering, University at Buffalo, SUNY. Work supported by NSF CAREER grant CCF-0844796.

Foreword

This book is based on lecture notes from coding theory courses taught by Venkatesan Guruswami at University at Washington and CMU; by Atri Rudra at University at Buffalo, SUNY and by Madhu Sudan at Harvard and MIT.

This version is dated **March 15, 2019**. For the latest version, please go to

http://www.cse.buffalo.edu/faculty/atri/courses/coding-theory/book/

The material in this book is supported in part by the National Science Foundation under CAREER grant CCF-0844796. Any opinions, findings and conclusions or recomendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

Contents

List of Figures

List of Tables

List of Algorithms

Part I

The Basics

Chapter 1

The Fundamental Question

1.1 Overview

Communication is a fundamental need of our modern lives. In fact, communication is something that humans have been doing for a long time. For simplicity, let us restrict ourselves to English. It is quite remarkable that different people speaking English can be understood pretty well: even if e.g. the speaker has an accent. This is because English has some built-in redundancy, which allows for “errors" to be tolerated. This came to fore for one of the authors when his two and a half year old son, Akash, started to speak his own version of English, which we will dub “Akash English." As an example,

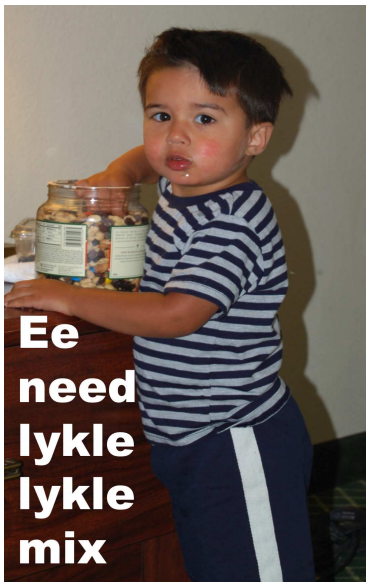


Figure 1.1: Decoding for Akash English, one gets “I need little little (trail)mix."

With some practice Akash’s parents were able to “decode" what Akash really meant. In fact,

Akash could communicate even if he did not say an entire word properly and gobbled up part(s) of word(s).

The above example shows that having redundancy in a language allows for communication even in the presence of (small amounts of) differences and errors. Of course in our modern digital world, all kinds of entities communicate (and most of the entities do not communicate in English or any natural language for that matter). Errors are also present in the digital world, so these digital communications also use redundancy.

Error-correcting codes (henceforth, just codes) are clever ways of representing data so that one can recover the original information even if parts of it are corrupted. The basic idea is to judiciously introduce redundancy so that the original information can be recovered even when parts of the (redundant) data have been corrupted.

For example, when packets are transmitted over the Internet, some of the packets get corrupted or dropped. Packet drops are resolved by the TCP layer by a combination of sequence numbers and ACKs. To deal with data corruption, multiple layers of the TCP/IP stack use a form of error correction called CRC Checksum [79]. From a theoretical point of view, the checksum is a terrible code (for that matter so is English). However, on the Internet, the current dominant mode of operation is to detect errors and if errors have occurred, then ask for retransmission. This is the reason why the use of checksum has been hugely successful in the Internet. However, there are other communication applications where re-transmission is not an option. Codes are used when transmitting data over the telephone line or via cell phones. They are also used in deep space communication and in satellite broadcast (for example, TV signals are transmitted via satellite). Indeed, asking the Mars Rover to re-send an image just because it got corrupted during transmission is not an option–this is the reason that for such applications, the codes used have always been very sophisticated.

Codes also have applications in areas not directly related to communication. In particular, in the applications above, we want to communicate over space. Codes can also be used to communicate over time. For example, codes are used heavily in data storage. CDs and DVDs work fine even in presence of scratches precisely because they use codes. Codes are used in Redundant Array of Inexpensive Disks (RAID) [15] and error correcting memory [14]. Sometimes, in the Blue Screen of Death displayed by Microsoft Windows family of operating systems, you might see a line saying something along the lines of “parity check failed"–this happens when the code used in the error-correcting memory cannot recover from error(s). Also, certain consumers of memory, e.g. banks, do not want to suffer from even one bit flipping (this e.g. could mean someone’s bank balance either got halved or doubled–neither of which are welcome[1]). Codes are also deployed in other applications such as paper bar codes; for example, the bar code used by UPS called MaxiCode [13]. Unlike the Internet example, in all of these applications, there is no scope for “re-transmission."

In this book, we will mainly think of codes in the communication scenario. In this framework, there is a sender who wants to send (say) k message symbols over a noisy channel. The sender first *encodes* the k message symbols into n symbols (called a *codeword*) and then sends

[1]This is a bit tongue-in-cheek: in real life banks have more mechanisms to prevent one bit flip from wreaking havoc.

it over the *channel*. The receiver gets a *received word* consisting of n symbols. The receiver then tries to *decode* and recover the original k message symbols. Thus, encoding is the process of adding redundancy and decoding is the process of removing errors.

Unless mentioned otherwise, in this book we will make the following assumption:

The sender and the receiver only communicate via the channel.[a] In other words, other than some setup information about the code, the sender and the receiver do not have any other information exchange (other than of course what was transmitted over the channel). In particular, no message is more likely to be transmitted over another.

[a]The scenario where the sender and receiver have a "side-channel" is an interesting topic that has been studied but is outside the scope of this book.

The fundamental question that will occupy our attention for almost the entire book is the tradeoff between the amount of redundancy used and the number of errors that can be corrected by a code. In particular, we would like to understand:

Question 1.1.1. *How much redundancy do we need to correct a given amount of errors? (We would like to correct as many errors as possible with as little redundancy as possible.)*

Intuitively, maximizing error correction and minimizing redundancy are contradictory goals: a code with higher redundancy should be able to tolerate more number of errors. By the end of this chapter, we will see a formalization of this question.

Once we determine the optimal tradeoff, we will be interested in achieving this optimal tradeoff with codes that come equipped with *efficient* encoding and decoding. (A DVD player that tells its consumer that it will recover from a scratch on a DVD by tomorrow is not exactly going to be a best-seller.) In this book, we will primarily define efficient algorithms to be ones that run in polynomial time.[2]

1.2 Some definitions and codes

To formalize Question 1.1.1, we begin with the definition of a code.

Definition 1.2.1 (Code)**.** A code of *block length* n over an *alphabet* Σ is a subset of Σ^n. Typically, we will use q to denote $|\Sigma|$.[3]

Remark 1.2.1. We note that the ambient space Σ^n can be viewed as a set of sequences, vectors or functions. In other words, we can think of a vector $(v_1, \ldots, v_n) \in \Sigma^n$ as just the sequence $v_1, \ldots, v_n$ (in order) or a vector tuple $(v_1, \ldots, v_n)$ or as the function $f : [n] \to \Sigma$ such that $f(i) = v_i$.

[2]We are not claiming that this is the correct notion of efficiency in practice. However, we believe that it is a good definition as the "first cut"– quadratic or cubic time algorithms are definitely more desirable than exponential time algorithms: see Section C.4 for more on this.

[3]Note that q need not be a constant and can depend on n: we'll see codes in this book where this is true.

Sequences assume least structure on Σ and hence are most generic. Vectors work well when Σ has some structure (and in particular is what is known as a *field*, which we will see next chapter). Functional representation will be convenient when the set of coordinates has structure (e.g., [n] may come from a finite field of size n). For now, however, the exact representation does not matter and the reader can work with representation as sequences.

We will also frequently use the following alternate way of looking at a code. Given a code $C \subseteq \Sigma^n$, with $|C| = M$, we will think of C as a mapping of the following form:

$$C : [M] \to \Sigma^n.$$

In the above, we have used to notation $[M]$ for any integer $M \geq 1$ to denote the set $\{1, 2, \ldots, M\}$.

We will also need the notion of *dimension* of a code.

Definition 1.2.2 (Dimension of a code)**.** Given a code $C \subseteq \Sigma^n$, its *dimension* is given by

$$k \stackrel{\text{def}}{=} \log_q |C|.$$

Let us begin by looking at two specific codes. Both codes are defined over $\Sigma = \{0, 1\}$ (also known as *binary codes*). In both cases $|C| = 2^4$ and we will think of each of the 16 messages as a 4 bit vector.

We first look at the so-called *parity code*, which we will denote by $C_\oplus$. Given a message $(x_1, x_2, x_3, x_4) \in \{0, 1\}^4$, its corresponding codeword is given by

$$C_\oplus(x_1,\ x_2,\ x_3,\ x_4) = (x_1,\ x_2,\ x_3,\ x_4,\ x_1 \oplus x_2 \oplus x_3 \oplus x_4),$$

where the $\oplus$ denotes the EXOR (also known as the XOR or Exclusive-OR) operator. In other words, the parity code appends the parity of the message bits (or takes the remainder of the sum of the message bits when divided by 2) at the end of the message. Note that such a code uses the minimum amount of non-zero redundancy.

The second code we will look at is the so-called *repetition code*. This is a very natural code (and perhaps the first code one might think of). The idea is to repeat every message bit a fixed number of times. For example, we repeat each of the 4 message bits 3 times and we use $C_{3,rep}$ to denote this code.

Let us now try to look at the tradeoff between the amount of redundancy and the number of errors each of these codes can correct. Even before we begin to answer the question, we need to define how we are going to measure the amount of redundancy. One natural way to define redundancy for a code with dimension k and block length n is by their difference $n - k$. By this definition, the parity code uses the least amount of redundancy. However, one "pitfall" of such a definition is that it does not distinguish between a code with $k = 100$ and $n = 102$ and another code with dimension and block length 2 and 4, respectively. Intuitively, the latter code is using more redundancy. This motivates the following notion of measuring redundancy.

Definition 1.2.3 (Rate of a code)**.** The *rate* of a code with dimension k and block length n is given by

$$R \stackrel{\text{def}}{=} \frac{k}{n}.$$

Note that the higher the rate, the lesser the amount of redundancy in the code. Also note that as $k \leq n$, $R \leq 1$.[4] Intuitively, the rate of a code is the average amount of real information in each of the n symbols transmitted over the channel. So, in some sense, rate captures the complement of redundancy. However, for historical reasons, we will deal with the rate R (instead of the more obvious $1-R$) as our notion of redundancy. Given the above definition, $C_{\oplus}$ and $C_{3,rep}$ have rates of $\frac{4}{5}$ and $\frac{1}{3}$. As was to be expected, the parity code has a higher rate than the repetition code.

We have formalized the notion of redundancy as the rate of a code as well as other parameters of a code. However, to formalize Question 1.1.1, we still need to formally define what it means to correct errors. We do so next.

1.3 Error correction

Before we formally define error correction, we will first formally define the notion of *encoding*.

Definition 1.3.1 (Encoding function)**.** Let $C \subseteq \Sigma^n$. An equivalent description of the code C is an injective mapping $E : [|C|] \to \Sigma^n$ called the encoding function.

Next we move to error correction. Intuitively, we can correct a received word if we can recover the transmitted codeword (or equivalently the corresponding message). This "reverse" process is called *decoding*.

Definition 1.3.2 (Decoding function)**.** Let $C \subseteq \Sigma^n$ be a code. A mapping $D : \Sigma^n \to [|C|]$ is called a decoding function for C.

The definition of a decoding function by itself does not give anything interesting. What we really need from a decoding function is that it recovers the transmitted message. This notion is captured next.

Definition 1.3.3 (Error Correction)**.** Let $C \subseteq \Sigma^n$ and let $t \geq 1$ be an integer. C is said to be *t-error-correcting* if there exists a decoding function D such that for every message $\mathbf{m} \in [|C|]$ and error pattern $\mathbf{e}$ with at most t errors, $D(C(\mathbf{m}) + \mathbf{e})) = \mathbf{m}$.

Figure 1.3 illustrates how the definitions we have examined so far interact.

We will also very briefly look at a weaker form of error recovery called *error detection*.

Definition 1.3.4 (Error detection)**.** Let $C \subseteq \Sigma^n$ and let $t \geq 1$ be an integer. C is said to be *t-error-detecting* if there exists a detecting procedure D such that for every message $\mathbf{m}$ and every error pattern $\mathbf{e}$ with at most t errors, D outputs a 1 if $(C(\mathbf{m}) + \mathbf{e}) \in C$ and 0 otherwise.

Note that a t-error correcting code is also a t-error detecting code (but not necessarily the other way round): see Exercise 1.1. Although error detection might seem like a weak error recovery model, it is useful in settings where the receiver can ask the sender to re-send the message. For example, error detection is used quite heavily in the Internet.

With the above definitions in place, we are now ready to look at the error correcting capabilities of the codes we looked at in the previous section.

[4] Further, in this book, we will always consider the case $k > 0$ and $n < \infty$ and hence, we can also assume that $R > 0$.

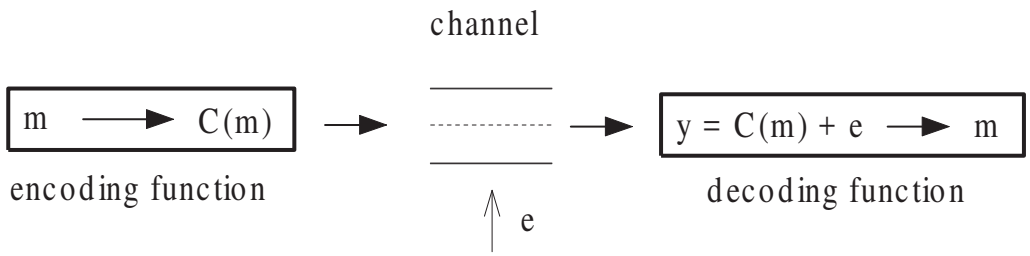


Figure 1.2: Coding process

1.3.1 Error-Correcting Capabilities of Parity and Repetition codes

In Section 1.2, we looked at examples of parity code and repetition code with the following properties:

$$C_{\oplus} : q = 2, k = 4, n = 5, R = 4/5.$$

$$C_{3,rep} : q = 2, k = 4, n = 12, R = 1/3.$$

We will start with the repetition code. To study its error-correcting capabilities, we will consider the following natural decoding function. Given a received word $\mathbf{y} \in \{0,1\}^{12}$, divide it up into four consecutive blocks (y_1, y_2, y_3, y_4) where every block consists of three bits. Then, for every block y_i $(1 \le i \le 4)$, output the majority bit as the message bit. We claim this decoding function can correct any error pattern with at most 1 error. (See Exercise 1.2.) For example, if a block of 010 is received, since there are two 0's we know the original message bit was 0. In other words, we have argued that

Proposition 1.3.1. *$C_{3,rep}$ is a 1-error correcting code.*

However, it is not too hard to see that $C_{3,rep}$ cannot correct two errors. For example, if both of the errors happen in the same block and a block in the received word is 010, then the original block in the codeword could have been either 111 or 000. Therefore in this case, no decoder can successfully recover the transmitted message.[5]

Thus, we have pin-pointed the error-correcting capabilities of the $C_{3,rep}$ code: it can correct one error, but not two or more. However, note that the argument assumed that the error positions can be located arbitrarily. In other words, we are assuming that the channel noise behaves arbitrarily (subject to a bound on the total number of errors). Obviously, we can model the noise differently. We now briefly digress to look at this issue in slightly more detail.

Digression: Channel Noise. As was mentioned above, until now we have been assuming the following noise model, which was first studied by Hamming:

[5] Recall we are assuming that the decoder has no side information about the transmitted message.

> Any error pattern can occur during transmission as long as the total number of errors is bounded. Note that this means that the location as well as the nature[6] of the errors is arbitrary.

We will frequently refer to Hamming's model as the Adversarial Noise Model. It is important to note that the atomic unit of error is a symbol from the alphabet. So for example, if the error pattern is $(1,0,1,0,0,0)$ and we consider the alphabet to be $\{0,1\}$, then the pattern has two errors. However, if our alphabet is $\{0,1\}^3$ (i.e. we think of the vector above as $((1,0,1),(0,0,0))$, with $(0,0,0)$ corresponding to the zero element in $\{0,1\}^3$), then the pattern has only one error. Thus, by increasing the alphabet size we can also change the adversarial noise model. As the book progresses, we will see how error correction over a larger alphabet is easier than error correction over a smaller alphabet.

However, the above is not the only way to model noise. For example, we could also have following error model:

> No more than 1 error can happen in any contiguous three-bit block.

First note that, for the channel model above, no more than four errors can occur when a codeword in $C_{3,rep}$ is transmitted. (Recall that in $C_{3,rep}$, each of the four bits is repeated three times.) Second, note that the decoding function that takes the majority vote of each block can successfully recover the transmitted codeword for *any* error pattern, while in the worst-case noise model it could only correct at most one error. This channel model is admittedly contrived, but it illustrates the point that the error-correcting capabilities of a code (and a decoding function) are crucially dependent on the noise model.

A popular alternate noise model is to model the channel as a stochastic process. As a concrete example, let us briefly mention the *binary symmetric channel with crossover probability* $0 \le p \le 1$, denoted by BSC_p, which was first studied by Shannon. In this model, when a (binary) codeword is transferred through the channel, every bit flips independently with probability p.

Note that the two noise models proposed by Hamming and Shannon are in some sense two extremes: Hamming's model assumes *no* knowledge about the channel (except that a bound on the total number of errors is known[7] while Shannon's noise model assumes *complete* knowledge about how noise is produced. In this book, we will consider only these two extreme noise models. In real life, the situation often is somewhere in between.

For real life applications, modeling the noise model correctly is an extremely important task, as we can tailor our codes to the noise model at hand. However, in this book we will not study this aspect of designing codes at all, and will instead mostly consider the worst-case noise model. Intuitively, if one can communicate over the worst-case noise model, then one could use the same code to communicate over nearly every other noise model with the same amount of noise.

[6] For binary codes, there is only one kind of error: a bit flip. However, for codes over a larger alphabet, say $\{0,1,2\}$, 0 being converted to a 1 and 0 being converted into a 2 are both errors, but are different kinds of errors.

[7] A bound on the total number of errors is necessary; otherwise, error correction would be impossible: see Exercise 1.3.

We now return to $C_\oplus$ and examine its error-correcting capabilities in the worst-case noise model. We claim that $C_\oplus$ cannot correct even one error. Suppose $\mathbf{y} = 10000$ is the received word. Then we know that an error has occurred, but we do not know which bit was flipped. This is because the two codewords $\mathbf{u} = 00000$ and $\mathbf{v} = 10001$ differ from the received word $\mathbf{y}$ in exactly one bit. As we are assuming that the receiver has no side information about the transmitted codeword, no decoder can know what the transmitted codeword was.

Thus, from an error-correction point of view, $C_\oplus$ is a terrible code (as it cannot correct even 1 error). However, we will now see that $C_\oplus$ can *detect* one error. Consider Algorithm 1. Note that

Algorithm 1 Error Detector for Parity Code

INPUT: Received word $\mathbf{y} = (y_1, y_2, y_3, y_4, y_5)$
OUTPUT: 1 if $\mathbf{y} \in C_\oplus$ and 0 otherwise

1: $b \leftarrow y_1 \oplus y_2 \oplus y_3 \oplus y_4 \oplus y_5$
2: RETURN $1 \oplus b$ ▷ If there is no error, then $b = 0$ and hence we need to "flip" the bit for the answer

when no error has occurred during transmission, $y_i = x_i$ for $1 \le i \le 4$ and $y_5 = x_1 \oplus x_2 \oplus x_3 \oplus x_4$, in which case $b = 0$ and we output $1 \oplus 0 = 1$ as required. If there is a single error then either $y_i = x_i \oplus 1$ (for exactly one $1 \le i \le 4$) or $y_5 = x_1 \oplus x_2 \oplus x_3 \oplus x_4 \oplus 1$. It is easy to check that in this case, $b = 1$. In fact, one can extend this argument to obtain the following result (see Exercise 1.4).

Proposition 1.3.2. *The parity code $C_\oplus$ can* detect *an odd number of errors.*

Let us now revisit the example that showed that one cannot correct one error using $C_\oplus$. Recall, we considered two codewords in $C_\oplus$, $\mathbf{u} = 00000$ and $\mathbf{v} = 10001$ (which are codewords corresponding to messages 0000 and 1000, respectively). Now consider the scenarios in which $\mathbf{u}$ and $\mathbf{v}$ are each transmitted and a single error occurs resulting in the received word $\mathbf{r} = 10000$. Thus, given the received word $\mathbf{r}$ and the fact that at most one error can occur, the decoder has no way of knowing whether the original transmitted codeword was $\mathbf{u}$ or $\mathbf{v}$. Looking back at the example, it is clear that the decoder is "confused" because the two codewords $\mathbf{u}$ and $\mathbf{v}$ do not differ in many positions. This notion is formalized in the next section.

1.4 Distance of a code

We now define a notion of distance that captures the concept that the two vectors $\mathbf{u}$ and $\mathbf{v}$ are "close-by."

Definition 1.4.1 (Hamming distance)**.** Given two vectors $\mathbf{u}, \mathbf{v} \in \Sigma^n$ the *Hamming distance* between $\mathbf{u}$ and $\mathbf{v}$, denoted by $\Delta(\mathbf{u}, \mathbf{v})$, is the number of positions in which $\mathbf{u}$ and $\mathbf{v}$ differ.

The Hamming distance is a distance in a very formal mathematical sense: see Exercise 1.5. Note that the definition of Hamming distance just depends on the *number* of differences and not the nature of the difference. For example, consider the vectors $\mathbf{u}$ and $\mathbf{v}$ from the previous

section. One can see that their Hamming distance is $\Delta(\mathbf{u},\mathbf{v}) = 2$. Now consider the vector $\mathbf{w} = 01010$. Note that even though $\mathbf{v} \neq \mathbf{w}$, we have that the Hamming distance $\Delta(\mathbf{u},\mathbf{w}) = 2$.

Armed with the notion of Hamming distance, we now define another important parameter of a code.

Definition 1.4.2 (Minimum distance)**.** Let $C \subseteq \Sigma^n$. The *minimum distance* (or just *distance*) of C is defined to be

$$d = \min_{\mathbf{c}_1 \neq \mathbf{c}_2 \in C} \Delta(\mathbf{c}_1, \mathbf{c}_2)$$

It is easy to check that the repetition code $C_{3,rep}$ has distance 3. Indeed, any two distinct messages will differ in at least one of the message bits. After encoding, the difference in one message bit will translate into a difference of three bits in the corresponding codewords. We now claim that the distance of $C_\oplus$ is 2. This is a consequence of the following observations. If two messages $\mathbf{m}_1$ and $\mathbf{m}_2$ differ in at least two places then $\Delta(C_\oplus(\mathbf{m}_1), C_\oplus(\mathbf{m}_2)) \geq 2$ (even if we just ignored the parity bits). If two messages differ in exactly one place then the parity bits in the corresponding codewords are different which implies a Hamming distance of 2 between the codewords. Thus, $C_\oplus$ has smaller distance than $C_{3,rep}$ and can correct less number of errors than $C_{3,rep}$. This suggests that a larger distance implies greater error-correcting capabilities. The next result formalizes this intuition.

Proposition 1.4.1. *Given a code C, the following are equivalent:*

1. *C has minimum distance $d \geq 2$,*
2. *If d is odd, C can correct $(d-1)/2$ errors.*
3. *C can detect $d-1$ errors.*
4. *C can correct $d-1$ erasures.*[8]

Remark 1.4.1. Property (2) above for even d is slightly different. In this case, one can correct up to $\frac{d}{2} - 1$ errors but cannot correct $\frac{d}{2}$ errors. (See Exercise 1.6.)

Before we prove Proposition 1.4.1, let us apply it to the codes $C_\oplus$ and $C_{3,rep}$ which have distances of 2 and 3 respectively. Proposition 1.4.1 implies the following facts that we have already proved:

- $C_{3,rep}$ can correct 1 error (Proposition 1.3.1).
- $C_\oplus$ can detect 1 error but cannot correct 1 error (Proposition 1.3.2).

The proof of Proposition 1.4.1 will need the following decoding function. *Maximum likelihood decoding* (MLD) is a well-studied decoding method for error correcting codes, which

[8] In the erasure noise model, the receiver *knows* where the errors have occurred. In this model, an erroneous symbol is denoted by "?", with the convention that any non-? symbol is a correct symbol.

outputs the codeword closest to the received word in Hamming distance (with ties broken arbitrarily). More formally, the MLD function denoted by $D_{MLD} : \Sigma^n \to C$ is defined as follows. For every $\mathbf{y} \in \Sigma^n$,

$$D_{MLD}(\mathbf{y}) = \arg\min_{\mathbf{c} \in C} \Delta(\mathbf{c}, \mathbf{y}).$$

Algorithm 2 is a naive implementation of the MLD.

Algorithm 2 Naive Maximum Likelihood Decoder

INPUT: Received word $\mathbf{y} \in \Sigma^n$
OUTPUT: $D_{MLD}(\mathbf{y})$

```
Pick an arbitrary c ∈ C and assign z ← c
FOR every c' ∈ C such that c ≠ c' DO
    IF Δ(c', y) < Δ(z, y) THEN
        z ← c'
RETURN z
```

Proof of Proposition 1.4.1 We will complete the proof in two steps. First, we will show that if property 1 is satisfied then so are properties 2,3 and 4. Then we show that if property 1 is not satisfied then none of properties 2,3 or 4 hold.

1. **implies** 2. Assume C has distance d. We first prove 2 (for this case assume that $d = 2t+1$). We now need to show that there exists a decoding function such that for all error patterns with at most t errors it always outputs the transmitted message. We claim that the MLD function has this property. Assume this is not so and let $\mathbf{c}_1$ be the transmitted codeword and let $\mathbf{y}$ be the received word. Note that

$$\Delta(\mathbf{y}, \mathbf{c}_1) \le t. \tag{1.1}$$

As we have assumed that MLD does not work, $D_{MLD}(\mathbf{y}) = \mathbf{c}_2 \neq \mathbf{c}_1$. Note that by the definition of MLD,

$$\Delta(\mathbf{y}, \mathbf{c}_2) \le \Delta(\mathbf{y}, \mathbf{c}_1). \tag{1.2}$$

Consider the following set of inequalities:

$$\begin{aligned}
\Delta(\mathbf{c}_1, \mathbf{c}_2) &\le \Delta(\mathbf{c}_2, \mathbf{y}) + \Delta(\mathbf{c}_1, \mathbf{y}) && (1.3)\\
&\le 2\Delta(\mathbf{c}_1, \mathbf{y}) && (1.4)\\
&\le 2t && (1.5)\\
&= d-1, && (1.6)
\end{aligned}$$

where (1.3) follows from the triangle inequality (see Exercise 1.5), (1.4) follows from (1.2) and (1.5) follows from (1.1). (1.6) implies that the distance of C is at most $d-1$, which is a contradiction.

1. **implies** 3. We now show that property 3 holds, that is, we need to describe an algorithm that can successfully detect whether errors have occurred during transmission (as long as the total number of errors is bounded by $d-1$). Consider the following error detection algorithm: check if the received word $\mathbf{y} = \mathbf{c}$ for some $\mathbf{c} \in C$ (this can be done via an exhaustive check). If no errors occurred during transmission, $\mathbf{y} = \mathbf{c}_1$, where $\mathbf{c}_1$ was the transmitted codeword and the algorithm above will accept (as it should). On the other hand if $1 \le \Delta(\mathbf{y}, \mathbf{c}_1) \le d-1$, then by the fact that the distance of C is d, $\mathbf{y} \notin C$ and hence the algorithm rejects, as required.

1. **implies** 4. Finally, we prove that property 4 holds. Let $\mathbf{y} \in (\Sigma \cup \{?\})^n$ be the received word. First we claim that there is a unique $\mathbf{c} = (c_1, \ldots, c_n) \in C$ that agrees with $\mathbf{y}$ (i.e. $y_i = c_i$ for every i such that $y_i \neq ?$). (For the sake of contradiction, assume that this is not true, i.e. there exists two distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in C$ such that both $\mathbf{c}_1$ and $\mathbf{c}_2$ agree with $\mathbf{y}$ in the unerased positions. Note that this implies that $\mathbf{c}_1$ and $\mathbf{c}_2$ agree in the positions i such that $y_i \neq ?$. Thus, $\Delta(\mathbf{c}_1, \mathbf{c}_2) \le |\{i | y_i = ?\}| \le d-1$, which contradicts the assumption that C has distance d.) Given the uniqueness of the codeword $\mathbf{c} \in C$ that agrees with $\mathbf{y}$ in the unerased position, an algorithm to find $\mathbf{c}$ is as follows: go through all the codewords in C and output the desired codeword.

$\neg$1. **implies** $\neg$2. For the other direction of the proof, assume that property 1 does not hold, that is, C has distance $d-1$. We now show that property 2 cannot hold: i.e., for every decoding function there exists a transmitted codeword $\mathbf{c}_1$ and a received word $\mathbf{y}$ (where $\Delta(\mathbf{y}, \mathbf{c}_1) \le (d-1)/2$) such that the decoding function cannot output $\mathbf{c}_1$. Let $\mathbf{c}_1 \neq \mathbf{c}_2 \in C$ be codewords such that $\Delta(\mathbf{c}_1, \mathbf{c}_2) = d-1$ (such a pair exists as C has distance $d-1$). Now consider a vector $\mathbf{y}$ such that $\Delta(\mathbf{y}, \mathbf{c}_1) = \Delta(\mathbf{y}, \mathbf{c}_2) = (d-1)/2$. Such a $\mathbf{y}$ exists as d is odd and by the choice of $\mathbf{c}_1$ and $\mathbf{c}_2$. Below is an illustration of such a $\mathbf{y}$ (matching color implies that the vectors agree on those positions):

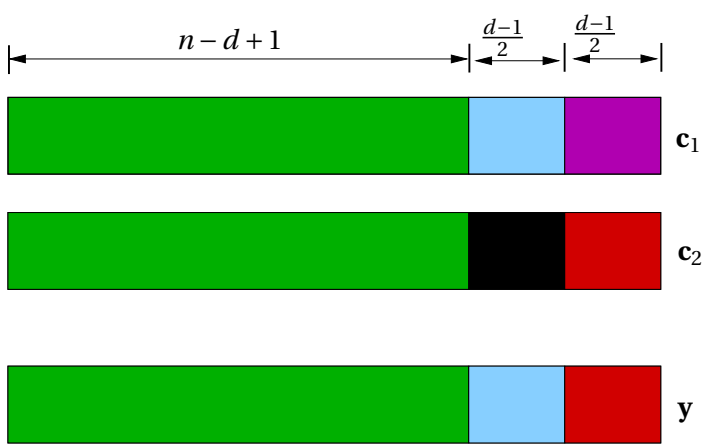


Figure 1.3: Bad example for unique decoding.

Now, since $\mathbf{y}$ could have been generated if *either* of $\mathbf{c}_1$ or $\mathbf{c}_2$ were the transmitted codeword, no decoding function can work in this case.[9]

[9] Note that this argument is just a generalization of the argument that $C_\oplus$ cannot correct 1 error.

$\neg 1$. **implies** $\neg 3$. For the remainder of the proof, assume that the transmitted word is $\mathbf{c}_1$ and there exists another codeword $\mathbf{c}_2$ such that $\Delta(\mathbf{c}_2, \mathbf{c}_1) = d - 1$. To see why property 3 is not true, let $\mathbf{y} = \mathbf{c}_2$. In this case, either the error detecting algorithm detects no error, or it declares an error when $\mathbf{c}_2$ is the transmitted codeword and no error takes place during transmission.

$\neg 1$. **implies** $\neg 4$. We finally argue that property 4 does not hold. Let $\mathbf{y}$ be the received word in which the positions that are erased are exactly those where $\mathbf{c}_1$ and $\mathbf{c}_2$ differ. Thus, given $\mathbf{y}$ both $\mathbf{c}_1$ and $\mathbf{c}_2$ could have been the transmitted codeword, and no algorithm for correcting (at most $d - 1$) erasures can work in this case. ■

Proposition 1.4.1 implies that Question 1.1.1 can be reframed as

Question 1.4.1. *What is the largest rate R that a code with distance d can have?*

We have seen that the repetition code $C_{3,rep}$ has distance 3 and rate $1/3$. A natural follow-up question (which is a special case of Question 1.4.1) is to ask

Question 1.4.2. *Can we have a code with distance* 3 *and rate* $R > \frac{1}{3}$*?*

1.5 Hamming Code

With the above question in mind, let us consider the so-called *Hamming code*, which we will denote by C_H. Given a message $(x_1, x_2, x_3, x_4) \in \{0,1\}^4$, its corresponding codeword is given by

$$C_H(x_1, x_2, x_3, x_4) = (x_1, x_2, x_3, x_4, x_2 \oplus x_3 \oplus x_4, x_1 \oplus x_3 \oplus x_4, x_1 \oplus x_2 \oplus x_4).$$

It is easy to check that this code has the following parameters:

$$C_H : q = 2, k = 4, n = 7, R = 4/7.$$

We will show shortly that C_H has a distance of 3. We would like to point out that we could have picked the three parities differently. The reason we mention the three particular parities above is due to historical reasons. We leave it as an exercise to define an alternate set of parities such that the resulting code still has a distance of 3: see Exercise 1.9.

Before we move on to determining the distance of C_H, we will need another definition.

Definition 1.5.1 (Hamming Weight)**.** Let $q \geq 2$. Given any vector $\mathbf{v} \in \{0, 1, 2, \ldots, q-1\}^n$, its Hamming weight, denoted by $wt(\mathbf{v})$ is the number of non-zero symbols in $\mathbf{v}$.

For example, if $\mathbf{v} = 01203400$, then $wt(\mathbf{v}) = 4$.

We now look at the distance of C_H.

Proposition 1.5.1. *C_H has a distance of 3.*

Proof. We will prove the claimed property by using two properties of C_H:

$$\min_{\mathbf{c}\in C_H, \mathbf{c}\neq \mathbf{0}} wt(\mathbf{c}) = 3, \tag{1.7}$$

and

$$\min_{\mathbf{c}\in C_H, \mathbf{c}\neq \mathbf{0}} wt(\mathbf{c}) = \min_{\mathbf{c}_1\neq \mathbf{c}_2\in C_H} \Delta(\mathbf{c}_1, \mathbf{c}_2) \tag{1.8}$$

The proof of (1.7) follows from a case analysis on the Hamming weight of the message bits. Let us use $\mathbf{x} = (x_1, x_2, x_3, x_4)$ to denote the message vector.

- Case 0: If $wt(\mathbf{x}) = 0$, then $C_H(\mathbf{x}) = \mathbf{0}$, which means we do not have to consider this codeword.
- Case 1: If $wt(\mathbf{x}) = 1$ then at least two parity check bits in $(x_2 \oplus x_3 \oplus x_4, x_1 \oplus x_2 \oplus x_4, x_1 \oplus x_3 \oplus x_4)$ are 1 (see Exercise 1.10). So in this case, $wt(C_H(\mathbf{x})) \geq 3$.
- Case 2: If $wt(\mathbf{x}) = 2$ then at least one parity check bit in $(x_2 \oplus x_3 \oplus x_4, x_1 \oplus x_2 \oplus x_4, x_1 \oplus x_3 \oplus x_4)$ is 1 (see Exercise 1.11). So in this case, $wt(C_H(\mathbf{x})) \geq 3$.
- Case 3: If $wt(\mathbf{x}) \geq 3$ then obviously $wt(C_H(\mathbf{x})) \geq 3$.

Thus, we can conclude that $\min_{\mathbf{c}\in C_H, \mathbf{c}\neq \mathbf{0}} wt(\mathbf{c}) \geq 3$. Further, note that $wt(C_H(1,0,0,0)) = 3$, which along with the lower bound that we just obtained proves (1.7).

We now turn to the proof of (1.8). For the rest of the proof, let $\mathbf{x} = (x_1, x_2, x_3, x_4)$ and $\mathbf{y} = (y_1, y_2, y_3, y_4)$ denote the two distinct messages. Using associativity and commutativity of the $\oplus$ operator, we obtain that

$$C_H(\mathbf{x}) + C_H(\mathbf{y}) = C_H(\mathbf{x} + \mathbf{y}),$$

where the "+" operator is just the bit-wise $\oplus$ of the operand vectors[10]. Further, it is easy to verify that for two vectors $\mathbf{u}, \mathbf{v} \in \{0,1\}^n$, $\Delta(\mathbf{u}, \mathbf{v}) = wt(\mathbf{u} + \mathbf{v})$ (see Exercise 1.12). Thus, we have

$$\begin{aligned}\min_{\mathbf{x}\neq\mathbf{y}\in\{0,1\}^4} \Delta(C_H(\mathbf{x}), C_H(\mathbf{y})) &= \min_{\mathbf{x}\neq\mathbf{y}\in\{0,1\}^4} wt(C_H(\mathbf{x}+\mathbf{y})) \\ &= \min_{\mathbf{x}\neq\mathbf{0}\in\{0,1\}^4} wt(C_H(\mathbf{x})),\end{aligned}$$

where the second equality follows from the observation that $\{\mathbf{x}+\mathbf{y} | \mathbf{x} \neq \mathbf{y} \in \{0,1\}^n\} = \{\mathbf{x} \in \{0,1\}^n | \mathbf{x} \neq \mathbf{0}\}$. Recall that $wt(C_H(\mathbf{x})) = 0$ if and only if $\mathbf{x} = \mathbf{0}$ and this completes the proof of (1.8). Combining (1.7) and (1.8), we conclude that C_H has a distance of 3. □

[10] E.g. $(0,1,1,0) + (1,1,1,0) = (1,0,0,0)$.

The second part of the proof could also be shown in the following manner. It can be verified easily that the Hamming code is the set $\{\mathbf{x} \cdot G_H | \mathbf{x} \in \{0,1\}^4\}$, where G_H is the following matrix (where we think $\mathbf{x}$ as a row vector).[11]

$$G_H = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

In fact, any binary code (of dimension k and block length n) that is generated[12] by a $k \times n$ matrix is called a *binary linear code*. (Both $C_\oplus$ and $C_{3,rep}$ are binary linear codes: see Exercise 1.13.) This implies the following simple fact.

Lemma 1.5.2. *For any binary linear code C and any two messages $\mathbf{x}$ and $\mathbf{y}$, $C(\mathbf{x}) + C(\mathbf{y}) = C(\mathbf{x}+\mathbf{y})$.*

Proof. For any binary linear code, we have a generator matrix G. The following sequence of equalities (which follow from the distributivity and associativity properties of the Boolean EXOR and AND operators) proves the lemma.

$$\begin{aligned} C(\mathbf{x}) + C(\mathbf{y}) &= \mathbf{x} \cdot G + \mathbf{y} \cdot G \\ &= (\mathbf{x} + \mathbf{y}) \cdot G \\ &= C(\mathbf{x} + \mathbf{y}) \end{aligned}$$

□

We stress that in the lemma above, $\mathbf{x}$ and $\mathbf{y}$ need *not* be distinct. Note that due to the fact that $b \oplus b = 0$ for every $b \in \{0,1\}$, $\mathbf{x}+\mathbf{x} = \mathbf{0}$, which along with the lemma above implies that $C(\mathbf{0}) = \mathbf{0}$.[13] We can infer the following result from the above lemma and the arguments used to prove (1.8) in the proof of Proposition 1.5.1.

Proposition 1.5.3. *For any binary linear code, its minimum distance is equal to minimum Hamming weight of any non-zero codeword.*

Thus, we have seen that C_H has distance $d = 3$ and rate $R = \frac{4}{7}$ while $C_{3,rep}$ has distance $d = 3$ and rate $R = \frac{1}{3}$. Thus, the Hamming code is provably better than the repetition code (in terms of the tradeoff between rate and distance) and thus, answers Question 1.4.2 in the affirmative. The next natural question is

Question 1.5.1. *Can we have a distance 3 code with a rate higher than that of C_H?*

[11] Indeed $(x_1, x_2, x_3, x_4) \cdot G_H = (x_1, x_2, x_3, x_4, x_2 \oplus x_3 \oplus x_4, x_1 \oplus x_3 \oplus x_4, x_1 \oplus x_2 \oplus x_4)$, as desired.

[12] That is, $C = \{\mathbf{x} \cdot G | \mathbf{x} \in \{0,1\}^k\}$, where addition is the $\oplus$ operation and multiplication is the AND operation.

[13] This of course should not be surprising as for any matrix G, we have $\mathbf{0} \cdot G = \mathbf{0}$.

We will address this question in the next section.

1.6 Hamming Bound

Now we switch gears to present our first tradeoff between redundancy (in the form of the dimension of a code) and its error-correction capability (in the form of its distance). In particular, we will first prove a special case of the so-called Hamming bound for a distance of 3.

We begin with another definition.

Definition 1.6.1 (Hamming Ball)**.** For any vector $\mathbf{x} \in [q]^n$,

$$B(\mathbf{x}, e) = \{\mathbf{y} \in [q]^n | \Delta(\mathbf{x}, \mathbf{y}) \leq e\}.$$

Next, we prove an upper bound on the dimension of *every* code with distance 3.

Theorem 1.6.1 (Hamming bound for $d = 3$)**.** *Every binary code with block length n, dimension k, distance $d = 3$ satisfies*

$$k \leq n - \log_2(n+1).$$

Proof. Given any two codewords, $\mathbf{c}_1 \neq \mathbf{c}_2 \in C$, the following is true (as C has distance[14] 3):

$$B(\mathbf{c}_1, 1) \cap B(\mathbf{c}_2, 1) = \emptyset. \tag{1.9}$$

See Figure 1.4 for an illustration. Note that for all $\mathbf{x} \in \{0,1\}^n$ (see Exercise 1.16),

$$|B(\mathbf{x}, 1)| = n + 1. \tag{1.10}$$

Now consider the union of all Hamming balls centered around some codeword. Obviously, their union is a subset of $\{0,1\}^n$. In other words,

$$\left| \bigcup_{\mathbf{c} \in C} B(\mathbf{c}, 1) \right| \leq 2^n. \tag{1.11}$$

As (1.9) holds for every pair of distinct codewords,

$$\begin{aligned} \left| \bigcup_{\mathbf{c} \in C} B(\mathbf{c}, 1) \right| &= \sum_{\mathbf{c} \in C} |B(\mathbf{c}, 1)| \\ &= \sum_{\mathbf{c} \in C} (n+1) && (1.12) \\ &= 2^k \cdot (n+1), && (1.13) \end{aligned}$$

[14] Assume that $\mathbf{y} \in B(\mathbf{c}_1, 1) \cap B(\mathbf{c}_2, 1)$, that is $\Delta(\mathbf{y}, \mathbf{c}_1) \leq 1$ and $\Delta(\mathbf{y}, \mathbf{c}_2) \leq 1$. Thus, by the triangle inequality $\Delta(\mathbf{c}_1, \mathbf{c}_2) \leq 2 < 3$, which is a contradiction.

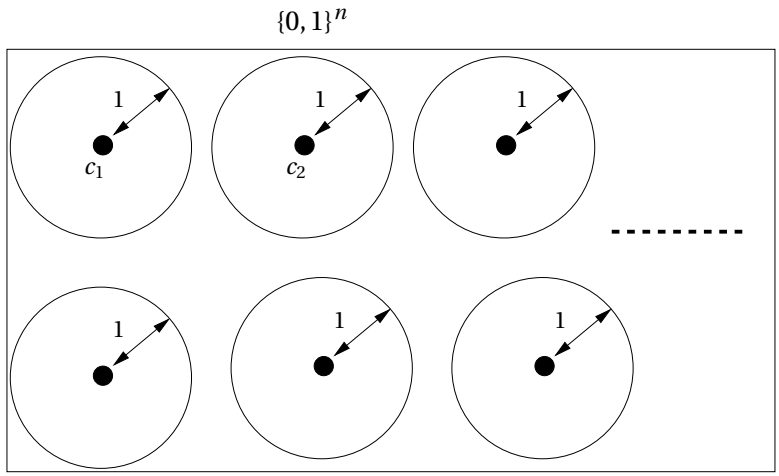


Figure 1.4: Hamming balls of radius 1 are disjoint. The figure is technically not correct: the balls above are actually balls in the Euclidean space, which is easier to visualize than the Hamming space.

where (1.12) follows from (1.10) and (1.13)) the fact that C has dimension k. Combining (1.13) and (1.11), we get

$$2^k(n+1) \leq 2^n,$$

or equivalently

$$2^k \leq \frac{2^n}{n+1}.$$

Taking $\log_2$ of both sides we get the desired bound:

$$k \leq n - \log_2(n+1).$$

□

Thus, Theorem 1.6.1 shows that for $n = 7$, C_H has the largest possible dimension for any binary code of block length 7 and distance 3 (as for $n = 7$, $n - \log_2(n+1) = 4$). In particular, it also answers Question 1.5.1 for $n = 7$ in the negative. Next, will present the general form of Hamming bound.

1.7 Generalized Hamming Bound

We start with a new notation.

Definition 1.7.1. A code $C \subseteq \Sigma^n$ with dimension k and distance d will be called a $(n,k,d)_\Sigma$ code. We will also refer it to as a $(n,k,d)_{|\Sigma|}$ code.

We now proceed to generalize Theorem 1.6.1 to any distance d.

Theorem 1.7.1 (Hamming Bound for any d)**.** *For every $(n,k,d)_q$ code*

$$k \le n - \log_q \left(\sum_{i=0}^{\left\lfloor \frac{(d-1)}{2} \right\rfloor} \binom{n}{i} (q-1)^i \right).$$

Proof. The proof is a straightforward generalization of the proof of Theorem 1.6.1. For notational convenience, let $e = \left\lfloor \frac{(d-1)}{2} \right\rfloor$. Given any two codewords, $\mathbf{c}_1 \neq \mathbf{c}_2 \in C$, the following is true (as C has distance[15] d):

$$B(\mathbf{c}_1, e) \cap B(\mathbf{c}_2, e) = \emptyset. \tag{1.14}$$

We claim that for all $\mathbf{x} \in [q]^n$,

$$|B(\mathbf{x}, e)| = \sum_{i=0}^{e} \binom{n}{i} (q-1)^i. \tag{1.15}$$

Indeed any vector in $B(\mathbf{x}, e)$ must differ from $\mathbf{x}$ in exactly $0 \le i \le e$ positions. In the summation $\binom{n}{i}$ is the number of ways of choosing the differing i positions and in each such position, a vector can differ from $\mathbf{x}$ in $q-1$ ways.

Now consider the union of all Hamming balls centered around some codeword. Obviously, their union is a subset of $[q]^n$. In other words,

$$\left| \bigcup_{\mathbf{c} \in C} B(\mathbf{c}, e) \right| \le q^n. \tag{1.16}$$

As (1.14) holds for every pair of distinct codewords,

$$\begin{aligned} \left| \bigcup_{\mathbf{c} \in C} B(\mathbf{c}, e) \right| &= \sum_{\mathbf{c} \in C} |B(\mathbf{c}, e)| \\ &= q^k \sum_{i=0}^{e} \binom{n}{i} (q-1)^i, \end{aligned} \tag{1.17}$$

where (1.17) follows from (1.15) and the fact that C has dimension k. Combining (1.17) and (1.16) and taking $\log_q$ of both sides we will get the desired bound:

$$k \le n - \log_q \left(\sum_{i=0}^{e} \binom{n}{i} (q-1)^i \right).$$

□

Note that the Hamming bound gives a partial answer to Question 1.4.1. In particular, any code of distance d can have rate R at most

$$1 - \frac{\log_q \left(\sum_{i=0}^{e} \binom{n}{i} (q-1)^i \right)}{n}.$$

Further, the Hamming bound also leads to the following definition:

[15] Assume that $\mathbf{y} \in B(\mathbf{c}_1, e) \cap B(\mathbf{c}_2, e)$, that is $\Delta(\mathbf{y}, \mathbf{c}_1) \le e$ and $\Delta(\mathbf{y}, \mathbf{c}_2) \le e$. Thus, by the triangle inequality, $\Delta(\mathbf{c}_1, \mathbf{c}_2) \le 2e \le d-1$, which is a contradiction.

Definition 1.7.2. Codes that meet Hamming bound are called *perfect codes*.

Intuitively, a perfect code leads to the following perfect "packing": if one constructs Hamming balls of radius $\left\lfloor \frac{d-1}{2} \right\rfloor$ around all the codewords, then we would cover the entire ambient space, i.e. every possible vector will lie in one of these Hamming balls.

One example of perfect code is the $(7,4,3)_2$ Hamming code that we have seen in this chapter (so is the family of general Hamming codes that we will see in the next chapter). A natural question to ask is if

Question 1.7.1. *Other than the Hamming codes, are there any other perfect (binary) codes?*

We will see the answer shortly.

1.8 Exercises

Exercise 1.1. Show that any t-error correcting code is also t-error detecting but not necessarily the other way around.

Exercise 1.2. Prove Proposition 1.3.1.

Exercise 1.3. Show that for every integer n, there is no code with block length n that can handle arbitrary number of errors.

Exercise 1.4. Prove Proposition 1.3.2.

Exercise 1.5. A distance function on Σ^n (i.e. $d : \Sigma^n \times \Sigma^n \to \mathbb{R}$) is called a *metric* if the following conditions are satisfied for every $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \Sigma^n$:

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$.
2. $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
3. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$.
4. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$. (This property is called the *triangle inequality*.)

Prove that the Hamming distance is a metric.

Exercise 1.6. Let C be a code with distance d for even d. Then argue that C can correct up to $d/2 - 1$ many errors but cannot correct $d/2$ errors. Using this or otherwise, argue that if a code C is t-error correctable then it either has a distance of $2t+1$ or $2t+2$.

Exercise 1.7. In this exercise, we will see that one can convert arbitrary codes into code with slightly different parameters:

1. Let C be an $(n,k,d)_2$ code with d odd. Then it can be converted into an $(n+1,k,d+1)_2$ code.

2. Let C be an $(n,k,d)_\Sigma$ code. Then it can be converted into an $(n-1,k,d-1)_\Sigma$ code.

Note: Other than the parameters of the code C, you should not assume anything else about the code. Also your conversion should work for every $n,k,d \geq 1$.

Exercise 1.8. In this problem we will consider a noise model that has both errors and erasures. In particular, let C be an $(n,k,d)_\Sigma$ code. As usual a codeword $\mathbf{c} \in C$ is transmitted over the channel and the received word is a vector $\mathbf{y} \in (\Sigma \cup \{?\})^n$, where as before a ? denotes an erasure. We will use s to denote the number of erasures in $\mathbf{y}$ and e to denote the number of (non-erasure) errors that occurred during transmission. To decode such a vector means to output a codeword $\mathbf{c} \in C$ such that the number of positions where $\mathbf{c}$ disagree with $\mathbf{y}$ in the $n-s$ non-erased positions is at most e. For the rest of the problem assume that

$$2e + s < d. \tag{1.18}$$

1. Argue that the output of the decoder for any C under (1.18) is unique.

2. Let C be a binary code (but not necessarily linear). Assume that there exists a decoder D that can correct from $< d/2$ many errors in $T(n)$ time. Then under (1.18) one can perform decoding in time $O(T(n))$.

Exercise 1.9. Define codes other than C_H with $k=4, n=7$ and $d=3$.
Hint: Refer to the proof of Proposition 1.5.1 to figure out the properties needed from the three parities.

Exercise 1.10. Argue that if $wt(\mathbf{x}) = 1$ then at least two parity check bits in $(x_2 \oplus x_3 \oplus x_4, x_1 \oplus x_2 \oplus x_4, x_1 \oplus x_3 \oplus x_4)$ are 1.

Exercise 1.11. Argue that if $wt(\mathbf{x}) = 2$ then at least one parity check bit in $(x_2 \oplus x_3 \oplus x_4, x_1 \oplus x_2 \oplus x_4, x_1 \oplus x_3 \oplus x_4)$ is 1.

Exercise 1.12. Prove that for any $\mathbf{u}, \mathbf{v} \in \{0,1\}^n$, $\Delta(\mathbf{u},\mathbf{v}) = wt(\mathbf{u}+\mathbf{v})$.

Exercise 1.13. Argue that $C_\oplus$ and $C_{3,rep}$ are binary linear codes.

Exercise 1.14. Let G be a generator matrix of an $(n,k,d)_2$ binary linear code. Then G has at least kd ones in it.

Exercise 1.15. Argue that in any binary linear code, either all all codewords begin with a 0 of exactly half of the codewords begin with a 0.

Exercise 1.16. Prove (1.10).

Exercise 1.17. Show that there is no binary code with block length 4 that achieves the Hamming bound.

Exercise 1.18. $^{(*)}$ There are n people in a room, each of whom is given a black/white hat chosen uniformly at random (and independent of the choices of all other people). Each person can see the hat color of all other people, but not their own. Each person is asked if (s)he wishes to guess their own hat color. They can either guess, or abstain. Each person makes their choice without knowledge of what the other people are doing. They either win collectively, or lose collectively. They win if all the people who don't abstain guess their hat color correctly *and* at least one person does not abstain. They lose if all people abstain, or if some person guesses their color incorrectly. Your goal below is to come up with a strategy that will allow the n people to win with pretty high probability. We begin with a simple warmup:

(a) Argue that the n people can win with probability at least $\frac{1}{2}$.

Next we will see how one can really bump up the probability of success with some careful modeling, and some knowledge of Hamming codes. (Below are assuming knowledge of the general Hamming code (see Section 2.4). If you do not want to skip ahead, you can assume that $n = 7$ in the last part of this problem.

(b) Lets say that a directed graph G is a subgraph of the n-dimensional hypercube if its vertex set is $\{0,1\}^n$ and if $u \to v$ is an edge in G, then u and v differ in at most one coordinate. Let $K(G)$ be the number of vertices of G with in-degree at least one, and out-degree zero. Show that the probability of winning the hat problem equals the maximum, over directed subgraphs G of the n-dimensional hypercube, of $K(G)/2^n$.

(c) Using the fact that the out-degree of any vertex is at most n, show that $K(G)/2^n$ is at most $\frac{n}{n+1}$ for any directed subgraph G of the n-dimensional hypercube.

(d) Show that if $n = 2^r - 1$, then there exists a directed subgraph G of the n-dimensional hypercube with $K(G)/2^n = \frac{n}{n+1}$.
Hint: This is where the Hamming code comes in.

1.9 Bibliographic Notes

Coding theory owes its origin to two remarkable papers: one by Shannon and the other by Hamming [52] both of which were published within a couple of years of each other. Shannon's paper defined the BSC_p channel (among others) and defined codes in terms of its encoding function. Shannon's paper also explicitly defined the decoding function. Hamming's work defined the notion of codes as in Definition 1.2.1 as well as the notion of Hamming distance. Both the Hamming bound and the Hamming code are (not surprisingly) due to Hamming. The specific definition of Hamming code that we used in this book was the one proposed by Hamming and is also mentioned in Shannon's paper (even though Shannon's paper pre-dates Hamming's).

The notion of erasures was defined by Elias.

One hybrid model to account for the fact that in real life the noise channel is somewhere in between the extremes of the channels proposed by Hamming and Shannon is the *Arbitrary Varying Channel* (the reader is referred to the survey by Lapidoth and Narayan [66]).

Chapter 2

A Look at Some Nicely Behaved Codes: Linear Codes

Let us now pause for a bit and think about how we can represent a code. In general, a code $C : [q]^k \longrightarrow [q]^n$ can be stored using nq^k symbols from $[q]$ (n symbols for each of the q^k codewords) or $nq^k \log q$ bits. For constant rate codes, this is exponential space, which is prohibitive even for modest values of k like $k = 100$. A natural question is whether we can do better. Intuitively, the code must have some extra structure that would facilitate a succinct representation. We will now look at a class of codes called *linear codes* that have more structure than general codes which leads to some other nice properties. We have already seen binary linear codes in Section 1.5, that is: $C \subseteq \{0,1\}^n$ is linear code if for all $\mathbf{c}_1, \mathbf{c}_2 \in C$, $\mathbf{c}_1 + \mathbf{c}_2 \in C$, where the "+" denotes bit-wise XOR.

Definition 2.0.1 (Linear Codes)**.** Let q be a prime power (i.e. $q = p^s$ for some prime p and integer $s \geq 1$). $C \subseteq \{0,1,...,q-1\}^n$ is a *linear code* if it is a *linear subspace* of $\{0,1,...,q-1\}^n$. If C has dimension k and distance d then it will be referred to as an $[n,k,d]_q$ or just an $[n,k]_q$ code.

Of course, the above definition is not complete because we have not yet defined a linear subspace. We do that next.

2.1 Finite Fields

To define linear subspaces, we will need to work with (finite) fields. At a high level, we need finite fields since when we talk about codes, we deal with finite symbols/numbers and we want to endow these symbols with the same math that makes arithmetic over real numbers work. Finite fields accomplish this precise task. We begin with a quick overview of fields.

Informally speaking, a field is a set of elements on which one can do addition, subtraction, multiplication and division and still stay in the set. More formally,

Definition 2.1.1. A field $\mathbb{F}$ is given by a triple $(S, +, \cdot)$, where S is the set of elements and $+, \cdot$ are functions $\mathbb{F} \times \mathbb{F} \to \mathbb{F}$ with the following properties:

- CLOSURE: For every $a, b \in S$, we have both $a + b \in S$ and $a \cdot b \in S$.
- ASSOCIATIVITY: $+$ and $\cdot$ are associative: that is, for every $a, b, c \in S$, $a + (b + c) = (a + b) + c$ and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.
- COMMUTATIVITY: $+$ and $\cdot$ are commutative: that is, for every $a, b \in S$, $a + b = b + a$ and $a \cdot b = b \cdot a$.
- DISTRIBUTIVITY: $\cdot$ distributes over $+$: that is, for every $a, b, c \in S$, $a \cdot (b + c) = a \cdot b + a \cdot c$.
- IDENTITIES: There exists special elements 0 and 1 in S such that the following holds for every $a \in S$: $a + 0 = a$ and $a \cdot 1 = a$.
- INVERSES: For every $a \in S$, there exists its unique *additive inverse* $-a$ such that $a + (-a) = 0$. Also for every $a \in S \setminus \{0\}$, there exists its unique *multiplicative inverse* a^{-1} such that $a \cdot a^{-1} = 1$.

With the usual semantics for $+$ and $\cdot$, $\mathbb{R}$ (set of real number) is a field, but $\mathbb{Z}$ (set of integers) is not a field as division of two integers results in a rational number that need not be an integer (the set of rational numbers itself is a field though: see Exercise 2.1). In this course, we will exclusively deal with *finite fields*. As the name suggests these are fields with a finite set of elements. (We will overload notation and denote the size of a field $|\mathbb{F}| = |S|$.) The following is a well known result.

Theorem 2.1.1 (Size of Finite Fields)**.** *The size of any finite field is p^s for prime p and integer $s \geq 1$.*

One example of a finite field that we have seen is the field with $S = \{0, 1\}$, which we will denote by $\mathbb{F}_2$ (we have seen this field in the context of binary linear codes). For $\mathbb{F}_2$, addition is the XOR operation, while multiplication is the AND operation. The additive inverse of an element in $\mathbb{F}_2$ is the number itself while the multiplicative inverse of 1 is 1 itself.

Let p be a prime number. Then the integers modulo p form a field, denoted by $\mathbb{F}_p$ (and also by $\mathbb{Z}_p$), where the addition and multiplication are carried out modulo p. For example, consider $\mathbb{F}_7$, where the elements are $\{0, 1, 2, 3, 4, 5, 6\}$. We have $(4 + 3) \mod 7 = 0$ and $4 \cdot 4 \mod 7 = 2$. Further, the additive inverse of 4 is 3 as $(3 + 4) \mod 7 = 0$ and the multiplicative inverse of 4 is 2 as $4 \cdot 2 \mod 7 = 1$.

More formally, we prove the following result.

Lemma 2.1.2. *Let p be a prime. Then $\mathbb{F}_p = (\{0, 1, \ldots, p-1\}, +_p, \cdot_p)$ is a field, where $+_p$ and $\cdot_p$ are addition and multiplication modulo p.*

Proof. The properties of associativity, commutativity, distributivity and identities hold for integers and hence, they hold for $\mathbb{F}_p$. The closure property follows since both the "addition" and "multiplication" are done modulo p, which implies that for any $a, b \in \{0, \ldots, p-1\}$, $a +_p b, a \cdot_p b \in \{0, \ldots, p-1\}$. Thus, to complete the proof, we need to prove the existence of unique additive and multiplicative inverses.

Fix an arbitrary $a \in \{0, \ldots, p-1\}$. Then we claim that its additive inverse is $p - a \mod p$. It is easy to check that $a + p - a = 0 \mod p$. Next we argue that this is the unique additive inverse. To see this note that the sequence $a, a+1, a+2, \ldots, a+p-1$ are p consecutive numbers and thus, exactly one of them is a multiple of p, which happens for $b = p - a \mod p$, as desired.

Now fix an $a \in \{1, \ldots, p-1\}$. Next we argue for the existence of a unique multiplicative universe a^{-1}. Consider the set of numbers $T = \{a \cdot_p b | b \in \{1, \ldots, p-1\}\}$. We claim that all these numbers are unique. To see this, note that if this is not the case, then there exist $b_1 \neq b_2 \in \{0, 1, \ldots, p-1\}$ such that $a \cdot b_1 = a \cdot b_2 \mod p$, which in turn implies that $a \cdot (b_1 - b_2) = 0 \mod p$. Since a and $b_1 - b_2$ are non-zero numbers, this implies that p divides $a \cdot (b_1 - b_2)$. Further, since a and $|b_1 - b_2|$ are both at most $p-1$, this implies that multiplying a and $(b_1 - b_2) \mod p$ results in p, which is a contradiction since p is prime. Thus, we have argued that $|T| = p - 1$ and since each number in T is in $[p-1]$, we have that $T = [p-1]$. Thus, we can conclude that there exists a unique element b such that $a \cdot b = 1 \mod p$ and thus, b is the required a^{-1}. □

One might think that there could be different finite fields with the same number of elements. However, this is not the case:

Theorem 2.1.3. *For every prime power q there is a* unique *finite field with q elements (up to isomorphism*[1]*).*

Thus, we are justified in just using $\mathbb{F}_q$ to denote a finite field on q elements.

2.2 Linear Subspaces

We are finally ready to define the notion of a linear subspace.

Definition 2.2.1 (Linear Subspace)**.** For any *non-empty* subset $S \subseteq \mathbb{F}_q^n$ is a linear subspace if the following properties hold:

1. For every $\mathbf{x}, \mathbf{y} \in S$, $\mathbf{x} + \mathbf{y} \in S$, where the addition is vector addition over $\mathbb{F}_q$ (that is, do addition component wise over $\mathbb{F}_q$).

2. For every $a \in \mathbb{F}_q$ and $\mathbf{x} \in S$, $a \cdot \mathbf{x} \in S$, where the multiplication is done component-wise over $\mathbb{F}_q$.

Here is a (trivial) example of a linear subspace of $\mathbb{F}_5^3$:

$$S_1 = \{(0,0,0), (1,1,1), (2,2,2), (3,3,3), (4,4,4)\}. \tag{2.1}$$

Note that for example $(1,1,1) + (3,3,3) = (4,4,4) \in S_1$ and $2 \cdot (4,4,4) = (3,3,3) \in S_1$ as required by the definition. Here is another somewhat less trivial example of a linear subspace over $\mathbb{F}_3^3$:

$$S_2 = \{(0,0,0), (1,0,1), (2,0,2), (0,1,1), (0,2,2), (1,1,2), (1,2,0), (2,1,0), (2,2,1). \tag{2.2}$$

Note that $(1,0,1) + (0,2,2) = (1,2,0) \in S_2$ and $2 \cdot (2,0,2) = (1,0,1) \in S_2$ as required.

[1] An isomorphism $\phi : S \to S'$ is a map (such that $\mathbb{F} = (S, +, \cdot)$ and $\mathbb{F}' = (S', \oplus, \circ)$ are fields) where for every $a_1, a_2 \in S$, we have $\phi(a_1 + a_2) = \phi(a_1) \oplus \phi(a_2)$ and $\phi(a_1 \cdot a_2) = \phi(a_1) \circ \phi(a_2)$.

Remark 2.2.1. Note that the second property implies that $\mathbf{0}$ is contained in every linear subspace. Further for any subspace over $\mathbb{F}_2$, the second property is redundant: see Exercise 2.4.

Before we state some properties of linear subspaces, we state some relevant definitions.

Definition 2.2.2 (Span)**.** Given a set $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$. The *span* of B is the set of vectors

$$\left\{ \sum_{i=1}^{\ell} a_i \cdot \mathbf{v}_i | a_i \in \mathbb{F}_q \text{ for every } i \in [\ell] \right\}.$$

Definition 2.2.3 (Linear independence of vectors)**.** We say that $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_k$ are *linearly independent* if for every $1 \le i \le k$ and for every $(k-1)$-tuple $(a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_k) \in \mathbb{F}_q^{k-1}$,

$$\mathbf{v}_i \neq a_1\mathbf{v}_1 + \ldots + a_{i-1}\mathbf{v}_{i-1} + a_{i+1}\mathbf{v}_{i+1} + \ldots + a_k\mathbf{v}_k.$$

In other words, $\mathbf{v}_i$ is not in the span of the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n\}$.

For example the vectors $(1,0,1), (1,1,1) \in S_2$ are linearly independent.

Definition 2.2.4 (Rank of a matrix)**.** The *rank* of matrix in $\mathbb{F}_q^{k\times k}$ is the maximum number of linearly independent rows (or columns). A matrix in $\mathbb{F}_q^{k\times n}$ with rank $\min(k,n)$ is said to have *full* rank.

One can define the *row (column) rank* of a matrix as the maximum number of linearly independent rows (columns). However, it is a well-known theorem that the row rank of a matrix is the same as its column rank. For example, the matrix below over $\mathbb{F}_3$ has full rank (see Exercise 2.5):

$$G_2 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}. \tag{2.3}$$

Any linear subspace satisfies the following properties (the full proof can be found in any standard linear algebra textbook).

Theorem 2.2.1. *If $S \subseteq {\mathbb{F}_q}^n$ is a linear subspace then*

1. *$|S| = q^k$ for some $k \ge 0$. The parameter k is called the* dimension *of S.*

2. *There exists at least one set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in S$ called* basis elements *such that every $\mathbf{x} \in S$ can be expressed as $\mathbf{x} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \ldots + a_n\mathbf{v}_n$ where $a_i \in \mathbb{F}_q$ for $1 \le i \le k$. In other words, there exists a full rank $k \times n$ matrix G (also known as a* generator matrix*) with entries from $\mathbb{F}_q$ such that every $\mathbf{x} \in S$, $\mathbf{x} = (a_1, a_2, \ldots a_k) \cdot G$ where*

$$G = \begin{pmatrix} \longleftarrow \mathbf{v}_1 \longrightarrow \\ \longleftarrow \mathbf{v}_2 \longrightarrow \\ \vdots \\ \longleftarrow \mathbf{v}_k \longrightarrow \end{pmatrix}.$$

3. *There exists a full rank* $(n-k) \times n$ *matrix* H *(called a* parity check matrix*) such that for every* $\mathbf{x} \in S$, $H\mathbf{x}^T = \mathbf{0}$.

4. G *and* H *are* orthogonal*, that is,* $G \cdot H^T = \mathbf{0}$.

Proof Sketch.

Property 1. We begin with the proof of the first property. For the sake of contradiction, let us assume that $q^k < |S| < q^{k+1}$, for some $k \geq 0$. Iteratively, we will construct a set of linearly independent vectors $B \subseteq S$ such that $|B| \geq k+1$. Note that by the definition of a linear subspace the span of B should be contained in S. However, this is a contradiction as the size of the span of B is at least[2] $q^{k+1} > |S|$.

To complete the proof, we show how to construct the set B in a greedy fashion. In the first step pick $\mathbf{v}_1$ to be any non-zero vector in S and set $B \leftarrow \{\mathbf{v}_1\}$ (we can find such a vector as $|S| > q^k \geq 1$). Now say after the step t (for some $t \leq k$), $|B| = t$. Now the size of the span of the current B is $q^t \leq q^k < |S|$. Thus there exists a vector $\mathbf{v}_{t+1} \in S \setminus B$ that is linearly independent of vectors in B. Set $B \leftarrow B \cup \{\mathbf{v}_{t+1}\}$. Thus, we can continue building B until $|B| = k+1$, as desired.

Property 2. We first note that we can pick $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ to be any set of k linearly independent vectors– this just follows from the argument above for **Property** 1.1. This is because the span of B is contained in S. However, since $|S| = q^k$ and the span of B has q^k vectors, the two have to be the same.

Property 3. Property 3 above follows from another fact that every linear subspace S has a null space $N \subseteq \mathbb{F}_q^n$ such that for every $\mathbf{x} \in S$ and $\mathbf{y} \in N$, $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Further, it is known that N itself is a linear subspace of dimension $n-k$. (The claim that N is also a linear subspace follows from the following two facts: for every $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{F}_q^n$, (i) $\langle \mathbf{x}, \mathbf{y}+\mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ and (ii) for any $a \in \mathbb{F}_q$, $\langle \mathbf{x}, a\mathbf{y} \rangle = a \cdot \langle \mathbf{x}, \mathbf{y} \rangle$.) In other words, there exists a generator matrix H for it. This matrix H is called the parity check matrix of S.

Property 4. See Exercise 2.8. □

As examples, the linear subspace S_1 in (2.1) has as one of its generator matrices

$$G_1 = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$$

and as one of its parity check matrices

$$H_1 = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 2 & 1 \end{pmatrix}.$$

Further, the linear subspace S_2 in (2.2) has G_2 as one of its generator matrices and has the following as one of its parity check matrices

$$H_2 = \begin{pmatrix} 1 & 1 & 2 \end{pmatrix}.$$

Finally, we state another property of linear subspaces that is useful.

[2] See Exercise 2.7.

Lemma 2.2.2. *Given matrix G of dimension $k \times n$ that is a generator matrix of subspace S_1 and matrix H of dimension $(n-k) \times n$ that is a parity check matrix of subspace S_2 such that $GH^T = \mathbf{0}$, then $S_1 = S_2$.*

Proof. We first prove that $S_1 \subseteq S_2$. Given any $\mathbf{c} \in S_1$, there exists $\mathbf{x} \in \mathbb{F}_q^k$ such that $\mathbf{c} = \mathbf{x}G$. Then,

$$H \cdot \mathbf{c}^T = H \cdot (\mathbf{x}G)^T = HG^T\mathbf{x}^T = \left(GH^T\right)^T \mathbf{x}^T = \mathbf{0},$$

which implies that $\mathbf{c} \in S_2$, as desired.

To complete the proof note that as H has full rank, its null space (or S_2) has dimension $n-(n-k) = k$ (this follows from a well known fact from linear algebra called the *rank-nullity theorem*). Now as G has full rank, the dimension of S_1 is also k. Thus, as $S_1 \subseteq S_2$, it has to be the case that $S_1 = S_2$.[3] □

2.3 Properties of Linear Codes

Theorem 2.2.1 gives two alternate characterizations of an $[n,k]_q$ linear code C:

- C is generated by its $k \times n$ generator matrix G. As an example that we have already seen, the $[7,4,3]_2$ Hamming code has the following generator matrix:

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

- C is also characterized by an $(n-k) \times n$ parity check matrix H. We claim that the following matrix is a parity check matrix of the $[7,4,3]_2$ Hamming code:

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

  Indeed, it can be easily verified that $G \cdot H^T = \mathbf{0}$. Then Lemma 2.2.2 proves that H is indeed a parity check matrix of the $[7,4,3]_2$ Hamming code.

We now look at some consequences of the above characterizations of an $[n,k]_q$ linear code C. We started this chapter with a quest for succinct representation of a code. Note that both the generator matrix and the parity check matrix can be represented using $O(n^2)$ symbols from $\mathbb{F}_q$ (which is much smaller than the exponential representation of a general code). More precisely (see Exercise 2.10),

[3] If not, $S_1 \subset S_2$ which implies that that $|S_2| \geq |S_1| + 1$. The latter is not possible if both S_1 and S_2 have the same dimension.

Proposition 2.3.1. *Any $[n,k]_q$ linear code can be represented with* $\min(nk, n(n-k))$ *symbols from* $\mathbb{F}_q$.

There is an encoding algorithm for C that runs in $O(n^2)$ (in particular $O(kn)$) time– given a message $\mathbf{m} \in \mathbb{F}_q^k$, the corresponding codeword $C(\mathbf{m}) = \mathbf{m} \cdot G$, where G is the generator matrix of C. (See Exercise 2.11.)

Proposition 2.3.2. *For any $[n,k]_q$ linear code, given its generator matrix, encoding can be done with $O(nk)$ operations over $\mathbb{F}_q$.*

There is an error-detecting algorithm for C that runs in $O(n^2)$. This is a big improvement over the naive brute force exponential time algorithm (that goes through all possible codewords $\mathbf{c} \in C$ and checks if $\mathbf{y} = \mathbf{c}$). (See Exercise 2.12.)

Proposition 2.3.3. *For any $[n,k]_q$ linear code, given its parity check matrix, error detection can be performed in $O(n(n-k))$ operations over $\mathbb{F}_q$.*

Next, we look at some alternate characterizations of the distance of a linear code.

2.3.1 On the Distance of a Linear Code

We start with the following property, which we have seen for the special case of binary linear codes (Proposition 1.5.3).

Proposition 2.3.4. *For a $[n,k,d]_q$ code C,*

$$d = \min_{\substack{\mathbf{c} \in C, \\ \mathbf{c} \neq \mathbf{0}}} wt(\mathbf{c}).$$

Proof. To show that d is the same as the minimum weight we show that d is no more than the minimum weight and d is no less than the minimum weight.

First, we show that d is no more than the minimum weight. We can see this by considering $\Delta(\mathbf{0}, \mathbf{c}')$ where $\mathbf{c}'$ is the non-zero codeword in C with minimum weight; its distance from $\mathbf{0}$ is equal to its weight. Thus, we have $d \leq wt(\mathbf{c}')$, as desired.

Now, to show that d is no less than the minimum weight, consider $\mathbf{c_1} \neq \mathbf{c_2} \in C$ such that $\Delta(\mathbf{c}_1, \mathbf{c}_2) = d$. Note that $\mathbf{c}_1 - \mathbf{c_2} \in C$ (this is because $-\mathbf{c_2} = -1 \cdot \mathbf{c}_2 \in C$, where -1 is the additive inverse of 1 in $\mathbb{F}_q$ and $\mathbf{c}_1 - \mathbf{c}_2 = \mathbf{c}_1 + (-\mathbf{c}_2)$, which in in C by the definition of linear codes). Now note that $wt(\mathbf{c_1} - \mathbf{c_2}) = \Delta(\mathbf{c_1}, \mathbf{c_2}) = d$, since the non-zero symbols in $\mathbf{c}_1 - \mathbf{c}_2$ occur exactly in the positions where the two codewords differ. Further, since $\mathbf{c}_1 \neq \mathbf{c}_2$, $\mathbf{c}_1 - \mathbf{c}_2 \neq \mathbf{0}$, which implies that the minimum Hamming weight of any non-zero codeword in C is at most d. □

Next, we look at another property implied by the parity check matrix of a linear code.

Proposition 2.3.5. *For any $[n,k,d]_q$ code C with parity check matrix H, d is the minimum number of linearly dependent*[4] *columns in H.*

[4] A set of vectors is linearly dependent if it is *not* linearly independent.

Proof. By Proposition 2.3.4, we need to show that the minimum weight of a non-zero codeword in C is the minimum number of linearly dependent columns. Let t be the minimum number of linearly dependent columns in H. To prove the claim we will show that $t \leq d$ and $t \geq d$.

For the first direction, Let $\mathbf{c} \neq \mathbf{0} \in C$ be a codeword with $wt(\mathbf{c}) = d$. Now note that, by the definition of the parity check matrix, $H \cdot \mathbf{c}^T = \mathbf{0}$. Working through the matrix multiplication, this gives us that $\sum_{i=1}^{n} c_i H^i$, where

$$H = \begin{pmatrix} \uparrow & \uparrow & & \uparrow & & \uparrow \\ H^1 & H^2 & \cdots & H^i & \cdots & H^n \\ \downarrow & \downarrow & & \downarrow & & \downarrow \end{pmatrix}$$

and $\mathbf{c} = (c_1, \ldots, c_n)$. Note that we can skip multiplication for those columns for which the corresponding bit c_i is zero, so for this to be zero, those H^i with $c_i \neq 0$ are linearly dependent. This means that $d \geq t$, as the columns corresponding to non-zero entries in $\mathbf{c}$ are one instance of linearly dependent columns.

For the other direction, consider the minimum set of columns from H, $H^{i_1}, H^{i_2}, \ldots, H^{i_t}$ that are linearly dependent. This implies that there exists non-zero elements $c'_{i_1}, \ldots, c'_{i_t} \in \mathbb{F}_q$ such that $c'_{i_i} H^{i_1} + \ldots + c'_{i_t} H^{i_t} = \mathbf{0}$. (Note that all the c'_{i_j} are non-zero as no set of less than t columns are linearly dependent.) Now extend $c'_{i_1}, \ldots, c'_{i_t}$ to the vector $\mathbf{c}'$ such that $c'_j = 0$ for $j \notin \{i_1, \ldots, i_t\}$. Note that we have $H \cdot \left(\mathbf{c}'\right)^T = \mathbf{0}$ and thus, we have $\mathbf{c}' \in C$. This in turn implies that $d \leq wt(\mathbf{c}') = t$ (where recall t is the minimum number of linearly independent columns in H). □

2.4 Hamming Codes

We now change gears and look at the general family of linear codes, which were discovered by Hamming. So far we have seen the $[7,4,3]_2$ Hamming code (in Section 1.5). In fact, for any $r \geq 2$ there is a $[2^r - 1, 2^r - r - 1, 3]_2$ Hamming code. Thus in Section 1.5, we have seen this code for $r = 3$.

Definition 2.4.1. Define the $r \times (2^r - 1)$ matrix $\mathbf{H}_r$ over $\mathbb{F}_2$, such that the ith column $\mathbf{H}_r^i$, $1 \leq i \leq 2^r - 1$, is the binary representation of i (note that such a representation is a vector in $\{0,1\}^r$).

For example, for the case we have seen ($r = 3$),

$$\mathbf{H}_3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Note that by its definition, the code that has $\mathbf{H}_r$ as its parity check matrix has block length $2^r - 1$ and dimension $2^r - r - 1$. This leads to the formal definition of the general Hamming code.

Definition 2.4.2. The $[2^r - 1, 2^r - r - 1]_2$ Hamming code, denoted by $C_{H,r}$ has parity check matrix $\mathbf{H}_r$.

In other words, the general $[2^r-1, 2^r-r-1]_2$ Hamming code is the code

$$\{\mathbf{c} \in \{0,1\}^{2^r-1} | \mathbf{H}_r \cdot \mathbf{c}^T = \mathbf{0}\}.$$

Next we argue that the above Hamming code has distance 3 (in Proposition 1.5.1, we argued this for $r = 3$).

Proposition 2.4.1. *The Hamming code* $[2^r-1, 2^r-r-1, 3]_2$ *has distance 3.*

Proof. No two columns in $\mathbf{H}_r$ are linearly dependent. If they were, we would have $\mathbf{H}_r^i + \mathbf{H}_r^j = \mathbf{0}$, but this is impossible since they differ in at least one bit (being binary representations of integers, $i \neq j$). Thus, by Proposition 2.3.5, the distance is at least 3. It is at most 3, since (e.g.) $\mathbf{H}_r^1 + \mathbf{H}_r^2 + \mathbf{H}_r^3 = \mathbf{0}$. □

Now note that under the Hamming bound for $d = 3$ (Theorem 1.6.1), $k \leq n - \log_2(n+1)$, so for $n = 2^r - 1$, $k \leq 2^r - r - 1$. Hence, the Hamming code is a perfect code. (See Definition 1.7.2.)

In Question 1.7.1, we asked which codes are perfect codes. Interestingly, the only perfect binary codes are the following:

- The Hamming codes which we just studied.
- The trivial $[n, 1, n]_2$ codes for odd n (which have 0^n and 1^n as the only codewords): see Exercise 2.22.
- Two codes due to Golay [36].

2.5 Family of codes

Until now, we have mostly studied specific codes, that is, codes with *fixed* block lengths and dimension. The only exception was the "family" of $[2^r-1, 2^r-r-1, 3]_2$ Hamming codes (for $r \geq 2$) that we studied in the last section. We will see shortly, when we do an asymptotic study of codes, that it makes more sense to talk about a family of codes. First, we define the notion of family of codes:

Definition 2.5.1 (Family of codes)**.** Let $q \geq 2$. Let $\{n_i\}_{i\geq 1}$ be an increasing sequence of block lengths and suppose there exists sequences $\{k_i\}_{i\geq 1}$ and $\{d_i\}_{i\geq 1}$ such that for all $i \geq 1$ there exists an $(n_i, k_i, d_i)_q$ code C_i. Then the sequence $C = \{C_i\}_{i\geq 1}$ is a family of codes. The rate of C is defined as

$$R(C) = \lim_{i\to\infty} \left\{ \frac{k_i}{n_i} \right\}.$$

The relative distance of C is defined as

$$\delta(C) = \lim_{i\to\infty} \left\{ \frac{d_i}{n_i} \right\}.$$

For example, C_H the family of Hamming code is a family of codes with $n_i = 2^i - 1, k_i = 2^i - i - 1, d_i = 3$ and thus,

$$R(C_H) = \lim_{i\to\infty} 1 - \frac{i}{2^i - 1} = 1,$$

and

$$\delta(C_H) = \lim_{i\to\infty} \frac{3}{2^i - 1} = 0.$$

We will mostly work with families of codes from now on. This is necessary as we will study the asymptotic behavior of algorithms for codes, which does not make sense for a fixed code. For example, when we say that a decoding algorithm for a code C takes $O(n^2)$ time, we would be implicitly assuming that C is a family of codes and that the algorithm has an $O(n^2)$ running time when the block length is large enough. From now on, unless mentioned otherwise, whenever we talk about a code, we will be implicitly assuming that we are talking about a family of codes.

Given that we can only formally talk about asymptotic run time of algorithms, we now also state our formal notion of efficient algorithms:

We'll call an algorithm related to a code of block length n to be efficient, if it runs in time polynomial in n.

For all the specific codes that we will study in this book, the corresponding family of codes will be a "family" in a more natural sense. By this we mean that all the specific codes in a family of codes will be the "same" code except with different parameters. A bit more formally, we will consider families $\{C_i\}_{i\geq 1}$, where given only the 'index' i, one can compute a sufficient description of C_i efficiently.[5]

Finally, the definition of a family of codes allows us to present the final version of the big motivating question for the book. The last formal version of the main question we considered was Question 1.4.1, where we were interested in the tradeoff of rate R and distance d. The comparison was somewhat unfair because R was a ratio while d was an integer. A more appropriate comparison should be between rate R and the relative distance δ. Further, we would be interested in tackling the main motivating question for families of codes, which results in the following final version:

Question 2.5.1. *What is the optimal tradeoff between $R(C)$ and $\delta(C)$ that can be achieved by some code family C?*

[5]We stress that this is not *always* going to be the case. In particular, we will consider "random" codes where this efficient constructibility will not be true.

2.6 Efficient Decoding of Hamming codes

We have shown that the Hamming code has a distance of 3 and thus, by Proposition 1.4.1, can correct one error. However, this is a *combinatorial* result and does not give us an efficient algorithm. One obvious candidate for decoding is the MLD function. Unfortunately, the only implementation of MLD that we know is the one in Algorithm 2, which will take time $2^{\Theta(n)}$, where n is the block length of the Hamming code. However, we can do much better. Consider the following simple algorithm: given the received word $\mathbf{y}$, first check if it is indeed a valid codeword. If it is, we are done. Otherwise, flip each of the n bits and check if the resulting vector is a valid codeword. If so, we have successfully decoded from one error. (If none of the checks are successful, then we declare a decoding failure.) Algorithm 3 formally presents this algorithm (where $C_{H,r}$ is the $[2^r-1, 2^r-r-1, 3]_2$ Hamming code).[6]

Algorithm 3 Naive Decoder for Hamming Code

INPUT: Received word $\mathbf{y}$
OUTPUT: $\mathbf{c}$ if $\Delta(\mathbf{y},\mathbf{c}) \le 1$ else `Fail`

```
IF y ∈ C_{H,r} THEN
    RETURN y
FOR i = 1...n DO
    y' ← y + e_i                    ▷ e_i is the ith standard basis vector
    IF y' ∈ C_{H,r} THEN
        RETURN y'
RETURN Fail
```

It is easy to check that Algorithm 3 can correct up to 1 error. If each of the checks $\mathbf{y}' \in C_{H,r}$ can be done in $T(n)$ time, then the time complexity of the proposed algorithm will be $O(nT(n))$. Note that since $C_{H,r}$ is a linear code (and dimension $k = n - O(\log n)$) by Proposition 2.3.3, we have $T(n) = O(n \log n)$. Thus, the proposed algorithm has running time $O(n^2 \log n)$.

Note that Algorithm 3 can be generalized to work for any linear code C with distance $2t+1$ (and hence, can correct up to t errors): go through all possible error vectors $\mathbf{z} \in [q]^n$ (with $wt(\mathbf{z}) \le t$) and check if $\mathbf{y}-\mathbf{z}$ is in the code or not. Algorithm 4 presents the formal algorithm (where C is an $[n,k,2t+1]_q$ code).

The number of error patterns $\mathbf{z}$ considered by Algorithm 4 is[7] $\sum_{i=0}^{t}\binom{n}{i}(q-1)^i \le O((nq)^t)$. Further by Proposition 2.3.3, Step 4 can be performed with $O(n^2)$ operations over $\mathbb{F}_q$. Thus, Algorithm 4 runs with $O(n^{t+2}q^t)$ operations over $\mathbb{F}_q$, which for q being a small polynomial in n, is $n^{O(t)}$ operations. In other words, the algorithm will have polynomial running time for codes with constant distance (though the running time would not be practical even for moderate values of t).

[6] Formally speaking, a decoding algorithm should return the transmitted message $\mathbf{x}$ but Algorithm 3 actually returns $C_{H,r}(\mathbf{x})$. However, since $C_{H,r}$ is a linear code, it is not too hard to see that one can obtain $\mathbf{x}$ from $C_{H,r}(\mathbf{x})$ in $O(n^3)$ time: see Exercise 2.23. Further, for $C_{H,r}$ one can do this in $O(n)$ time: see Exercise 2.24.

[7] Recall (1.15).

Algorithm 4 Decoder for Any Linear Code

INPUT: Received word $\mathbf{y}$
OUTPUT: $\mathbf{c} \in C$ if $\Delta(\mathbf{y}, \mathbf{c}) \leq t$ else `Fail`

```
FOR i = 0 ... t DO
    FOR S ⊆ [n] such that |S| = i DO
        FOR z ∈ F_q^n such that wt(z_S) = wt(z) = i DO
            IF y − z ∈ C THEN
                RETURN y − z
RETURN Fail
```

However, it turns out that for Hamming codes there exists a decoding algorithm with an $O(n^2)$ running time. To see this, first note that if the received word $\mathbf{y}$ has no errors, then $\mathbf{H}_r \cdot \mathbf{y}^T = \mathbf{0}$. If not, then $\mathbf{y} = \mathbf{c} + \mathbf{e}_i$, where $\mathbf{c} \in C$ and $\mathbf{e}_i$ is the unit vector with the only nonzero element at the i-th position. Thus, if $\mathbf{H}_r^i$ stands for the i-th column of $\mathbf{H}_r$,

$$\mathbf{H}_r \cdot \mathbf{y}^T = \mathbf{H}_r \cdot \mathbf{c}^T + \mathbf{H}_r \cdot (\mathbf{e}_i)^T = \mathbf{H}_r \cdot (\mathbf{e}_i)^T = \mathbf{H}_r^i,$$

where the second equality follows as $\mathbf{H}_r \cdot \mathbf{c}^T = 0$, which in turn follows from the fact that $\mathbf{c} \in C$. In other words, $\mathbf{H}_r \cdot \mathbf{y}^T$ gives the *location* of the error. This leads to Algorithm 5.

Algorithm 5 Efficient Decoder for Hamming Code

INPUT: Received word $\mathbf{y}$
OUTPUT: $\mathbf{c}$ if $\Delta(\mathbf{y}, \mathbf{c}) \leq 1$ else `Fail`

```
1: b ← H_r · y^T.
2: Let i ∈ [n] be the number whose binary representation is b
3: IF y − e_i ∈ C_H THEN
4:     RETURN y − e_i
5: RETURN Fail
```

Note that $\mathbf{H}_r$ is an $r \times n$ matrix where $n = 2^r - 1$ and thus, $r = \Theta(\log n)$. This implies Step 1 in Algorithm 5, which is a matrix vector multiplication can be done in time $O(n \log n)$. By a similar argument and by Proposition 2.3.3 Step 3 can be performed in $O(n \log n)$ time, and therefore Algorithm 5 overall runs in $O(n \log n)$ time. Thus,

Theorem 2.6.1. *The $[n = 2^r - 1, 2^r - r - 1, 3]_2$ Hamming code is 1-error correctable. Furthermore, decoding can be performed in time $O(n \log n)$.*

2.7 Dual of a Linear Code

Until now, we have thought of parity check matrix as defining a code via its null space. However, we are not beholden to think of the parity check matrix in this way. A natural alternative is to use the parity check matrix as a generator matrix. The following definition addresses this question.

Definition 2.7.1 (Dual of a code)**.** Let H be a parity check matrix of a code C, then the code generated by H is called the dual of C. For any code C, its dual is denoted by $C^{\perp}$.

It is obvious from the definition that if C is an $[n,k]_q$ code, then $C^{\perp}$ is an $[n,n-k]_q$ code. The first example that might come to mind is $C^{\perp}_{H,r}$, which is also known as the *Simplex code* (we will denote it by $C_{Sim,r}$). Adding an all 0's column to H_r and using the resulting matrix as a generating matrix, we get the *Hadamard* code (we will denote it by C_{Had}, r). We claim that $C_{Sim,r}$ and $C_{Had,r}$ are $[2^r-1, r, 2^{r-1}]_2$ and $[2^r, r, 2^{r-1}]_2$ codes respectively. The claimed block length and dimension follow from the definition of the codes, while the distance follows from the following result.

Proposition 2.7.1. *$C_{Sim,r}$ and $C_{Had,r}$ both have distances of 2^{r-1}.*

Proof. We first show the result for $C_{Had,r}$. In fact, we will show something stronger: every non-zero codeword in $C_{Had,r}$ has weight exactly equal to 2^{r-1} (the claimed distance follows from Proposition 2.3.4). Consider a message $\mathbf{x} \neq 0$. Let its ith entry be $x_i = 1$. $\mathbf{x}$ is encoded as

$$\mathbf{c} = (x_1, x_2, \ldots, x_r)(H_r^0, H_r^1, \ldots, H_r^{2^r-1}),$$

where H_r^j is the binary representation of $0 \leq j \leq 2^r - 1$ (that is, it contains all the vectors in $\{0,1\}^r$). Further note that the jth bit of the codeword $\mathbf{c}$ is $\langle \mathbf{x}, H_r^j \rangle$. Group all the columns of the generator matrix into pairs $(\mathbf{u}, \mathbf{v})$ such that $\mathbf{v} = \mathbf{u} + \mathbf{e}_i$ (i.e. $\mathbf{v}$ and $\mathbf{u}$ are the same except in the ith position). Notice that this partitions all the columns into 2^{r-1} disjoint pairs. Then,

$$\langle \mathbf{x}, \mathbf{v} \rangle = \langle \mathbf{x}, \mathbf{u} + \mathbf{e}_i \rangle = \langle \mathbf{x}, \mathbf{u} \rangle + \langle \mathbf{x}, \mathbf{e}_i \rangle = \langle \mathbf{x}, \mathbf{u} \rangle + x_i = \langle \mathbf{x}, \mathbf{u} \rangle + 1.$$

Thus we have that exactly one of $\langle \mathbf{x}, \mathbf{v} \rangle$ and $\langle \mathbf{x}, \mathbf{u} \rangle$ is 1. As the choice of the pair $(\mathbf{u}, \mathbf{v})$ was arbitrary, we have proved that for any non-zero codeword $\mathbf{c}$ such that $\mathbf{c} \in C_{Had}$, $wt(\mathbf{c}) = 2^{r-1}$.

For the simplex code, we observe that all codewords of $C_{Had,3}$ are obtained by padding a 0 to the beginning of the codewords in $C_{Sim,r}$, which implies that all non-zero codewords in $C_{Sim,r}$ also have a weight of 2^{r-1}, which completes the proof. □

We remark that the family of Hamming code has a rate of 1 and a (relative) distance of 0 while the families of Simplex/Hadamard codes have a rate of 0 and a relative distance of 1/2. Notice that both code families either have rate or relative distance equal to 0. Given this, the following question is natural special case of Question 2.5.1:

Question 2.7.1. *Does there exist a family of codes C such that $R(C) > 0$ and $\delta(C) > 0$ hold* simultaneously*?*

Codes that have the above property are called *asymptotically good*.

2.8 Exercises

Exercise 2.1. Prove that the set of rationals (i.e. the set of reals of the form $\frac{a}{b}$, where both a and $b \neq 0$ are integers), denoted by $\mathbb{Q}$, is a field.

Exercise 2.2. Let q be a prime power. Let $x \in \mathbb{F}_q$ such that $x \notin \{0,1\}$. Then prove that for any $n \leq q-1$:

$$\sum_{i=0}^{n} x^i = \frac{x^{n+1}-1}{x-1}.$$

Exercise 2.3. The main aim of this exercise is to prove the following identity that is true for any $\alpha \in \mathbb{F}_q$:

$$\alpha^q = \alpha \tag{2.4}$$

To make progress towards the above we will prove a sequence of properties of *groups*. A group G is a pair $(S, \circ)$ where the operator $\circ : G \times G \to G$ such that $\circ$ is commutative[8] and the elements of S are closed under $\circ$. Further, there is a special element $\iota \in S$ that is the identity element and every element $a \in S$ has an inverse element $b \in S$ such that $a \circ b = \iota$. Note that a finite field $\mathbb{F}_q$ consists of an *additive group* with the + operator (and 0 as additive identity) and a *multiplicative* group on the non-zero elements of $\mathbb{F}_q$ (which is also denoted by $\mathbb{F}_q^*$) with the $\cdot$ operator (and 1 as the multiplicative identity).[9]

For the rest of the problem let $G = (S, \cdot)$ be a multiplicative group with $|G| = m$. Prove the following statements.

1. For any $\beta \in G$, let $o(\beta)$ be the smallest integer o such that $\beta^o = 1$. Prove that such an $o \leq m$ always exists. Further, argue that $T = \{1, \beta, \ldots, \beta^{o-1}\}$ also forms a group. $(T, \cdot)$ is called a *sub-group* of G and $o(\beta)$ is called the *order* of β.

2. For any $g \in G$, define the *coset* (w.r.t. T) as

$$gT = \{g \cdot \beta | \beta \in T\}.$$

   Prove that if $g \cdot h^{-1} \in T$ then $gT = hT$ and $gT \cap hT = \emptyset$ otherwise. Further argue that these cosets partition the group G into disjoint sets.

3. Argue that for any $g \in G$, we have $|gT| = |T|$.

4. Using the above results or otherwise, argue that for any $\beta \in G$, we have

$$\beta^m = 1.$$

5. Prove (2.4).

Exercise 2.4. Prove that for $q = 2$, the second condition in Definition 2.2.1 is implied by the first condition.

[8] Technically, G is an *abelian* group.

[9] Recall Definition 2.1.1.

Exercise 2.5. Prove that G_2 from (2.3) has full rank.

Exercise 2.6. In this problem we will look at the problem of solving a system of linear equations over $\mathbb{F}_q$. That is, one needs to solve for unknowns $x_1, \ldots, x_n$ given the following m linear equations (where $a_{i,j}, b_i \in \mathbb{F}_q$ for $1 \le i \le m$ and $1 \le j \le n$):

$$a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = b_1.$$

$$a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n = b_2.$$

$$\vdots$$

$$a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n = b_m.$$

1. (*Warm-up*) Convince yourself that the above problem can be stated as $A \cdot \mathbf{x}^T = \mathbf{b}^T$, where A is an $m \times n$ matrix over $\mathbb{F}_q$, $\mathbf{x} \in \mathbb{F}_q^n$ and $\mathbf{b} \in \mathbb{F}_q^m$.

2. (*Upper Triangular Matrix*) Assume $n = m$ and that A is upper triangular, i.e. all diagonal elements ($a_{i,i}$) are non-zero and all lower triangular elements ($a_{i,j}$, $i > j$) are 0. Then present an $O(n^2)$ time[10] algorithm to compute the unknown vector $\mathbf{x}$.

3. (*Gaussian Elimination*) Assume that A has *full rank* (or equivalently a *rank* of n.)

   (a) Prove that the following algorithm due to Gauss converts A into an upper triangular matrix. By permuting the columns if necessary make sure that $a_{1,1} \neq 0$. (Why can one assume w.l.o.g. that this can be done?) Multiply all rows $1 < i \le n$ with $\frac{a_{1,1}}{a_{i,1}}$ and then subtract $a_{1,j}$ from the (i,j)th entry $1 \le j \le n$. Recurse with the same algorithm on the $(n-1) \times (n-1)$ matrix A' obtained by removing the first row and column from A. (Stop when $n = 1$.)

   (b) What happens if A does not have full rank? Show how one can modify the algorithm above to either upper triangulate a matrix or report that it does not have full rank. (Convince yourself that your modification works.)

   (c) Call a system of equations $A \cdot \mathbf{x}^T = \mathbf{b}^T$ *consistent* if there exists a solution to $\mathbf{x} \in \mathbb{F}_q^n$. Show that there exists an $O(n^3)$ algorithm that finds the solution if the system of equations is consistent and A has full rank (and report "fail" otherwise).

4. ($m < n$ *case*) Assume that A has full rank, i.e. has a rank of m. In this scenario either the system of equations is inconsistent or there are q^{n-m} solutions to $\mathbf{x}$. Modify the algorithm from above to design an $O(m^2 n)$ time algorithm to output the solutions (or report that the system is inconsistent).

   - Note that in case the system is consistent there will be q^{n-m} solutions, which might be much bigger than $O(m^2 n)$. Show that this is not a problem as one can represent the solutions as system of linear equations. (I.e. one can have $n - m$ "free" variables and m "bound" variables.)

[10] For this problem, any basic operation over $\mathbb{F}_q$ takes unit time.

5. ($m > n$ *case*) Assume that A has full rank, i.e. a rank of n. In this scenario either the system of equations is inconsistent or there is a unique solution to $\mathbf{x}$. Modify the algorithm from above to design an $O(m^2 n)$ time algorithm to output the solution (or report that the system is inconsistent).

6. (*Non-full rank case*) Give an $O(m^2 n)$ algorithm for the general case, i.e. the $m \times n$ matrix A need not have full rank. (The algorithm should either report that the system of equations is inconsistent or output the solution(s) to $\mathbf{x}$.)

Exercise 2.7. Prove that the span of k linearly independent vectors over $\mathbb{F}_q$ has size exactly q^k.

Exercise 2.8. Let G and H be a generator and parity check matrix of the same linear code of dimension k and block length n. Then $G \cdot H^T = \mathbf{0}$.

Exercise 2.9. Let C be an $[n,k]_q$ linear code with a generator matrix with no all zeros columns. Then for every position $i \in [n]$ and $\alpha \in \mathbb{F}_q$, the number of codewords $\mathbf{c} \in C$ such that $c_i = \alpha$ is exactly q^{k-1}.

Exercise 2.10. Prove Proposition 2.3.1.

Exercise 2.11. Prove Proposition 2.3.2.

Exercise 2.12. Prove Proposition 2.3.3.

Exercise 2.13. A set of vector $S \subseteq \mathbb{F}_q^n$ is called t-wise independent if for every set of positions I with $|I| = t$, the set S projected to I has each of the vectors in $\mathbb{F}_q^t$ appear the same number of times. (In other words, if one picks a vector $(s_1, \ldots, s_n)$ from S at random then any of the t random variables are uniformly and independently random over $\mathbb{F}_q$).

Prove that any linear code C whose dual $C^\perp$ has distance $d^\perp$ is $(d^\perp - 1)$-wise independent.

Exercise 2.14. A set of vectors $S \subseteq \mathbb{F}_2^k$ is called ε-biased sample space if the following property holds. Pick a vector $X = (x_1, \ldots, x_k)$ uniformly at random from S. Then X has *bias* at most ε, that is, for every $I \subseteq [k]$,

$$\left| \Pr\left(\sum_{i \in I} x_i = 0 \right) - \Pr\left(\sum_{i \in I} x_i = 1 \right) \right| \leq \varepsilon.$$

We will look at some connections of such sets to codes.

1. Let C be an $[n,k]_2$ code such that all non-zero codewords have Hamming weight in the range $\left[\left(\frac{1}{2} - \varepsilon\right) n, \left(\frac{1}{2} + \varepsilon\right) n\right]$. Then there exists an ε-biased space of size n.

2. Let C be an $[n,k]_2$ code such that all non-zero codewords have Hamming weight in the range $\left[\left(\frac{1}{2} - \gamma\right) n, \left(\frac{1}{2} + \gamma\right) n\right]$ for some constant $0 < \gamma < 1/2$. Then there exists an ε-biased space of size $n^{O(\gamma^{-1} \cdot \log(1/\varepsilon))}$.

Exercise 2.15. Let C be an $[n,k,d]_q$ code. Let $\mathbf{y} = (y_1, \ldots, y_n) \in (\mathbb{F}_q \cup \{?\})^n$ be a received word[11] such that $y_i = ?$ for at most $d-1$ values of i. Present an $O(n^3)$ time algorithm that outputs a codeword $\mathbf{c} = (c_1, \ldots, c_n) \in C$ that agrees with y in all un-erased positions (i.e., $c_i = y_i$ if $y_i \neq ?$) or states that no such $\mathbf{c}$ exists. (Recall that if such a $\mathbf{c}$ exists then it is unique.)

[11] A ? denotes an erasure.

Exercise 2.16. In the chapter, we did not talk about how to obtain the parity check matrix of a linear code from its generator matrix. In this problem, we will look at this "conversion" procedure.

(a) Prove that any generator matrix $\mathbf{G}$ of an $[n,k]_q$ code C (recall that $\mathbf{G}$ is a $k \times n$ matrix) can be converted into another equivalent generator matrix of the form $\mathbf{G}' = [\mathbf{I}_k|\mathbf{A}]$, where $\mathbf{I}_k$ is the $k \times k$ identity matrix and $\mathbf{A}$ is some $k \times (n-k)$ matrix. By "equivalent," we mean that the code generated by $\mathbf{G}'$ has a linear bijective map to C.

Note that the code generated by $\mathbf{G}'$ has the message symbols as its first k symbols in the corresponding codeword. Such codes are called *systematic codes*. In other words, every linear code can be converted into a systematic code. Systematic codes are popular in practice as they allow for immediate access to the message symbols.

(b) Given an $k \times n$ generator matrix of the form $[\mathbf{I}_k|\mathbf{A}]$, give a corresponding $(n-k) \times n$ parity check matrix. Briefly justify why your construction of the parity check matrix is correct.

Hint: Try to think of a parity check matrix that can be decomposed into two submatrices: one will be closely related to $\mathbf{A}$ and the other will be an identity matrix, though the latter might not be a $k \times k$ matrix).

(c) Use part (b) to present a generator matrix for the $[2^r-1, 2^r-r-1, 3]_2$ Hamming code.

Exercise 2.17. So far in this book we have seen that one can modify one code to get another code with interesting properties (for example, the construction of the Hadamard code from the Simplex code from Section 2.7 and Exercise 1.7). In this problem you will need to come up with more ways of constructing new codes from existing ones.

Prove the following statements (recall that the notation $(n,k,d)_q$ code is used for general codes with q^k codewords where k need not be an integer, whereas the notation $[n,k,d]_q$ code stands for a *linear code* of dimension k):

1. If there exists an $(n,k,d)_{2^m}$ code, then there also exists an $(nm, km, d' \geq d)_2$ code.

2. If there exists an $[n,k,d]_{2^m}$ code, then there also exists an $[nm, km, d' \geq d]_2$ code.

3. If there exists an $[n,k,d]_q$ code, then there also exists an $[n-d, k-1, d' \geq \lceil d/q \rceil]_q$ code.

4. If there exists an $[n,k,\delta n]_q$ code, then for every $m \geq 1$, there also exists an $\left(n^m, k/m, \left(1-(1-\delta)^m\right) \cdot n^m\right)_{q^m}$ code.

5. If there exists an $[n,k,\delta n]_2$ code, then for every *odd* $m \geq 1$, there also exists an $\left[n^m, k, \frac{1}{2} \cdot \left(1-(1-2\delta)^m\right) \cdot n^m\right.$ code.

Note: In all the parts, the only things that you can assume about the original code are only the parameters given by its definition– nothing else!

Exercise 2.18. Let C_1 be an $[n, k_1, d_1]_q$ code and C_2 be an $[n, k_2, d_2]_q$ code. Then define a new code as follows:

$$C_1 \ominus C_2 = \{(\mathbf{c}_1, \mathbf{c}_1 + \mathbf{c}_2) | \mathbf{c}_1 \in C_1, \mathbf{c}_2 \in C_2\}.$$

Next we will prove interesting properties of this operations on codes:

1. If G_i is the generator matrix for C_i for $i \in [2]$, what is a generator matrix for $C_1 \ominus C_2$?

2. Argue that $C_1 \ominus C_2$ is an $[2n, k_1 + k_2, d \stackrel{\text{def}}{=} \min(2d_1, d_2)]_q$ code.

3. Assume there exists algorithms $\mathscr{A}_i$ for code C_i for $i \in [2]$ such that: (i) $\mathscr{A}_1$ can decode from e errors and s erasures such that $2e + s < d_1$ and (ii) $\mathscr{A}_2$ can decode from $\lfloor (d_2 - 1)/2 \rfloor$ errors. Then argue that one can correct $\lfloor (d-1)/2 \rfloor$ errors for $C_1 \ominus C_2$.
   Hint: Given a received word $(\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{F}_q^n \times \mathbb{F}_q^n$, first apply $\mathscr{A}_2$ on $\mathbf{y}_2 - \mathbf{y}_1$. Then create an intermediate received word for $\mathscr{A}_1$.

4. We will now consider a recursive construction of a binary linear code that uses the $\ominus$ operator. For integers $0 \le r \le m$, we define the code $C(r, m)$ as follows:

   - $C(r, r) = \mathbb{F}_2^r$ and $C(0, r)$ is the code with only two codewords: the all ones and all zeroes vector in $\mathbb{F}_2^r$.
   - For $1 < r < m$, $C(r, m) = C(r, m-1) \ominus C(r-1, m-1)$.

   Determine the parameters of the code $C(r, m)$.

Exercise 2.19. Let C_1 be an $[n_1, k_1, d_1]_2$ binary linear code, and C_2 an $[n_2, k_2, d_2]$ binary linear code. Let $C \subseteq \mathbb{F}_2^{n_1 \times n_2}$ be the subset of $n_1 \times n_2$ matrices whose columns belong to C_1 and whose rows belong to C_2. C is called the tensor of C_1 and C_2 and is denoted by $C_1 \otimes C_2$.

Prove that C is an $[n_1 n_2, k_1 k_2, d_1 d_2]_2$ binary linear code.

Exercise 2.20. In Section 2.4 we considered the *binary* Hamming code. In this problem we will consider the more general q-ary Hamming code. In particular, let q be a prime power and $r \ge 1$ be an integer. Define the following $r \times n$ matrix $H_{q,r}$, where each column is an non-zero vector from $\mathbb{F}_q^r$ such that the first non-zero entry is 1. For example,

$$H_{3,2} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 \end{pmatrix}$$

In this problem we will derive the parameters of the code. Define the generalized Hamming code $C_{H,r,q}$ to be the linear code whose parity check matrix is $H_{q,r}$. Argue that

1. The block length of $C_{H,r,q}$ is $n = \frac{q^r - 1}{q - 1}$.

2. $C_{H,q,r}$ has dimension $n - r$.

3. $C_{H,q,r}$ has distance 3.

Exercise 2.21. Design the best 6-ary code (family) with distance 3 that you can.

Hint: Start with a 7-ary Hamming code.

Exercise 2.22. Prove that the $[n,1,n]_2$ code for odd n (i.e. the code with the all zeros and all ones vector as it only two codewords) attains the Hamming bound (Theorem 1.7.1).

Exercise 2.23. Let C be an $[n,k]_q$ code with generator matrix G. Then given a codeword $\mathbf{c} \in C$ one can compute the corresponding message in time $O(kn^2)$.

Exercise 2.24. Given a $\mathbf{c} \in C_{H,r}$, one can compute the corresponding message in time $O(n)$.

Exercise 2.25. Let C be an $(n,k)_q$ code. Prove that if C can be decoded from e errors in time $T(n)$, then it can be decoded from $n+c$ errors in time $O((nq)^c \cdot T(n))$.

Exercise 2.26. Show that the bound of kd of the number of ones in the generator matrix of any binary linear code (see Exercise 1.14) cannot be improved for every code.

Exercise 2.27. Let C be a linear code. Then prove that $\left(C^{\perp}\right)^{\perp} = C$.

Exercise 2.28. Note that for any linear code C, the codewords $\mathbf{0}$ is in both C and $C^{\perp}$. Show that there exists a linear code C such that it shares a non-zero codeword with $C^{\perp}$.

Exercise 2.29. We go into a bit of diversion and look at how finite fields are different from infinite fields (e.g. $\mathbb{R}$). Most of the properties of linear subspaces that we have used for linear codes (e.g. notion of dimension, the existence of generator and parity check matrices, notion of duals) also hold for linear subspaces over $\mathbb{R}$.[12] One trivial property that holds for linear subspaces over finite fields that does not hold over $\mathbb{R}$ is that linear subspaces over $\mathbb{F}_q$ with dimension k has size q^k (though this is a trivial consequence that F_q are finite field while $\mathbb{R}$ is an infinite field). Next, we consider a more subtle distinction.

Let $S \subseteq \mathbb{R}^n$ be a linear subspace over $\mathbb{R}$ and let $S^{\perp}$ is the dual of S. Then show that

$$S \cap S^{\perp} = \{\mathbf{0}\}.$$

By contrast, linear subspaces over finite fields can have non-trivial intersection with their duals (see e.g. Exercise 2.28).

Exercise 2.30. A linear code C is called *self-orthogonal* if $C \subseteq C^{\perp}$. Show that

1. The binary repetition code with even number of repetitions is self-orthogonal.
2. The Hadamard code $C_{Had,r}$ is self-orthogonal.

Exercise 2.31. A linear code C is called self dual if $C = C^{\perp}$. Show that for

1. Any self dual code has dimension $n/2$.
2. Prove that the following code is self-dual

$$\{(\mathbf{x},\mathbf{x}) | \mathbf{x} \in \mathbb{F}_2^k\}.$$

[12] A linear subspace $S \subseteq \mathbb{R}^n$ is the same as in Definition 2.2.1 where all occurrences of the finite field $\mathbb{F}_q$ is replaced by $\mathbb{R}$.

Exercise 2.32. Given a code C a *puncturing* of C is another code C' where the same set of positions are dropped in all codewords of C. More precisely, if $C \subseteq \Sigma^n$ and the set of punctured positions is $P \subseteq [n]$, then the punctured code is $\{(c_i)_{i \notin P} | (c_1, \ldots, c_n) \in C\}$.

Prove that a linear code with no repetitions (i.e. there are no two positions $i \neq j$ such that for every codeword $\mathbf{c} \in C$, $c_i = c_i$) is a puncturing of the Hadamard code. Hence, Hadamard code is the "longest" linear code that does not repeat.

Exercise 2.33. In this problem we will consider the *long code*. For the definition, we will use the functional way of looking at the ambient space as mentioned in Remark 1.2.1. A long code of dimension k is a binary code such that the codeword corresponding to $\mathbf{x} = \mathbb{F}_2^k$, is the function $f : \{0,1\}^{2^k} \to \{0,1\}$ defined as follows. For any $\mathbf{m} \in \{0,1\}^{\mathbb{F}_2^k}$, we have $f((m_\alpha)_{\alpha \in \mathbb{F}_2^k}) = m_{\mathbf{x}}$. Derive the parameters of the long code.

Finally, argue that the long code is the code with the longest block length such that the codewords do not have a repeated coordinate (i.e. there does not exists $i \neq j$ such that for every codeword $\mathbf{c}$, $c_i = c_j$). (Contrast this with the property of Hadamard code above.)

2.9 Bibliographic Notes

Finite fields are also called Galois fields (another common notation for $\mathbb{F}_q$ is $GF(q)$), named after Évariste Galois, whose worked laid the foundations of their theory. (Galois led an extremely short and interesting life, which ended in death from a duel.) For a more thorough treatment refer to any standard text on algebra or the book on finite fields by Lidl and Niederreiter [69].

The answer to Question 1.7.1 was proved by van Lint [101] and Tietavainen [100].

Chapter 3

Probability as Fancy Counting and the q-ary Entropy Function

In the first half of this chapter, we will develop techniques that will allow us to answer questions such as

Question 3.0.1. *Does there exist a* $[2,2,1]_2$ *code?*

We note that the answer to the above question is trivially yes: just pick the generator matrix to be the 2×2 identity matrix. However, we will use the above as a simple example to illustrate a powerful technique called the *probabilistic method*.

As the name suggests, the method uses probability. Before we talk more about the probabilistic method, we do a quick review of the basics of probability that we will need in this book.

3.1 A Crash Course on Probability

In this book, we will only consider probability distributions defined over finite spaces. In particular, given a finite domain $\mathbb{D}$, a probability distribution is defined as a function

$$p : \mathbb{D} \to [0,1] \text{ such that } \sum_{x \in \mathbb{D}} p(x) = 1,$$

where $[0,1]$ is shorthand for the interval of all real numbers between 0 and 1. In this book, we will primarily deal with the following special distribution:

Definition 3.1.1 (Uniform Distribution)**.** The *uniform distribution* over $\mathbb{D}$, denoted by $\mathscr{U}_{\mathbb{D}}$, is given by

$$\mathscr{U}_{\mathbb{D}}(x) = \frac{1}{|\mathbb{D}|} \text{ for every } x \in \mathbb{D}.$$

Typically we will drop the subscript when the domain $\mathbb{D}$ is clear from the context.

G	$\mathscr{U}(G)$	V_{00}	V_{01}	V_{10}	V_{11}
$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$	$\frac{1}{16}$	0	0	0	0
$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\frac{1}{16}$	0	1	0	1
$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$	$\frac{1}{16}$	0	1	0	1
$\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$	$\frac{1}{16}$	0	2	0	2
$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$	$\frac{1}{16}$	0	0	1	1
$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$	$\frac{1}{16}$	0	1	1	0
$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$	$\frac{1}{16}$	0	1	1	2
$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$	$\frac{1}{16}$	0	2	1	1

G	$\mathscr{U}(G)$	V_{00}	V_{01}	V_{10}	V_{11}
$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\frac{1}{16}$	0	0	1	1
$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\frac{1}{16}$	0	1	1	2
$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$	$\frac{1}{16}$	0	1	1	0
$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$	$\frac{1}{16}$	0	2	1	1
$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$	$\frac{1}{16}$	0	0	2	2
$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$	$\frac{1}{16}$	0	1	2	1
$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$	$\frac{1}{16}$	0	1	2	1
$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$	$\frac{1}{16}$	0	2	2	0

Table 3.1: Uniform distribution over $\mathbb{F}_2^{2\times 2}$ along with values of four random variables.

For example, consider the domain $\mathbb{D} = \mathbb{F}_2^{2\times 2}$, i.e. the set of all 2×2 matrices over $\mathbb{F}_2$. (Note that each such matrix is a generator matrix of some $[2,2]_2$ code.) The first two columns of Table 3.1 list the elements of this $\mathbb{D}$ along with the corresponding probabilities for the uniform distribution.

Typically, we will be interested in a real-valued function defined on $\mathbb{D}$ and how it behaves under a probability distribution defined over $\mathbb{D}$. This is captured by the notion of a random variable:

Definition 3.1.2 (Random Variable)**.** Let $\mathbb{D}$ be a finite domain and $I \subset \mathbb{R}$ be a finite[1] subset. Let p be a probability distribution defined over $\mathbb{D}$. A *random variable* is a function:

$$V : \mathbb{D} \to I.$$

The *expectation* of V is defined as

$$\mathbb{E}[V] = \sum_{x \in \mathbb{D}} p(x) \cdot V(x).$$

For example, given $(i, j) \in \{0,1\}^2$, let V_{ij} denote the random variable $V_{ij}(G) = wt\big((i, j) \cdot G\big)$, for any $G \in \mathbb{F}_2^{2\times 2}$. The last four columns of Table 3.1 list the values of these four random variables.

In this book, we will mainly consider binary random variables, i.e., with $I = \{0,1\}$. In particular, given a predicate or *event* E over $\mathbb{D}$, we will define its *indicator variable* to be a function

[1] In general, I need not be finite. However, for this book this definition suffices.

$\mathbb{1}_E : \mathbb{D} \to \{0,1\}$ such that for any $x \in \mathbb{D}$:

$$\mathbb{1}_E(x) = \begin{cases} 1 & \text{if } E(x) = \text{true} \\ 0 & \text{otherwise.} \end{cases}$$

For example,

$$\mathbb{1}_{V_{01}=0}\left(\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right) = 1 \text{ and } \mathbb{1}_{V_{01}=0}\left(\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}\right) = 0.$$

In most cases we will shorten this notation to $\mathbb{1}_{E(x)}$ or simply $\mathbb{1}_E$. Finally, sometimes we will abuse notation and use E instead of $\mathbb{1}_E$.

As a further use of indicator variables, consider the expectations of the four indicator variables:

$$\mathbb{E}\left[\mathbb{1}_{V_{00}=0}\right] = 16 \cdot \frac{1}{16} = 1.$$

$$\mathbb{E}\left[\mathbb{1}_{V_{01}=0}\right] = 4 \cdot \frac{1}{16} = \frac{1}{4}. \tag{3.1}$$

$$\mathbb{E}\left[\mathbb{1}_{V_{10}=0}\right] = 4 \cdot \frac{1}{16} = \frac{1}{4}. \tag{3.2}$$

$$\mathbb{E}\left[\mathbb{1}_{V_{11}=0}\right] = 4 \cdot \frac{1}{16} = \frac{1}{4}. \tag{3.3}$$

3.1.1 Some Useful Results

Before we proceed, we record a simple property of indicator variables that will be useful. (See Exercise 3.1.)

Lemma 3.1.1. *Let E be any event. Then*

$$\mathbb{E}[\mathbb{1}_E] = \Pr[E \textit{ is true}].$$

Next, we state a simple yet useful property of expectation of a sum of random variables:

Proposition 3.1.2 (Linearity of Expectation)**.** *Given random variables $V_1, \ldots, V_m$ defined over the same domain $\mathbb{D}$ and with the same probability distribution p, we have*

$$\mathbb{E}\left[\sum_{i=1}^{m} V_i\right] = \sum_{i=1}^{m} \mathbb{E}[V_i].$$

Proof. For notational convenience, define $V = V_1 + \cdots + V_m$. Thus, we have

$$\mathbb{E}[V] = \sum_{x \in \mathbb{D}} V(x) \cdot p(x) \tag{3.4}$$

$$= \sum_{x \in \mathbb{D}} \left(\sum_{i=1}^{m} V_i(x)\right) \cdot p(x) \tag{3.5}$$

$$= \sum_{i=1}^{m} \sum_{x \in \mathbb{D}} V_i(x) \cdot p(x) \tag{3.6}$$

$$= \sum_{i=1}^{m} \mathbb{E}[V_i]. \tag{3.7}$$

In the equalities above, (3.4) and (3.7) follow from the definition of expectation of a random variable. (3.5) follows from the definition of V and (3.6) follows by switching the order of the two summations. □

As an example, we have

$$\mathbb{E}\left[\mathbb{1}_{V_{01}=0} + \mathbb{1}_{V_{10}=0} + \mathbb{1}_{V_{11}=0}\right] = \frac{3}{4} \tag{3.8}$$

Frequently, we will need to deal with the probability of the union of events. We will use the following result to upper bound such probabilities:

Proposition 3.1.3 (Union Bound)**.** *Given m binary random variables $A_1, \ldots, A_m$, we have*

$$\Pr\left[\left(\bigvee_{i=1}^{m} A_i\right) = 1\right] \leq \sum_{i=1}^{m} \Pr\left[A_i = 1\right].$$

Proof. For every $i \in [m]$, define

$$S_i = \{x \in \mathbb{D} | A_i(x) = 1\}.$$

Then we have

$$\Pr\left[\left(\bigvee_{i=1}^{m} A_i\right) = 1\right] = \sum_{x \in \cup_{i=1}^{m} S_i} p(x) \tag{3.9}$$

$$\leq \sum_{i=1}^{m} \sum_{x \in S_i} p(x) \tag{3.10}$$

$$= \sum_{i=1}^{m} \Pr[A_i = 1]. \tag{3.11}$$

In the above, (3.9) and (3.11) follow from the definition of S_i. (3.10) follows from the fact that some of the $x \in \cup_i S_i$ get counted more than once. □

We remark that the union bound is tight when the events are *disjoint*. (In other words, using the notation in the proof above, when $S_i \cap S_j = \emptyset$ for every $i \neq j$.)

As an example, let $A_1 = \mathbb{1}_{V_{01}=0}$, $A_2 = \mathbb{1}_{V_{10}=0}$ and $A_3 = \mathbb{1}_{V_{11}=0}$. Note that in this case the event $A_1 \vee A_2 \vee A_3$ is the same as the event that there exists a non-zero $\mathbf{m} \in \{0,1\}^2$ such that $wt(\mathbf{m} \cdot G) = 0$. Thus, the union bound implies (that under the uniform distribution over $\mathbb{F}_2^{2\times 2}$)

$$\Pr\left[\text{There exists an } \mathbf{m} \in \{0,1\}^2 \setminus \{(0,0)\}, \text{ such that } wt(\mathbf{m}G) = 0\right] \leq \frac{3}{4}. \tag{3.12}$$

Finally, we present two bounds on the probability of a random variable deviating significantly from its expectation. The first bound holds for any random variable:

Lemma 3.1.4 (Markov Bound)**.** *Let V be a non-zero random variable. Then for any $t > 0$,*

$$\Pr[V \geq t] \leq \frac{\mathbb{E}[V]}{t}.$$

In particular, for any $a \geq 1$,

$$\Pr[V \geq a \cdot \mathbb{E}[V]] \leq \frac{1}{a}.$$

Proof. The second bound follows from the first bound by substituting $t = a \cdot \mathbb{E}[V]$. Thus, to complete the proof, we argue the first bound. Consider the following sequence of relations:

$$\mathbb{E}[V] = \sum_{i \in [0,t)} i \cdot \Pr[V = i] + \sum_{i \in [t,\infty)} i \cdot \Pr[V = i] \tag{3.13}$$

$$\geq \sum_{i \geq t} i \cdot \Pr[V = i] \tag{3.14}$$

$$\geq t \cdot \sum_{i \geq t} \Pr[V = i] \tag{3.15}$$

$$= t \cdot \Pr[V \geq t]. \tag{3.16}$$

In the above relations, (3.13) follows from the definition of expectation of a random variable and the fact that V is positive. (3.14) follows as we have dropped some non-negative terms. (3.15) follows by noting that in the summands $i \geq t$. (3.16) follows from the definition of $\Pr[V \geq t]$.

The proof is complete by noting that (3.16) implies the claimed bound. □

The second bound works only for sums of *independent* random variables. We begin by defining independent random variables:

Definition 3.1.3 (Independence)**.** Two random variables A and B are called *independent* if for every a and b in the ranges of A and B respectively, we have

$$\Pr[A = a \wedge B = b] = \Pr[A = a] \cdot \Pr[B = b].$$

For example, for the uniform distribution in Table 3.1, let A denote the bit $G_{0,0}$ and B denote the bit $G_{0,1}$. It can be verified that these two random variables are independent. In fact, it can be verified all the random variables corresponding to the four bits in G are independent random variables. (We'll come to a related comment shortly.)

Another related concept that we will use is that of probability of an event happening conditioned on another event happening:

Definition 3.1.4 (Conditional Probability)**.** Given two events A and B defined over the same domain and probability distribution, we define the probability of A *conditioned on* B as

$$\Pr[A|B] = \frac{\Pr[A \text{ and } B]}{\Pr[B]}.$$

For example, note that

$$\Pr[\mathbb{1}_{V_{01}=1} | G_{0,0} = 0] = \frac{4/16}{1/2} = \frac{1}{2}.$$

The above definition implies that two events A and B are independent if and only if $\Pr[A] = \Pr[A|B]$. We will also use the following result later on in the book (see Exercise 3.2):

Lemma 3.1.5. *For any two events A and B defined on the same domain and the probability distribution:*

$$\Pr[A] = \Pr[A|B] \cdot \Pr[B] + \Pr[A|\neg B] \cdot \Pr[\neg B].$$

Next, we state the deviation bound. (We only state it for sums of binary random variables, which is the form that will be needed in the book.)

Theorem 3.1.6 (Chernoff Bound)**.** *Let $X_1, \ldots, X_m$ be independent binary random variables and define $X = \sum X_i$. Then the* multiplicative *Chernoff bound sates that for* $0 < \varepsilon \leq 1$,

$$\Pr\left[|X - \mathbb{E}(X)| > \varepsilon \mathbb{E}(X)\right] < 2e^{-\varepsilon^2 \mathbb{E}(X)/3},$$

and the additive *Chernoff bound states that*

$$\Pr\left[|X - \mathbb{E}(X)| > \varepsilon m\right] < 2e^{-\varepsilon^2 m/2}.$$

We omit the proof, which can be found in any standard textbook on randomized algorithms.

Finally, we present an alternate view of uniform distribution over product spaces and then use that view to prove a result that we will use later in the book. Given probability distributions p_1 and p_2 over domains $\mathbb{D}_1$ and $\mathbb{D}_2$ respectively, we define the product distribution $p_1 \times p_2$ over $\mathbb{D}_1 \times \mathbb{D}_2$ as follows: every element $(x, y) \in \mathbb{D}_1 \times \mathbb{D}_2$ under $p_1 \times p_2$ is picked by choosing x from $\mathbb{D}_1$ according to p_1 and y is picked *independently* from $\mathbb{D}_2$ under p_2. This leads to the following observation (see Exercise 3.4).

Lemma 3.1.7. *For any $m \geq 1$, the distribution $\mathcal{U}_{\mathbb{D}_1 \times \mathbb{D}_2 \times \cdots \times \mathbb{D}_m}$ is identical*[2] *to the distribution $\mathcal{U}_{\mathbb{D}_1} \times \mathcal{U}_{\mathbb{D}_2} \times \cdots \times \mathcal{U}_{\mathbb{D}_m}$.*

For example, the uniform distribution in Table 3.1 can be described equivalently as follows: pick each of the four bits in G independently and uniformly at random from $\{0, 1\}$.

We conclude this section by proving the following result:

Lemma 3.1.8. *Given a non-zero vector $\mathbf{m} \in \mathbb{F}_q^k$ and a uniformly random $k \times n$ matrix G over $\mathbb{F}_q$, the vector $\mathbf{m} \cdot G$ is uniformly distributed over $\mathbb{F}_q^n$.*

Proof. Let the (j, i)th entry in G ($1 \leq j \leq k, 1 \leq i \leq n$) be denoted by g_{ji}. Note that as G is a random $k \times n$ matrix over $\mathbb{F}_q$, by Lemma 3.1.7, each of the g_{ji} is an independent uniformly random element from $\mathbb{F}_q$. Now, note that we would be done if we can show that for every $1 \leq i \leq n$, the ith entry in $\mathbf{m} \cdot G$ (call it b_i) is an independent uniformly random element from $\mathbb{F}_q$. To finish the proof, we prove this latter fact. If we denote $\mathbf{m} = (m_1, \ldots, m_k)$, then $b_i = \sum_{j=1}^k m_j g_{ji}$. Note that the disjoint entries of G participate in the sums for b_i and b_j for $i \neq j$. Given our choice of G, this implies that the random variables b_i and b_j are independent. Hence, to complete the proof we need to prove that b_i is a uniformly independent element of $\mathbb{F}_q$. The rest of the proof is a generalization of the argument we used in the proof of Proposition 2.7.1.

Note that to show that b_i is uniformly distributed over $\mathbb{F}_q$, it is sufficient to prove that b_i takes every value in $\mathbb{F}_q$ equally often over all the choices of values that can be assigned to $g_{1i}, g_{2i}, \ldots, g_{ki}$. Now, as $\mathbf{m}$ is non-zero, at least one of the its element is non-zero. Without loss of generality assume that $m_1 \neq 0$. Thus, we can write $b_i = m_1 g_{1i} + \sum_{j=2}^k m_j g_{ji}$. Now, for every fixed assignment of values to $g_{2i}, g_{3i}, \ldots, g_{ki}$ (note that there are q^{k-1} such assignments), b_i takes a different value for each of the q distinct possible assignments to g_{1i} (this is where we use the assumption that $m_1 \neq 0$). Thus, over all the possible assignments of $g_{1i}, \ldots, g_{ki}$, b_i takes each of the values in $\mathbb{F}_q$ exactly q^{k-1} times, which proves our claim. $\square$

[2] We say two distributions p_1 and p_2 on $\mathbb{D}$ are identical if for every $x \in \mathbb{D}$, $p_1(x) = p_2(x)$.

3.2 The Probabilistic Method

The *probabilistic method* is a very powerful method in combinatorics which can be used to show the existence of objects that satisfy certain properties. In this course, we will use the probabilistic method to prove existence of a code $\mathscr{C}$ with certain property $\mathscr{P}$. Towards that end, we define a distribution $\mathscr{D}$ over all possible codes and prove that when $\mathscr{C}$ is chosen according to $\mathscr{D}$:

$$\Pr\left[\mathscr{C} \text{ has property } \mathscr{P}\right] > 0 \text{ or equivalently } \Pr\left[\mathscr{C} \text{ doesn't have property } \mathscr{P}\right] < 1.$$

Note that the above inequality proves the existence of $\mathscr{C}$ with property $\mathscr{P}$.

As an example consider Question 3.0.1. To answer this in the affirmative, we note that the set of all $[2,2]_2$ linear codes is covered by the set of all 2×2 matrices over $\mathbb{F}_2$. Then, we let $\mathscr{D}$ be the uniform distribution over $\mathbb{F}_2^{2\times 2}$. Then by Proposition 2.3.4 and (3.12), we get that

$$\Pr_{\mathscr{U}_{\mathbb{F}_2^{2\times 2}}}[\text{There is no } [2,2,1]_2 \text{ code}] \le \frac{3}{4} < 1,$$

which by the probabilistic method answers the Question 3.0.1 in the affirmative.

For the more general case, when we apply the probabilistic method, the typical approach will be to define (sub-)properties $P_1, \ldots, P_m$ such that $\mathscr{P} = P_1 \wedge P_2 \wedge P_3 \ldots \wedge P_m$ and show that for every $1 \le i \le m$:

$$\Pr\left[\mathscr{C} \text{ doesn't have property } P_i\right] = \Pr\left[\overline{P_i}\right] < \frac{1}{m}.$$

Finally, by the union bound, the above will prove that[3] $\Pr\left[\mathscr{C} \text{ doesn't have property } \mathscr{P}\right] < 1$, as desired.

As an example, an alternate way to answer Question 3.0.1 in the affirmative is the following. Define $P_1 = \mathbb{1}_{V_{01}\ge 1}$, $P_2 = \mathbb{1}_{V_{10}\ge 1}$ and $P_3 = \mathbb{1}_{V_{11}\ge 1}$. (Note that we want a $[2,2]_2$ code that satisfies $P_1 \wedge P_2 \wedge P_3$.) Then, by (3.1), (3.2) and (3.3), we have for $i \in [3]$,

$$\Pr\left[\mathscr{C} \text{ doesn't have property } P_i\right] = \Pr\left[\overline{P_i}\right] = \frac{1}{4} < \frac{1}{3},$$

as desired.

Finally, we mention a special case of the general probabilistic method that we outlined above. In particular, let $\mathscr{P}$ denote the property that the randomly chosen $\mathscr{C}$ satisfies $f(\mathscr{C}) \le b$. Then we claim (see Exercise 3.5) that $\mathbb{E}[f(C)] \le b$ implies that $\Pr[\mathscr{C} \text{ has property } \mathscr{P}] > 0$. Note that this implies that $\mathbb{E}[f(C)] \le b$ implies that there exists a code $\mathscr{C}$ such that $f(C) \le b$.

3.3 The q-ary Entropy Function

We begin with the definition of a function that will play a central role in many of our combinatorial results.

[3] Note that $\overline{P} = \overline{P_1} \vee \overline{P_2} \vee \cdots \vee \overline{P_m}$.

Definition 3.3.1 (q-ary Entropy Function)**.** Let q be an integer and x be a real number such that $q \geq 2$ and $0 \leq x \leq 1$. Then the *q-ary entropy function* is defined as follows:

$$H_q(x) = x\log_q(q-1) - x\log_q(x) - (1-x)\log_q(1-x).$$

Figure 3.1 presents a pictorial representation of the H_q function for the first few values of q. For the special case of $q = 2$, we will drop the subscript from the entropy function and denote

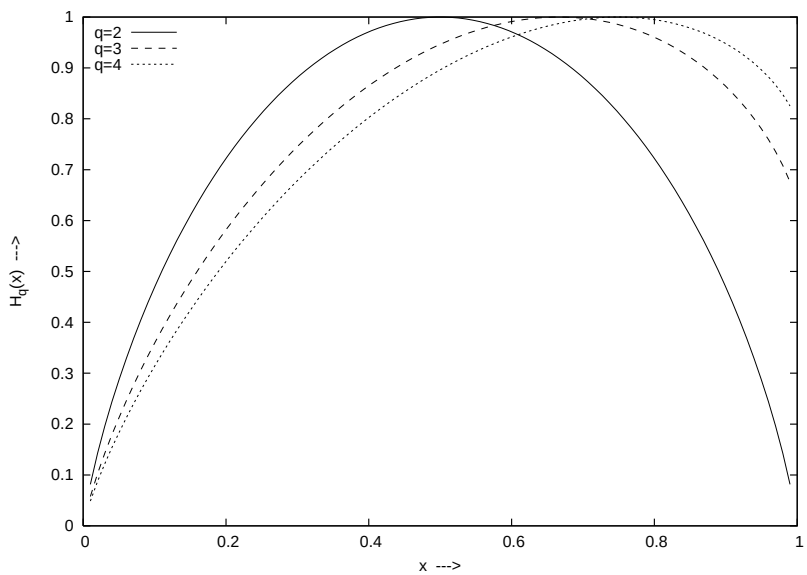


Figure 3.1: A plot of $H_q(x)$ for $q = 2, 3$ and 4. The maximum value of 1 is achieved at $x = 1 - 1/q$.

$H_2(x)$ by just $H(x)$, that is, $H(x) = -x\log x - (1-x)\log(1-x)$, where $\log x$ is defined as $\log_2(x)$ (we are going to follow this convention for the rest of the book).

Under the lens of Shannon's entropy function, $H(x)$ denotes the entropy of the distribution over $\{0,1\}$ that selects 1 with probability x and 0 with probability $1-x$. However, there is no similar analogue for the more general $H_q(x)$. The reason why this quantity will turn out to be so central in this book is that it is very closely related to the "volume" of a Hamming ball. We make this connection precise in the next subsection.

3.3.1 Volume of Hamming Balls

It turns out that in many of our combinatorial results, we will need good upper and lower bounds on the volume of a Hamming ball. Next we formalize the notion of the volume of a Hamming ball:

Definition 3.3.2 (Volume of a Hamming Ball)**.** Let $q \geq 2$ and $n \geq r \geq 1$ be integers. Then the volume of a Hamming ball of radius r is given by

$$Vol_q(r,n) = |B_q(\mathbf{0}, r)| = \sum_{i=0}^{r} \binom{n}{i} (q-1)^i.$$

The choice of $\mathbf{0}$ as the center for the Hamming ball above was arbitrary: since the volume of a Hamming ball is independent of its center (as is evident from the last equality above), we could have picked any point as the center.

We will prove the following result:

Proposition 3.3.1. *Let $q \geq 2$ be an integer and $0 \leq p \leq 1 - \frac{1}{q}$ be a real number. Then:*

(i) $Vol_q(pn, n) \leq q^{H_q(p)n}$*; and*

(ii) for large enough n, $Vol_q(pn, n) \geq q^{H_q(p)n - o(n)}$.

Proof. We start with the proof of *(i)*. Consider the following sequence of relations:

$$\begin{aligned}
1 &= (p + (1-p))^n \\
&= \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} && (3.17)\\
&= \sum_{i=0}^{pn} \binom{n}{i} p^i (1-p)^{n-i} + \sum_{i=pn+1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\
&\geq \sum_{i=0}^{pn} \binom{n}{i} p^i (1-p)^{n-i} && (3.18)\\
&= \sum_{i=0}^{pn} \binom{n}{i} (q-1)^i \left(\frac{p}{q-1}\right)^i (1-p)^{n-i} \\
&= \sum_{i=0}^{pn} \binom{n}{i} (q-1)^i (1-p)^n \left(\frac{p}{(q-1)(1-p)}\right)^i \\
&\geq \sum_{i=0}^{pn} \binom{n}{i} (q-1)^i (1-p)^n \left(\frac{p}{(q-1)(1-p)}\right)^{pn} && (3.19)\\
&= \sum_{i=0}^{pn} \binom{n}{i} (q-1)^i \left(\frac{p}{q-1}\right)^{pn} (1-p)^{(1-p)n} && (3.20)\\
&\geq Vol_q(pn, n) q^{-H_q(p)n}. && (3.21)
\end{aligned}$$

In the above, (3.17) follows from the binomial expansion. (3.18) follows by dropping the second sum and (3.19) follows from the facts that $\frac{p}{(q-1)(1-p)} \leq 1$ (as[4] $p \leq 1 - 1/q$). Rest of the steps except (3.21) follow from rearranging the terms. (3.21) follows as $q^{-H_q(p)n} = \left(\frac{p}{q-1}\right)^{pn} (1-p)^{(1-p)n}$.

(3.21) implies that

$$1 \geq Vol_q(pn, n) q^{-H_q(p)n},$$

which proves *(i)*.

[4]Indeed, note that $\frac{p}{(q-1)(1-p)} \leq 1$ is true if $\frac{p}{1-p} \leq \frac{q-1}{1}$, which in turn is true if $p \leq \frac{q-1}{q}$, where the last step follows from Lemma B.2.1.

We now turn to the proof of part *(ii)*. For this part, we will need Stirling's approximation for $n!$ (Lemma B.1.2).

By the Stirling's approximation, we have the following inequality:

$$\begin{aligned}\binom{n}{pn} &= \frac{n!}{(pn)!((1-p)n)!} \\ &> \frac{(n/e)^n}{(pn/e)^{pn}((1-p)n/e)^{(1-p)n}} \cdot \frac{1}{\sqrt{2\pi p(1-p)n}} \cdot e^{\lambda_1(n)-\lambda_2(pn)-\lambda_2((1-p)n)} \\ &= \frac{1}{p^{pn}(1-p)^{(1-p)n}} \cdot \ell(n), \end{aligned} \tag{3.22}$$

where $\ell(n) = \frac{e^{\lambda_1(n)-\lambda_2(pn)-\lambda_2((1-p)n)}}{\sqrt{2\pi p(1-p)n}}$.

Now consider the following sequence of relations that complete the proof:

$$Vol_q(pn, n) \geq \binom{n}{pn}(q-1)^{pn} \tag{3.23}$$

$$> \frac{(q-1)^{pn}}{p^{pn}(1-p)^{(1-p)n}} \cdot \ell(n) \tag{3.24}$$

$$\geq q^{H_q(p)n-o(n)}. \tag{3.25}$$

In the above (3.23) follows by only looking at one term. (3.24) follows from (3.22) while (3.25) follows from the definition of $H_q(\cdot)$ and the fact that for large enough n, $\ell(n)$ is $q^{-o(n)}$. □

Next, we consider how the q-ary entropy function behaves for various ranges of its parameters.

3.3.2 Other Properties of the q-ary Entropy function

We begin by recording the behavior of q-ary entropy function for large q.

Proposition 3.3.2. *For small enough ε, $1 - H_q(\rho) \geq 1 - \rho - \varepsilon$ for every $0 < \rho \leq 1 - 1/q$ if and only if q is $2^{\Omega(1/\varepsilon)}$.*

Proof. We first note that by definition of $H_q(\rho)$ and $H(\rho)$,

$$\begin{aligned} H_q(\rho) &= \rho \log_q(q-1) - \rho \log_q \rho - (1-\rho)\log_q(1-\rho) \\ &= \rho \log_q(q-1) + H(\rho)/\log_2 q. \end{aligned}$$

Now if $q \geq 2^{1/\varepsilon}$, we get that

$$H_q(\rho) \leq \rho + \varepsilon$$

as $\log_q(q-1) \leq 1$ and $H(\rho) \leq 1$. Thus, we have argued that for $q \geq 2^{1/\varepsilon}$, we have $1 - H_q(\rho) \geq 1 - \rho - \varepsilon$, as desired.

Next, we consider the case when $q = 2^{o(1/\varepsilon)}$. We begin by claiming that for small enough ε,

$$\text{if } q \geq 1/\varepsilon^2 \text{ then } \log_q(q-1) \geq 1-\varepsilon.$$

Indeed, $\log_q(q-1) = 1 + (1/\ln q)\ln(1-1/q) = 1 - O\left(\frac{1}{q \ln q}\right)$,[5] which is at least $1-\varepsilon$ for $q \geq 1/\varepsilon^2$ (and small enough ε).

Finally, if $q = 2^{o\left(\frac{1}{\varepsilon}\right)}$, then for fixed ρ,

$$H(\rho)/\log q = \varepsilon \cdot \omega(1).$$

Then for $q = 2^{o\left(\frac{1}{\varepsilon}\right)}$ (but $q \geq 1/\varepsilon^2$) we have

$$\rho \log_q(q-1) + H(\rho)/\log q \geq \rho - \varepsilon + \varepsilon \cdot \omega(1) > \rho + \varepsilon,$$

which implies that

$$1 - H_q(\rho) < 1 - \rho - \varepsilon,$$

as desired. For $q \leq 1/\varepsilon^2$, Lemma 3.3.3 shows that $1 - H_q(\rho) \leq 1 - H_{1/\varepsilon^2}(\rho) < 1 - \rho - \varepsilon$, as desired. □

We will also be interested in how $H_q(x)$ behaves for fixed x and increasing q:

Lemma 3.3.3. *Let $q \geq 2$ be an integer and let $0 \leq \rho \leq 1 - 1/q$, then for any real $m \geq 1$ such that*

$$q^{m-1} \geq \left(1 + \frac{1}{q-1}\right)^{q-1}, \tag{3.26}$$

we have

$$H_q(\rho) \geq H_{q^m}(\rho).$$

Proof. Note that $H_q(0) = H_{q^m}(0) = 0$. Thus, for the rest of the proof we will assume that $\rho \in (0, 1-1/q]$.

As observed in the proof of Proposition 3.3.2, we have

$$H_q(\rho) = \rho \cdot \frac{\log(q-1)}{\log q} + H(\rho) \cdot \frac{1}{\log q}.$$

Using this, we obtain

$$H_q(\rho) - H_{q^m}(\rho) = \rho\left(\frac{\log(q-1)}{\log q} - \frac{\log(q^m-1)}{m \log q}\right) + H(\rho)\left(\frac{1}{\log q} - \frac{1}{m \log q}\right).$$

The above in turn implies that

$$\frac{1}{\rho} \cdot m \log q \cdot (H_q(\rho) - H_{q^m}(\rho)) = \log(q-1)^m - \log(q^m-1) + \frac{H(\rho)}{\rho}(m-1)$$

[5] The last equality follows from the fact that by Lemma B.2.2, for $0 < x < 1$, $\ln(1-x) = -O(x)$.

$$\geq \log(q-1)^m - \log(q^m-1) + \frac{H(1-1/q)}{1-1/q}(m-1) \tag{3.27}$$

$$= \log(q-1)^m - \log(q^m-1) + (m-1)\left(\log\frac{q}{q-1} + \frac{\log q}{q-1}\right)$$

$$= \log\left(\frac{(q-1)^m}{q^m-1}\cdot\left(\frac{q}{q-1}\right)^{m-1}\cdot q^{\frac{m-1}{q-1}}\right)$$

$$= \log\left(\frac{(q-1)\cdot q^{m-1}\cdot q^{\frac{m-1}{q-1}}}{q^m-1}\right)$$

$$\geq 0 \tag{3.28}$$

In the above (3.27) follows from the fact that $H(\rho)/\rho$ is decreasing[6] in ρ and that $\rho \leq 1-1/q$. (3.28) follows from the claim that

$$(q-1)\cdot q^{\frac{m-1}{q-1}} \geq q.$$

Indeed the above follows from (3.26).

Finally, note that (3.28) completes the proof. □

Since $(1+1/x)^x \leq e$ (by Lemma B.2.5), we also have that (3.26) is also satisfied for $m \geq 1 + \frac{1}{\ln q}$. Further, we note that (3.26) is satisfied for every $m \geq 2$ (for any $q \geq 3$), which leads to the following (also see Exercise 3.6):

Corollary 3.3.4. *Let $q \geq 3$ be an integer and let $0 \leq \rho \leq 1-1/q$, then for any $m \geq 2$, we have*

$$H_q(\rho) \geq H_{q^m}(\rho).$$

Next, we look at the entropy function when its input is very close to 1.

Proposition 3.3.5. *For small enough $\varepsilon > 0$,*

$$H_q\left(1-\frac{1}{q}-\varepsilon\right) \leq 1 - c_q\varepsilon^2,$$

where c_q is a constant that only depends on q.

Proof. The intuition behind the proof is the following. Since the derivative of $H_q(x)$ is zero at $x = 1-1/q$, in the Taylor expansion of $H_q(1-1/q-\varepsilon)$ the ε term will vanish. We will now make this intuition more concrete. We will think of q as fixed and $1/\varepsilon$ as growing. In particular, we will assume that $\varepsilon < 1/q$. Consider the following equalities:

$$H_q(1-1/q-\varepsilon) = -\left(1-\frac{1}{q}-\varepsilon\right)\log_q\left(\frac{1-1/q-\varepsilon}{q-1}\right) - \left(\frac{1}{q}+\varepsilon\right)\log_q\left(\frac{1}{q}+\varepsilon\right)$$

[6] Indeed, $H(\rho)/\rho = \log(1/\rho) - (1/\rho - 1)\log(1-\rho)$. Note that the first term is deceasing in ρ. We claim that the second term is also decreasing in ρ– this e.g. follows from the observation that $-(1/\rho-1)\ln(1-\rho) = (1-\rho)(1+\rho/2! + \rho^2/3! + \cdots) = 1 - \rho/2 - \rho^2(1/2 - 1/3!) - \cdots$ is also decreasing in ρ.

$$
\begin{aligned}
&= -\log_q\left(\frac{1}{q}\left(1-\frac{\varepsilon q}{q-1}\right)\right)+\left(\frac{1}{q}+\varepsilon\right)\log_q\left(\frac{1-(\varepsilon q)/(q-1)}{1+\varepsilon q}\right)\\
&= 1-\frac{1}{\ln q}\left[\ln\left(1-\frac{\varepsilon q}{q-1}\right)-\left(\frac{1}{q}+\varepsilon\right)\ln\left(\frac{1-(\varepsilon q)/(q-1)}{1+\varepsilon q}\right)\right]\\
&= 1+o(\varepsilon^2)-\frac{1}{\ln q}\left[-\frac{\varepsilon q}{q-1}-\frac{\varepsilon^2 q^2}{2(q-1)^2}-\left(\frac{1}{q}+\varepsilon\right)\left(-\frac{\varepsilon q}{q-1}\right.\right.\\
&\qquad\left.\left.-\frac{\varepsilon^2 q^2}{2(q-1)^2}-\varepsilon q+\frac{\varepsilon^2 q^2}{2}\right)\right] \qquad (3.29)\\
&= 1+o(\varepsilon^2)-\frac{1}{\ln q}\left[-\frac{\varepsilon q}{q-1}-\frac{\varepsilon^2 q^2}{2(q-1)^2}\right.\\
&\qquad\left.-\left(\frac{1}{q}+\varepsilon\right)\left(-\frac{\varepsilon q^2}{q-1}+\frac{\varepsilon^2 q^3(q-2)}{2(q-1)^2}\right)\right]\\
&= 1+o(\varepsilon^2)-\frac{1}{\ln q}\left[-\frac{\varepsilon^2 q^2}{2(q-1)^2}+\frac{\varepsilon^2 q^2}{q-1}-\frac{\varepsilon^2 q^2(q-2)}{2(q-1)^2}\right] \qquad (3.30)\\
&= 1-\frac{\varepsilon^2 q^2}{2\ln q(q-1)}+o(\varepsilon^2)\\
&\le 1-\frac{\varepsilon^2 q^2}{4\ln q(q-1)}
\end{aligned}
$$

(3.31)

(3.29) follows from the fact that for $|x|<1$, $\ln(1+x)=x-x^2/2+x^3/3-\dots$ (Lemma B.2.2) and by collecting the ε^3 and smaller terms in $o(\varepsilon^2)$. (3.30) follows by rearranging the terms and by absorbing the ε^3 terms in $o(\varepsilon^2)$. The last step is true assuming ε is small enough. □

Next, we look at the entropy function when its input is very close to 0.

Proposition 3.3.6. *For small enough $\varepsilon>0$,*

$$H_q(\varepsilon)=\Theta\left(\frac{1}{\log q}\cdot\varepsilon\log\left(\frac{1}{\varepsilon}\right)\right).$$

Proof. By definition

$$H_q(\varepsilon)=\varepsilon\log_q(q-1)+\varepsilon\log_q(1/\varepsilon)+(1-\varepsilon)\log_q(1/(1-\varepsilon)).$$

Since all the terms in the RHS are positive we have

$$H_q(\varepsilon)\ge\varepsilon\log(1/\varepsilon)/\log q. \qquad (3.32)$$

Further, by Lemma B.2.2, $(1-\varepsilon)\log_q(1/(1-\varepsilon))\le 2\varepsilon/\ln q$ for small enough ε. Thus, this implies that

$$H_q(\varepsilon)\le\frac{2+\ln(q-1)}{\ln q}\cdot\varepsilon+\frac{1}{\ln q}\cdot\varepsilon\ln\left(\frac{1}{\varepsilon}\right). \qquad (3.33)$$

(3.32) and (3.33) proves the claimed bound.

□

We will also work with the inverse of the q-ary entropy function. Note that $H_q(\cdot)$ on the domain $[0, 1-1/q]$ is a bijective map into $[0,1]$. Thus, we define $H_q^{-1}(y) = x$ such that $H_q(x) = y$ and $0 \le x \le 1-1/q$. Finally, we will need the following lower bound:

Lemma 3.3.7. *For every $0 \le y \le 1-1/q$ and for every small enough $\varepsilon > 0$,*

$$H_q^{-1}(y - \varepsilon^2/c'_q) \ge H_q^{-1}(y) - \varepsilon,$$

where $c'_q \ge 1$ is a constant that depends only on q.

Proof. It is easy to check that $H_q^{-1}(y)$ is a strictly increasing convex function when $y \in [0,1]$. This implies that the derivative of $H_q^{-1}(y)$ increases with y. In particular, $(H_q^{-1})'(1) \ge (H_q^{-1})'(y)$ for every $0 \le y \le 1$. In other words, for every $0 < y \le 1$, and (small enough) $\delta > 0$, $\frac{H_q^{-1}(y) - H_q^{-1}(y-\delta)}{\delta} \le \frac{H_q^{-1}(1) - H_q^{-1}(1-\delta)}{\delta}$. Proposition 3.3.5 along with the facts that $H_q^{-1}(1) = 1-1/q$ and H_q^{-1} is increasing completes the proof if one picks $c'_q = \max(1, 1/c_q)$ and $\delta = \varepsilon^2/c'_q$. □

3.4 Exercises

Exercise 3.1. Prove Lemma 3.1.1.

Exercise 3.2. Prove Lemma 3.1.5.

Exercise 3.3. In this exercise, we will see a common use of the Chernoff bound (Theorem 3.1.6). Say we are trying to determine an (unknown) value $x \in \mathbb{F}$ to which we have access to via a randomized algorithm $\mathscr{A}$ that on input (random) input $\mathbf{r} \in \{0,1\}^m$ outputs an estimate $\mathscr{A}(\mathbf{r})$ of x such that

$$\Pr_{\mathbf{r}}[\mathbf{A}(\mathbf{r}) = x] \ge \frac{1}{2} + \gamma,$$

for some $0 < \gamma < \frac{1}{2}$. Then show that for any $t \ge 1$ with $O\left(\frac{t}{\gamma^2}\right)$ calls to $\mathscr{A}$ one can determine x with probability at least $1 - e^{-t}$.

Hint: Call $\mathscr{A}$ with independent random bits and take majority of the answer and then use the Chernoff bound.

Exercise 3.4. Prove Lemma 3.1.7.

Exercise 3.5. Let $\mathscr{P}$ denote the property that the randomly chosen $\mathscr{C}$ satisfies $f(\mathscr{C}) \le b$. Then $\mathbb{E}[f(C)] \le b$ implies that $\Pr[\mathscr{C}$ has property $\mathscr{P}] > 0$.

Exercise 3.6. Show that for any $Q \ge q \ge 2$ and $\rho \le 1-1/q$, we have $H_Q(\rho) \le H_q(\rho)$.

3.5 Bibliographic Notes

Shannon was one of the very early adopters of probabilistic method (and we will see one such use in Chapter 6). Later, the probabilistic method was popularized Erdős. For more on probabilistic method, see the book by Alon and Spencer [3].

Proofs of various concentration bounds can e.g. be found in [23].

Part II

The Combinatorics

Chapter 4

What Can and Cannot Be Done-I

In this chapter, we will try to tackle Question 2.5.1. We will approach this trade-off in the following way:

> If we fix the relative distance of the code to be δ, what is the best rate R that we can achieve?

Note that an upper bound on R is a *negative* result, while a lower bound on R is a *positive* result.

In this chapter, we will consider only one positive result, i.e. a lower bound on R called the Gilbert-Varshamov bound in Section 4.2. In Section 4.1, we recall a negative result that we have already seen– Hamming bound and state its asymptotic version to obtain an upper bound on R. We will consider two other upper bounds: the Singleton bound (Section 4.3), which gives a tight upper bound for large enough alphabets (but not binary codes) and the Plotkin bound (Section 4.4).

4.1 Asymptotic Version of the Hamming Bound

We have already seen an upper bound in Section 1.7 due to Hamming. However, we had stated this as an upper bound on the dimension k in terms of n, q and d. We begin by considering the trade-off between R and δ given by the Hamming bound. Recall that Theorem 1.7.1 states the following:

$$\frac{k}{n} \leq 1 - \frac{\log_q Vol_q\left(\left\lfloor \frac{d-1}{2} \right\rfloor, n\right)}{n}$$

Recall that Proposition 3.3.1 states following lower bound on the volume of a Hamming ball:

$$Vol_q\left(\left\lfloor \frac{d-1}{2} \right\rfloor, n\right) \geq q^{H_q\left(\frac{\delta}{2}\right)n - o(n)},$$

which implies the following asymptotic version of the Hamming bound:

$$R \leq 1 - H_q\left(\frac{\delta}{2}\right) + o(1).$$

See Figure 4.1 for a pictorial description of the Hamming bound for binary codes.

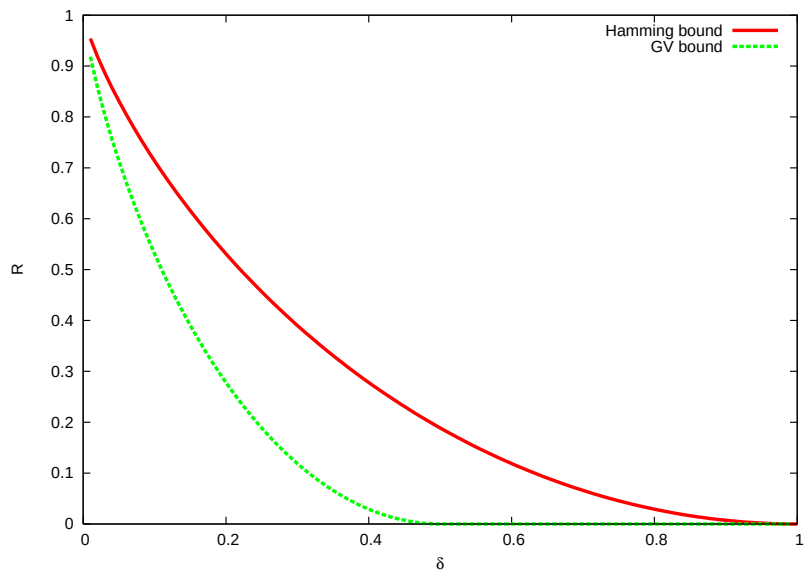


Figure 4.1: The Hamming and GV bounds for binary codes. Note that any point below the GV bound is achievable by some code while no point above the Hamming bound is achievable by any code. In this part of the book we would like to push the GV bound as much up as possible while at the same time try and push down the Hamming bound as much as possible.

4.2 Gilbert-Varshamov Bound

Next, we will switch gears by proving our first non-trivial lower bound on R in terms of δ. (In fact, this is the only positive result on the R vs δ tradeoff question that we will see in this book.) In particular, we will prove the following result:

Theorem 4.2.1 (Gilbert-Varshamov Bound)**.** *Let $q \geq 2$. For every $0 \leq \delta < 1 - \frac{1}{q}$, and $0 < \varepsilon \leq 1 - H_q(\delta)$, there exists a code with rate $R \geq 1 - H_q(\delta) - \varepsilon$ and relative distance δ.*

The bound is generally referred to as the GV bound. For a pictorial description of the GV bound for binary codes, see Figure 4.1. We will present the proofs for general codes and linear codes in Sections 4.2.1 and 4.2.2 respectively.

4.2.1 Greedy Construction

We will prove Theorem 4.2.1 for general codes by the following greedy construction (where $d = \delta n$): start with the empty code C and then keep on adding vectors not in C that are at Hamming distance at least d from all the existing codewords in C. Algorithm 6 presents a formal description of the algorithm and Figure 4.2 illustrates the first few executions of this algorithm.

We claim that Algorithm 6 terminates and the C that it outputs has distance d. The latter is true by step 2, which makes sure that in Step 3 we never add a vector $\mathbf{c}$ that will make the distance of C fall below d. For the former claim, note that, if we cannot add $\mathbf{v}$ at some point, we

Algorithm 6 Gilbert's Greedy Code Construction

INPUT: n, q, d

OUTPUT: A code $C \subseteq [q]^n$ of distance $d \geq\geq 11$

1: $C \leftarrow \emptyset$
2: WHILE there exists a $\mathbf{v} \in [q]^n$ such that $\Delta(\mathbf{v}, \mathbf{c}) \geq d$ for every $\mathbf{c} \in C$ DO
3: Add $\mathbf{v}$ to C
4: RETURN C

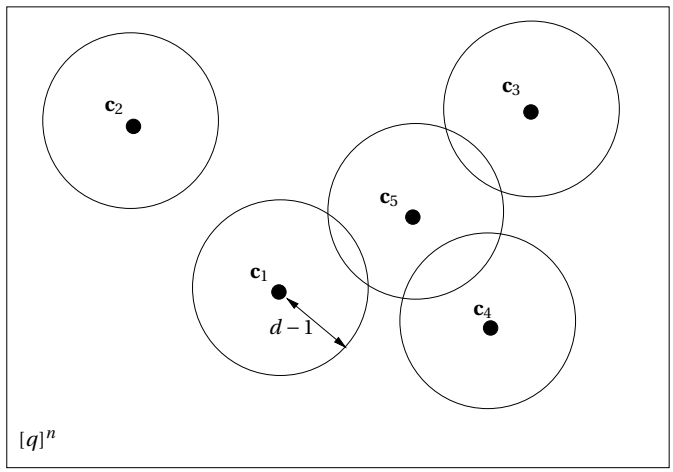


Figure 4.2: An illustration of Gilbert's greedy algorithm (Algorithm 6) for the first five iterations.

cannot add it later. Indeed, since we only add vectors to C, if a vector $\mathbf{v} \in [q]^n$ is ruled out in a certain iteration of Step 2 because $\Delta(\mathbf{c}, \mathbf{v}) < d$, then in all future iterations, we have $\Delta(\mathbf{v}, \mathbf{c}) < d$ and thus, this $\mathbf{v}$ will never be added in Step 3 in any future iteration.

The running time of Algorithm 6 is $q^{O(n)}$. To see this, note that Step 2 in the worst-case could be repeated for every vector in $[q]^n$, that is at most q^n times. In a naive implementation, for each iteration, we cycle through all vectors in $[q]^n$ and for each vector $\mathbf{v} \in [q]^n$, iterate through all (at most q^n) vectors $\mathbf{c} \in C$ to check whether $\Delta(\mathbf{c}, \mathbf{v}) < d$. If no such $\mathbf{c}$ exists, then we add $\mathbf{v}$ to C. Otherwise, we move to the next $\mathbf{v}$. However, note that we can do slightly better– since we know that once a $\mathbf{v}$ is "rejected" in an iteration, it'll keep on being rejected in the future iterations, we can fix up an ordering of vectors in $[q]^n$ and for each vector $\mathbf{v}$ in this order, check whether it can be added to C or not. If so, we add $\mathbf{v}$ to C, else we move to the next vector in the order. This algorithm has time complexity $O(nq^{2n})$, which is still $q^{O(n)}$.

Further, we claim that after termination of Algorithm 6

$$\bigcup_{\mathbf{c} \in C} B(\mathbf{c}, d-1) = [q]^n.$$

This is because if the above is not true, then there exists a vector $\mathbf{v} \in [q]^n \setminus C$, such that $\Delta(\mathbf{v}, \mathbf{c}) \geq d$ and hence $\mathbf{v}$ can be added to C. However, this contradicts the fact that Algorithm 6 has termi-

nated. Therefore,

$$\left|\bigcup_{\mathbf{c}\in C} B(\mathbf{c}, d-1)\right| = q^n. \tag{4.1}$$

It is not too hard to see that

$$\sum_{\mathbf{c}\in C} |B(\mathbf{c}, d-1)| \geq \left|\bigcup_{\mathbf{c}\in C} B(\mathbf{c}, d-1)\right|,$$

which by (4.1) implies that

$$\sum_{\mathbf{c}\in C} |B(\mathbf{c}, d-1)| \geq q^n$$

or since the volume of a Hamming ball is translation invariant,

$$\sum_{\mathbf{c}\in C} Vol_q(d-1, n) \geq q^n.$$

Since $\sum_{\mathbf{c}\in C} Vol_q(d-1, n) = Vol_q(d-1, n) \cdot |C|$, we have

$$\begin{aligned} |C| &\geq \frac{q^n}{Vol_q(d-1, n)} \\ &\geq \frac{q^n}{q^{nH_q(\delta)}} \\ &= q^{n(1-H_q(\delta))}, \end{aligned} \tag{4.2}$$

as desired. In the above, (4.2) follows from the fact that

$$\begin{aligned} Vol_q(d-1, n) &\leq Vol_q(\delta n, n) \\ &\leq q^{nH_q(\delta)}, \end{aligned} \tag{4.3}$$

where the second inequality follows from the upper bound on the volume of a Hamming ball in Proposition 3.3.1.

It is worth noting that the code from Algorithm 6 is not guaranteed to have any special structure. In particular, even storing the code can take exponential space. We have seen in Proposition 2.3.1 that linear codes have a much more succinct representation. Thus, a natural question is:

Question 4.2.1. *Do linear codes achieve the $R \geq 1 - H_q(\delta)$ tradeoff that the greedy construction achieves?*

Next, we will answer the question in the affirmative.

4.2.2 Linear Code Construction

Now we will show that a random linear code, with high probability, lies on the GV bound. The construction is a use of the probabilistic method (Section 3.2).

By Proposition 2.3.4, we are done if we can show that there exists a $k \times n$ matrix $\mathbf{G}$ of full rank (for $k = (1 - H_q(\delta) - \varepsilon)n$) such that

$$\text{For every } \mathbf{m} \in \mathbb{F}_q^k \setminus \{\mathbf{0}\}, wt(\mathbf{mG}) \geq d.$$

We will prove the existence of such a $\mathbf{G}$ by the probabilistic method. Pick a random linear code by picking a random $k \times n$ matrix $\mathbf{G}$ where each of kn entries is chosen uniformly and independently at random from $\mathbb{F}_q$. Fix $\mathbf{m} \in \mathbb{F}_q^k \setminus \{\mathbf{0}\}$. Recall that by Lemma 3.1.8, for a random $\mathbf{G}$, $\mathbf{mG}$ is a uniformly random vector from $\mathbb{F}_q^n$. Thus, we have

$$\begin{aligned} Pr[wt(\mathbf{mG}) < d] &= \frac{Vol_q(d-1,n)}{q^n} \\ &\leq \frac{q^{nH_q(\delta)}}{q^n}, \end{aligned} \tag{4.4}$$

where (4.4) follows from (4.3). Thus, by the union bound (Lemma 3.1.3)

$$\begin{aligned} Pr[\text{There exists a non-zero } \mathbf{m}, wt(\mathbf{mG}) < d] &\leq q^k q^{-n(1-H_q(\delta))} \\ &= q^{-\varepsilon \cdot n}, \end{aligned}$$

where the equality follows by choosing $k = (1 - H_q(\delta) - \varepsilon)n$. Since $q^{-\varepsilon n} \ll 1$, by the probabilistic method, there exists a linear code C with relative distance δ.

All that's left is to argue that the code C has dimension at least $k = (1 - H_q(\delta) - \varepsilon)n$. To show this, we need to show that the chosen generator matrix $\mathbf{G}$ has full rank. Note that there is a non-zero probability that a uniformly matrix $\mathbf{G}$ does not have full rank. There are two ways to deal with this. First, we can show that with high probability a random $\mathbf{G}$ does have full rank, so that $|C| = q^k$. Further, the proof above has already shown that, with high probability, the distance is greater than zero, which implies that distinct messages will be mapped to distinct codewords and thus $|C| = q^k$. In other words, C does indeed have dimension k, as desired

Discussion. We now digress a bit to discuss some consequences of the proofs of the GV bound.

We first note that the probabilistic method proof shows something stronger than Theorem 4.2.1: *most* linear codes (with appropriate parameters) meet the Gilbert-Varshamov bound.

Note that we can also pick a random linear code by picking a random $(n-k) \times n$ parity check matrix. This also leads to a proof of the GV bound: see Exercise 4.1.

Finally, we note that Theorem 4.2.1 requires $\delta < 1 - \frac{1}{q}$. An inspection of Gilbert and Varshamov's proofs shows that the only reason the proof required that $\delta \leq 1 - \frac{1}{q}$ is because it is needed for the volume bound (recall the bound in Proposition 3.3.1)– $Vol_q(\delta n, n) \leq q^{H_q(\delta)n}$– to hold. It is natural to wonder if the above is just an artifact of the proof or, for example:

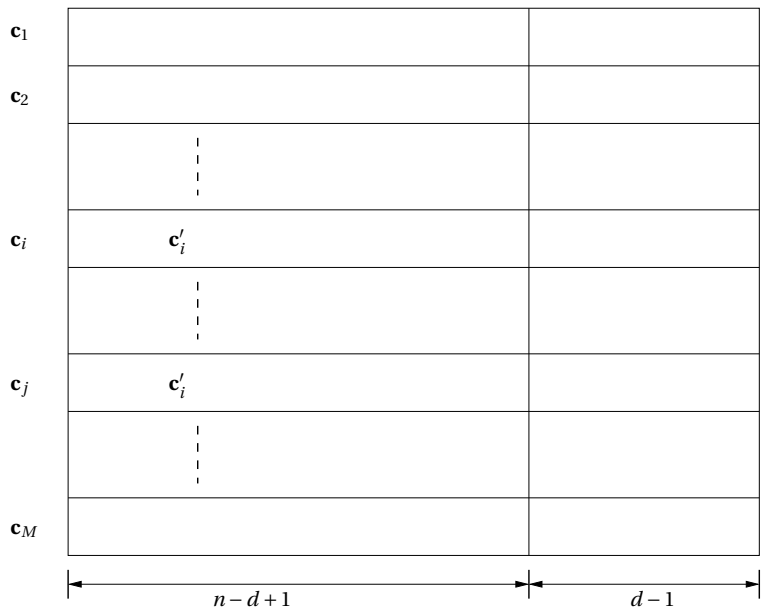


Figure 4.3: Construction of a new code in the proof of the Singleton bound.

Question 4.2.2. *Does there exists a code with $R > 0$ and $\delta > 1 - \frac{1}{q}$?*

We will return to this question in Section 4.4.

4.3 Singleton Bound

We will now change gears again and prove an upper bound on R (for fixed δ). We start by proving the Singleton bound.

Theorem 4.3.1 (Singleton Bound)**.** *For every $(n,k,d)_q$ code,*

$$k \le n - d + 1.$$

Proof. Let $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M$ be the codewords of an $(n,k,d)_q$ code C. Note that we need to show $M \le q^{n-d+1}$. To this end, we define $\mathbf{c}'_i$ to be the prefix of the codeword $\mathbf{c}_i$ of length $n-d+1$ for every $i \in [M]$. See Figure 4.3 for a pictorial description.

We now claim that for every $i \ne j$, $\mathbf{c}'_i \ne \mathbf{c}'_j$. For the sake of contradiction, assume that there exits an $i \ne j$ such that $\mathbf{c}'_i = \mathbf{c}'_j$. Notice this implies that $\mathbf{c}_i$ and $\mathbf{c}_j$ agree in all the first $n-d+1$ positions, which in turn implies that $\Delta(\mathbf{c}_i, \mathbf{c}_j) \le d-1$. This contradicts the fact that C has distance d. Thus, M is the number of prefixes of codewords in C of length $n-d+1$, which implies that $M \le q^{n-d+1}$ as desired. □

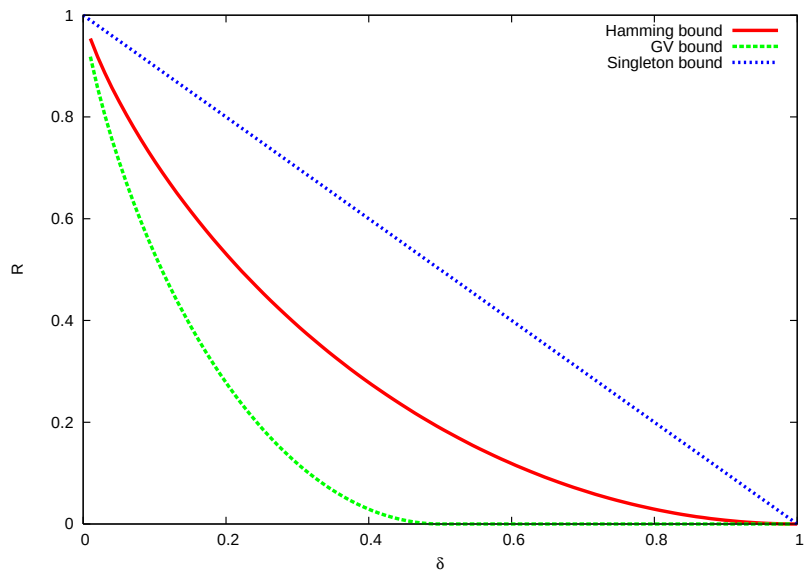


Figure 4.4: The Hamming, GV and Singleton bound for binary codes.

Note that the asymptotic version of the Singleton bound states that $k/n \leq 1 - d/n + 1/n$. In other words,

$$R \leq 1 - \delta + o(1).$$

Figure 4.4 presents a pictorial description of the asymptotic version of the Singleton bound. It is worth noting that the bound is *independent* of the alphabet size. As is evident from Figure 4.4, the Singleton bound is worse than the Hamming bound for binary codes. However, this bound is better for larger alphabet sizes. In fact, we will look at a family of codes called Reed-Solomon codes in Chapter 5 that meets the Singleton bound. However, the alphabet size of the Reed-Solomon codes increases with the block length n. Thus, a natural follow-up question is the following:

Question 4.3.1. *Given a fixed $q \geq 2$, does there exist a q-ary code that meets the Singleton bound?*

We'll see an answer to this question in the next section.

4.4 Plotkin Bound

In this section, we will study the Plotkin bound, which will answer Questions 4.2.2 and 4.3.1. We start by stating the bound.

Theorem 4.4.1 (Plotkin bound)**.** *The following holds for any code $C \subseteq [q]^n$ with distance d:*

1. *If $d = \left(1 - \frac{1}{q}\right) n$, $|C| \leq 2qn$.*
2. *If $d > \left(1 - \frac{1}{q}\right) n$, $|C| \leq \frac{qd}{qd-(q-1)n}$.*

Note that the Plotkin bound (Theorem 4.4.1) implies that a code with relative distance $\delta \geq 1 - \frac{1}{q}$, must necessarily have $R = 0$, which answers Question 4.2.2 in the negative.

Before we prove Theorem 4.4.1, we make a few remarks. We first note that the upper bound in the first part of Theorem 4.4.1 can be improved to $2n$ for $q = 2$. (See Exercise 4.12.) Second, it can be shown that this bound is tight. (See Exercise 4.13.) Third, the statement of Theorem 4.4.1 gives a trade-off only for relative distance greater than $1 - 1/q$. However, as the following corollary shows, the result can be extended to work for $0 \leq \delta \leq 1 - 1/q$. (See Figure 4.5 for an illustration for binary codes.)

Corollary 4.4.2. *For any q-ary code with relative distance $0 \leq \delta \leq 1 - \frac{1}{q}$,*

$$R \leq 1 - \left(\frac{q}{q-1}\right)\delta + o(1).$$

Proof. Define $d = \delta n$. The proof proceeds by shortening the codewords. We group the codewords so that they agree on the first $n - n'$ symbols, where $n' = \left\lfloor \frac{qd}{q-1} \right\rfloor - 1$. (We will see later why this choice of n' makes sense.) In particular, for any $\mathbf{x} \in [q]^{n-n'}$, define the 'prefix code'

$$C_{\mathbf{x}} = \{(c_{n-n'+1}, \ldots c_n) \mid (c_1 \ldots c_N) \in C, (c_1 \ldots c_{n-n'}) = \mathbf{x}\}.$$

For all $\mathbf{x}$, $C_{\mathbf{x}}$ has distance d as C has distance d.[1] Additionally, it has block length $n' < \left(\frac{q}{q-1}\right) d$ and thus, $d > \left(1 - \frac{1}{q}\right) n'$. By Theorem 4.4.1, this implies that

$$|C_{\mathbf{x}}| \leq \frac{qd}{qd - (q-1)n'} \leq qd, \tag{4.5}$$

where the second inequality follows from the fact that $qd - (q-1)n'$ is an integer.

Note that by the definition of $C_{\mathbf{x}}$:

$$|C| = \sum_{\mathbf{x} \in [q]^{n-n'}} |C_{\mathbf{x}}|,$$

[1] If for some $\mathbf{x}$, $\mathbf{c}_1 \neq \mathbf{c}_2 \in C_{\mathbf{x}}$, $\Delta(\mathbf{c}_1, \mathbf{c}_2) < d$, then $\Delta((\mathbf{x}, \mathbf{c}_1), (\mathbf{x}, \mathbf{c}_2)) < d$, which implies that the distance of C is less than d (as by definition of $C_{\mathbf{x}}$, both $(\mathbf{x}, \mathbf{c}_1), (\mathbf{x}, \mathbf{c}_2) \in C$), which in turn is a contradiction.

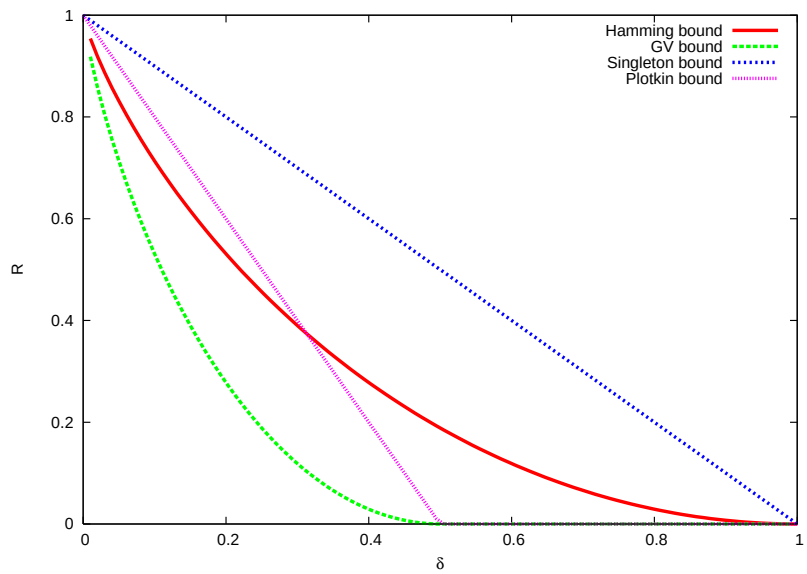


Figure 4.5: The current bounds on the rate R vs. relative distance δ for binary codes. The GV bound is a lower bound on R while the other three bounds are upper bounds on R.

which by (4.5) implies that

$$|C| \leq \sum_{\mathbf{x} \in [q]^{n-n'}} qd = q^{n-n'} \cdot qd \leq q^{n - \frac{q}{q-1} d + o(n)} = q^{n\left(1 - \delta \cdot \frac{q}{q-1} + o(1)\right)}.$$

In other words, $R \leq 1 - \left(\frac{q}{q-1}\right)\delta + o(1)$ as desired. □

Note that Corollary 4.4.2 implies that for any q-ary code of rate R and relative distance δ (where q is a *constant* independent of the block length of the code), $R < 1 - \delta$. In other words, this answers Question 4.3.1 in the negative.

Let us pause for a bit at this point and recollect the bounds on R versus δ that we have proved till now. Figure 4.5 depicts all the bounds we have seen till now (for $q = 2$). The GV bound is the best known lower bound at the time of writing of this book. Better upper bounds are known and we will see one such trade-off (called the Elias-Bassalygo bound) in Section 8.1.

Now, we turn to the proof of Theorem 4.4.1, for which we will need two more lemmas. The first lemma deals with vectors over real spaces. We quickly recap the necessary definitions. Consider a vector $\mathbf{v}$ in $\mathbb{R}^n$, that is, a tuple of n real numbers. This vector has (Euclidean) norm $\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}$, and is a unit vector if and only if its norm is 1. The inner product of two vectors, $\mathbf{u}$ and $\mathbf{v}$, is $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_i u_i \cdot v_i$. The following lemma gives a bound on the number of vectors that can exist such that every pair is at an obtuse angle with each other.

Lemma 4.4.3 (Geometric Lemma)**.** *Let* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \in \mathbb{R}^N$ *be non-zero vectors.*

1. *If* $\langle \mathbf{v}_i, \mathbf{v}_j \rangle \le 0$ *for all* $i \ne j$*, then* $m \le 2N$.

2. *Let* $\mathbf{v}_i$ *be unit vectors for* $1 \le i \le m$*. Further, if* $\langle \mathbf{v}_i, \mathbf{v}_j \rangle \le -\varepsilon < 0$ *for all* $i \ne j$*, then* $m \le 1 + \frac{1}{\varepsilon}$.[2]

(Item 1 is tight: see Exercise 4.14.) The proof of the Plotkin bound will need the existence of a map from codewords to real vectors with certain properties, which the next lemma guarantees.

Lemma 4.4.4 (Mapping Lemma)**.** *Let* $C \subseteq [q]^n$*. Then there exists a function* $f : C \longrightarrow \mathbb{R}^{nq}$ *such that*

1. *For every* $\mathbf{c} \in C$, $\|f(\mathbf{c})\| = 1$.

2. *For every* $\mathbf{c}_1 \ne \mathbf{c}_2$ *such that* $\mathbf{c}_1, \mathbf{c}_2 \in C$, $\langle f(\mathbf{c}_1), f(\mathbf{c}_2) \rangle = 1 - \left(\frac{q}{q-1}\right)\left(\frac{\Delta(\mathbf{c}_1,\mathbf{c}_2)}{n}\right)$.

We defer the proofs of the lemmas above to the end of the section. We are now in a position to prove Theorem 4.4.1.

Proof of Theorem 4.4.1 Let $C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m\}$. For all $i \ne j$,

$$\left\langle f(\mathbf{c}_i), f(\mathbf{c}_j) \right\rangle \le 1 - \left(\frac{q}{q-1}\right)\frac{\Delta(\mathbf{c}_i, \mathbf{c}_j)}{n} \le 1 - \left(\frac{q}{q-1}\right)\frac{d}{n}.$$

The first inequality holds by Lemma 4.4.4, and the second holds as C has distance d.

For part 1, if $d = \left(1 - \frac{1}{q}\right)n = \frac{(q-1)n}{q}$, then for all $i \ne j$,

$$\left\langle f(\mathbf{c}_i), f(\mathbf{c}_j) \right\rangle \le 0$$

and so by the first part of Lemma 4.4.3, $m \le 2nq$, as desired.

For part 2, $d > \left(\frac{q-1}{q}\right)n$ and so for all $i \ne j$,

$$\langle f(\mathbf{c}_i), f(\mathbf{c}_j) \rangle \le 1 - \left(\frac{q}{q-1}\right)\frac{d}{n} = -\left(\frac{qd - (q-1)n}{(q-1)n}\right)$$

and, since $\varepsilon \stackrel{\text{def}}{=} \left(\frac{qd-(q-1)n}{(q-1)n}\right) > 0$, we can apply the second part of Lemma 4.4.3. Thus, $m \le 1 + \frac{(q-1)n}{qd-(q-1)n} = \frac{qd}{qd-(q-1)n}$, as desired □

4.4.1 Proof of Geometric and Mapping Lemmas

Next, we prove Lemma 4.4.3.

[2] Note that since $\mathbf{v}_i$ and $\mathbf{v}_j$ are both unit vectors, $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ is the cosine of the angle between them.

Proof of Lemma 4.4.3. We begin with a proof of the first result. The proof is by induction on n. Note that in the base case of $N = 0$, we have $m = 0$, which satisfies the claimed inequality $m \leq 2N$.

In the general case, we have $m \geq 1$ non-zero vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^N$ such that for every $i \neq j$,

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle \leq 0. \tag{4.6}$$

Since rotating all the vectors by the same amount does not change the sign of the inner product (nor does scaling any of the vectors), w.l.o.g. we can assume that $\mathbf{v}_m = \langle 1, 0, \ldots, 0 \rangle$. For $1 \leq i \leq m-1$, denote the vectors as $\mathbf{v}_i = \langle \alpha_i, \mathbf{y}_i \rangle$, for some $\alpha_i \in \mathbb{R}$ and $\mathbf{y}_i \in \mathbb{R}^{N-1}$. Now, for any $i \neq 1$, $\langle \mathbf{v}_1, \mathbf{v}_i \rangle = 1 \cdot \alpha_i + \sum_{i=2}^{m} 0 = \alpha_i$. However, note that (4.6) implies that $\langle \mathbf{v}_1, \mathbf{v}_i \rangle \leq 0$, which in turn implies that

$$\alpha_i \leq 0. \tag{4.7}$$

Next, we claim that at most one of $\mathbf{y}_1, \ldots, \mathbf{y}_{m-1}$ can be the all zeroes vector, $\mathbf{0}$. If not, assume w.l.o.g., that $\mathbf{y}_1 = \mathbf{y}_2 = \mathbf{0}$. This in turn implies that

$$\begin{aligned}
\langle \mathbf{v}_1, \mathbf{v}_2 \rangle &= \alpha_1 \cdot \alpha_2 + \langle \mathbf{y}_1, \mathbf{y}_2 \rangle \\
&= \alpha_1 \cdot \alpha_2 + 0 \\
&= \alpha_1 \cdot \alpha_2 \\
&> 0,
\end{aligned}$$

where the last inequality follows from the subsequent argument. As $\mathbf{v}_1 = \langle \alpha_1, \mathbf{0} \rangle$ and $\mathbf{v}_2 = \langle \alpha_2, \mathbf{0} \rangle$ are non-zero, this implies that $\alpha_1, \alpha_2 \neq 0$. (4.7) then implies that $\alpha_1, \alpha_2 < 0$. However, $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle > 0$ contradicts (4.6).

Thus, w.l.o.g., assume that $\mathbf{v}_1, \ldots, \mathbf{v}_{m-2}$ are all non-zero vectors. Further, note that for every $i \neq j \in [m-2]$, $\langle \mathbf{y}_i, \mathbf{y}_j \rangle = \langle \mathbf{v}_i, \mathbf{v}_j \rangle - \alpha_i \cdot \alpha_j \leq \langle \mathbf{v}_i, \mathbf{v}_j \rangle \leq 0$. Thus, we have reduced problem on m vectors with dimension N to an equivalent problem on $m-2$ vectors with dimension dimension $N-1$. If we continue this process, we can conclude that every loss in dimension of the vector results in twice in loss in the numbers of the vectors in the set. Induction then implies that $m \leq 2N$, as desired.

We now move on to the proof of the second part. Define $\mathbf{z} = \mathbf{v}_1 + \ldots + \mathbf{v}_m$. Now consider the following sequence of relationships:

$$\|\mathbf{z}\|^2 = \sum_{i=1}^{m} \|\mathbf{v}_i\|^2 + 2 \sum_{i<j} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \leq m + 2 \cdot \binom{m}{2} \cdot (-\varepsilon) = m(1 - \varepsilon m + \varepsilon).$$

The inequality follows from the facts that each $\mathbf{v}_i$ is a unit vector and the assumption that for every $i \neq j$, $\langle \mathbf{v}_i . \mathbf{v}_j \rangle \leq -\varepsilon$. As $\|\mathbf{z}\|^2 \geq 0$,

$$m(1 - \varepsilon m + \varepsilon) \geq 0.$$

Since $m \geq 1$, we have that

$$1 - \varepsilon m + \varepsilon \geq 0$$

or

$$\varepsilon m \leq 1 + \varepsilon.$$

Thus, we have $m \leq 1 + \frac{1}{\varepsilon}$, as desired. □

Finally, we prove Lemma 4.4.4.

Proof of Lemma 4.4.4. We begin by defining a map $\phi : [q] \to \mathbb{R}^q$ with certain properties. Then we apply ϕ to all the coordinates of a codeword to define the map $f : \mathbb{R}^q \to \mathbb{R}^{nq}$ that satisfies the claimed properties. We now fill in the details.

Define $\phi : [q] \to \mathbb{R}^q$ as follows. For every $i \in [q]$, we define

$$\phi(i) = \left\langle \frac{1}{q}, \frac{1}{q}, \ldots, \underbrace{\frac{-(q-1)}{q}}_{i^{\text{th}}\text{position}}, \ldots \frac{1}{q} \right\rangle.$$

That is, all but the i'th position in $\phi(i) \in \mathbb{R}^q$ has a value of $1/q$ and the ith position has value $-(q-1)/q$.

Next, we record two properties of ϕ that follow immediately from its definition. For every $i \in [q]$,

$$\phi(i)^2 = \frac{(q-1)}{q^2} + \frac{(q-1)^2}{q^2} = \frac{(q-1)}{q}. \tag{4.8}$$

Also for every $i \neq j \in [q]$,

$$\langle \phi(i), \phi(j) \rangle = \frac{(q-2)}{q^2} - \frac{2(q-1)}{q^2} = -\frac{1}{q}. \tag{4.9}$$

We are now ready to define our final map $f : C \to \mathbb{R}^{nq}$. For every $\mathbf{c} = (c_1, \ldots, c_n) \in C$, define

$$f(\mathbf{c}) = \sqrt{\frac{q}{n(q-1)}} \cdot \big(\phi(c_1), \phi(c_2), \ldots, \phi(c_n)\big).$$

(The multiplicative factor $\sqrt{\frac{q}{n(q-1)}}$ is to ensure that $f(\mathbf{c})$ for any $\mathbf{c} \in C$ is a unit vector.)

To complete the proof, we will show that f satisfies the claimed properties. We begin with condition 1. Note that

$$\|f(\mathbf{c})\|^2 = \frac{q}{(q-1)n} \cdot \sum_{i=1}^{n} \left|\phi(i)\right|^2 = 1,$$

where the first equality follows from the definition of f and the second equality follows from (4.8).

We now turn to the second condition. For notational convenience, define $\mathbf{c}_1 = (x_1, \ldots, x_n)$ and $\mathbf{c}_2 = (y_1, \ldots, y_n)$. Consider the following sequence of relations:

$$\big\langle f(\mathbf{c}_1), f(\mathbf{c}_2) \big\rangle = \sum_{\ell=1}^{n} \big\langle f(x_\ell), f(y_\ell) \big\rangle$$

$$
\begin{aligned}
&= \left[\sum_{\ell : x_\ell \neq y_\ell} \langle \phi(x_\ell), \phi(y_\ell) \rangle + \sum_{\ell : x_\ell = y_\ell} \langle \phi(x_\ell), \phi(y_\ell) \rangle \right] \cdot \left(\frac{q}{n(q-1)} \right) \\
&= \left[\sum_{\ell : x_\ell \neq y_\ell} \left(\frac{-1}{q} \right) + \sum_{\ell : x_\ell = y_\ell} \left(\frac{q-1}{q} \right) \right] \cdot \left(\frac{q}{n(q-1)} \right) && (4.10) \\
&= \left[\Delta(\mathbf{c}_1, \mathbf{c}_2) \left(\frac{-1}{q} \right) + (n - \Delta(\mathbf{c}_1, \mathbf{c}_2)) \left(\frac{q-1}{q} \right) \right] \cdot \left(\frac{q}{n(q-1)} \right) && (4.11) \\
&= 1 - \Delta(\mathbf{c}_1, \mathbf{c}_2) \left(\frac{q}{n(q-1)} \right) \left[\frac{1}{q} + \frac{q-1}{q} \right] \\
&= 1 - \left(\frac{q}{q-1} \right) \left(\frac{\Delta(\mathbf{c}_1, \mathbf{c}_2)}{n} \right),
\end{aligned}
$$

as desired. In the above, (4.10) is obtained using (4.9) and (4.8) while (4.11) follows from the definition of the Hamming distance. □

4.5 Exercises

Exercise 4.1. Pick a $(n-k) \times n$ matrix H over $\mathbb{F}_q$ at random. Show that with high probability the code whose parity check matrix is H achieves the GV bound.

Exercise 4.2. Recall the definition of an ε-biased space from Exercise 2.14. Show that there exists an ε-biased space of size $O(k/\varepsilon^2)$.

Hint: Recall part 1 of Exercise 2.14.

Exercise 4.3. Argue that a random linear code as well as its dual both lie on the corresponding GV bound.

Exercise 4.4. In Section 4.2.2, we saw that random *linear* code meets the GV bound. It is natural to ask the question for general random codes. (By a random $(n,k)_q$ code, we mean the following: for each of the q^k messages, pick a random vector from $[q]^n$. Further, the choices for each codeword is independent.) We will do so in this problem.

1. Prove that a random q-ary code with rate $R > 0$ with high probability has relative distance $\delta \geq H_q^{-1}(1 - 2R - \varepsilon)$. Note that this is worse than the bound for random linear codes in Theorem 4.2.1.

2. Prove that with high probability the relative distance of a random q-ary code of rate R is at most $H_q^{-1}(1-2R) + \varepsilon$. In other words, general random codes are worse than random linear codes in terms of their distance.

   Hint: Use Chebyshev's inequality.

Exercise 4.5. We saw that Algorithm 6 can compute an $(n,k)_q$ code on the GV bound in time $q^{O(n)}$. Now the construction for linear codes is a randomized construction and it is natural to ask how quickly can we compute an $[n,k]_q$ code that meets the GV bound. In this problem,

we will see that this can also be done in $q^{O(n)}$ deterministic time, though the deterministic algorithm is not that straight-forward anymore.

1. Argue that Theorem 4.2.1 gives a $q^{O(kn)}$ time algorithm that constructs an $[n,k]_q$ code on the GV bound. (Thus, the goal of this problem is to "shave" off a factor of k from the exponent.)

2. A $k \times n$ *Toeplitz Matrix* $A = \{A_{i,j}\}_{i=1,\ j=1}^{k\ ,\ n}$ satisfies the property that $A_{i,j} = A_{i-1,j-1}$. In other words, any diagonal has the same value. For example, the following is a 4×6 Toeplitz matrix:
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 1 & 2 & 3 & 4 & 5 \\ 8 & 7 & 1 & 2 & 3 & 4 \\ 9 & 8 & 7 & 1 & 2 & 3 \end{pmatrix}$$
A *random* $k \times n$ Toeplitz matrix $T \in \mathbb{F}_q^{k \times n}$ is chosen by picking the entries in the first row and column uniformly (and independently) at random.

   Prove the following claim: For any non-zero $\mathbf{m} \in \mathbb{F}_q^k$, the vector $\mathbf{m} \cdot T$ is uniformly distributed over $\mathbb{F}_q^n$, that is for every $\mathbf{y} \in \mathbb{F}_q^n$, $\Pr\left[\mathbf{m} \cdot T = \mathbf{y}\right] = q^{-n}$.

3. Briefly argue why the claim in part 2 implies that a random code defined by picking its generator matrix as a random Toeplitz matrix with high probability lies on the GV bound.

4. Conclude that an $[n,k]_q$ code on the GV bound can be constructed in time $q^{O(k+n)}$.

Exercise 4.6. Show that one can construct the parity check matrix of an $[n,k]_q$ code that lies on the GV bound in time $q^{O(n)}$.

Exercise 4.7. So far in Exercises 4.5 and 4.6, we have seen two constructions of $[n,k]_q$ code on the GV bound that can be constructed in $q^{O(n)}$ time. For constant rate codes, at the time of writing of this book, this is fastest known construction of any code that meets the GV bound. For $k = o(n)$, there is a better construction known, which we explore in this exercise.

We begin with some notation. For the rest of the exercise we will target a distance of $d = \delta n$. Given a message $\mathbf{m} \in \mathbb{F}_q^k$ and an $[n,k]_q$ code C, define the indicator variable:
$$W_{\mathbf{m}}(C) = \begin{cases} 1 & \text{if } wt(C(\mathbf{m})) < d \\ 0 & \text{otherwise.} \end{cases}$$
Further, define
$$D(C) = \sum_{\mathbf{m} \in \mathbb{F}_q^k \setminus \{\mathbf{0}\}} W_{\mathbf{m}}(C).$$
We will also use $D(G)$ and $W_{\mathbf{m}}(G)$ to denote the variables above for the code C generated by G.

Given an $k \times n$ matrix M, we will use M^i to denote the ith column of M and $M^{\leq i}$ to denote the column submatrix of M that contains the first i columns. Finally below we will use $\mathcal{G}$ to denote a uniformly random $k \times n$ generator matrix and G to denote a specific instantiation of the generator matrix. We will arrive at the final construction in a sequence of steps. In what follows define $k < (1 - H_q(\delta))n$ for large enough n.

1. Argue that C has a distance d if and only if $D(C) < 1$.

2. Argue that $\mathbb{E}[D(\mathscr{G})] < 1$.

3. Argue that for any $1 \le i < n$ and fixed $k \times n$ matrix G,

$$\min_{\mathbf{v} \in \mathbb{F}_q^k} \mathbb{E}\left[D(\mathscr{G}) | \mathscr{G}^{\le i} = G^{\le i}, \mathscr{G}^{i+1} = \mathbf{v}\right] \le \mathbb{E}\left[D(\mathscr{G}) | \mathscr{G}^{\le i} = G^{\le i}\right].$$

4. We are now ready to define the algorithm to compute the final generator matrix G: see Algorithm 7. Prove that Algorithm 7 outputs a matrix G such that the linear code generated

Algorithm 7 $q^{O(k)}$ time algorithm to compute a code on the GV bound

INPUT: Integer parameters $1 \le k \ne n$ such that $k < (1 - H_q(\delta)n)$
OUTPUT: An $k \times n$ generator matrix G for a code with distance δn

1: Initialize G to be the all 0s matrix ▷ This initialization is arbitrary
2: FOR every $1 \le i \le n$ DO
3: $\quad G^i \leftarrow \arg\min_{\mathbf{v} \in \mathbb{F}_q^k} \mathbb{E}\left[D(\mathscr{G}) | \mathscr{G}^{\le i} = G^{\le i}, \mathscr{G}^{i+1} = \mathbf{v}\right]$
4: RETURN G

by G is an $[n, k, \delta n]_q$ code. Conclude that this code lies on the GV bound.

5. Finally, we will analyze the run time of Algorithm 7. Argue that Step 2 can be implemented in $\text{poly}(n, q^k)$ time. Conclude Algorithm 7 can be implemented in time $\text{poly}(n, q^k)$.

Hint: It might be useful to maintain a data structure that keeps track of one number for every non-zero $\mathbf{m} \in \mathbb{F}_q^k$ throughout the run of Algorithm 7.

Exercise 4.8. In this problem we will derive the GV bound using a graph-theoretic proof, which is actually equivalent to the greedy proof we saw in Section 4.2.1. Let $1 \le d \le n$ and $q \ge 1$ be integers. Now consider the graph $G_{n,d,q} = (V, E)$, where the vertex set is the set of all vectors in $[q]^n$. Given two vertices $\mathbf{u} \ne \mathbf{v} \in [q]^n$, we have the edge $(u, v) \in E$ if and only if $\Delta(\mathbf{u}, \mathbf{v}) < d$. An *independent set* of a graph $G = (V, E)$ is a subset $I \subseteq V$ such that for every $u \ne v \in I$, we have that (u, v) is *not* an edge. We now consider the following sub-problems:

1. Argue that any independent set C of $G_{n,d,q}$ is a q-ary code of distance d.

2. The *degree* of a vertex in a graph G is the number of edges incident on that vertex. Let Δ be the maximum degree of any vertex in $G = (V, E)$.Then argue that G has an independent set of size at least $\frac{|V|}{\Delta+1}$.

3. Using parts 1 and 2 argue the GV bound.

Exercise 4.9. In this problem we will improve slightly on the GV bound using a more sophisticated graph-theoretic proof. Let $G_{n,d,q}$ and N and Δ be as in the previous exercise (Exercise 4.8). So far we used the fact that $G_{n,d,q}$ has many vertices and small degree to prove it has a large independent set, and thus to prove there is a large code of minimum distance d. In this exercise we will see how a better result can be obtained by counting the number of "triangles" in the graph. A triangle in a graph $G = (V, E)$ is a set $\{u, v, w\} \subset V$ of three vertices such that all three vertices are adjancent, i.e., $(u, v), (v, w), (w, u) \in E$. For simplicity we will focus on the case where $q = 2$ and $d = n/5$, and consider the limit as $n \to \infty$.

1. Prove that a graph on N vertices of maximum degree Δ has at most $O(N\Delta^2)$ triangles.

2. Prove that the number of triangle in graph $G_{n,d,2}$ is at most

$$2^n \cdot \sum_{0 \le e \le 3d/2} \binom{n}{e} \cdot 3^e.$$

   Hint: Fix u and let e count the number of coordinates where at least one of v or w disagree with u. Prove that e is at most $3d/2$.

3. Simplify the expression in the case where $d = n/5$ to show that the number of triangles in $G_{n,n/5,2}$ is $O(N \cdot \Delta^{2-\eta})$ for some $\eta > 0$.

4. A famous result in the "probabilistic method" shows (and you don't have to prove this), that if a graph on N vertices of maximum degree Δ has at most $O(N \cdot \Delta^{2-\eta})$ triangles, then it has an independent set of size $\Omega(\frac{N}{\Delta} \log \Delta)$. Use this result to conclude that there is a binary code of block length n and distance $n/5$ of size $\Omega(n2^n/\binom{n}{n/5})$. (Note that this improves over the GV-bound by an $\Omega(n)$ factor.)

Exercise 4.10. Use part 2 from Exercise 1.7 to prove the Singleton bound.

Exercise 4.11. Let C be an $(n, k, d)_q$ code. Then prove that fixing any $n - d + 1$ positions uniquely determines the corresponding codeword.

Exercise 4.12. Let C be a binary code of block length n and distance $n/2$. Then $|C| \le 2n$. (Note that this is a factor 2 better than part 1 in Theorem 4.4.1.)

Exercise 4.13. Prove that the bound in Exercise 4.12 is tight– i.e. there exists binary codes C with block length n and distance $n/2$ such that $|C| = 2n$.

Exercise 4.14. Prove that part 1 of Lemma 4.4.3 is tight.

Exercise 4.15. In this exercise we will prove the Plotkin bound (at least part 2 of Theorem 4.4.1) via a purely combinatorial proof.

Given an $(n, k, d)_q$ code C with $d > \left(1 - \frac{1}{q}\right) n$ define

$$S = \sum_{\mathbf{c}_1 \neq \mathbf{c}_2 \in C} \Delta(\mathbf{c}_1, \mathbf{c}_2).$$

For the rest of the problem think of C has an $|C| \times n$ matrix where each row corresponds to a codeword in C. Now consider the following:

1. Looking at the contribution of each column in the matrix above, argue that

$$S \leq \left(1 - \frac{1}{q}\right) \cdot n|C|^2.$$

2. Look at the contribution of the rows in the matrix above, argue that

$$S \geq |C|\,(|C| - 1) \cdot d.$$

3. Conclude part 2 of Theorem 4.4.1.

Exercise 4.16. In this exercise, we will prove the so called *Griesmer Bound*. For any $[n, k, d]_q$, prove that

$$n \geq \sum_{i=0}^{k-1} \left\lceil \frac{d}{q^i} \right\rceil.$$

Hint: Recall Exercise 2.17.

Exercise 4.17. Use Exercise 4.16 to prove part 2 of Theorem 4.4.1 for linear codes.

Exercise 4.18. Use Exercise 4.16 to prove Theorem 4.3.1 for linear code.

4.6 Bibliographic Notes

Theorem 4.2.1 was proved for general codes by Edgar Gilbert ([35]) and for linear codes by Rom Varshamov ([102]). Hence, the bound is called the Gilbert-Varshamov bound. The Singleton bound (Theorem 4.3.1) is due to Richard C. Singleton [92]. For larger (but still constant) values of q, better lower bounds than the GV bound are known. In particular, for any prime power $q \geq 49$, there exist linear codes, called *algebraic geometric* (or AG) codes that outperform the corresponding GV bound[3]. AG codes out of the scope of this book. One starting point could be the following [56].

The proof method illustrated in Exercise 4.15has a name– *double counting*: in this specific case this follows since we count S in two different ways.

[3]The lower bound of 49 comes about as AG codes are only defined for q being a square (i.e. $q = (q')^2$) and it turns out that $q' = 7$ is the smallest value where AG bound beats the GV bound.

Chapter 5

The Greatest Code of Them All: Reed-Solomon Codes

In this chapter, we will study the Reed-Solomon codes. Reed-Solomon codes have been studied a lot in coding theory. These codes are optimal in the sense that they meet the Singleton bound (Theorem 4.3.1). We would like to emphasize that these codes meet the Singleton bound not just asymptotically in terms of rate and relative distance but also in terms of the dimension, block length and distance. As if this were not enough, Reed-Solomon codes turn out to be more versatile: they have many applications outside of coding theory. (We will see some applications later in the book.)

These codes are defined in terms of univariate polynomials (i.e. polynomials in one unknown/variable) with coefficients from a finite field $\mathbb{F}_q$. It turns out that polynomials over $\mathbb{F}_p$, for prime p, also help us define finite fields $\mathbb{F}_{p^s}$, for $s > 1$. To kill two birds with one stone[1], we first do a quick review of polynomials over finite fields. Then we will define and study some properties of Reed-Solomon codes.

5.1 Polynomials and Finite Fields

We begin with the formal definition of a (univariate) polynomial.

Definition 5.1.1. Let $\mathbb{F}_q$ be a finite field with q elements. Then a function $F(X) = \sum_{i=0}^{\infty} f_i X^i, f_i \in \mathbb{F}_q$ is called a polynomial.

For our purposes, we will only consider the finite case; that is, $F(X) = \sum_{i=0}^{d} f_i X^i$ for some integer $d > 0$, with coefficients $f_i \in \mathbb{F}_q$, and $f_d \neq 0$. For example, $2X^3 + X^2 + 5X + 6$ is a polynomial over $\mathbb{F}_7$.

Next, we define some useful notions related to polynomials. We begin with the notion of degree of a polynomial.

[1] No birds will be harmed in this exercise.

Definition 5.1.2. For $F(X) = \sum_{i=0}^{d} f_i X^i$ with $f_d \neq 0$, we call d the *degree* of $F(X)$. We denote the degree of the polynomial $F(X)$ by $\deg(F)$.

For example, $2X^3 + X^2 + 5X + 6$ has degree 3.

Let $\mathbb{F}_q[X]$ be the set of polynomials over $\mathbb{F}_q$, that is, with coefficients from $\mathbb{F}_q$. Let $F(X), G(X) \in \mathbb{F}_q[X]$ be polynomials. Then $\mathbb{F}_q[X]$ has the following natural operations defined on it:

Addition:

$$F(X) + G(X) = \sum_{i=0}^{\max(\deg(F),\deg(G))} (f_i + g_i) X^i,$$

where the addition on the coefficients is done over $\mathbb{F}_q$. For example, over $\mathbb{F}_2$, $X + (1 + X) = X \cdot (1 + 1) + 1 \cdot (0 + 1) = 1$ (recall that over $\mathbb{F}_2$, $1 + 1 = 0$).[2]

Multiplication:

$$F(X) \cdot G(X) = \sum_{i=0}^{\deg(F)+\deg(G)} \left(\sum_{j=0}^{\min(i,\deg(F))} p_j \cdot q_{i-j} \right) X^i,$$

where all the operations on the coefficients are over $\mathbb{F}_q$. For example, over $\mathbb{F}_2$, $X(1 + X) = X + X^2$; $(1 + X)^2 = 1 + 2X + X^2 = 1 + X^2$, where the latter equality follows since $2 \equiv 0 \bmod 2$.

Next, we define a root of a polynomial.

Definition 5.1.3. $\alpha \in \mathbb{F}_q$ is a root of a polynomial $F(X)$, if $F(\alpha) = 0$.

For instance, 1 is a root of $1 + X^2$ over $\mathbb{F}_2$.

We will also need the notion of a special class of polynomials, which are analogous to how prime numbers are special for natural numbers.

Definition 5.1.4. A polynomial $F(X)$ is irreducible if for every $G_1(X), G_2(X)$ such that $F(X) = G_1(X)G_2(X)$, we have $\min(\deg(G_1), \deg(G_2)) = 0$

For example, $1 + X^2$ is not irreducible over $\mathbb{F}_2$, as $(1 + X)(1 + X) = 1 + X^2$. However, $1 + X + X^2$ is irreducible, since its non-trivial factors have to be from the linear terms X or $X + 1$. However, it is easy to check that neither is a factor of $1 + X + X^2$. (In fact, one can show that $1 + X + X^2$ is the only irreducible polynomial of degree 2 over $\mathbb{F}_2$– see Exercise 5.1.) A word of caution: if a polynomial $E(X) \in \mathbb{F}_q[X]$ has no root in $\mathbb{F}_q$, it does *not* mean that $E(X)$ is irreducible. For example consider the polynomial $(1 + X + X^2)^2$ over $\mathbb{F}_2$– it does not have any root in $\mathbb{F}_2$ but it obviously is not irreducible.

Just as the set of integers modulo a prime is a field, so is the set of polynomials modulo an irreducible polynomial:

Theorem 5.1.1. *Let $E(X)$ be an irreducible polynomial with degree at least 2 over $\mathbb{F}_p$, p prime. Then the set of polynomials in $\mathbb{F}_p[X]$ modulo $E(X)$, denoted by $\mathbb{F}_p[X]/E(X)$, is a field.*

[2] This will be a good time to remember that operations over a finite field are much different from operations over integers/reals. For example, over reals/integers $X + (X + 1) = 2X + 1$.

The proof of the theorem above is similar to the proof of Lemma 2.1.2, so we only sketch the proof here. In particular, we will explicitly state the basic tenets of $\mathbb{F}_p[X]/E(X)$.

- Elements are polynomials in $\mathbb{F}_p[X]$ of degree at most $s-1$. Note that there are p^s such polynomials.
- Addition: $(F(X)+G(X)) \mod E(X) = F(X) \mod E(X) + G(X) \mod E(X) = F(X)+G(X)$. (Since $F(X)$ and $G(X)$ are of degree at most $s-1$, addition modulo $E(X)$ is just plain polynomial addition.)
- Multiplication: $(F(X)\cdot G(X)) \mod E(X)$ is the unique polynomial $R(X)$ with degree at most $s-1$ such that for some $A(X)$, $R(X)+A(X)E(X) = F(X)\cdot G(X)$
- The additive identity is the zero polynomial, and the additive inverse of any element $F(X)$ is $-F(X)$.
- The multiplicative identity is the constant polynomial 1. It can be shown that for every element $F(X)$, there exists a unique multiplicative inverse $(F(X))^{-1}$.

For example, for $p=2$ and $E(X) = 1+X+X^2$, $\mathbb{F}_2[X]/(1+X+X^2)$ has as its elements $\{0,1,X,1+X\}$. The additive inverse of any element in $\mathbb{F}_2[X]/(1+X+X^2)$ is the element itself while the multiplicative inverses of $1, X$ and $1+X$ in $\mathbb{F}_2[X]/(1+X+X^2)$ are $1, 1+X$ and X respectively.

A natural question to ask is if irreducible polynomials exist for every degree. Indeed, they do:

Theorem 5.1.2. *For all $s \geq 2$ and $\mathbb{F}_p$, there exists an irreducible polynomial of degree s over $\mathbb{F}_p$. In fact, the number of such irreducible polynomials is $\Theta\left(\frac{p^s}{s}\right)$.*

The result is true even for general finite fields $\mathbb{F}_q$ and not just prime fields but we stated the version over prime fields for simplicity. Given any monic [3] polynomial $E(X)$ of degree s, it can be verified whether it is an irreducible polynomial by checking if $\gcd(E(X), X^{q^s} - X) = E(X)$. This is true as the product of all monic irreducible polynomials in $\mathbb{F}_q[X]$ of degree exactly s is known to be the polynomial $X^{q^s} - X$. Since Euclid's algorithm for computing the $\gcd(F(X), G(X))$ can be implemented in time polynomial in the minimum of $\deg(F)$ and $\deg(G)$ and $\log q$ (see Section D.7.2), this implies that checking whether a given polynomial of degree s over $\mathbb{F}_q[X]$ is irreducible can be done in time $\mathrm{poly}(s, \log q)$.

This implies an efficient Las Vegas algorithm[4] to generate an irreducible polynomial of degree s over $\mathbb{F}_q$. Note that the algorithm is to keep on generating random polynomials until it comes across an irreducible polynomial (Theorem 5.1.2 implies that the algorithm will check $O(p^s)$ polynomials in expectation). Algorithm 8 presents the formal algorithm.

The above discussion implies the following:

[3] I.e. the coefficient of the highest degree term is 1. It is easy to check that if $E(X) = e_s X^s + e_{s-1}X^{s-1} + \cdots + 1$ is irreducible, then $e_s^{-1}\cdot E(X)$ is also an irreducible polynomial.

[4] A Las Vegas algorithm is a randomized algorithm which always succeeds and we consider its time complexity to be its expected worst-case run time.

Algorithm 8 Generating Irreducible Polynomial

INPUT: Prime power q and an integer $s > 1$
OUTPUT: A monic irreducible polynomial of degree s over $\mathbb{F}_q$

```
b ← 0
WHILE b = 0 DO
    F(X) ← X^s + Σ_{i=0}^{s-1} f_i X^i, where each f_i is chosen uniformly at random from F_q.
    IF gcd(F(X), X^{q^s} − X) = F(X) THEN
        b ← 1.
RETURN F(X)
```

Corollary 5.1.3. *There is a Las Vegas algorithm to generate an irreducible polynomial of degree s over any $\mathbb{F}_q$ in expected time* $\mathrm{poly}(s, \log q)$.

Now recall that Theorem 2.1.3 states that for every prime power p^s, there a unique field $\mathbb{F}_{p^s}$. This along with Theorems 5.1.1 and 5.1.2 imply that:

Corollary 5.1.4. *The field $\mathbb{F}_{p^s}$ is $\mathbb{F}_p[X]/E(X)$, where $E(X)$ is an irreducible polynomial of degree s.*

5.2 Reed-Solomon Codes

Recall that the Singleton bound (Theorem 4.3.1) states that for any $(n,k,d)_q$ code, $k \le n-d+1$. Next, we will study Reed-Solomon codes, which meet the Singleton bound, i.e. satisfy $k = n-d+1$ (but have the unfortunate property that $q \ge n$). Note that this implies that the Singleton bound is tight, at least for $q \ge n$.

We begin with the definition of Reed-Solomon codes.

Definition 5.2.1 (Reed-Solomon code)**.** Let $\mathbb{F}_q$ be a finite field. Let $\alpha_1, \alpha_2, \ldots \alpha_n$ be distinct elements (also called *evaluation points*) from $\mathbb{F}_q$ and choose n and k such that $k \le n \le q$. We define an encoding function for Reed-Solomon code as $RS : \mathbb{F}_q^k \to \mathbb{F}_q^n$ as follows. A message $\mathbf{m} = (m_0, m_1, \ldots, m_{k-1})$ with $m_i \in \mathbb{F}_q$ is mapped to a degree $k-1$ polynomial.

$$\mathbf{m} \mapsto f_{\mathbf{m}}(X),$$

where

$$f_{\mathbf{m}}(X) = \sum_{i=0}^{k-1} m_i X^i. \tag{5.1}$$

Note that $f_{\mathbf{m}}(X) \in \mathbb{F}_q[X]$ is a polynomial of degree at most $k-1$. The encoding of $\mathbf{m}$ is the evaluation of $f_{\mathbf{m}}(X)$ at all the α_i's :

$$RS(\mathbf{m}) = \left(f_{\mathbf{m}}(\alpha_1), f_{\mathbf{m}}(\alpha_2), \ldots, f_{\mathbf{m}}(\alpha_n)\right).$$

We call this image Reed-Solomon code or RS code. A common special case is $n = q-1$ with the set of evaluation points being $\mathbb{F}^* \stackrel{\text{def}}{=} \mathbb{F} \setminus \{0\}$.

For example, the first row below are all the codewords in the $[3,2]_3$ Reed-Solomon codes where the evaluation points are $\mathbb{F}_3$ (and the codewords are ordered by the corresponding messages from $\mathbb{F}_3^2$ in lexicographic order where for clarity the second row shows the polynomial $f_{\mathbf{m}}(X)$ for the corresponding $\mathbf{m} \in \mathbb{F}_3^2$ in gray):

(0,0,0),	(1,1,1),	(2,2,2),	(0,1,2),	(1,2,0),	(2,0,1),	(0,2,1),	(1,0,2),	(2,1,0)
0,	1,	2,	X,	X+1,	X+2,	2X,	2X+1,	2X+2

Notice that by definition, the entries in $\{\alpha_1, ..., \alpha_n\}$ are distinct and thus, must have $n \leq q$.

We now turn to some properties of Reed-Solomon codes.

Claim 5.2.1. *RS codes are linear codes.*

Proof. The proof follows from the fact that if $a \in \mathbb{F}_q$ and $f(X), g(X) \in \mathbb{F}_q[X]$ are polynomials of degree $\leq k-1$, then $af(X)$ and $f(X)+g(X)$ are also polynomials of degree $\leq k-1$. In particular, let messages $\mathbf{m}_1$ and $\mathbf{m}_2$ be mapped to $f_{\mathbf{m}_1}(X)$ and $f_{\mathbf{m}_2}(X)$ where $f_{\mathbf{m}_1}(X), f_{\mathbf{m}_2}(X) \in \mathbb{F}_q[X]$ are polynomials of degree at most $k-1$ and because of the mapping defined in (5.1), it is easy to verify that:

$$f_{\mathbf{m}_1}(X) + f_{\mathbf{m}_2}(X) = f_{\mathbf{m}_1+\mathbf{m}_2}(X),$$

and

$$af_{\mathbf{m}_1}(X) = f_{a\mathbf{m}_1}(X).$$

In other words,

$$RS(\mathbf{m}_1) + RS(\mathbf{m}_2) = RS(\mathbf{m}_1 + \mathbf{m}_2)$$

$$aRS(\mathbf{m}_1) = RS(a\mathbf{m}_1).$$

Therefore RS is a $[n,k]_q$ linear code. □

The second and more interesting claim is the following:

Claim 5.2.2. *RS is a $[n,k,n-k+1]_q$ code. That is, it matches the Singleton bound.*

The claim on the distance follows from the fact that every non-zero polynomial of degree $k-1$ over $\mathbb{F}_q[X]$ has at most $k-1$ (not necessarily distinct) roots, which we prove first (see Proposition 5.2.3 below). This implies that if two polynomials agree on more than $k-1$ places then they must be the same polynomial– note that this implies two polynomials when evaluated at the same n points must differ in at least $n-(k-1) = n-k+1$ positions, which is what we want.

Proposition 5.2.3 ("Degree Mantra")**.** *A nonzero polynomial $f(X)$ of degree t over a field $\mathbb{F}_q$ has at most t roots in $\mathbb{F}_q$*

Proof. We will prove the theorem by induction on t. If $t = 0$, we are done. Now, consider $f(X)$ of degree $t > 0$. Let $\alpha \in \mathbb{F}_q$ be a root such that $f(\alpha) = 0$. If no such root α exists, we are done. If there is a root α, then we can write

$$f(X) = (X - \alpha)g(X)$$

where $\deg(g) = \deg(f) - 1$ (i.e. $X - \alpha$ divides $f(X)$). Note that $g(X)$ is non-zero since $f(X)$ is non-zero. This is because by the fundamental rule of division of polynomials:

$$f(X) = (X - \alpha)g(X) + R(X)$$

where $\deg(R) \leq 0$ (as the degree cannot be negative this in turn implies that $\deg(R) = 0$) and since $f(\alpha) = 0$,

$$f(\alpha) = 0 + R(\alpha),$$

which implies that $R(\alpha) = 0$. Since $R(X)$ has degree zero (i.e. it is a constant polynomial), this implies that $R(X) \equiv 0$.

Finally, as $g(X)$ is non-zero and has degree $t-1$, by induction, $g(X)$ has at most $t-1$ roots, which implies that $f(X)$ has at most t roots. □

We are now ready to prove Claim 5.2.2

Proof of Claim 5.2.2. We start by proving the claim on the distance. Fix arbitrary $\mathbf{m}_1 \neq \mathbf{m}_2 \in \mathbb{F}_q^k$. Note that $f_{\mathbf{m}_1}(X), f_{\mathbf{m}_2}(X) \in \mathbb{F}_q[X]$ are distinct polynomials of degree at most $k-1$ since $\mathbf{m}_1 \neq \mathbf{m}_2 \in \mathbb{F}_q^k$. Then $f_{\mathbf{m}_1}(X) - f_{\mathbf{m}_2}(X) \neq 0$ also has degree at most $k-1$. Note that $wt(RS(\mathbf{m}_2) - RS(\mathbf{m}_1)) = \Delta(RS(\mathbf{m}_1), RS(\mathbf{m}_2))$. The weight of $RS(\mathbf{m}_2) - RS(\mathbf{m}_1)$ is n minus the number of zeroes in $RS(\mathbf{m}_2) - RS(\mathbf{m}_1)$, which is equal to n minus the number of roots that $f_{\mathbf{m}_1}(X) - f_{\mathbf{m}_2}(X)$ has among $\{\alpha_1, ..., \alpha_n\}$. That is,

$$\Delta(RS(\mathbf{m}_1), RS(\mathbf{m}_2)) = n - |\{\alpha_i \mid f_{\mathbf{m}_1}(\alpha_i) = f_{\mathbf{m}_2}(\alpha_i)\}|.$$

By Proposition 5.2.3, $f_{\mathbf{m}_1}(X) - f_{\mathbf{m}_2}(X)$ has at most $k-1$ roots. Thus, the weight of $RS(\mathbf{m}_2) - RS(\mathbf{m}_1)$ is at least $n - (k-1) = n - k + 1$. Therefore $d \geq n - k + 1$, and since the Singleton bound (Theorem 4.3.1) implies that $d \leq n-k+1$, we have $d = n-k+1$.[5] The argument above also shows that distinct polynomials $f_{\mathbf{m}_1}(X), f_{\mathbf{m}_2}(X) \in \mathbb{F}_q[X]$ are mapped to distinct codewords. (This is because the Hamming distance between any two codewords is at least $n - k + 1 \geq 1$, where the last inequality follows as $k \leq n$.) Therefore, the code contains q^k codewords and has dimension k. The claim on linearity of the code follows from Claim 5.2.1. □

Recall that the Plotkin bound (Corollary 4.4.2) implies that to achieve the Singleton bound, the alphabet size cannot be a constant. Thus, some dependence of q on n in Reed-Solomon codes is unavoidable.

Let us now find a generator matrix for RS codes (which exists by Claim 5.2.1). By Definition 5.2.1, any basis $f_{\mathbf{m}_1}, ..., f_{\mathbf{m}_k}$ of polynomial of degree at most $k-1$ gives rise to a basis $RS(\mathbf{m}_1), ..., RS(\mathbf{m}_k)$ of the code. A particularly nice polynomial basis is the set of monomials $1, X, ..., X^i, ..., X^{k-1}$. The corresponding generator matrix, whose ith row (numbering rows from 0 to $k-1$) is

$$(\alpha_1^i, \alpha_2^i, ..., \alpha_j^i, ..., \alpha_n^i)$$

and this generator matrix is called the *Vandermonde* matrix of size $k \times n$:

[5] See Exercise 5.3 for an alternate direct argument.

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_j & \cdots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_j^2 & \cdots & \alpha_n^2 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \alpha_1^i & \alpha_2^i & \cdots & \alpha_j^i & \cdots & \alpha_n^i \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \cdots & \alpha_j^{k-1} & \cdots & \alpha_n^{k-1} \end{pmatrix}$$

The class of codes that match the Singleton bound have their own name, which we define and study next.

5.3 A Property of MDS Codes

Definition 5.3.1 (MDS codes)**.** An $(n,k,d)_q$ code is called *Maximum Distance Separable (MDS)* if $d = n-k+1$.

Thus, Reed-Solomon codes are MDS codes.

Next, we prove an interesting property of an MDS code $C \subseteq \Sigma^n$ with integral dimension k. We begin with the following notation.

Definition 5.3.2. For any subset of indices $S \subseteq [n]$ of size exactly k and a code $C \subseteq \Sigma^n$, C_S is the set of all codewords in C projected onto the indices in S.

MDS codes have the following nice property that we shall prove for the special case of Reed-Solomon codes first and subsequently for the general case as well.

Proposition 5.3.1. *Let $C \subseteq \Sigma^n$ of integral dimension k be an MDS code, then for all $S \subseteq [n]$ such that $|S| = k$, we have $|C_S| = \Sigma^k$.*

Before proving Proposition 5.3.1 in its full generality, we present its proof for the special case of Reed-Solomon codes.

Consider any $S \subseteq [n]$ of size k and fix an arbitrary $\mathbf{v} = (v_1, \ldots, v_k) \in \mathbb{F}_q^k$, we need to show that there exists a codeword $\mathbf{c} \in RS$ (assume that the RS code evaluates polynomials of degree at most $k-1$ over $\alpha_1, \ldots, \alpha_n \subseteq \mathbb{F}_q$) such that $\mathbf{c}_S = \mathbf{v}$. Consider a generic degree $k-1$ polynomial $F(X) = \sum_{i=0}^{k-1} f_i X^i$. Thus, we need to show that there exists $F(X)$ such that $F(\alpha_i) = v_i$ for all $i \in S$, where $|S| = k$.

For notational simplicity, assume that $S = [k]$. We think of f_i's as unknowns in the equations that arise out of the relations $F(\alpha_i) = v_i$. Thus, we need to show that there is a solution to the following system of linear equations:

$$\begin{pmatrix} p_0 & p_1 & \cdots & p_{k-1} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ \alpha_1 & \alpha_i & \alpha_k \\ \alpha_1^2 & \alpha_i^2 & \alpha_k^2 \\ \vdots & \vdots & \vdots \\ \alpha_1^{k-1} & \alpha_i^{k-1} & \alpha_k^{k-1} \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_k \end{pmatrix}$$

The above constraint matrix is a Vandermonde matrix and is known to have full rank (see Exercise 5.7). Hence, by Exercise 2.6, there always exists a unique solution for $(p_0, \ldots, p_{k-1})$. This completes the proof for Reed-Solomon codes.

Next, we prove the property for the general case which is presented below

Proof of Proposition 5.3.1. Consider a $|C| \times n$ matrix where each row represents a codeword in C. Hence, there are $|C| = |\Sigma|^k$ rows in the matrix. The number of columns is equal to the block length n of the code. Since C is Maximum Distance Separable, its distance $d = n - k + 1$.

Let $S \subseteq [n]$ be of size exactly k. It is easy to see that for any $\mathbf{c}^i \neq \mathbf{c}^j \in C$, the corresponding projections $\mathbf{c}^i_S$ and $\mathbf{c}^j_S \in C_S$ are not the same. As otherwise $\triangle(\mathbf{c}^i, \mathbf{c}^j) \leq d - 1$, which is not possible as the minimum distance of the code C is d. Therefore, every codeword in C gets mapped to a distinct codeword in C_S. As a result, $|C_S| = |C| = |\Sigma|^k$. As $C_S \subseteq \Sigma^k$, this implies that $C_S = \Sigma^k$, as desired. □

Proposition 5.3.1 implies an important property in pseudorandomness: see Exercise 5.8 for more.

5.4 Exercises

Exercise 5.1. Prove that $X^2 + X + 1$ is the unique irreducible polynomial of degree two over $\mathbb{F}_2$.

Exercise 5.2. Argue that any function $f : \mathbb{F}_q \to \mathbb{F}_q$ is equivalent to a polynomial $P(X) \in \mathbb{F}_q[X]$ of degree at most $q - 1$: that is, for every $\alpha \in \mathbb{F}_q$

$$f(\alpha) = P(\alpha).$$

Exercise 5.3. For any $[n, k]_q$ Reed-Solomon code, exhibit two codewords that are at Hamming distance exactly $n - k + 1$.

Exercise 5.4. Let $\mathrm{RS}_{\mathbb{F}_q^*}[n, k]$ denote the Reed-Solomon code over $\mathbb{F}_q$ where the evaluation points is $\mathbb{F}_q$ (i.e. $n = q$). Prove that

$$\left(\mathrm{RS}_{\mathbb{F}_q}[n, k]\right)^{\perp} = \mathrm{RS}_{\mathbb{F}_q}[n, n - k],$$

that is, the dual of these Reed-Solomon codes are Reed-Solomon codes themselves. Conclude that Reed-Solomon codes contain self-dual codes (see Exercise 2.31 for a definition).

Hint: Exercise 2.2 might be useful.

Exercise 5.5. Since Reed-Solomon codes are linear codes, by Proposition 2.3.3, one can do error detection for Reed-Solomon codes in quadratic time. In this problem, we will see that one can design even more efficient error detection algorithm for Reed-Solomon codes. In particular, we will consider *data streaming algorithms* (see Section 19.5 for more motivation on this class of algorithms). A data stream algorithm makes a sequential pass on the input, uses poly-logarithmic space and spend only poly-logarithmic time on each location in the input. In this problem we show that there exists a randomized data stream algorithm to solve the error detection problem for Reed-Solomon codes.

1. Give a randomized data stream algorithm that given as input a sequence $(i_1, \alpha_1), \ldots, (i_n, \alpha_n) \in [m] \times \mathbb{F}_q$ that *implicitly* defines $\mathbf{y} \in \mathbb{F}_q^m$, where for any $\ell \in [m]$, $y_\ell = \sum_{j \in [n] | i_j = \ell} \alpha_\ell$, decides whether $\mathbf{y} = \mathbf{0}$ with probability at least 2/3. Your algorithm should use $O(\log q(m+n))$ space and polylog$(q(m+n))$ time per position of $\mathbf{y}$. For simplicity, you can assume that given an integer $t \geq 1$ and prime power q, the algorithm has oracle access to an irreducible polynomial of degree t over $\mathbb{F}_q$.
   Hint: Use Reed-Solomon codes.

2. Given $[q, k]_q$ Reed-Solomon code C (i.e. with the evaluation points being $\mathbb{F}_q$), present a data stream algorithm for error detection of C with $O(\log q)$ space and polylog$q$ time per position of the received word. The algorithm should work correctly with probability at least 2/3. You should assume that the data stream algorithm has access to the values of k and q (and knows that C has $\mathbb{F}_q$ as its evaluation points).
   Hint: Part 1 and Exercise 5.4 should be helpful.

Exercise 5.6. We have defined Reed-Solomon in this chapter and Hadamard codes in Section 2.7. In this problem we will prove that certain alternate definitions also suffice.

1. Consider the Reed-Solomon code over a field $\mathbb{F}_q$ and block length $n = q - 1$ defined as

$$\mathrm{RS}_{\mathbb{F}_q^*}[n, k, n-k+1] = \{(p(1), p(\alpha), \ldots, p(\alpha^{n-1})) \mid p(X) \in \mathbb{F}[X] \text{ has degree} \leq k-1\}$$

   where α is the generator of the multiplicative group $\mathbb{F}^*$ of $\mathbb{F}$.[6]

   Prove that

$$\begin{aligned} \mathrm{RS}_{\mathbb{F}_q^*}[n, k, n-k+1] &= \{(c_0, c_1, \ldots, c_{n-1}) \in \mathbb{F}^n \mid c(\alpha^\ell) = 0 \text{ for } 1 \leq \ell \leq n-k\,, \\ &\qquad \text{where } c(X) = c_0 + c_1 X + \cdots + c_{n-1} X^{n-1}\}\,. \end{aligned} \tag{5.2}$$

   Hint: Exercise 2.2 might be useful.

2. Recall that the $[2^r, r, 2^{r-1}]_2$ Hadamard code is generated by the $r \times 2^r$ matrix whose ith (for $0 \leq i \leq 2^r - 1$) column is the binary representation of i. Briefly argue that the Hadamard

[6] This means that $\mathbb{F}_q^* = \{1, \alpha, \ldots, \alpha^{n-1}\}$. Further, $\alpha^n = 1$.

codeword for the message $(m_1, m_2, \ldots, m_r) \in \{0,1\}^r$ is the evaluation of the (multivariate) polynomial $m_1 X_1 + m_2 X_2 + \cdots + m_r X_r$ (where $X_1, \ldots, X_r$ are the r variables) over all the possible assignments to the variables $(X_1, \ldots, X_r)$ from $\{0,1\}^r$.

Using the definition of Hadamard codes above (re)prove the fact that the code has distance 2^{r-1}.

Exercise 5.7. Prove that the $k \times k$ Vandermonde matrix (where the (i,j)th entry is α_j^i) has full rank (where $\alpha_1, \ldots, \alpha_k$ are distinct).

Exercise 5.8. A set $S \subseteq \mathbb{F}_q^n$ is said to be a t-wise independent source (for some $1 \le t \le n$) if given a uniformly random sample $(X_1, \ldots, X_n)$ from S, the n random variables are t-wise independent: i.e. any subset of t variables are uniformly independent random variables over $\mathbb{F}_q$. We will explore properties of these objects in this exercise.

1. Argue that the definition of t-wise independent source is equivalent to the definition in Exercise 2.13.

2. Argue that for any $k \ge 1$, any $[n,k]_q$ code C is a 1-wise independent source.

3. Prove that any $[n,k]_q$ MDS code is a k-wise independent source.

4. Using part 3 or otherwise prove that there exists a k-wise independent source over $\mathbb{F}_2$ of size at most $(2n)^k$. Conclude that $k(\log_2 n + 1)$ uniformly and independent random bits are enough to compute n random bits that are k-wise independent. Improve the bound slightly to show that $k(\log_2 n - \log_2 \log_2 n + O(1))$-random bits are enough to generate k-wise independent source over $\mathbb{F}_2$.

5. For $0 < p \le 1/2$, we say the n binary random variables $X_1, \ldots, X_n$ are p-biased and t-wise independent if any of the t random variables are independent and $\Pr[X_i = 1] = p$ for every $i \in [n]$. For the rest of the problem, let p be a power of $1/2$. Then show that any $t \cdot \log_2(1/p)$-wise independent random variables can be converted into t-wise independent p-biased random variables. Conclude that one can construct such sources with $t \log_2(1/p)(1 + \log_2 n)$ uniformly random bits. Then improve this bound to $t(1 + \max(\log_2(1/p), \log_2 n))$ uniformly random bits.

Exercise 5.9. In many applications, errors occur in "bursts"– i.e. all the error locations are contained in a contiguous region (think of a scratch on a DVD or disk). In this problem we will use how one can use Reed-Solomon codes to correct bursty errors.

An error vector $\mathbf{e} \in \{0,1\}^n$ is called a *t-single burst* error pattern if all the non-zero bits in $\mathbf{e}$ occur in the range $[i, i+t-1]$ for some $1 \le i \le n = t+1$. Further, a vector $\mathbf{e} \in \{0,1\}^n$ is called a (s,t)-burst error pattern if it is the union of at most s t-single burst error pattern (i.e. all non-zero bits in $\mathbf{e}$ are contained in one of at most s contiguous ranges in $[n]$).

We call a binary code $C \subseteq \{0,1\}^n$ to be (s,t)-*burst error correcting* if one can uniquely decode from any (s,t)-burst error pattern. More precisely, given an (s,t)-burst error pattern $\mathbf{e}$ and any codeword $\mathbf{c} \in C$, the only codeword $\mathbf{c}' \in C$ such that $(\mathbf{c} + \mathbf{e}) - \mathbf{c}'$ is an (s,t)-burst error pattern satisfies $\mathbf{c}' = \mathbf{c}$.

1. Argue that if C is (st)-error correcting (in the sense of Definition 1.3.3), then it is also (s,t)-burst error correcting. Conclude that for any $\varepsilon > 0$, there exists code with rate $\Omega(\varepsilon^2)$ and block length n that is (s,t)-burst error correcting for any s,t such that $s \cdot t \le \left(\frac{1}{4} - \varepsilon\right) \cdot n$.

2. Argue that for any rate $R > 0$ and for large enough n, there exist (s,t)-burst error correcting as long as $s \cdot t \le \left(\frac{1-R-\varepsilon}{2}\right) \cdot n$ and $t \ge \Omega\left(\frac{\log n}{\varepsilon}\right)$. In particular, one can correct from $\frac{1}{2} - \varepsilon$ fraction of burst-errors (as long as each burst is "long enough") with rate $\Omega(\varepsilon)$ (compare this with item 1).

   Hint: Use Reed-Solomon codes.

Exercise 5.10. In this problem we will look at a very important class of codes called BCH codes[7]. Let $\mathbb{F} = \mathbb{F}_{2^m}$. Consider the binary code C_{BCH} defined as $\mathrm{RS}_{\mathbb{F}}[n, k, n-k+1] \cap \mathbb{F}_2^n$.

1. Prove that C_{BCH} is a binary linear code of distance at least $d = n - k + 1$ and dimension at least $n - (d-1)\log_2(n+1)$.

   Hint: Use the characterization (5.2) of the Reed-Solomon code from Exercise 5.6.

2. Prove a better lower bound of $n - \left\lceil \frac{d-1}{2} \right\rceil \log_2(n+1)$ on the dimension of C_{BCH}.

   Hint: Try to find redundant checks among the "natural" parity checks defining C_{BCH}).

3. For $d = 3$, C_{BCH} is the same as another code we have seen. What is that code?

4. For constant d (and growing n), prove that C_{BCH} have nearly optimal dimension for distance d, in that the dimension cannot be $n - t\log_2(n+1)$ for $t < \frac{d-1}{2}$.

Exercise 5.11. Show that for $1 \le k \le n$, $\left\lceil \frac{k}{2} \right\rceil \log_2(n+1)$ random bits are enough to compute n-bits that are k-wise independent. Note that this is an improvement of almost a factor of 2 from the bound from Exercise 5.8 part 4 (and this new bound is known to be optimal).

Hint: Use Exercises 2.13 and 5.10.

Exercise 5.12. In this exercise, we continue in the theme of Exercise 5.10 and look at the intersection of a Reed-Solomon code with $\mathbb{F}_2^n$ to get a binary code. Let $\mathbb{F} = \mathbb{F}_{2^m}$. Fix positive integers d, n with $(d-1)m < n < 2^m$, and a set $S = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ of n distinct nonzero elements of $\mathbb{F}$. For a vector $\mathbf{v} = (v_1, \ldots, v_n) \in (\mathbb{F}^*)^n$ of n not necessarily distinct nonzero elements from $\mathbb{F}$, define the *Generalized Reed-Solomon code* $\mathrm{GRS}_{S,\mathbf{v},d}$ as follows:

$$\mathrm{GRS}_{S,\mathbf{v},d} = \{(v_1 p(\alpha_1), v_2 p(\alpha_2), \ldots, v_n p(\alpha_n)) \mid p(X) \in \mathbb{F}[X] \text{ has degree} \le n-d\}.$$

1. Prove that $\mathrm{GRS}_{S,\mathbf{v},d}$ is an $[n, n-d+1, d]_{\mathbb{F}}$ linear code.

2. Argue that $\mathrm{GRS}_{S,\mathbf{v},d} \cap \mathbb{F}_2^n$ is a binary linear code of rate at least $1 - \frac{(d-1)m}{n}$.

[7] The acronym BCH stands for Bose-Chaudhuri-Hocquenghem, the discoverers of this family of codes.

3. Let $\mathbf{c} \in \mathbb{F}_2^n$ be a nonzero binary vector. Prove that (for every choice of d, S) there are at most $(2^m - 1)^{n-d+1}$ choices of the vector $\mathbf{v}$ for which $\mathbf{c} \in \mathrm{GRS}_{S,\mathbf{v},d}$.

4. Using the above, prove that if the integer D satisfies $\mathrm{Vol}_2(n, D-1) < (2^m - 1)^{d-1}$ (where $\mathrm{Vol}_2(n, D-1) = \sum_{i=0}^{D-1} \binom{n}{i}$), then there exists a vector $\mathbf{v} \in (\mathbb{F}^*)^n$ such that the minimum distance of the binary code $\mathrm{GRS}_{S,\mathbf{v},d} \cap \mathbb{F}_2^n$ is at least D.

5. Using parts 2 and 4 above (or otherwise), argue that the family of codes $\mathrm{GRS}_{S,\mathbf{v},d} \cap \mathbb{F}_2^n$ contains binary linear codes that meet the Gilbert-Varshamov bound.

Exercise 5.13. In this exercise we will show that the dual of a GRS code is a GRS itself with different parameters. First, we state the obvious definition of GRS codes over a general finite field $\mathbb{F}_q$ (as opposed to the definition over fields of characteristic two in Exercise 5.12). In particular, define the code $\mathrm{GRS}_{S,\mathbf{v},d,q}$ as follows:

$$\mathrm{GRS}_{S,\mathbf{v},d,q} = \{(v_1 p(\alpha_1), v_2 p(\alpha_2), \ldots, v_n p(\alpha_n)) \mid p(X) \in \mathbb{F}_q[X] \text{ has degree} \leq n-d\} .$$

Then show that

$$\left(\mathrm{GRS}_{S,\mathbf{v},d,q}\right)^{\perp} = \mathrm{GRS}_{S,\mathbf{v}',n-d+2,q},$$

where $\mathbf{v}' \in \mathbb{F}_q^n$ is a vector with all non-zero components.

Exercise 5.14. In Exercise 2.16, we saw that any linear code can be converted in to a systematic code. In other words, there is a map to convert Reed-Solomon codes into a systematic one. In this exercise the goal is to come up with an explicit encoding function that results in a systematic Reed-Solomon code.

In particular, given the set of evaluation points $\alpha_1, \ldots, \alpha_n$, design an explicit map f from $\mathbb{F}_q^k$ to a polynomial of degree at most $k-1$ such that the following holds. For every message $\mathbf{m} \in \mathbb{F}_q^k$, if the corresponding polynomial is $f_{\mathbf{m}}(X)$, then the vector $\left(f_{\mathbf{m}}(\alpha_i)\right)_{i \in [n]}$ has the message $\mathbf{m}$ appear in the corresponding codeword (say in its first k positions). Further, argue that this map results in an $[n, k, n-k+1]_q$ code.

Exercise 5.15. In this problem, we will consider the number-theoretic counterpart of Reed-Solomon codes. Let $1 \leq k < n$ be integers and let $p_1 < p_2 < \cdots < p_n$ be n distinct primes. Denote $K = \prod_{i=1}^{k} p_i$ and $N = \prod_{i=1}^{n} p_i$. The notation $\mathbb{Z}_M$ stands for integers modulo M, i.e., the set $\{0, 1, \ldots, M-1\}$. Consider the *Chinese Remainder code* defined by the encoding map $E : \mathbb{Z}_K \to \mathbb{Z}_{p_1} \times \mathbb{Z}_{p_2} \times \cdots \times \mathbb{Z}_{p_n}$ defined by:

$$E(m) = (m \mod p_1,\ m \mod p_2,\ \cdots,\ m \mod p_n) .$$

(Note that this is not a code in the usual sense we have been studying since the symbols at different positions belong to different alphabets. Still notions such as distance of this code make sense and are studied in the question below.)

Suppose that $m_1 \neq m_2$. For $1 \leq i \leq n$, define the indicator variable $b_i = 1$ if $E(m_1)_i \neq E(m_2)_i$ and $b_i = 0$ otherwise. Prove that $\prod_{i=1}^{n} p_i^{b_i} > N/K$.

Use the above to deduce that when $m_1 \neq m_2$, the encodings $E(m_1)$ and $E(m_2)$ differ in at least $n-k+1$ locations.

Exercise 5.16. In this problem, we will consider derivatives over a finite field $\mathbb{F}_q$. Unlike the case of derivatives over reals, derivatives over finite fields do not have any physical interpretation but as we shall see shortly, the notion of derivatives over finite fields is still a useful concept. In particular, given a polynomial $f(X) = \sum_{i=0}^{t} f_i X^i$ over $\mathbb{F}_q$, we define its derivative as

$$f'(X) = \sum_{i=0}^{t-1} (i+1) \cdot f_{i+1} \cdot X^i.$$

Further, we will denote by $f^{(i)}(X)$, the result of applying the derivative on f i times. In this problem, we record some useful facts about derivatives.

1. Define $R(X,Z) = f(X+Z) = \sum_{i=0}^{t} r_i(X) \cdot Z^i$. Then for any $j \geq 1$,

$$f^{(j)}(X) = j! \cdot r_j(X).$$

2. Using part 1 or otherwise, show that for any $j \geq \mathrm{char}(\mathbb{F}_q)$,[8] $f^{(j)}(X) \equiv 0$.

3. Let $j \leq \mathrm{char}(\mathbb{F}_q)$. Further, assume that for every $0 \leq i < j$, $f^{(i)}(\alpha) = 0$ for some $\alpha \in \mathbb{F}_q$. Then prove that $(X-\alpha)^j$ divides $f(X)$.

4. Finally, we will prove the following generalization of the degree mantra (Proposition 5.2.3). Let $f(X)$ be a non-zero polynomial of degree t and $m \leq \mathrm{char}(\mathbb{F}_q)$. Then there exists at most $\left\lfloor \frac{t}{m} \right\rfloor$ distinct elements $\alpha \in \mathbb{F}_q$ such that $f^{(j)}(\alpha) = 0$ for every $0 \leq j < m$.

Exercise 5.17. In this exercise, we will consider a code that is related to Reed-Solomon codes and uses derivatives from Exercise 5.16. These codes are called *derivative codes*.

Let $m \geq 1$ be an integer parameter and consider parameters $k > \mathrm{char}(\mathbb{F}_q)$ and n such that $m < k < nm$. Then the derivative code with parameters (n,k,m) is defined as follow. Consider any message $\mathbf{m} \in \mathbb{F}_q^k$ and let $f_{\mathbf{m}}(X)$ be the message polynomial as defined for the Reed-Solomon code. Let $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$ be distinct elements. Then the codeword for $\mathbf{m}$ is given by

$$\begin{pmatrix} f_{\mathbf{m}}(\alpha_1) & f_{\mathbf{m}}(\alpha_2) & \cdots & f_{\mathbf{m}}(\alpha_n) \\ f_{\mathbf{m}}^{(1)}(\alpha_1) & f_{\mathbf{m}}^{(1)}(\alpha_2) & \cdots & f_{\mathbf{m}}^{(1)}(\alpha_n) \\ \vdots & \vdots & \vdots & \vdots \\ f_{\mathbf{m}}^{(m-1)}(\alpha_1) & f_{\mathbf{m}}^{(m-1)}(\alpha_2) & \cdots & f_{\mathbf{m}}^{(m-1)}(\alpha_n) \end{pmatrix}.$$

Prove that the above code is an $\left[n, \frac{k}{m}, n - \left\lfloor \frac{k-1}{m} \right\rfloor\right]_{q^m}$-code (and is thus MDS).

Exercise 5.18. In this exercise, we will consider another code related to Reed-Solomon codes that are called *Folded Reed-Solomon* codes. We will see a lot more of these codes in Chapter 16.

Let $m \geq 1$ be an integer parameter and let $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$ are distinct elements such that for some element $\gamma \in \mathbb{F}_q^*$, the sets

$$\{\alpha_i, \alpha_i\gamma, \alpha_i\gamma^2, \ldots, \alpha_i\gamma^{m-1}\}, \tag{5.3}$$

[8] $\mathrm{char}(\mathbb{F}_q)$ denotes the *characteristic* of $\mathbb{F}_q$. That is, if $q = p^s$ for some prime p, then $\mathrm{char}(\mathbb{F}_q) = p$. Any natural number i in $\mathbb{F}_q$ is equivalent to $i \mod \mathrm{char}(\mathbb{F}_q)$.

are pair-wise disjoint for different $i \in [n]$. Then the folded Reed-Solomon code with parameters $(m, k, n, \gamma, \alpha_1, \ldots, \alpha_n)$ is defined as follows. Consider any message $\mathbf{m} \in \mathbb{F}_q^k$ and let $f_{\mathbf{m}}(X)$ be the message polynomial as defined for the Reed-Solomon code. Then the codeword for $\mathbf{m}$ is given by:

$$\begin{pmatrix} f_{\mathbf{m}}(\alpha_1) & f_{\mathbf{m}}(\alpha_2) & \cdots & f_{\mathbf{m}}(\alpha_n) \\ f_{\mathbf{m}}(\alpha_1 \cdot \gamma) & f_{\mathbf{m}}(\alpha_2 \cdot \gamma) & \cdots & f_{\mathbf{m}}(\alpha_n \cdot \gamma) \\ \vdots & \vdots & \vdots & \vdots \\ f_{\mathbf{m}}(\alpha_1 \cdot \gamma^{m-1}) & f_{\mathbf{m}}(\alpha_2 \cdot \gamma^{m-1}) & \cdots & f_{\mathbf{m}}(\alpha_n \cdot \gamma^{m-1}) \end{pmatrix}.$$

Prove that the above code is an $\left[n, \frac{k}{m}, n - \left\lfloor \frac{k-1}{m} \right\rfloor\right]_{q^m}$-code (and is thus, MDS).

Exercise 5.19. In this problem we will see that Reed-Solomon codes, derivative codes (Exercise 5.17) and folded Reed-Solomon codes (Exercise 5.18) are all essentially special cases of a large family of codes that are based on polynomials. We begin with the definition of these codes.

Let $m \geq 1$ be an integer parameter and define $m < k \leq n$. Further, let $E_1(X), \ldots, E_n(X)$ be n polynomials over $\mathbb{F}_q$, each of degree m. Further, these polynomials pair-wise do not have any non-trivial factors (i.e. $\gcd(E_i(X), E_j(X))$ has degree 0 for every $i \neq j \in [n]$.) Consider any message $\mathbf{m} \in \mathbb{F}_q^k$ and let $f_{\mathbf{m}}(X)$ be the message polynomial as defined for the Reed-Solomon code. Then the codeword for $\mathbf{m}$ is given by:

$$\left(f_{\mathbf{m}}(X) \mod E_1(X), f_{\mathbf{m}}(X) \mod E_2(X), \ldots, f_{\mathbf{m}}(X) \mod E_n(X)\right).$$

In the above we think of $f_{\mathbf{m}}(X) \mod E_i(X)$ as an element of $\mathbb{F}_{q^m}$. In particular, given given a polynomial of degree at most $m-1$, we will consider any bijection between the q^m such polynomials and $\mathbb{F}_{q^m}$. We will first see that this code is MDS and then we will see why it contains Reed-Solomon and related codes as special cases.

1. Prove that the above code is an $\left[n, \frac{k}{m}, n - \left\lfloor \frac{k-1}{m} \right\rfloor\right]_{q^m}$-code (and is thus, MDS).

2. Let $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$ be distinct elements. Define $E_i(X) = X - \alpha_i$. Argue that for this special case the above code (with $m = 1$) is the Reed-Solomon code.

3. Let $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$ be distinct elements. Define $E_i(X) = (X - \alpha_i)^m$. Argue that for this special case the above code is the derivative code (with an appropriate mapping from polynomials of degree at most $m-1$ and $\mathbb{F}_q^m$, where the mapping could be different for each $i \in [n]$ and can depend on $E_i(X)$).

4. Let $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$ be distinct elements and $\gamma \in \mathbb{F}_q^*$ such that (5.3) is satisfied. Define $E_i(X) = \prod_{j=0}^{m-1}(X - \alpha_i \cdot \gamma^j)$. Argue that for this special case the above code is the folded Reed-Solomon code (with an appropriate mapping from polynomials of degree at most $m-1$ and $\mathbb{F}_q^m$, where the mapping could be different for each $i \in [n]$ and can depend on $E_i(X)$).

Exercise 5.20. In this exercise we will develop a sufficient condition to determine the irreducibility of certain polynomials called the *Eisenstein's criterion*.

Let $F(X,Y)$ be a polynomial of $\mathbb{F}_q$. Think of this polynomial as over X with coefficients as polynomials in Y over $\mathbb{F}_q$. Technically, we think of the coefficients as coming from the ring of polynomials in Y over $\mathbb{F}_q$. We will denote the ring of polynomials in Y over $\mathbb{F}_q$ as $\mathbb{F}_q(Y)$ and we will denote the polynomials in X with coefficients from $\mathbb{F}_q(Y)$ as $\mathbb{F}_q(Y)[X]$.

In particular, let

$$F(X,Y) = X^t + f_{t-1}(Y)\cdot X^{t-1} + \cdots + f_0(Y),$$

where each $f_i(Y) \in \mathbb{F}_q(Y)$. Let $P(Y)$ be a *prime* for $\mathbb{F}_q(Y)$ (i.e. $P(Y)$ has degree at least one and if $P(Y)$ divides $A(Y)\cdot B(Y)$ then $P(Y)$ divides at least one of $A(Y)$ or $B(Y)$). If the following conditions hold:

(i) $P(Y)$ divides $f_i(Y)$ for every $0 \le i < t$; but

(ii) $P^2(Y)$ does not divide $f_0(Y)$

then $F(X,Y)$ does not have any non-trivial factors over $\mathbb{F}_q(Y)[X]$ (i.e. all factors have either degree t or 0 in X).

In the rest of the problem, we will prove this result in a sequence of steps:

1. For the sake of contradiction assume that $F(X,Y) = G(X,Y)\cdot H(X,Y)$ where

$$G(X,Y) = \sum_{i=0}^{t_1} g_i(Y)\cdot X^I \text{ and } H(X,Y) = \sum_{i=0}^{t_2} h_i(Y)\cdot X^i,$$

   where $0 < t_1, t_2 < t$. Then argue that $P(Y)$ does not divide both of $g_0(Y)$ and $h_0(Y)$.

   For the rest of the problem WLOG assume that $P(Y)$ divides $g_0(Y)$ (and hence does not divide $h_0(Y)$).

2. Argue that there exists an i^* such that $P(Y)$ divide $g_i(Y)$ for every $0 \le i < i^*$ but $P(Y)$ does not divide $g_{i^*}(Y)$ (define $g_t(Y) = 1$).

3. Argue that $P(Y)$ does not divide $f_i(Y)$. Conclude that $F(X,Y)$ does not have any non-trivial factors, as desired.

Exercise 5.21. We have mentioned objects called algebraic-geometric (AG) codes, that generalize Reed-Solomon codes and have some amazing properties: see for example, Section 4.6. The objective of this exercise is to construct one such AG code, and establish its rate vs distance trade-off.

Let p be a prime and $q = p^2$. Consider the equation

$$Y^p + Y = X^{p+1} \tag{5.4}$$

over $\mathbb{F}_q$.

1. Prove that there are exactly p^3 solutions in $\mathbb{F}_q \times \mathbb{F}_q$ to (5.4). That is, if $S \subseteq \mathbb{F}_q^2$ is defined as

$$S = \left\{(\alpha, \beta) \in \mathbb{F}_q^2 \mid \beta^p + \beta = \alpha^{p+1}\right\}$$

then $|S| = p^3$.

2. Prove that the polynomial $F(X, Y) = Y^p + Y - X^{p+1}$ is irreducible over $\mathbb{F}_q$.
Hint: Exercise 5.20 could be useful.

3. Let $n = p^3$. Consider the evaluation map $\mathrm{ev} : \mathbb{F}_q[X, Y] \to \mathbb{F}_q^n$ defined by

$$\mathrm{ev}(f) = (f(\alpha, \beta) : (\alpha, \beta) \in S) .$$

Argue that if $f \neq 0$ and is not divisible by $Y^p + Y - X^{p+1}$, then $\mathrm{ev}(f)$ has Hamming weight at least $n - \deg(f)(p+1)$, where $\deg(f)$ denotes the *total* degree of f.
Hint: You are allowed to make use of *Bézout's theorem*, which states that if $f, g \in \mathbb{F}_q[X, Y]$ are nonzero polynomials *with no common factors*, then they have at most $\deg(f)\deg(g)$ common zeroes.

4. For an integer parameter $\ell \geq 1$, consider the set $\mathscr{F}_\ell$ of bivariate polynomials

$$\mathscr{F}_\ell = \left\{f \in \mathbb{F}_q[X, Y] \mid \deg(f) \leq \ell, \deg_X(f) \leq p\right\}$$

where $\deg_X(f)$ denotes the degree of f in X.

Argue that $\mathscr{F}_\ell$ is an $\mathbb{F}_q$-linear space of dimension $(\ell+1)(p+1) - \frac{p(p+1)}{2}$.

5. Consider the code $C \subseteq \mathbb{F}_q^n$ for $n = p^3$ defined by

$$C = \left\{\mathrm{ev}(f) \mid f \in \mathscr{F}_\ell\right\} .$$

Prove that C is a linear code with minimum distance at least $n - \ell(p+1)$.

6. Deduce a construction of an $[n, k]_q$ code with distance $d \geq n - k + 1 - p(p-1)/2$.

(Note that Reed-Solomon codes have $d = n - k + 1$, whereas these codes are off by $p(p-1)/2$ from the Singleton bound. However they are much longer than Reed-Solomon codes, with a block length of $n = q^{3/2}$, and the deficiency from the Singleton bound is only $o(n)$.)

5.5 Bibliographic Notes

Reed-Solomon codes were invented by Irving Reed and Gus Solomon [84]. Even though Reed-Solomon codes need $q \geq n$, they are used widely in practice. For example, Reed-Solomon codes are used in storage of information in CDs and DVDs. This is because they are robust against burst-errors that come in contiguous manner. In this scenario, a large alphabet is then a good thing since bursty errors will tend to corrupt the entire symbol in $\mathbb{F}_q$ unlike partial errors, e.g. errors over bits. (See Exercise 5.9.)

It is a big open question to present a deterministic algorithm to compute an irreducible polynomial of a given degree with the same time complexity as in Corollary 5.1.3. Such results are known in general if one is happy with polynomial dependence on q instead of $\log q$. See the book by Shoup [91] for more details.

Chapter 6

What Happens When the Noise is Stochastic: Shannon's Theorem

Shannon was the first to present a rigorous mathematical framework for communication, which (as we have already seen) is the problem of reproducing at one point (typically called the "receiver" of the channel) a message selected at another point (called the "sender" to the channel). Unlike Hamming, Shannon modeled the noise stochastically, i.e. as a well defined random process. He proved a result that pin-pointed the best possible rate of transmission of information over a very wide range of stochastic channels. In fact, Shannon looked at the communication problem at a higher level, where he allowed for compressing the data first (before applying any error-correcting code), so as to minimize the amount of symbols transmitted over the channel.

In this chapter, we will study some stochastic noise models most of which were proposed by Shannon. We then prove an optimal tradeoff between the rate and fraction of errors that are correctable for a specific stochastic noise model called the Binary Symmetric Channel.

6.1 Overview of Shannon's Result

Shannon introduced the notion of reliable communication[1] over noisy channels. Broadly, there are two types of channels that were studied by Shannon:

- (Noisy Channel) This type of channel introduces errors during transmission, which result in an incorrect reception of the transmitted signal by the receiver. Redundancy is added at the transmitter to increase reliability of the transmitted data. The redundancy is taken off at the receiver. This process is termed as *Channel Coding*.

- (Noise-free Channel) As the name suggests, this channel does not introduce any type of error in transmission. Redundancy in source data is used to compress the source data at the transmitter. The data is decompressed at the receiver. The process is popularly known as *Source Coding*.

[1] That is, the ability to successfully send the required information over a channel that can lose or corrupt data.

Figure 6.1 presents a generic model of a communication system, which combines the two concepts we discussed above.

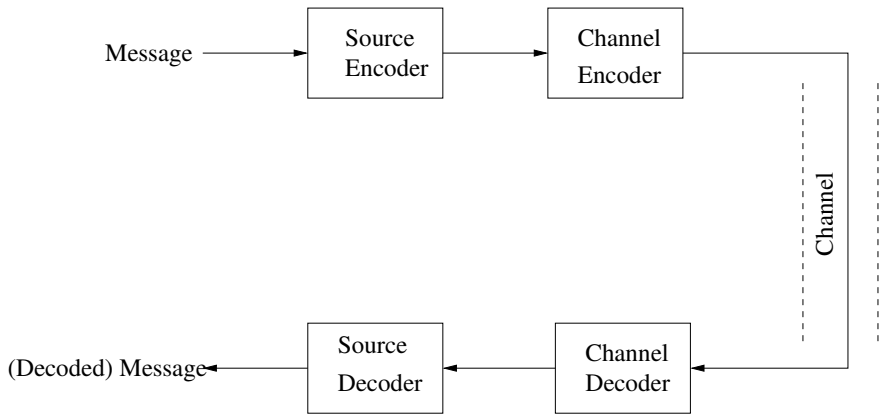


Figure 6.1: The communication process

In Figure 6.1, source coding and channel coding are coupled. In general, to get the optimal performance, it makes sense to design both the source and channel coding schemes simultaneously. However, Shannon's source coding theorem allows us to decouple both these parts of the communication setup and study each of these parts separately. Intuitively, this makes sense: if one can have reliable communication over the channel using channel coding, then for the source coding the resulting channel effectively has no noise.

For source coding, Shannon proved a theorem that precisely identifies the amount by which the message can be compressed: this amount is related to the *entropy* of the message. We will not talk much more about source coding in in this book. (However, see Exercises 6.10, 6.11 and 6.12.) From now on, we will exclusively focus on the channel coding part of the communication setup. Note that one aspect of channel coding is how we model the channel noise. So far we have seen Hamming's worst case noise model in some detail. Next, we will study some specific stochastic channels.

6.2 Shannon's Noise Model

Shannon proposed a stochastic way of modeling noise. The input symbols to the channel are assumed to belong to some *input alphabet* $\mathscr{X}$, while the channel outputs symbols from its *output alphabet* $\mathscr{Y}$. The following diagram shows this relationship:

$$\mathscr{X} \ni x \rightarrow \boxed{\text{channel}} \rightarrow y \in \mathscr{Y}$$

The channels considered by Shannon are also *memoryless*, that is, noise acts independently on each transmitted symbol. In this book, we will only study *discrete* channels where both the alphabets $\mathscr{X}$ and $\mathscr{Y}$ are finite. For the sake of variety, we will define one channel that is continuous, though we will not study it in any detail later on.

The final piece in specification of a channel is the *transition matrix* $\mathbf{M}$ that governs the process of how the channel introduces error. In particular, the channel is described in the form of a matrix with entries as the crossover probability over all combination of the input and output alphabets. For any pair $(x, y) \in \mathscr{X} \times \mathscr{Y}$, let $\Pr(y|x)$ denote the probability that y is output by the channel when x is input to the channel. Then the transition matrix is given by $\mathbf{M}(x, y) = \Pr(y|x)$. The specific structure of the matrix is shown below.

$$\mathbf{M} = \begin{pmatrix} & \vdots & \\ \cdots & \Pr(y|x) & \cdots \\ & \vdots & \end{pmatrix}$$

Next, we look at some specific instances of channels.

Binary Symmetric Channel (BSC)**.** Let $0 \le p \le 1$. The Binary Symmetric Channel with *crossover probability* p or BSC_p is defined as follows. $\mathscr{X} = \mathscr{Y} = \{0, 1\}$. The 2×2 transition matrix can naturally be represented as a bipartite graph where the left vertices correspond to the rows and the right vertices correspond to the columns of the matrix, where $\mathbf{M}(x, y)$ is represented as the weight of the corresponding (x, y) edge. For BSC_p, the graph is illustrated in Figure 6.2.

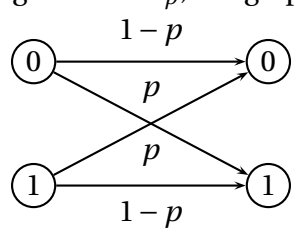


Figure 6.2: Binary Symmetric Channel BSC_p

The corresponding transition matrix would look like this:

$$\begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \end{array}.$$

In other words, every bit is flipped with probability p. We claim that we need to only consider the case when $p \le \frac{1}{2}$, i.e. if we know how to ensure reliable communication over BSC_p for $p \le \frac{1}{2}$, then we can also handle the case of $p > \frac{1}{2}$. (See Exercise 6.1.)

q-ary Symmetric Channel (qSC)**.** We now look at the generalization of BSC_p to alphabets of size $q \ge 2$. Let $0 \le p \le 1 - \frac{1}{q}$. (As with the case of BSC_p, we can assume that $p \le 1 - \frac{1}{q}$– see Exercise 6.2.) The q-ary Symmetric Channel with crossover probability p, or $q\mathrm{SC}_p$, is defined as follows. $\mathscr{X} = \mathscr{Y} = [q]$. The transition matrix $\mathbf{M}$ for $q\mathrm{SC}_p$ is defined as follows.

$$M(x, y) = \begin{cases} 1-p & \text{if } y = x \\ \frac{p}{q-1} & \text{if } y \ne x \end{cases}$$

In other words, every symbol is retained as is at the output with probability $1-p$ and is distorted to each of the $q-1$ possible different symbols with equal probability of $\frac{p}{q-1}$.

Binary Erasure Channel (BEC) In the previous two examples that we saw, $\mathscr{X} = \mathscr{Y}$. However, this might not always be the case.

Let $0 \le \alpha \le 1$. The Binary Erasure Channel with *erasure probability* α (denoted by BEC_α) is defined as follows. $\mathscr{X} = \{0,1\}$ and $\mathscr{Y} = \{0,1,?\}$, where ? denotes an *erasure*. The transition matrix is as follows:

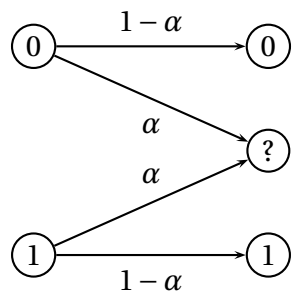


Figure 6.3: Binary Erasure Channel BEC_α

In Figure 6.3 any missing edge represents a transition that occurs with 0 probability. In other words, every bit in BEC_α is erased with probability α (and is left unchanged with probability $1-\alpha$).

Binary Input Additive Gaussian White Noise Channel (BIAGWN)**.** We now look at a channel that is continuous. Let $\sigma \ge 0$. The Binary Input Additive Gaussian White Noise Channel with standard deviation σ or BIAGWN_σ is defined as follows. $\mathscr{X} = \{-1,1\}$ and $\mathscr{Y} = \mathbb{R}$. The noise is modeled by the continuous Gaussian probability distribution function. The Gaussian distribution has lots of nice properties and is a popular choice for modeling noise continuous in nature. Given $(x,y) \in \{-1,1\} \times \mathbb{R}$, the noise $y - x$ is distributed according to the Gaussian distribution of mean of zero and standard deviation of σ. In other words,

$$\Pr\left(y \mid x\right) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\left(\frac{(y-x)^2}{2\sigma^2}\right)\right)$$

6.2.1 Error Correction in Stochastic Noise Models

We now need to revisit the notion of error correction from Section 1.3. Note that unlike Hamming's noise model, we cannot hope to *always* recover the transmitted codeword. As an example, in BSC_p there is always some positive probability that a codeword can be distorted into another codeword during transmission. In such a scenario no decoding algorithm can hope to recover the transmitted codeword. Thus, in some stochastic channels there is always some *decoding error probability* (where the randomness is from the channel noise): see Exercise 6.14 for example channels where one can have zero decoding error probability. However, we would like this error probability to be small for every possible transmitted codeword. More precisely, for every message, we would like the decoding algorithm to recover the transmitted message with probability $1 - f(n)$, where $\lim_{n\to\infty} f(n) \to 0$; that is, $f(n)$ is $o(1)$. Ideally, we would like to have $f(n) = 2^{-\Omega(n)}$. We will refer to $f(n)$ as the decoding error probability.

6.2.2 Shannon's General Theorem

Recall that the big question of interest in this book is the tradeoff between the rate of the code and the fraction of errors that can be corrected. For stochastic noise models that we have seen, it is natural to think of the fraction of errors to be the parameter that governs the amount of error that is introduced by the channel. For example, for BSC_p, we will think of p as the fraction of errors.

Shannon's remarkable theorem on channel coding was to *precisely* identify when reliable transmission is possible over the stochastic noise models that he considered. In particular, for the general framework of noise models, Shannon defined the notion of *capacity*, which is a real number such that reliable communication is possible if and only if the rate is less than the capacity of the channel. In other words, given a noisy channel with capacity C, if information is transmitted at rate R for any $R < C$, then there exists a coding scheme that guarantees negligible probability of miscommunication. On the other hand if $R > C$, then regardless of the chosen coding scheme there will be some message for which the decoding error probability is bounded from below by some constant.

In this chapter, we are going to state (and prove) Shannon's general result for the special case of BSC_p.

6.3 Shannon's Result for BSC_p

We begin with a notation. For the rest of the chapter, we will use the notation $\mathbf{e} \sim \mathrm{BSC}_p$ to denote an error pattern $\mathbf{e}$ that is drawn according to the error distribution induced by BSC_p. We are now ready to state the theorem.

Theorem 6.3.1 (Shannon's Capacity Theorem for BSC)**.** *For real numbers p, ε such that $0 \le p < \frac{1}{2}$ and $0 \le \varepsilon \le \frac{1}{2} - p$, the following statements are true for large enough n:*

1. *There exists a real $\delta > 0$, an encoding function $E : \{0,1\}^k \to \{0,1\}^n$ and a decoding function $D : \{0,1\}^n \to \{0,1\}^k$ where $k \le \lfloor (1 - H(p+\varepsilon))\, n \rfloor$, such that the following holds for every $\mathbf{m} \in \{0,1\}^k$:*
$$\Pr_{\mathbf{e} \sim \mathrm{BSC}_p} [D(E(\mathbf{m}) + \mathbf{e})) \neq \mathbf{m}] \le 2^{-\delta n}.$$

2. *If $k \ge \lceil (1 - H(p) + \varepsilon) n \rceil$ then for every pair of encoding and decoding functions, $E : \{0,1\}^k \to \{0,1\}^n$ and $D : \{0,1\}^n \to \{0,1\}^k$, there exists $\mathbf{m} \in \{0,1\}^k$ such that*
$$\Pr_{\mathbf{e} \sim \mathrm{BSC}_p} [D(E(\mathbf{m}) + \mathbf{e})) \neq \mathbf{m}] \ge \frac{1}{2}.$$

Note that Theorem 6.3.1 implies that the capacity of BSC_p is $1 - H(p)$. It can also be shown that the capacity of $q\mathrm{SC}_p$ and BEC_α are $1 - H_q(p)$ and $1 - \alpha$ respectively. (See Exercises 6.6 and 6.7.)

The entropy function appears in Theorem 6.3.1 due to the same technical reason that it appears in the GV bound: the entropy function allows us to use sufficiently tight bounds on the volume of a Hamming ball (Proposition 3.3.1).

6.3.1 Proof of Converse of Shannon's Capacity Theorem for BSC

We start with the proof of part (2) of Theorem 6.3.1. (Proof of part (1) follows in the next section.)

For the proof we will assume that $p > 0$ (since when $p = 0$, $1 - H(p) + \varepsilon > 1$ and so we have nothing to prove). For the sake of contradiction, assume that the following holds for every $\mathbf{m} \in \{0,1\}^k$:

$$\Pr_{\mathbf{e} \sim \mathrm{BSC}_p} [D(E(\mathbf{m}) + \mathbf{e}) \neq \mathbf{m}] \leq \frac{1}{2}.$$

Define $D_{\mathbf{m}}$ to be the set of received words $\mathbf{y}$ that are decoded to $\mathbf{m}$ by D, that is,

$$D_{\mathbf{m}} = \left\{\mathbf{y} | D(\mathbf{y}) = \mathbf{m}\right\}.$$

The main idea behind the proof is the following: first note that the sets $D_{\mathbf{m}}$ partition the entire space of received words $\{0,1\}^n$ (see Figure 6.4 for an illustration) since D is a function.

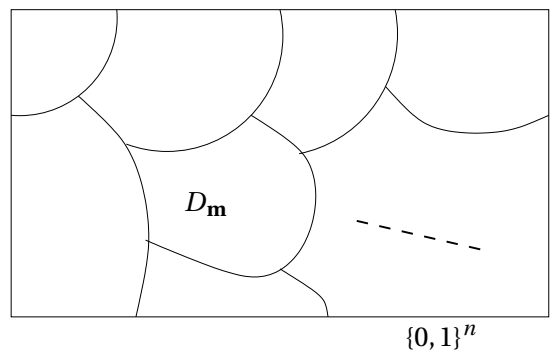


Figure 6.4: The sets $D_{\mathbf{m}}$ partition the ambient space $\{0,1\}^n$.

Next we will argue that since the decoding error probability is at most a $1/2$, then $D_{\mathbf{m}}$ for every $\mathbf{m} \in \{0,1\}^k$ is "large." Then by a simple packing argument, it follows that we cannot have too many distinct $\mathbf{m}$, which we will show implies that $k < (1 - H(p) + \varepsilon)n$: a contradiction. Before we present the details, we outline how we will argue that $D_{\mathbf{m}}$ is large. Let $S_{\mathbf{m}}$ be the shell of radius $[(1-\gamma)pn, (1+\gamma)pn]$ around $E(\mathbf{m})$, that is,

$$S_{\mathbf{m}} = B\left(E(\mathbf{m}), (1+\gamma)pn\right) \setminus B\left(E(\mathbf{m}), (1-\gamma)pn\right).$$

We will set $\gamma > 0$ in terms of ε and p at the end of the proof. (See Figure 6.5 for an illustration.) Then we argue that because the decoding error probability is bounded by $1/2$, most of the received words in the shell $S_{\mathbf{m}}$ are decoded correctly, i.e. they fall in $D_{\mathbf{m}}$. To complete the argument, we show that number of such received words is indeed large enough.

Fix an arbitrary message $\mathbf{m} \in \{0,1\}^k$. Note that by our assumption, the following is true (where from now on we omit the explicit dependence of the probability on the BSC_p noise for clarity):

$$\Pr\left[E(\mathbf{m}) + \mathbf{e} \notin D_{\mathbf{m}}\right] \leq \frac{1}{2}. \tag{6.1}$$

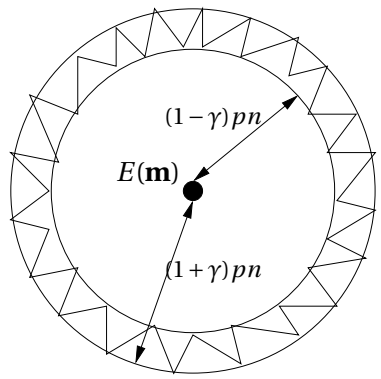


Figure 6.5: The shell $S_{\mathbf{m}}$ of inner radius $(1-\gamma)pn$ and outer radius $(1+\gamma)pn$.

Further, by the (multiplicative) Chernoff bound (Theorem 3.1.6),

$$\Pr\left[E(\mathbf{m})+\mathbf{e}\notin S_{\mathbf{m}}\right]\le 2^{-\Omega(\gamma^2 n)}. \tag{6.2}$$

(6.1) and (6.2) along with the union bound (Proposition 3.1.3) imply the following:

$$\Pr\left[E(\mathbf{m})+\mathbf{e}\notin D_{\mathbf{m}}\cap S_{\mathbf{m}}\right]\le \frac{1}{2}+2^{-\Omega(\gamma^2 n)}.$$

The above in turn implies that

$$\Pr\left[E(\mathbf{m})+\mathbf{e}\in D_{\mathbf{m}}\cap S_{\mathbf{m}}\right]\ge \frac{1}{2}-2^{-\Omega(\gamma^2 n)}\ge \frac{1}{4}, \tag{6.3}$$

where the last inequality holds for large enough n. Next we upper bound the probability above to obtain a lower bound on $|D_{\mathbf{m}}\cap S_{\mathbf{m}}|$.

It is easy to see that

$$\Pr\left[E(\mathbf{m})+\mathbf{e}\in D_{\mathbf{m}}\cap S_{\mathbf{m}}\right]\le |D_{\mathbf{m}}\cap S_{\mathbf{m}}|\cdot p_{\max},$$

where

$$p_{\max}=\max_{\mathbf{y}\in S_{\mathbf{m}}}\Pr\left[E(\mathbf{m})+\mathbf{e}=\mathbf{y}\right]=\max_{d\in[(1-\gamma)pn,(1+\gamma)pn]} p^d(1-p)^{n-d}.$$

In the above, the second equality follows from the fact that all error patterns with the same Hamming weight appear with the same probability when chosen according to BSC_p. Next, note that $p^d(1-p)^{n-d}$ is decreasing in d for $p\le\frac{1}{2}$.[2] Thus, we have

$$p_{\max}=p^{(1-\gamma)pn}(1-p)^{n-(1-\gamma)pn}=\left(\frac{1-p}{p}\right)^{\gamma pn}\cdot p^{pn}(1-p)^{(1-p)n}=\left(\frac{1-p}{p}\right)^{\gamma pn}2^{-nH(p)}.$$

[2]Indeed $p^d(1-p)^{n-d}=(p/(1-p))^d(1-p)^n$ and the bound $p\le\frac{1}{2}$ implies that the first exponent is at most 1, which implies that the expression is decreasing in d.

Thus, we have shown that

$$\Pr\left[E(\mathbf{m})+\mathbf{e}\in D_{\mathbf{m}}\cap S_{\mathbf{m}}\right]\le |D_{\mathbf{m}}\cap S_{\mathbf{m}}|\cdot\left(\frac{1-p}{p}\right)^{\gamma pn}2^{-nH(p)},$$

which, by (6.3), implies that

$$|D_{\mathbf{m}}\cap S|\ge\frac{1}{4}\cdot\left(\frac{1-p}{p}\right)^{-\gamma pn}2^{nH(p)}. \tag{6.4}$$

Next, we consider the following sequence of relations:

$$\begin{aligned}
2^n &= \sum_{\mathbf{m}\in\{0,1\}^k}|D_{\mathbf{m}}| && (6.5)\\
&\ge \sum_{\mathbf{m}\in\{0,1\}^k}|D_{\mathbf{m}}\cap S_{\mathbf{m}}|\\
&\ge \frac{1}{4}\left(\frac{1}{p}-1\right)^{-\gamma pn}\sum_{\mathbf{m}\in\{0,1\}^k}2^{H(p)n} && (6.6)\\
&= 2^{k-2}\cdot 2^{H(p)n-\gamma p\log(1/p-1)n}\\
&> 2^{k+H(p)n-\varepsilon n}. && (6.7)
\end{aligned}$$

In the above, (6.5) follows from the fact that for $\mathbf{m}_1\ne\mathbf{m}_2$, $D_{\mathbf{m}_1}$ and $D_{\mathbf{m}_2}$ are disjoint. (6.6) follows from (6.4). (6.7) follows for large enough n and if we pick $\gamma=\frac{\varepsilon}{2p\log\left(\frac{1}{p}-1\right)}$. (Note that as $0<p<\frac{1}{2}$, $\gamma=\Theta(\varepsilon)$.)

(6.7) implies that $k<(1-H(p)+\varepsilon)n$, which is a contradiction. The proof of part (2) of Theorem 6.3.1 is complete.

Remark 6.3.1. It can be verified that the proof above can also work if the decoding error probability is bounded by $1-2^{-\beta n}$ (instead of the $1/2$ in part (2) of Theorem 6.3.1) for small enough $\beta=\beta(\varepsilon)>0$.

Next, we will prove part (1) of Theorem 6.3.1.

6.3.2 Proof of Positive Part of Shannon's Theorem

Proof Overview. The proof of part (1) of Theorem 6.3.1 will be done by the probabilistic method (Section 3.2). In particular, we randomly select an encoding function $E:\{0,1\}^k\to\{0,1\}^n$. That is, for every $\mathbf{m}\in\{0,1\}^k$ pick $E(\mathbf{m})$ uniformly and independently at random from $\{0,1\}^n$. D will be the maximum likelihood decoding (MLD) function. The proof will have the following two steps:

- (Step 1) For any arbitrary $\mathbf{m}\in\{0,1\}^k$, we will show that for a random choice of E, the probability of failure, over BSC_p noise, is small. This implies the existence of a good encoding function for any arbitrary message.

- (Step 2) We will show a similar result for *all* $\mathbf{m}$. This involves dropping half of the code words.

Note that there are two sources of randomness in the proof:

1. Randomness in the choice of encoding function E and
2. Randomness in the noise.

We stress that the first kind of randomness is for the probabilistic method while the second kind of randomness will contribute to the decoding error probability.

"Proof by picture" of Step 1**.** Before proving part (1) of Theorem 6.3.1, we will provide a pictorial proof of Step 1. We begin by fixing $\mathbf{m} \in \{0,1\}^k$. In Step 1, we need to estimate the following quantity:

$$\mathbb{E}_E\left[\Pr_{\mathbf{e}\sim \mathrm{BSC}_p}\left[D\left(E\left(\mathbf{m}\right)+\mathbf{e}\right)\neq \mathbf{m}\right]\right].$$

By the additive Chernoff bound (Theorem 3.1.6), with all but an exponentially small probability, the received word will be contained in a Hamming ball of radius $\left(p+\varepsilon'\right)n$ (for some $\varepsilon'>0$ that we will choose appropriately). So one can assume that the received word $\mathbf{y}$ with high probability satisfies $\Delta(E(\mathbf{m}),\mathbf{y}) \leq (p+\varepsilon')n$. Given this, pretty much the only thing to do is to estimate the decoding error probability for such a $\mathbf{y}$. Note that by the fact that D is MLD, an error can happen only if there exists another message $\mathbf{m}'$ such that $\Delta(E(\mathbf{m}'),\mathbf{y}) \leq \Delta(E(\mathbf{m}),\mathbf{y})$. The latter event implies that $\Delta(E(\mathbf{m}'),\mathbf{y}) \leq (p+\varepsilon')n$ (see Figure 6.6).

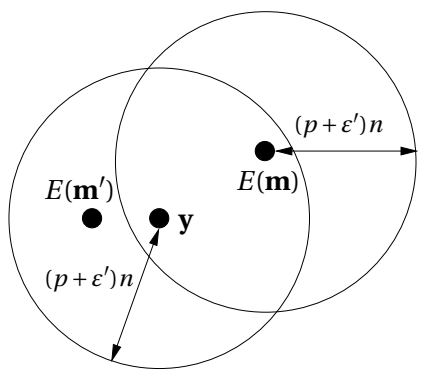


Figure 6.6: Hamming balls of radius $\left(p+\varepsilon'\right)n$ and centers $E\left(\mathbf{m}\right)$ and $\mathbf{y}$) illustrates Step 1 in the proof of part (1) of Shannon's capacity theorem for the BSC.

Thus, the decoding error probability is upper bounded by

$$\Pr_{\mathbf{e}\sim \mathrm{BSC}_p}\left[E\left(\mathbf{m}'\right)\in B\left(\mathbf{y},\left(p+\varepsilon'\right)n\right)\right]=\frac{Vol_2\left(\left(p+\varepsilon'\right)n,n\right)}{2^n}\approx\frac{2^{H(p)n}}{2^n},$$

where the last step follows from Proposition 3.3.1. Finally, by the union bound (Proposition 3.1.3), the existence of such a "bad" $\mathbf{m}'$ is upper bounded by $\approx \frac{2^k 2^{nH(p)}}{2^n}$, which by our choice of k is $2^{-\Omega(n)}$, as desired.

The Details. For notational convenience, we will use $\mathbf{y}$ and $E(\mathbf{m}) + \mathbf{e}$ interchangeably:

$$\mathbf{y} = E(\mathbf{m}) + \mathbf{e}.$$

That is, $\mathbf{y}$ is the received word when $E(\mathbf{m})$ is transmitted and $\mathbf{e}$ is the error pattern.

We start the proof by restating the decoding error probability in part (1) of Shannon's capacity theorem for BSC_p (Theorem 6.3.1) by breaking up the quantity into two sums:

$$\begin{aligned}\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}[D(E(\mathbf{m})+\mathbf{e}) \neq \mathbf{m}] = &\sum_{\mathbf{y}\in B(E(\mathbf{m}),(p+\varepsilon')n)} \Pr\left[\mathbf{y}|E(\mathbf{m})\right] \cdot \mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}} \\ &+ \sum_{\mathbf{y}\notin B(E(\mathbf{m}),(p+\varepsilon')n)} \Pr\left[\mathbf{y}|E(\mathbf{m})\right] \cdot \mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}},\end{aligned}$$

where $\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}$ is the indicator function for the event that $D(\mathbf{y}) \neq \mathbf{m}$ *given* that $E(\mathbf{m})$ was the transmitted codeword and we use $\mathbf{y}|E(\mathbf{m})$ as a shorthand for "$\mathbf{y}$ is the received word given that $E(\mathbf{m})$ was the transmitted codeword." As $\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}} \leq 1$ (since it takes a value in $\{0,1\}$) and by the (additive) Chernoff bound (Theorem 3.1.6) we have

$$\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}[D(E(\mathbf{m})+\mathbf{e}) \neq \mathbf{m}] \leq \sum_{\mathbf{y}\in B(E(\mathbf{m}),(p+\varepsilon')n)} \Pr\left[\mathbf{y}|E(\mathbf{m})\right] \cdot \mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}} + e^{-(\varepsilon')^2 n/2}.$$

In order to apply the probabilistic method (Section 3.2), we will analyze the expectation (over the random choice of E) of the decoding error probability, which by the upper bound above satisfies

$$\mathbb{E}_E\left[\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}[D(E(\mathbf{m})+\mathbf{e}) \neq \mathbf{m}]\right] \leq e^{-(\varepsilon')^2 n/2} + \sum_{\mathbf{y}\in B(E(\mathbf{m}),(p+\varepsilon')n)} \Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[\mathbf{y}|E(\mathbf{m})\right] \cdot \mathbb{E}_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}\right]. \quad (6.8)$$

In the above, we used linearity of expectation (Proposition 3.1.2) and the fact that the distributions on $\mathbf{e}$ and E are independent.

Next, for a fixed received word $\mathbf{y}$ and the transmitted codeword $E(\mathbf{m})$ such that $\Delta(\mathbf{y}, E(\mathbf{m})) \leq (p+\varepsilon')n$ we estimate $\mathbb{E}_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}\right]$. Since D is MLD, we have

$$\mathbb{E}_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}\right] = \Pr_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}|E(\mathbf{m})\right] \leq \sum_{\mathbf{m}'\neq\mathbf{m}} \Pr\left[\Delta\left(E\left(\mathbf{m}'\right),\mathbf{y}\right) \leq \Delta\left(E(\mathbf{m}),\mathbf{y}\right)|E(\mathbf{m})\right], \quad (6.9)$$

where in the above "$|E(\mathbf{m})$" is short for "being conditioned on $E(\mathbf{m})$ being transmitted" and the inequality follows from the union bound (Proposition 3.1.3) and the fact that D is MLD.

Noting that $\Delta(E(\mathbf{m}'),\mathbf{y}) \le \Delta(E(\mathbf{m}),\mathbf{y}) \le (p+\varepsilon')n$ (see Figure 6.6), by (6.9) we have

$$\begin{aligned}
\mathbb{E}_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}\right] &\le \sum_{\mathbf{m}'\neq\mathbf{m}} \Pr\left[E\left(\mathbf{m}'\right)\in B\left(\mathbf{y},\left(p+\varepsilon'\right)n\right)|E(\mathbf{m})\right] \\
&= \sum_{\mathbf{m}'\neq\mathbf{m}} \frac{\left|B\left(\mathbf{y},\left(p+\varepsilon'\right)n\right)\right|}{2^n} && (6.10)\\
&\le \sum_{\mathbf{m}'\neq\mathbf{m}} \frac{2^{H(p+\varepsilon')n}}{2^n} && (6.11)\\
&< 2^k\cdot 2^{-n(1-H(p+\varepsilon'))} \\
&\le 2^{n(1-H(p+\varepsilon))-n(1-H(p+\varepsilon'))} && (6.12)\\
&= 2^{-n(H(p+\varepsilon)-H(p+\varepsilon'))}. && (6.13)
\end{aligned}$$

In the above, (6.10) follows from the fact that the choice for $E(\mathbf{m}')$ is independent of $E(\mathbf{m})$. (6.11) follows from the upper bound on the volume of a Hamming ball (Proposition 3.3.1), while (6.12) follows from our choice of k.

Using (6.13) in (6.8), we get

$$\begin{aligned}
\mathbb{E}_E\left[\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[D(E(\mathbf{m})+\mathbf{e})\neq\mathbf{m}\right]\right] &\le e^{-(\varepsilon')^2 n/2} + 2^{-n(H(p+\varepsilon)-H(p+\varepsilon'))} \sum_{\mathbf{y}\in B(E(\mathbf{m}),(p+\varepsilon')n)} \Pr\left[\mathbf{y}|E(\mathbf{m})\right] \\
&\le e^{-(\varepsilon')^2 n/2} + 2^{-n(H(p+\varepsilon)-H(p+\varepsilon'))} \le 2^{-\delta' n}, && (6.14)
\end{aligned}$$

where the second inequality follows from the fact that

$$\sum_{\mathbf{y}\in B(E(\mathbf{m}),(p+\varepsilon')n)} \Pr\left[\mathbf{y}|E(\mathbf{m})\right] \le \sum_{\mathbf{y}\in\{0,1\}^n} \Pr\left[\mathbf{y}|E(\mathbf{m})\right] = 1$$

and the last inequality follows for large enough n, say $\varepsilon' = \varepsilon/2$ and by picking $\delta' > 0$ to be small enough. (See Exercise 6.3.)

Thus, we have shown that for any arbitrary $\mathbf{m}$ the average (over the choices of E) decoding error probability is small. However, we still need to show that the decoding error probability is exponentially small for *all* messages *simultaneously*. Towards this end, as the bound holds for each $\mathbf{m}$, we have

$$\mathbb{E}_{\mathbf{m}}\left[\mathbb{E}_E\left[\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[D\left(E\left(\mathbf{m}\right)+\mathbf{e}\right)\neq\mathbf{m}\right]\right]\right] \le 2^{-\delta' n}.$$

The order of the summation in the expectation with respect to $\mathbf{m}$ and the summation in the expectation with respect to the choice of E can be switched (as the probability distributions are defined over different domains), resulting in the following expression:

$$\mathbb{E}_E\left[\mathbb{E}_{\mathbf{m}}\left[\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[D\left(E\left(\mathbf{m}\right)+\mathbf{e}\right)\neq\mathbf{m}\right]\right]\right] \le 2^{-\delta' n}.$$

By the probabilistic method, there exists an encoding function E^* (and a corresponding decoding function D^*) such that

$$\mathbb{E}_{\mathbf{m}}\left[\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[D^*\left(E^*\left(\mathbf{m}\right)+\mathbf{e}\right)\neq\mathbf{m}\right]\right]\leq 2^{-\delta' n}. \tag{6.15}$$

(6.15) implies that the *average* decoding error probability is exponentially small. However, recall we need to show that the *maximum* decoding error probability is small. To achieve such a result, we will throw away half of the messages, i.e. *expurgate* the code. In particular, we will order the messages in decreasing order of their decoding error probability and then drop the top half. We claim that the maximum decoding error probability for the remaining messages is $2\cdot 2^{-\delta' n}$. Next, we present the details.

From Average to Worst-Case Decoding Error Probability. We begin with the following "averaging" argument.

Claim 6.3.2. *Let the messages be ordered as* $\mathbf{m}_1,\mathbf{m}_2,\ldots,\mathbf{m}_{2^k}$ *and define*

$$P_i=\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[D(E(\mathbf{m}_i)+\mathbf{e})\neq\mathbf{m}_i\right].$$

Assume that $P_1\leq P_2\leq\ldots\leq P_{2^k}$ *and (6.15) holds, then* $P_{2^{k-1}}\leq 2\cdot 2^{-\delta' n}$

Proof. By the definition of P_i,

$$\begin{aligned}\frac{1}{2^k}\sum_{i=1}^{2^k}P_i&=\mathbb{E}_{\mathbf{m}}\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[D(E(\mathbf{m})+\mathbf{e})\neq\mathbf{m}\right]\\&\leq 2^{-\delta' n},\end{aligned} \tag{6.16}$$

where (6.16) follows from (6.15). For the sake of contradiction assume that

$$P_{2^{k-1}}>2\cdot 2^{-\delta' n}. \tag{6.17}$$

So,

$$\frac{1}{2^k}\sum_{i=1}^{2^k}P_i\geq\frac{1}{2^k}\sum_{i=2^{k-1}+1}^{2^k}P_i \tag{6.18}$$

$$>\frac{2\cdot 2^{-\delta' n}\cdot 2^{k-1}}{2^k} \tag{6.19}$$

$$>2^{-\delta' n}, \tag{6.20}$$

where (6.18) follows by dropping half the summands from the sum. (6.19) follows from (6.17) and the assumption on the sortedness of P_i. The proof is now complete by noting that (6.20) contradicts (6.16). □

Thus, our final code will have $\mathbf{m}_1, \ldots, \mathbf{m}_{2^{k-1}}$ as its messages and hence, has dimension $k' = k-1$. Define $\delta = \delta' + \frac{1}{n}$. In the new code, maximum error probability is at most $2^{-\delta n}$. Also if we picked $k \le \lfloor (1 - H(p+\varepsilon))\, n \rfloor + 1$, then $k' \le \lfloor (1 - H(p+\varepsilon))\, n \rfloor$, as required. This completes the proof of Theorem 6.3.1.

We have shown that a random code can achieve capacity. However, we do not know of even a succinct representation of general codes. A natural question to ask is if random linear codes can achieve the capacity of BSC_p. The answer is yes: see Exercise 6.4.

For linear code, representation and encoding are efficient. But the proof does not give an explicit construction. Intuitively, it is clear that since Shannon's proof uses a random code it does not present an 'explicit' construction. Below, we formally define what we mean by an explicit construction.

Definition 6.3.1. A code C of block length n is called *explicit* if there exists a poly(n)-time algorithm that computes a succinct description of C given n. For linear codes, such a succinct description could be a generator matrix or a parity check matrix.

We will also need the following stronger notion of an explicitness:

Definition 6.3.2. A linear $[n, k]$ code C is called *strongly explicit*, if given any index pair $(i, j) \in [k] \times [n]$, there is a poly($\log n$) time algorithm that outputs $G_{i,j}$, where G is a generator matrix of C.

Further, Shannon's proof uses MLD for which only exponential time implementations are known. Thus, the biggest question left unsolved by Shannon's work is the following.

Question 6.3.1. *Can we come up with an explicit construction of a code of rate* $1 - H(p+\varepsilon)$ *with efficient decoding and encoding algorithms that achieves reliable communication over* BSC_p*?*

As a baby step towards the resolution of the above question, one can ask the following question:

Question 6.3.2. *Can we come up with an explicit construction with* $R > 0$ *and* $p > 0$*?*

Note that the question above is similar to Question 2.7.1 in Hamming's world. See Exercise 6.13 for an affirmative answer.

6.4 Hamming vs. Shannon

As a brief interlude, let us compare the salient features of the works of Hamming and Shannon that we have seen so far:

HAMMING	SHANNON
Focus on codewords itself	Directly deals with encoding and decoding functions
Looked at explicit codes	Not explicit at all
Fundamental trade off: rate vs. distance (easier to get a handle on this)	Fundamental trade off: rate vs. error
Worst case errors	Stochastic errors

Intuitively achieving positive results in the Hamming world is harder than achieving positive results in Shannon's world. The reason is that the adversary in Shannon's world (e.g. BSC_p) is much weaker than the worst-case adversary in Hamming's world (say for bits). We make this intuition (somewhat) precise as follows:

Proposition 6.4.1. *Let $0 \le p < \frac{1}{2}$ and $0 < \varepsilon \le \frac{1}{2} - p$. If an algorithm A can handle $p + \varepsilon$ fraction of worst case errors, then it can be used for reliable communication over* BSC_p

Proof. By the additive Chernoff bound (Theorem 3.1.6), with probability $\ge 1 - e^{\frac{-\varepsilon^2 n}{2}}$, the fraction of errors in BSC_p is $\le p + \varepsilon$. Then by assumption on A, it can be used to recover the transmitted message. □

Note that the above result implies that one can have reliable transmission over BSC_p with any code of relative distance $2p + \varepsilon$ (for any $\varepsilon > 0$).

A much weaker converse of Proposition 6.4.1 is also true. More precisely, if the decoding error probability is exponentially small for the BSC, then the corresponding code must have constant relative distance (though this distance does not come even close to achieving say the Gilbert-Varshamov bound). For more see Exercise 6.5.

6.5 Exercises

Exercise 6.1. Let (E, D) be a pair of encoder and decoder that allows for successful transmission over BSC_p for every $p \le \frac{1}{2}$. Then there exists a pair (E', D') that allows for successful transmission over $\mathrm{BSC}_{p'}$ for any $p' > 1/2$. If D is (deterministic) polynomial time algorithm, then D' also has to be a (deterministic) polynomial time algorithm.

Exercise 6.2. Let (E, D) be a pair of encoder and decoder that allows for successful transmission over $q\mathrm{SC}_p$ for every $p \le 1 - \frac{1}{q}$. Then there exists a pair (E', D') that allows for successful transmission over $q\mathrm{SC}_{p'}$ for any $p' > 1 - \frac{1}{2}$. If D is polynomial time algorithm, then D' also has to be a polynomial time algorithm though D' can be a randomized algorithm even if D is deterministic.[3]

Exercise 6.3. Argue that in the positive part of Theorem 6.3.1, one can pick $\delta = \Theta(\varepsilon^2)$. That is, for $0 \le p < 1/2$ and small enough ε, there exist codes of rate $1 - H(p) - \varepsilon$ and block length n that can be decoded with error probability at most $2^{-\Theta(\varepsilon^2)n}$ over BSC_p.

[3] A randomized D' means that given a received word $\mathbf{y}$ the algorithm can use random coins and the decoding error probability is over both the randomness from its internal coin tosses as well as the randomness from the channel.

Exercise 6.4. Prove that there exists linear codes that achieve the BSC_p capacity. (Note that in Section 6.3 we argued that there exists not necessarily a linear code that achieves the capacity.)

Hint: Modify the argument in Section 6.3: in some sense the proof is easier.

Exercise 6.5. Prove that for communication on BSC_p, if an encoding function E achieves a maximum decoding error probability (taken over all messages) that is exponentially small, i.e., at most $2^{-\gamma n}$ for some $\gamma > 0$, then there exists a $\delta = \delta(\gamma, p) > 0$ such that the code defined by E has relative distance at least δ. In other words, good distance is *necessary* for exponentially small maximum decoding error probability.

Exercise 6.6. Prove that the capacity of the $q\mathrm{SC}_p$ is $1 - H_q(p)$.

Exercise 6.7. The binary erasure channel with erasure probability α has capacity $1-\alpha$. In this problem, you will prove this result (and its generalization to larger alphabets) via a sequence of smaller results.

1. For positive integers $k \le n$, show that less than a fraction q^{k-n} of the $k \times n$ matrices G over $\mathbb{F}_q$ fail to generate a linear code of block length n and dimension k. (Or equivalently, except with probability less than q^{k-n}, the rank of a random $k \times n$ matrix G over $\mathbb{F}_q$ is k.)

   Hint: Try out the obvious greedy algorithm to construct a $k \times n$ matrix of rank k. You will see that you will have many choices every step: from this compute (a lower bound on) the number of full rank matrices that can be generated by this algorithm.

2. Consider the q-ary erasure channel with erasure probability α ($q\mathrm{EC}_\alpha$, for some α, $0 \le \alpha \le 1$): the input to this channel is a field element $x \in \mathbb{F}_q$, and the output is x with probability $1-\alpha$, and an erasure '?' with probability α. For a linear code C generated by an $k \times n$ matrix G over $\mathbb{F}_q$, let $D : (\mathbb{F}_q \cup \{?\})^n \to C \cup \{\mathsf{fail}\}$ be the following decoder:

   $$D(\mathbf{y}) = \begin{cases} \mathbf{c} & \text{if } \mathbf{y} \text{ agrees with exactly one } \mathbf{c} \in C \text{ on the unerased entries in } \mathbb{F}_q \\ \mathsf{fail} & \text{otherwise} \end{cases}$$

   For a set $J \subseteq \{1, 2, \ldots, n\}$, let $P_{\mathrm{err}}(G|J)$ be the probability (over the channel noise and choice of a random message) that D outputs fail conditioned on the erasures being indexed by J. Prove that the average value of $P_{\mathrm{err}}(G|J)$ taken over all $G \in \mathbb{F}_q^{k \times n}$ is less than $q^{k-n+|J|}$.

3. Let $P_{\mathrm{err}}(G)$ be the decoding error probability of the decoder D for communication using the code generated by G on the $q\mathrm{EC}_\alpha$. Show that when $k = Rn$ for $R < 1-\alpha$, the average value of $P_{\mathrm{err}}(G)$ over all $k \times n$ matrices G over $\mathbb{F}_q$ is exponentially small in n.

4. Conclude that one can reliably communicate on the $q\mathrm{EC}_\alpha$ at any rate less than $1-\alpha$ using a linear code.

Exercise 6.8. Consider a binary channel whose input/output alphabet is $\{0, 1\}$, where a 0 is transmitted faithfully as a 0 (with probability 1), but a 1 is transmitted as a 0 with probability $\frac{1}{2}$ and a 1 with probability $1/2$. Compute the capacity of this channel.

Hint: This can be proved from scratch using only simple probabilistic facts already stated/used in the book.

Exercise 6.9. Argue that Reed-Solomon codes from Chapter 5 are strongly explicit codes (as in Definition 6.3.2).

Exercise 6.10. In this problem we will prove a special case of the source coding theorem. For any $0 \le p \le 1/2$, let $\mathscr{D}(p)$ be the distribution on $\{0,1\}^n$, where each of the n bits are picked independently to be 1 with probability p and 0 otherwise. Argue that for every $\varepsilon > 0$, strings from $\mathscr{D}(p)$ can be compressed with $H(p+\varepsilon)\cdot n$ bits for large enough n.

More precisely show that for any constant $0 \le p \le 1/2$ and every $\varepsilon > 0$, for large enough n there exists an encoding (or compression) function $E : \{0,1\}^n \to \{0,1\}^*$ and a decoding (or decompression) function $D : \{0,1\}^* \to \{0,1\}^n$ such that[4]

1. For every $\mathbf{x} \in \{0,1\}^n$, $D(E(\mathbf{x})) = \mathbf{x}$, and

2. $\mathbb{E}_{\mathbf{x} \leftarrow \mathscr{D}(p)}[|E(\mathbf{x})|] \le H(p+\varepsilon)\cdot n$, where we use $|E(\mathbf{x})|$ to denote the length of the string $E(\mathbf{x})$. In other words, the *compression rate* is $H(p+\varepsilon)$.

Hint: Handle the "typical" strings from $\mathscr{D}$ and non-typical strings separately.

Exercise 6.11. Show that if there is a constructive solution to Shannon's channel coding theorem with E being a linear map, then there is a constructive solution to Shannon's source coding theorem in the case where the source produces a sequence of independent bits of bias p.

More precisely, let (E, D) be an encoding and decoding pairs that allows for reliable communication over BSC_p with exponentially small decoding error and E is a linear map with rate $1 - H(p) - \varepsilon$. Then there exists a compressing and decompressing pair (E', D') that allows for compression rate $H(p) + \varepsilon$ (where compression rate is as defined in part 2 in Exercise 6.10). The decompression algorithm D' can be randomized and is allowed exponentially small error probability (where the probability can be taken over both the internal randomness of D' and $\mathscr{D}(p)$). Finally if (E, D) are both polynomial time algorithms, then (E', D') have to be polynomial time algorithms too.

Exercise 6.12. Consider a Markovian source of bits, where the source consists of a 6-cycle with three successive vertices outputting 0, and three successive vertices outputting 1, with the probability of either going left (or right) from any vertex is exactly 1/2. More precisely, consider a graph with six vertices $v_0, v_1, \ldots, v_5$ such that there exists an edge $(v_i, v_{(i+1) \bmod 6})$ for every $0 \le i \le 5$. Further the vertices v_i for $0 \le i < 3$ are labeled $\ell(v_i) = 0$ and vertices v_j for $3 \le j < 6$ are labeled $\ell(v_j) = 1$. Strings are generated from this source as follows: one starts with some *start* vertex u_0 (which is one of the v_i's): i.e. the start *state* is u_0. Any any point of time if the current state if u, then the source outputs $\ell(u)$. Then with probability 1/2 the states moves to each of the two neighbors of u.

Compute the optimal compression rate of this source.

[4] We use $\{0,1\}^*$ to denote the set of all binary strings.

Hint: Compress "state diagram" to a minimum and then make some basic observations to compress the source information.

Exercise 6.13. Given codes C_1 and C_2 with encoding functions $E_1 : \{0,1\}^{k_1} \to \{0,1\}^{n_1}$ and $E_2 : \{0,1\}^{k_2} \to \{0,1\}^{n_2}$ let $E_1 \otimes E_2 : \{0,1\}^{k_1 \times k_2} \to \{0,1\}^{n_1 \times n_2}$ be the encoding function obtained as follows: view a message $\mathbf{m}$ as a $k_1 \times k_2$ matrix. Encode the columns of $\mathbf{m}$ individually using the function E_1 to get an $n_1 \times k_2$ matrix $\mathbf{m}'$. Now encode the rows of $\mathbf{m}'$ individually using E_2 to get an $n_1 \times n_2$ matrix that is the final encoding under $E_1 \otimes E_2$ of $\mathbf{m}$. Let $C_1 \otimes C_2$ be the code associated with $E_1 \otimes E_2$ (recall Exercise 2.19).

For $i \geq 3$, let H_i denote the $[2^i - 1, 2^i - i - 1, 3]_2$-Hamming code. Let $C_i = H_i \otimes C_{i-1}$ with $C_3 = H_3$ be a new family of codes.

1. Give a lower bound on the relative minimum distance of C_i. Does it go to zero as $i \to \infty$?

2. Give a lower bound on the rate of C_i. Does it go to zero as $i \to \infty$?

3. Consider the following simple decoding algorithm for C_i: Decode the rows of the rec'd vector recursively using the decoding algorithm for C_{i-1}. Then decode each column according to the Hamming decoding algorithm (e.g. Algorithm 5). Let δ_i denote the probability of decoding error of this algorithm on the BSC_p. Show that there exists a $p > 0$ such that $\delta_i \to 0$ as $i \to \infty$.

   Hint: First show that $\delta_i \leq 4^i \delta_{i-1}^2$.

Exercise 6.14. We consider the problem of determining the best possible rate of transmission on a stochastic memoryless channel with *zero decoding error probability*. Recall that a memoryless stochastic channel is specified by a transition matrix $\mathbf{M}$ s.t. $\mathbf{M}(x, y)$ denotes the probability of y being received if x was transmitted over the channel. Further, the noise acts independently on each transmitted symbol. Let $\mathbb{D}$ denote the input alphabet. Let $R(\mathbf{M})$ denote the best possible rate for a code C such that there exists a decoder D such that for every $\mathbf{c} \in C$, $\Pr[D(\mathbf{y}) \neq \mathbf{c}] = 0$, where $\mathbf{y}$ is picked according to the distribution induced by $\mathbf{M}$ when $\mathbf{c}$ is transmitted over the channel (i.e. the probability that $\mathbf{y}$ is a received word is exactly $\prod_{i=1}^{n} \mathbf{M}(c_i, y_i)$ where C has block length n). In this exercise we will derive an alternate characterization of $R(\mathbf{M})$.

We begin with some definitions related to graphs $\mathcal{G} = (V, E)$. An *independent set* S of $\mathcal{G}$ is a subset $S \subseteq V$ such that there is no edge contained in S, i.e. for every $u \neq v \in S$, $(u, v) \notin E$. For a given graph $\mathcal{G}$, we use $\alpha(\mathcal{G})$ to denote the size of largest independent set in $\mathcal{G}$. Further, given an integer $n \geq 1$, the *n-fold product* of $\mathcal{G}$, which we will denote by $\mathcal{G}^n$, is defined as follows: $\mathcal{G}^n = (V^n, E')$, where $((u_1, \ldots, u_n), (v_1, \ldots, v_n)) \in E'$ if and only if for every $i \in [n]$ either $u_i = v_i$ or $(u_i, v_i) \in E$.

Finally, define a *confusion graph* $\mathcal{G}_{\mathbf{M}} = (V, E)$ as follows. The set of vertices $V = \mathbb{D}$ and for every $x_1 \neq x_2 \in \mathbb{D}$, $(x, y) \in E$ if and only if there exists a y such that $\mathbf{M}(x_1, y) \neq 0$ and $\mathbf{M}(x_2, y) \neq 0$.

1. Prove that
$$R(\mathbf{M}) = \lim_{n\to\infty} \frac{1}{n} \cdot \log_{|\mathbb{D}|} \left(\alpha\left(\mathscr{G}_{\mathbf{M}}^n\right)\right).^5 \tag{6.21}$$

2. A *clique cover* for a graph $\mathscr{G} = (V, E)$ is a partition of the vertices $V = \{V_1, \ldots, V_c\}$ (i.e. V_i and V_j are disjoint for every $i \neq j \in [c]$ and $\cup_i V_i = V$) such that the graph induced on V_i is a *complete graph* (i.e. for every $i \in [c]$ and $x \neq y \in V_i$, we have $(x, y) \in E$). We call c to be the *size* of the clique cover $V_1, \ldots, V_c$. Finally, define $\nu(\mathscr{G})$ to be the size of the smallest clique cover for $\mathscr{G}$. Argue that
$$\alpha(\mathscr{G})^n \leq \alpha(\mathscr{G}^n) \leq \nu(\mathscr{G})^n.$$
Conclude that
$$\log_{|\mathbb{D}|} \alpha(\mathscr{G}) \leq R(\mathbf{M}) \leq \log_{|\mathbb{D}|} \nu(\mathscr{G}). \tag{6.22}$$

3. Consider any transition matrix $\mathbf{M}$ such that the corresponding graph $\mathscr{C}_4 = \mathscr{G}_{\mathbf{M}}$ is a 4-cycle (i.e. the graph $(\{0,1,2,3\}, E)$ where $(i, i+1 \mod 4) \in E$ for every $0 \leq i \leq 3$). Using part 2 or otherwise, argue that $R(\mathbf{M}) = \frac{1}{2}$.

4. Consider any transition matrix $\mathbf{M}$ such that the corresponding graph $\mathscr{C}_5 = \mathscr{G}_{\mathbf{M}}$ is a 5-cycle (i.e. the graph $(\{0,1,2,4\}, E)$ where $(i, i+1 \mod 5) \in E$ for every $0 \leq i \leq 4$). Using part 2 or otherwise, argue that $R(\mathbf{M}) \geq \frac{1}{2} \cdot \log_5 5$. (This lower bound is known to be tight: see Section 6.6 for more.)

6.6 Bibliographic Notes

Shannon's results that were discussed in this chapter appeared in his seminal 1948 paper [89]. All the channels mentioned in this chapter were considered by Shannon except for the *BEC* channel, which was introduced by Elias.

The proof method used to prove Shannon's result for BSC_p has its own name– "random coding with expurgation."

Elias [27] answered Question 6.3.2 (the argument in Exercise 6.13 is due to him).

[5] In literature, $R(\mathbf{M})$ is defined with $\log_{|\mathbb{D}|}$ replaced by $\log_2$. We used the definition in (6.21) to be consistent with our definition of capacity of a noisy channel. See Section 6.6 for more.

Chapter 7

Bridging the Gap Between Shannon and Hamming: List Decoding

In Section 6.4, we made a qualitative comparison between Hamming and Shannon's world. We start this chapter by doing a more quantitative comparison between the two threads of coding theory. In Section 7.2 we introduce the notion of list decoding, which potentially allows us to go beyond the (quantitative) results of Hamming and approach those of Shannon's. Then in Section 7.3, we show how list decoding allows us to go beyond half the distance bound for any code. Section 7.4 proves the optimal trade-off between rate and fraction of correctable errors via list decoding. Finally, in Section 7.5, we formalize why list decoding could be a useful primitive in practical communication setups.

7.1 Hamming versus Shannon: part II

Let us compare Hamming and Shannon theories in terms of the asymptotic bounds we have seen so far (recall rate $R = \frac{k}{n}$ and relative distance $\delta = \frac{d}{n}$).

- Hamming theory: Can correct $\leq \frac{\delta}{2}$ fraction of worse case errors for codes of relative distance δ. By the Singleton bound (Theorem 4.3.1),

  $$\delta \leq 1 - R,$$

  which by Proposition 1.4.1 implies that p fraction of errors can be corrected has to satisfy

  $$p \leq \frac{1-R}{2}.$$

  The above can be achieved via efficient decoding algorithms for example for Reed-Solomon codes (we'll see this later in the book).

- Shannon theory: In $q\mathrm{SC}_p$, for $0 \leq p < 1 - 1/q$, we can have reliable communication with $R < 1 - H_q(p)$. It can be shown that

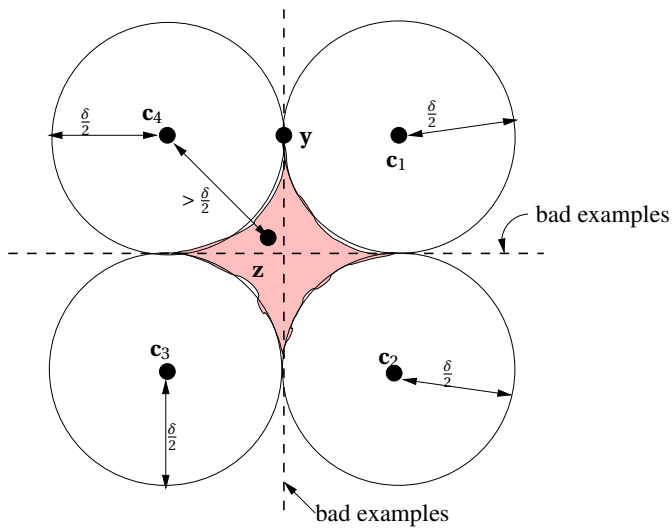


Figure 7.1: In this example vectors are embedded into Euclidean space such that the Euclidean distance between two mapped points is the same as the Hamming distance between vectors. $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4$ are codewords. The dotted lines contain the "bad examples," that is, the received words for which unique decoding is not possible.

1. $1 - H_q(p) \leq 1 - p$ (this is left as an exercise); and
2. $1 - H_q(p) \geq 1 - p - \varepsilon$, for large enough q– in particular, $q = 2^{\Omega(1/\varepsilon)}$ (Proposition 3.3.2).

Thus, we can have reliable communication with $p \sim 1 - R$ on $q\mathrm{SC}_p$ for large enough q.

There is a gap between Shannon and Hamming world: one can correct twice as many errors in Shannon's world. One natural question to ask is whether we can somehow "bridge" this gap. Towards this end, we will now re-visit the the bad example for unique decoding (Figure 1.3) and consider an extension of the bad example as shown in Figure 7.1.

Recall that $\mathbf{y}$ and the codewords $\mathbf{c}_1$ and $\mathbf{c}_2$ form the bad example for unique decoding that we have already seen before. Recall that for this particular received word we can not do error recovery by unique decoding since there are two codewords $\mathbf{c}_1$ and $\mathbf{c}_2$ having the same distance $\frac{\delta}{2}$ from vector $\mathbf{y}$. On the other hand, the received word $\mathbf{z}$ has an unique codeword $\mathbf{c}_1$ with distance $p > \frac{\delta}{2}$. However, unique decoding does not allow for error recovery from $\mathbf{z}$. This is because by definition of unique decoding, the decoder must be able to recover from *every* error pattern (with a given Hamming weight bound). Thus, by Proposition 1.4.1, the decoded codeword cannot have relative Hamming distance larger than $\delta/2$ from the received word. In this example, because of the received word $\mathbf{y}$, unique decoding gives up on the opportunity to decode $\mathbf{z}$.

Let us consider the example in Figure 7.1 for the binary case. It can be shown that the number of vectors in dotted lines is insignificant compared to volume of shaded area (for large enough block length of the code). The volume of all Hamming balls of radius $\frac{\delta}{2}$ around all the

2^k codewords is roughly equal to:

$$2^k 2^{nH(\frac{\delta}{2})},$$

which implies that the volume of the shaded area (without the dotted lines) is approximately equal to:

$$2^n - 2^k 2^{nH(\frac{\delta}{2})}.$$

In other words, the volume when expressed as a fraction of the volume of the ambient space is roughly:

$$1 - 2^{-n(1-H(\frac{\delta}{2})-R)}, \tag{7.1}$$

where $k = Rn$ and by the Hamming bound (Theorem 1.3) $R \leq 1 - H(\frac{\delta}{2})$. If $R < 1 - H(\frac{\delta}{2})$ then second term of (7.1) is very small. Therefore the number of vectors in shaded area (without the bad examples) is almost all of the ambient space. Note that by the stringent condition on unique decoding none of these received words can be decoded (even though for such received words there is a unique closest codeword). Thus, in order to be able to decode such received vectors, we need to relax the notion of unique decoding. We will consider such a relaxation called *list decoding* next.

7.2 List Decoding

The new notion of decoding that we will discuss is called *list decoding* as the decoder is allowed to output a list of answers. We now formally define (the combinatorial version of) list decoding:

Definition 7.2.1. Given $0 \leq \rho \leq 1, L \geq 1$, a code $C \subseteq \Sigma^n$ is (ρ, L)-list decodable if for every received word $\mathbf{y} \in \Sigma^n$,

$$\left|\{c \in C | \Delta(\mathbf{y}, c) \leq \rho n\}\right| \leq L$$

Given an error parameter ρ, a code C and a received word $\mathbf{y}$, a list-decoding algorithm should output all codewords in C that are within (relative) Hamming distance ρ from $\mathbf{y}$. Note that if the fraction of errors that occurred during transmission is at most ρ then the transmitted codeword is *guaranteed* to be in the output list. Further, note that if C is (ρ, L)-list decodable then the algorithm will always output at most L codewords for any received word. In other words, for efficient list-decoding algorithm, L should be a polynomial in the block length n (as otherwise the algorithm will have to output a super-polynomial number of codewords and hence, cannot have a polynomial running time). Thus, the restriction of L being at most some polynomial in n is an *a priori* requirement enforced by the fact that we are interested in efficient polynomial time decoding algorithms. Another reason for insisting on a bound on L is that otherwise the decoding problem can become trivial: for example, one can output all the codewords in the code. Finally, it is worthwhile to note that one can always have an exponential time list-decoding algorithm: go through all the codewords in the code and pick the ones that are within ρ (relative) Hamming distance of the received word.

Note that in the communication setup, we need to recover the transmitted message. In such a scenario, outputting a list might not be useful. There are two ways to get around this "problem":

1. Declare a decoding error if list size > 1. Note that this generalizes unique decoding (as when the number of errors is at most half the distance of the code then there is a unique codeword and hence, the list size will be at most one). However, the gain over unique decoding would be substantial only if for most error patterns (of weight significantly more than half the distance of the code) the output list size is at most one. Fortunately, it can be show that:

   - For random codes, with high probability, for most error patterns, the list size is at most one. In other words, for *most* codes, we can hope to see a gain over unique decoding. The proof of this fact follows from Shannon's proof for the capacity for qSC: the details are left as an exercise.
   - In Section 7.5, we show that the above behavior is in fact general: i.e. for *any* code (over a large enough alphabet) it is true that with high probability, for most error patterns, the list size is at most one.

   Thus, using this option to deal with multiple answers, we still deal with worse case errors but can correct more error patterns than unique decoding.

2. If the decoder has access to some side information, then it can use that to prune the list. Informally, if the worst-case list size is L, then the amount of extra information one needs is $O(\log L)$. This will effectively decrease[1] the dimension of the code by $O(\log L)$, so if L is small enough, this will have negligible effect on the rate of the code. There are also application (especially in complexity theory) where one does not really care about the rate being the best possible.

Recall that Proposition 1.4.1 implies that $\delta/2$ is the maximum fraction of errors correctable with unique decoding. Since list decoding is a relaxation of unique decoding, it is natural to wonder

Question 7.2.1. *Can we correct more than $\delta/2$ fraction of errors using list decoding?*

and if so

Question 7.2.2. *What is the maximum fraction of errors correctable using list decoding?*

In particular, note that the intuition from Figure 7.1 states that the answer to Question 7.2.1 should be yes.

[1] Note that side information effectively means that not all possible vectors are valid messages.

7.3 Johnson Bound

In this section, we will indeed answer Question 7.2.1 in the affirmative by stating a bound due to Johnson. To setup the context again, recall that Proposition 1.4.1 implies that any code with relative distance δ is $(\delta/2, 1)$-list decodable.

Notice that if we can show a code for some $e > \left\lfloor \frac{d-1}{2} \right\rfloor$ is $(e/n, n^{O(1)})$-list decodable, then (combinatorially) it is possible to list decode that code up to e errors. We'll show by proving the Johnson bound that this is indeed the case for any code.

Theorem 7.3.1 (Johnson Bound)**.** *Let $C \subseteq [q]^n$ be a code of distance d. If $\rho < J_q\left(\frac{d}{n}\right)$, then C is a (ρ, qdn)-list decodable code, where the function $J_q(\delta)$ is defined as*

$$J_q(\delta) = \left(1 - \frac{1}{q}\right)\left(1 - \sqrt{1 - \frac{q\delta}{q-1}}\right).$$

Proof (for $q = 2$). The proof technique that we will use has a name: double counting. The main idea is to count the same quantity in two different ways to get both an upper and lower bound on the same quantity. These bounds then imply an inequality and we will derive our desired bound from this inequality.

We have to show that for every binary code $C \subseteq \{0,1\}^n$ with distance d (i.e. for every $\mathbf{c}_1 \neq \mathbf{c}_2 \in C$, $\Delta(\mathbf{c}_1, \mathbf{c}_2) \geq d$) and every $\mathbf{y} \in \{0,1\}^n$, $|B(\mathbf{y}, e) \bigcap C| \leq 2dn$.

Fix arbitrary C and $\mathbf{y}$. Let $\mathbf{c}_1, \ldots, \mathbf{c}_M \in B(\mathbf{y}, e)$. We need to show that $M \leq 2dn$. Define $\mathbf{c}'_i = \mathbf{c}_i - \mathbf{y}$ for $1 \leq i \leq M$. Then we have the following:

(i) $wt(\mathbf{c}'_i) \leq e$ for $1 \leq i \leq M$ because $\mathbf{c}_i \in B(\mathbf{y}, e)$.

(ii) $\Delta(\mathbf{c}'_i, \mathbf{c}'_j) \geq d$ for every $i \neq j$ because $\Delta(\mathbf{c}_i, \mathbf{c}_j) \geq d$.

Define

$$S = \sum_{i<j} \Delta(\mathbf{c}'_i, \mathbf{c}'_j).$$

We will prove both an upper and a lower bound on S from which we will extract the required upper bound on M. Then from (ii) we have

$$S \geq \binom{M}{2} d \tag{7.2}$$

Consider the $n \times M$ matrix $(\mathbf{c}'^T_1, \cdots, \mathbf{c}'^T_M)$. Define m_i as the number of 1's in the i-th row for $1 \leq i \leq n$. Then the i-th row of the matrix contributes the value $m_i(M - m_i)$ to S because this is the number of 0-1 pairs in that row. (Note that each such pair contributes one to S.) This implies that

$$S = \sum_{i=1}^{n} m_i(M - m_i). \tag{7.3}$$

Define

$$\bar{e} = \sum_i \frac{m_i}{M}.$$

Note that

$$\sum_{i=1}^{n} m_i = \sum_{j=1}^{M} wt(\mathbf{c}_i) \le eM,$$

where the inequality follows From (i) above. Thus, we have

$$\bar{e} \le e.$$

Using the Cauchy-Schwartz inequality (i.e., $\langle \mathbf{x}, \mathbf{y} \rangle \le ||\mathbf{x}|| \cdot ||\mathbf{y}||$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$) by taking $\mathbf{x} = (m_1, \cdots, m_n)$, $\mathbf{y} = (1/n, \cdots, 1/n)$, we have

$$\left(\frac{\sum_{i=1}^{n} m_i}{n} \right)^2 \le \left(\sum_{i=1}^{n} m_i^2 \right) \frac{1}{n}. \tag{7.4}$$

Thus, from (7.3)

$$S = \sum_{i=1}^{n} m_i (M - m_i) = M^2 \bar{e} - \sum_{i=1}^{n} m_i^2 \le M^2 \bar{e} - \frac{(M\bar{e})^2}{n} = M^2 (\bar{e} - \frac{\bar{e}^2}{n}), \tag{7.5}$$

where the inequality follows from (7.4). By (7.2) and (7.5),

$$M^2 \left(\bar{e} - \frac{\bar{e}^2}{n} \right) \ge \frac{M(M-1)}{2} d,$$

which implies that

$$\begin{aligned} M \le \frac{dn}{dn - 2n\bar{e} + 2\bar{e}^2} = \frac{2dn}{2dn - n^2 + n^2 - 4n\bar{e} + 4\bar{e}^2} &= \frac{2dn}{(n-2\bar{e})^2 - n(n-2d)} \\ &\le \frac{2dn}{(n-2e)^2 - n(n-2d)}, \end{aligned} \tag{7.6}$$

where the last inequality follows from the fact that $\bar{e} \le e$. Then from

$$\frac{e}{n} < \frac{1}{2} \left(1 - \sqrt{1 - \frac{2d}{n}} \right),$$

we get

$$n - 2e > \sqrt{n(n-2d)}.$$

In other words

$$(n-2e)^2 > n(n-2d).$$

Thus, $(n-2e)^2 - n(n-2d) \ge 1$ because n, e are all integers and therefore, from (7.6), we have $M \le 2dn$ as desired. □

Next, we prove the following property of the function $J_q(\cdot)$, which along with the Johnson bound answers Question 7.2.1 in the affirmative.

Lemma 7.3.2. *Let $q \geq 2$ be an integer and let $0 \leq x \leq 1 - \frac{1}{q}$. Then the following inequalities hold:*

$$J_q(x) \geq 1 - \sqrt{1-x} \geq \frac{x}{2},$$

where the second inequality is tight for $x > 0$.

Proof. We start with by proving the inequality

$$\left(1 - \frac{1}{q}\right)\left(1 - \sqrt{1 - \frac{xq}{q-1}}\right) \geq 1 - \sqrt{1-x}.$$

Indeed, both the LHS and RHS of the inequality are zero at $x = 0$. Further, it is easy to check that the derivatives of the LHS and RHS are $\frac{1}{\sqrt{1-\frac{xq}{q-1}}}$ and $\frac{1}{\sqrt{1-x}}$ respectively. The former is always larger than the latter quantity. This implies that the LHS increases more rapidly than the RHS, which in turn proves the required inequality.

The second inequality follows from the subsequent relations. As $x \geq 0$,

$$1 - x + \frac{x^2}{4} \geq 1 - x,$$

which implies that

$$\left(1 - \frac{x}{2}\right)^2 \geq 1 - x,$$

which in turn implies the required inequality. (Note that the two inequalities above are strict for $x > 0$, which implies that $1 - \sqrt{1-x} > x/2$ for every $x > 0$, as desired.) □

Theorem 7.3.1 and Lemma 7.3.2 imply that for *any* code, list decoding can potentially correct strictly more errors than unique decoding in polynomial time, as long as q is at most some polynomial in n (which will be true of all the codes that we discuss in this book). This answers Question 7.2.1 in the affirmative. See Figure 7.2 for an illustration of the gap between the Johnson bound and the unique decoding bound.

Theorem 7.3.1 and Lemma 7.3.2 also implies the following "alphabet-free" version of the Johnson bound.

Theorem 7.3.3 (Alphabet-Free Johnson Bound)**.** *If $e \leq n - \sqrt{n(n-d)}$, then any code with distance d is $(e/n, qnd)$-list decodable for all the q.*

A natural question to ask is the following:

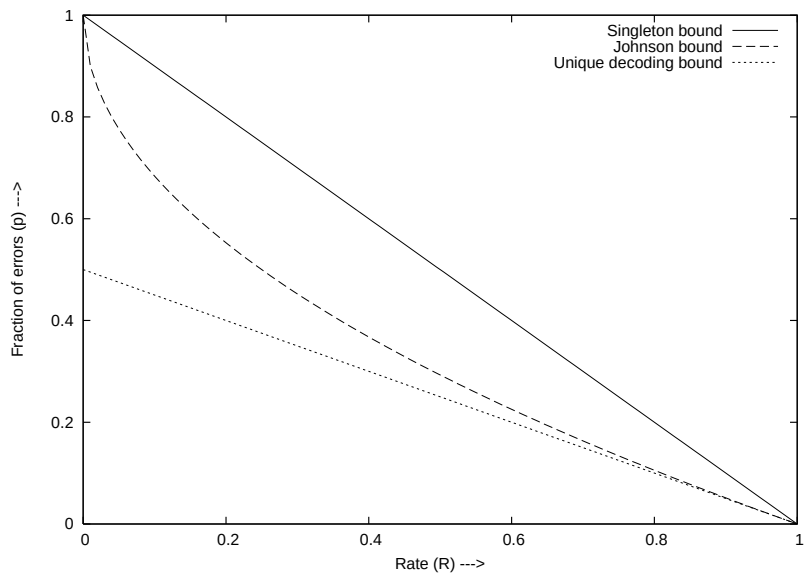


Figure 7.2: The trade-off between rate R and the fraction of errors that can be corrected. $1-\sqrt{R}$ is the trade-off implied by the Johnson bound. The bound for unique decoding is $(1-R)/2$ while $1-R$ is the Singleton bound (and the list decoding capacity for codes over large alphabets).

Question 7.3.1. *Is the Johnson bound tight?*

The answer is yes in the sense that there exist linear codes with relative distance δ such that we can find Hamming ball of radius larger than $J_q(\delta)$ with super-polynomially many codewords. On the other hand, in the next section, we will show that, in some sense, it is not tight.

7.4 List-Decoding Capacity

In the previous section, we saw what can one achieve with list decoding in terms of distance of a code. In this section, let us come back to Question 7.2.2. In particular, we will consider the trade-off between rate and the fraction of errors correctable by list decoding. Unlike the case of unique decoding and like the case of BSC_p, we will be able to prove an optimal trade-off.

Next, we will prove the following result regarding the optimal trade-off between rate of a code and the fraction of errors that can be corrected via list decoding.

Theorem 7.4.1. *Let $q \geq 2$, $0 \leq p < 1-\frac{1}{q}$, and $\varepsilon > 0$. Then the following holds for codes of large enough block length n:*

(i) If $R \leq 1 - H_q(p) - \varepsilon$, then there exists a $\left(p, O\left(\frac{1}{\varepsilon}\right)\right)$-list decodable code.

(ii) If $R > 1 - H_q(p) + \varepsilon$, every (ρ, L)-list decodable code has $L \geq q^{\Omega(n)}$.

Thus, the *List-decoding capacity*[2] is $1 - H_q(p)$ (where p is the fraction of errors). Further, this fully answers Question 7.2.2. Finally, note that this exactly matches capacity for $q\text{SC}_p$ and hence, list decoding can be seen as a bridge between Shannon's world and Hamming's world. The remarkable aspect of this result is that we bridge the gap between these worlds by allowing the decoder to output at most $O(1/\varepsilon)$ many codewords.

7.4.1 Proof of Theorem 7.4.1

We begin with the basic idea behind the proof of part (i) of the theorem.

As in Shannon's proof for capacity of BSC_p, we will use the probabilistic method (Section 3.2). In particular, we will pick a random code and show that it satisfies the required property with non-zero probability. In fact, we will show that a random code is (ρ, L)-list decodable with high probability as long as:

$$R \leq 1 - H_q(p) - \frac{1}{L}$$

The analysis will proceed by proving that probability of a "bad event" is small. "Bad event" means there exist messages $\mathbf{m}_0, \mathbf{m}_1, \cdots, \mathbf{m}_L \in [q]^{Rn}$ and a received code $\mathbf{y} \in [q]^n$ such that:

$$\Delta\left(C(\mathbf{m}_i), \mathbf{y}\right) \leq \rho n, \text{for every } 0 \leq i \leq L.$$

Note that if a bad event occurs, then the code is not a (ρ, L)-list decodable code. The probability of the occurrence of any bad event will then be calculated by an application of the union bound.

Next, we restate Theorem 7.4.1 and prove a stronger version of part (i). (Note that $L = \lceil \frac{1}{\varepsilon} \rceil$ in Theorem 7.4.2 implies Theorem 7.4.1.)

Theorem 7.4.2 (List-Decoding Capacity)**.** *Let $q \geq 2$ be an integer, and $0 < \rho < 1 - \frac{1}{q}$ be a real number.*

(i) Let $L \geq 1$ be an integer, then there exists an (ρ, L)-list decodable code with rate

$$R \leq 1 - H_q(\rho) - \frac{1}{L}$$

(ii) For every (ρ, L) code of rate $1 - H_q(\rho) + \varepsilon$, it is necessary that $L \geq 2^{\Omega(\varepsilon n)}$.

Proof. We start with the proof of (i). Pick a code C at random where

$$|C| = q^k, \text{ where } k \leq \left(1 - H_q(\rho) - \frac{1}{L}\right) n.$$

That is, as in Shannon's proof, for every message $\mathbf{m}$, pick $C(\mathbf{m})$ uniformly and independently at random from $[q]^n$.

[2] Actually the phrase should be something like "capacity of worst case noise model under list decoding" as the capacity is a property of the channel. However, in the interest of brevity we will only use the term list-decoding capacity.

Given $\mathbf{y} \in [q]^n$, and $\mathbf{m}_0, \cdots, \mathbf{m}_L \in [q]^k$, the tuple $(\mathbf{y}, \mathbf{m}_0, \cdots, \mathbf{m}_L)$ defines a *bad event* if

$$C(\mathbf{m}_i) \in B(\mathbf{y}, \rho n), 0 \le i \le L.$$

Note that a code is (ρ, L)-list decodable if and only if there does not exist any bad event.

Fix $\mathbf{y} \in [q]^n$ and $\mathbf{m}_0, \cdots, \mathbf{m}_L \in [q]^k$.
Note that for fixed i, by the choice of C, we have:

$$\Pr[C(\mathbf{m}_i) \in B(\mathbf{y}, \rho n)] = \frac{Vol_q(\rho n, n)}{q^n} \le q^{-n(1-H_q(\rho))}, \tag{7.7}$$

where the inequality follows from the upper bound on the volume of a Hamming ball (Proposition 3.3.1). Now the probability of a bad event given $(\mathbf{y}, \mathbf{m}_0, \cdots, \mathbf{m}_L)$ is

$$\Pr\left[\bigwedge_{i=0}^{L} C(\mathbf{m}_i) \in B(\mathbf{y}, \rho n)\right] = \prod_{0}^{L} \Pr[C(\mathbf{m}_i) \in B(\mathbf{y}, \rho n)] \le q^{-n(L+1)(1-H_q(\rho))}, \tag{7.8}$$

where the equality follows from the fact that the random choices of codewords for distinct messages are independent and the inequality follows from (7.7). Then,

$$\begin{aligned}
\Pr[\text{ There is a bad event}] &\le q^n \binom{q^k}{L+1} q^{-n(L+1)(1-H_q(\rho))} && (7.9)\\
&\le q^n q^{Rn(L+1)} q^{-n(L+1)(1-H_q(\rho))} && (7.10)\\
&= q^{-n(L+1)[1-H_q(\rho)-\frac{1}{L+1}-R]} \\
&\le q^{-n(L+1)[1-H_q(\rho)-\frac{1}{L+1}-1+H_q(\rho)+\frac{1}{L}]} && (7.11)\\
&= q^{-\frac{n}{L}} \\
&< 1
\end{aligned}$$

In the above, (7.9) follows by the union bound (Lemma 3.1.3) with (7.8) and by counting the number of $\mathbf{y}$'s (which is q^n), and the number of $L+1$ tuples (which is $\binom{q^k}{L+1}$). (7.10) follows from the fact that $\binom{a}{b} \le a^b$ and $k = Rn$. (7.11) follows by assumption $R \le 1 - H_q(\rho) - \frac{1}{L}$. The rest of the steps follow from rearranging and canceling the terms. Therefore, by the probabilistic method, there exists C such that it is (ρ, L)-list decodable.

Now we turn to the proof of part (ii). For this part, we need to show the existence of a $\mathbf{y} \in [q]^n$ such that $|C \cap B(\mathbf{y}, \rho n)|$ is exponentially large for every C of rate $R \ge 1 - H_q(\rho) + \varepsilon$. We will again use the probabilistic method to prove this result.

Pick $\mathbf{y} \in [q]^n$ uniformly at random. Fix $\mathbf{c} \in C$. Then

$$\begin{aligned}
\Pr[\mathbf{c} \in B(\mathbf{y}, \rho n)] &= \Pr[\mathbf{y} \in B(\mathbf{c}, \rho n)] \\
&= \frac{Vol_q(\rho n, n)}{q^n} && (7.12)\\
&\ge q^{-n(1-H_q(\rho))-o(n)}, && (7.13)
\end{aligned}$$

where (7.12) follows from the fact that $\mathbf{y}$ is chosen uniformly at random from $[q]^n$ and (7.13) follows by the lower bound on the volume of the Hamming ball (Proposition 3.3.1).

We have

$$\begin{aligned}
E[|C \cap B(\mathbf{y}, \rho n)|] &= \sum_{\mathbf{c} \in C} E[\mathbb{1}_{\mathbf{c} \in B(\mathbf{y}, \rho n)}] && (7.14)\\
&= \sum_{\mathbf{c} \in C} \Pr[\mathbf{c} \in B(\mathbf{y}, \rho n)] \\
&\geq \sum_{\mathbf{c} \in C} q^{-n(1-H_q(\rho)+o(n))} && (7.15)\\
&= q^{n[R-1+H_q(\rho)-o(1)]} \\
&\geq q^{\Omega(\varepsilon n)} && (7.16)
\end{aligned}$$

In the above, (7.14) follows by the linearity of expectation (Proposition 3.1.2), (7.15) follows from (7.13), and (7.16) follows by choice of R. Hence, by the probabilistic method, there exists $\mathbf{y}$ such that $|B(\mathbf{y}, \rho n) \cap C|$ is $q^{\Omega(n)}$, as desired. □

The above proof can be modified to work for random linear codes. (See Exercise **??**.)

We now return to Question 7.3.1. Note that by the Singleton bound, the Johnson bound implies that for any code one can hope to list decode from about $p \leq 1 - \sqrt{R}$ fraction of errors. However, this trade-off between p and R is not tight. Note that Lemma 3.3.2 along with Theorem 7.4.1 implies that for large q, the list decoding capacity is $1 - R > 1 - \sqrt{R}$. Figure 7.2 plots and compares the relevant trade-offs.

Finally, we have shown that the list decoding capacity is $1 - H_q(p)$. However, we showed the existence of a code that achieves the capacity by the probabilistic method. This then raises the following question:

Question 7.4.1. *Do there exist explicit codes that achieve list decoding capacity?*

Also the only list decoding algorithm that we have seen so far is the brute force algorithm that checks every codeword to see if they need to be output. This also leads to the follow-up question

Question 7.4.2. *Can we achieve list decoding capacity with efficient list decoding algorithms?*

A more modest goal related to the above would be the following:

Question 7.4.3. *Can we design an efficient list decoding algorithm that can achieve the Johnson bound? In particular, can we efficiently list decode a code of rate R from $1-\sqrt{R}$ fraction of errors?*

7.5 List Decoding from Random Errors

In this section, we formalize the intuition we developed from Figure 7.1. In particular, recall that we had informally argued that for most error patterns we can correct beyond the $\delta/2$ bound for unique decoding (Proposition 1.4.1). Johnson bound (Theorem 7.3.1) tells us that one can indeed correct beyond $\delta/2$ fraction of errors. However, there are two shortcomings. The first is that the Johnson bounds tells us that the output list size is qdn but it does not necessarily imply that for most error patterns, there is unique by closest codewords (i.e. one can uniquely recover the transmitted codeword). In other words, Johnson bound is a "true" list decoding result and tells us nothing about the behavior of codes on the "average." The second aspect is that the Johnson bound holds for up to $1-\sqrt{1-\delta}$ fraction of errors. Even though it is more than $\delta/2$ for every $\delta > 0$, the bound e.g. is not say twice the unique decoding bound for every $\delta > 0$.

Next we show that for *any* code with relative distance δ (over a large enough alphabet size) for most error patterns, the output of a list decoder for any fraction of errors arbitrarily close to δ will have size one. In fact, the result is somewhat stronger: it show that even if one fixes the error locations *arbitrarily*, for most error patterns the output list size is one.

Theorem 7.5.1. *Let $\varepsilon > 0$ be a real and $q \geq 2^{\Omega(1/\varepsilon)}$ be an integer. Then the following is true for any $0 < \delta < 1 - 1/q$ and large enough n. Let $C \subseteq \{0,1,\ldots q-1\}^n$ be a code with relative distance δ and let $S \subseteq [n]$ such that $|S| = (1-\rho)n$, where $(0 < \rho \leq \delta - \varepsilon)$.*

Then, for all $\mathbf{c} \in C$ and all but a $q^{-\Omega(\varepsilon n)}$ fraction of error patterns, $\mathbf{e} \in \{0,1\ldots q-1\}^n$ such that

$$\mathbf{e}_S = \mathbf{0} \text{ and } wt(\mathbf{e}) = \rho n \tag{7.17}$$

the only codeword within Hamming distance ρn of $\mathbf{c}+\mathbf{e}$ is $\mathbf{c}$ itself.

For illustration of the kinds of error pattern we will deal with, see Figure 7.3.

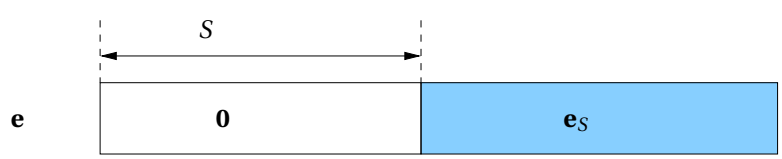


Figure 7.3: Illustration of the kind of error patterns we are trying to count.

Before we present the proof, we present certain corollaries (the proofs of which we leave as exercises). First the result above implies a similar result of the output list size being one for the

following two random noise models: (i) uniform distribution over *all* error patterns of weight ρn and (ii) $q\mathrm{SC}_p$. In fact, we claim that the result also implies that any code with distance at least $p+\varepsilon$ allows for reliable communication over $q\mathrm{SC}_p$. (Contrast the $2p+\varepsilon$ distance that was needed for a similar result that was implied by Proposition 6.4.1.)

Finally, we present a lemma (the proof is left as an exercise) that will be crucial to the proof of Theorem 7.5.1.

Lemma 7.5.2. *Let be C be an $(n,k,d)_q$ code. If we fix the values in $n-d+1$ out of the n positions in a possible codeword, then at most one codeword in C can agree with the fixed values.*

Proof of Theorem 7.5.1. For the rest of the proof, fix a $\mathbf{c} \in C$. For notational convenience define $\mathscr{E}_S$ to be the set of all error patterns $\mathbf{e}$ such that $\mathbf{e}_S = 0$ and $wt(\mathbf{e}) = \rho n$. Note that as every error position has $(q-1)$ non-zero choices and there are ρn such positions in $[n] \setminus S$, we have

$$|\mathscr{E}_s| = (q-1)^{\rho n}. \tag{7.18}$$

Call an error pattern $\mathbf{e} \in \mathscr{E}_s$ as *bad* if there exists another codeword $\mathbf{c}' \neq \mathbf{c}$ such that

$$\triangle(\mathbf{c}', \mathbf{c}+\mathbf{e}) \leq \rho n.$$

Now, we need to show that the number of bad error patterns is at most

$$q^{-\Omega(\varepsilon n)}|\mathscr{E}_s|.$$

We will prove this by a somewhat careful counting argument.

We begin with a definition.

Definition 7.5.1. Every error pattern $\mathbf{e}$ is associated with a codeword $c(\mathbf{e})$, which is the closest codeword which lies within Hamming distance ρn from it.

For a bad error pattern we insist on having $c(\mathbf{e}) \neq \mathbf{c}$– note that for a bad error pattern such a codeword always exists. Let A be the set of positions where $c(\mathbf{e})$ agrees with $\mathbf{c}+e$.

The rest of the argument will proceed as follows. For each possible A, we count how many bad patterns $\mathbf{e}$ are associated with it (i.e. $\mathbf{c}+\mathbf{e}$ and $c(\mathbf{e})$ agree exactly in the positions in A). To bound this count non-trivially, we will use Lemma 7.5.2.

Define a real number α such that $|A| = \alpha n$. Note that since $c(\mathbf{e})$ and $\mathbf{c}+\mathbf{e}$ agree in at least $1-\rho$ positions,

$$\alpha \geq 1-\rho \geq 1-\delta+\varepsilon. \tag{7.19}$$

For now let us fix A with $|A| = \alpha n$ and to expedite the counting of the number of bad error patterns, let us define two more sets:

$$A_1 = A \cap S,$$

and

$$A_2 = A \setminus A_1.$$

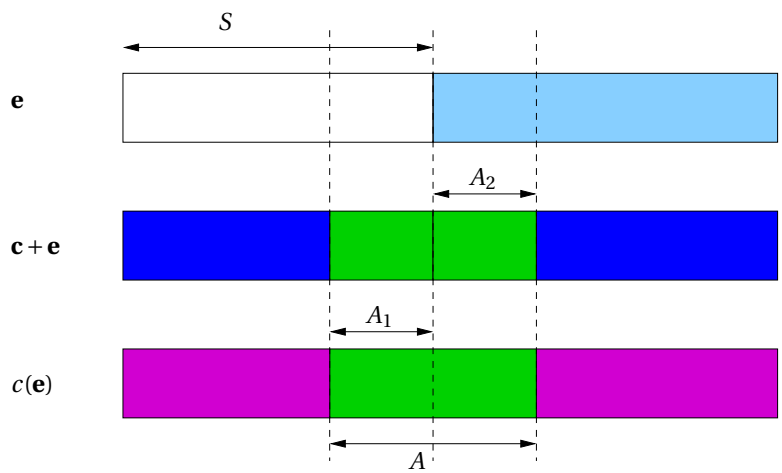


Figure 7.4: Illustration of notation used in the proof of Theorem 7.5.1. Positions in two different vectors that agree have the same color.

.

See Figure 7.4 for an illustration of the notation that we have fixed so far.

Define β such that

$$|A_1| = \beta n. \tag{7.20}$$

Note that this implies that

$$|A_2| = (\alpha - \beta)n. \tag{7.21}$$

Further, since $A_1 \subseteq A$, we have

$$\beta \leq \alpha.$$

To recap, we have argued that every bad error pattern $\mathbf{e}$ corresponds to a codeword $c(\mathbf{e}) \neq \mathbf{c}$ and is associated with a pair of subsets (A_1, A_2). So, we fix (A_1, A_2) and then count the number of bad $\mathbf{e}$'s that map to (A_1, A_2). (Later on we will aggregate this count over all possible choices of (A_1, A_2).)

Towards this end, first we overestimate the number of error patterns $\mathbf{e}$ that map to (A_1, A_2) by allowing such $\mathbf{e}$ to have arbitrary values in $[n] \setminus (S \cup A_2)$. Note that all such values have to be non-zero (because of (7.17). This implies that the number of possible distinct $\mathbf{e}_{[n]\setminus(S\cup A_2)}$ is at most

$$(q-1)^{n-|S|-|A_2|} = q^{n-(1-\rho)n-(\alpha-\beta)n}, \tag{7.22}$$

where the equality follows from the given size of S and (7.21). Next fix a non-zero $\mathbf{x}$ and let us only consider error patterns $\mathbf{e}$ such that

$$\mathbf{e}_{[n]\setminus(S\cup A_2)} = \mathbf{x}.$$

Note that at this stage we have an error pattern $\mathbf{e}$ as depicted in Figure 7.5.

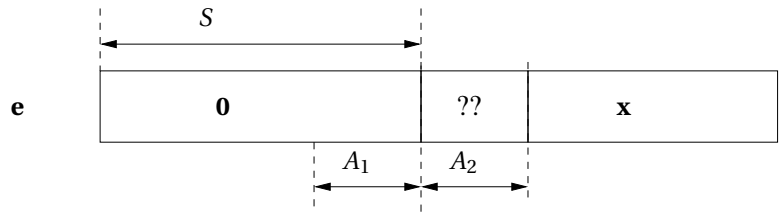


Figure 7.5: Illustration of the kind of error patterns we are trying to count now. The ? denote values that have not been fixed yet.

Now note that if we fix $c(\mathbf{e})_{A_2}$, then we would also fix $\mathbf{e}_{A_2}$ (as $(\mathbf{c}+\mathbf{e})_{A_2} = (c(\mathbf{e}))_{A_2}$). Recall that $\mathbf{c}$ is already fixed and hence, this would fix $\mathbf{e}$ as well. Further, note that

$$c(\mathbf{e})_{A_1} = (\mathbf{c}+\mathbf{e})_{A_1} = \mathbf{c}_{A_1}.$$

This implies that $c(\mathbf{e})_{A_1}$ is already fixed and hence, by Lemma 7.5.2 we would fix $c(\mathbf{e})$ if we fix (say the first) $(1-\delta)n+1-|A_1|$ positions in $c(\mathbf{e})_{A_2}$. Or in other words, by fixing the first $(1-\delta)n+1-|A_1|$ positions in $\mathbf{e}_{A_2}$, $\mathbf{e}$ would be completely determined. Thus, the number of choices for $\mathbf{e}$ that have the pattern in Figure 7.5 is upper bounded by

$$q^{(1-\delta)n+1-|A_1|} = (q-1)^{(1-\delta)n+1-\beta n}, \tag{7.23}$$

where the equality follows from (7.20).

Thus, by (7.22) and (7.23) the number of possible bad error patterns $\mathbf{e}$ that map to (A_1, A_2) is upper bounded by

$$(q-1)^{n-(1-\rho)n-\alpha n+\beta n+(1-\delta)n+1-\beta n} \le (q-1)^{\rho n-\varepsilon n+1} = (q-1)^{-\varepsilon n+1}|\mathscr{E}_s|,$$

where the inequality follows from (7.19) and the equality follows from (7.18).

Finally, summing up over all choices of $A = (A_1, A_2)$ (of which there are at most 2^n), we get that the total number of bad patterns is upper bounded by

$$2^n \cdot (q-1)^{-\varepsilon n+1} \cdot |\mathscr{E}_S| \le q^{\frac{n}{\log_2 q} - \frac{\varepsilon n}{2} + \frac{1}{2}} \cdot |\mathscr{E}_A| \le q^{-\varepsilon n/4} \cdot |\mathscr{E}_S|,$$

where the first inequality follows from $q-1 \ge \sqrt{q}$ (which in turn is true for $q \ge 3$) while the last inequality follows from the fact that for $q \ge \Omega(1/\varepsilon)$ and large enough n, $\frac{n+1/2}{\log_2 q} < \frac{\varepsilon n}{4}$. This completes the proof. □

It can be shown that Theorem 7.5.1 is not true for $q = 2^{o(1/\varepsilon)}$. The proof is left as an exercise.

7.6 Exercises

Exercise 7.1. Show that with high probability, a random linear code is (ρ, L)-list decodable code as long as

$$R \le 1 - H_q(\rho) - \frac{1}{\lceil \log_q(L+1) \rceil}. \tag{7.24}$$

Hint: Think how to fix (7.8) for random linear code.

Exercise 7.2. In this exercise we will see how we can "fix" the dependence on L is the rate of random linear codes from Exercise 7.1. In particular, we will consider the following family of codes that are somewhere between linear and general codes and are called pseudolinear codes, which are defined as follows.

Let q be a prime power and let $1 \le k \le n$ and $L \ge 1$ be integers. Then an (n,k,L,r,q)-family of pseudolinear codes is defined as follows. Let $\mathbf{H}$ be the parity check matrix of an $[q^k-1, q^k-1-r, L+1]_q$ linear code and $\mathbf{H}'$ be an extension of $\mathbf{H}$ with the first column being $\mathbf{0}$ (and the rest being $\mathbf{H}$). Every code in the family is indexed by a matrix $\mathbf{A} \in \mathbb{F}_q^{n\times r}$. Fix such a $\mathbf{A}$. Then the corresponding code $C_{\mathbf{A}}$ is defined as follows. For any $\mathbf{x} \in \mathbb{F}_q^k$, we have

$$C_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\cdot\mathbf{H}'_{\mathbf{x}},$$

where $\mathbf{H}'_{\mathbf{x}}$ is the column corresonding to $\mathbf{x}$, when though of as an integer between 0 and q^k-1.

Next, we will argue that random pseudolinear codes have near optimal list decodability:

1. Fix non-zero messages $\mathbf{m}_1,\ldots\mathbf{m}_L$. Then for a random code $C_{\mathbf{A}}$ from an (n,k,L,r,q)-family of pseudolinear code family, the codewords $C_{\mathbf{A}}(\mathbf{m}_1),\ldots,C_{\mathbf{A}}(\mathbf{m}_L)$ are *independent* random vectors in $\mathbb{F}_q^n$.

2. Define (n,k,L,q)-family of pseudolinear codes to be $(n,k,L,O(kL),q)$-family of pseudolinear codes. Argue that (n,k,L,q)-family of pseudolinear codes exist.
   Hint: Exercise 5.10 might be helpful.

3. Let $\varepsilon > 0$ and $q \ge 2$ be a prime power. Further let $0 \le \rho < 1-1/q$. Then for a large enough n and k such that
   $$\frac{k}{n} \ge 1 - H_q(\rho) - \frac{1}{L} - \varepsilon,$$
   a random (n,k,L,q)-pesudolinear code is (ρ,L)-list decodable.

4. Show that one can construct a (ρ,L)-list decodable pseudolinear code with rate at least $1-H_q(\rho)-\frac{1}{L}-\varepsilon$ in $q^{O(kL+n)}$ time.
   Hint: Use method of conditional expectations.

Exercise 7.3. In this exercise we will consider a notion of “average" list decoding that is closely related to our usual notion of list decoding. As we will see in some subsequent exercises, sometimes it is easier to work with this average list decoding notion.

1. We begin with an equivalent definition of our usual notion of list decoding. Argue that a code C is (ρ,L) list decodable if and only if for every $\mathbf{y} \in [q]^n$ and every subset of $L+1$ codewords $\mathbf{c}_0,\ldots,\mathbf{c}_L$ we have that
   $$\max_{0\le i\le L} \Delta(\mathbf{y},\mathbf{c}_i) > \rho n.$$

2. We define a code C to be (ρ, L)-*average list decodable* if for every $\mathbf{y} \in [q]^n$ and $L+1$ codewords $\mathbf{c}_0, \ldots, \mathbf{c}_L$ we have
$$\frac{1}{L} \cdot \sum_{i=0}^{L} \Delta(\mathbf{y}, \mathbf{c}_i) > \rho n.$$
Argue that if C is (ρ, L)-average list decodable then it is also (ρ, L)-list decodable.

3. Argue that if C is (ρ, L)-list decodable then it is also $(\rho(1-\gamma), \lceil L/\gamma \rceil)$-average list decodable (for any $0 < \gamma < \rho$).

Exercise 7.4. In Section 7.5 we saw that for any code one can correct arbitrarily close to relative distance fraction of random errors. In this exercise we will see that one can prove a weaker result. In particular let $\mathscr{D}$ be an arbitrary distribution on $B_q(\mathbf{0}, \rho n)$. Then argue that for most codes, the list size with high probability is 1. In other words, show that for $1 - o(1)$, fraction of codes C we have that for every codeword $\mathbf{c} \in C$
$$\Pr_{\mathbf{e} \leftarrow \mathscr{D}} \left[|B_q(\mathbf{c} + \mathbf{e}, \rho n) \cap C| > 1 \right] = o(1).$$

Hint: Adapt the proof of Theorem 6.3.1 from Section 6.3.2.

Exercise 7.5. We call a code (ρ, L)-*erasure list-decodable* is informally for any received word with at most ρ fraction of erasures at most L codewords agree with it in the unerased positions. More formally, an $(n, k)_q$-code C is (ρ, L)-erasure list-decodable if for every $\mathbf{y} \in [q]^{(1-\rho)n}$ and every subset $T \subseteq [n]$ with $|T| = (1-\rho)n$, we have that
$$\left| \{ \mathbf{c} \in C | \mathbf{c}_T = \mathbf{y} \} \right| \leq L.$$

In this exercise you will prove some simple bounds on the best possible rate for erasure-list decodable code.

1. Argue that if C has distance d then it is $\left(\frac{d-1}{n}, 1\right)$-erasure list decodable.

2. Show that there exists a (ρ, L)-erasure list decodable code of rate
$$\frac{L}{L+1} \cdot (1-\rho) - \frac{H_q(\rho)}{L} - \gamma,$$
for any $\gamma > 0$.

3. Argue that there exists a *linear* (ρ, L)-erasure list decodable code with rate
$$\frac{J-1}{J} \cdot (1-\rho) - \frac{H_q(\rho)}{J-1} - \gamma,$$
where $J = \left\lceil \log_q(L+1) \right\rceil$ and $\gamma > 0$.

4. Argue that the bound in item 2 is tight for large enough L by showing that if a code of rate $1 - \rho + \varepsilon$ is (ρ, L)-erasure list decodable then L is $2^{\Omega_\varepsilon(n)}$.

Exercise 7.6. In this exercise we will see an alternate characterization of erasure list-decodable code for linear codes, which we will use to show separation between linear and non-linear code in the next exercise.

Given a linear code $C \subseteq \mathbb{F}_q^n$ and an integer $1 \le r \le n$, define the r'th *generalized Hamming distance*, denoted by $d_r(C)$, as follows. First given a set $D \subseteq \mathbb{F}_q^N$, we define the *support* of D as the union of the supports of vectors in D. More precisely

$$\mathrm{supp}(D) = \{i \mid \text{there exists } (u_1, \ldots, u_n) \in S \text{ such that } u_i \neq 0\}.$$

Then $d_r(C)$ is size of the smallest support of all r-dimensional subcodes of C.

Argue the following:

1. (Warmup) Convince yourself that $d_1(C)$ is the usual Hamming distance of C.

2. Prove that C is $(\rho n, L)$-erasure list-decodable if and only if $d_{1+\lfloor \log_q L \rfloor}(C) > \rho n$.

Exercise 7.7. In this exercise we use the connection between generalized Hamming distance and erasure list decodability from Exercise 7.6 to show an "exponential separation" between linear and non-linear codes when it comes to list decoding from erasure.

Argue the following:

1. Let C be an $[n,k]_q$ code. Then show that the average support size of r-dimensional subcodes of C is exactly
$$\frac{q^r - 1}{q^r} \cdot \frac{|C|}{|C| - 1} \cdot n.$$

2. From previous part or otherwise, conclude that if for an $[n,k]_q$ code C we have $d_r(C) > n(1 - q^{-r})$, then we have
$$|C| \le \frac{d_r(C)}{d_r(C) - n(1 - q^{-r})},$$
Note that the above bound for $r = 1$ recovers the Plotkin bound (second part of Theorem 4.4.1).

3. Argue that any (family) of code C with $d_r(C) = \delta_r \cdot n$, its rate satisfies:
$$R(C) \le 1 - \frac{q^r}{q^r - 1} \cdot \delta_r + o(1).$$

   Hint: Use a the result from previous part on a code related to C.

4. Argue that for small enough $\varepsilon > 0$, any linear $(1 - \varepsilon, L)$-erasure list decodable code with positive rate must have $L \ge \Omega(1/\varepsilon)$.

5. Argue that there exist $(1 - \varepsilon, O(\log(1/\varepsilon)))$-erasure list decodable code with positive (in fact $\Omega(\varepsilon)$) rate. Conclude that there exists non-linear codes that have the same erasure list decodability but with exponentially smaller list sizes than linear codes.

Exercise 7.8. In this exercise we will prove an analog of the Johnson bound (Theorem 7.3.1) but for erasure list-decodable codes. In particular, let C be an $(n,k,\delta n)_q$ code. Then show that for any $\varepsilon > 0$, C is an $\left(\left(\frac{q}{q-1}-\varepsilon\right)\delta, \frac{q}{(q-1)\varepsilon}\right)$-erasure list decodable.

Hint: The Plotkin bound (Theorem 4.4.1) might be useful.

Exercise 7.9. Let C be a q-ary (ρ, L)-(average) list decodable of rate R, then show that there exists another (ρ, L)-(average) list decodable code with rate at least

$$R + H_q(\lambda) - 1 - o(1),$$

for any $\lambda \in (\rho, 1-1/q]$ such that *all* codewords in C' have Hamming weight *exactly* λn.

Hint: Try to translate C.

Exercise 7.10. In this exercise, we will prove a lower bound on the list size of list decodable codes that have optimal rate. We do this via a sequence of following steps:

1. Let $C \subseteq [q]^n$ be a $(\rho, L-1)$-list decodable code such that all codewords have Hamming weight exactly λn for

$$\lambda = \rho + \frac{1}{2L} \cdot \rho^L.$$

   Then prove that

$$|C| < \frac{2L^2}{\lambda^L}.$$

   Hint: It might be useful to use the following result due to Erdös [29] (where we choose the variables to match the relevant ones in the problem). Let $\mathscr{A}$ be a family of subsets of $[n]$. Then if every $A \in \mathscr{A}$ has size at least $2L^2/\lambda^L$, then there exist distinct $A_1, \ldots, A_L \in \mathscr{A}$ such that $\cap_{i=1}^{L} A_i$ has size at least $\frac{n\lambda^L}{2}$.

2. Argue that any q-ary $(\rho, L-1)$-list decodable code C (for large enough block length) has rate at most $1 - H_q(\rho) - b_{\rho,q} \cdot \frac{\rho^L}{L}$ for some constant $b_{\rho,q}$ that only depends on ρ and q.
   Hint: Use the previous part and Exercise 7.9.

3. Argue that any q-ary (ρ, L)-list decodable C with rate $1-H_q(\rho)-\varepsilon$ mush satisfy $L \geq \Omega_{\rho,q}(\log(1/\varepsilon))$.

Exercise 7.11. It follows from Theorem 7.4.1 that a random code of rate $1 - H_q(\rho) - \varepsilon$ with high probability is $(\rho, O(1/\varepsilon))$-list decodable. On the other hand, the best lower bound on the list size for codes of rate $1 - H_q(\rho) - \varepsilon$ (for constant p, q) is $\Omega(\log(1/\varepsilon))$ (as we just showed in Exercise 7.10). It is natural to wonder if one can perhaps do a better argument for random codes. In this exercise, we will show that our argument for random codes is the best possible (for random codes). We will show this via the following sequence of steps:

1. Let C be a random $(n,k)_q$ code of rate $1 - H_q(\rho) - \varepsilon$. For any $\mathbf{y} \in [q]^n$ and any subset $S \subseteq [q]^k$ of size $L+1$, define the random event $\mathscr{E}(\mathbf{y}, S)$ that for every $\mathbf{m} \in S$, $C(\mathbf{m})$ is at Hamming distance at most ρn from $\mathbf{y}$. Define

$$W = \sum_{\mathbf{y},S} \mathscr{E}(\mathbf{y}, S).$$

   Argue that C is (ρ, L)-list decodable if and only if $W = 0$.

2. Define
$$\mu = q^{-n} \cdot Vol_q(\rho n, n).$$
Argue that
$$\mathbb{E}[W] \geq \frac{1}{(L+1)^{L+1}} \cdot \mu^{L+1} \cdot q^n \cdot q^{k(L+1)}.$$

3. Argue that
$$\sigma^2(Z) \leq q^{2n} \cdot \sum_{\ell=1}^{L+1} (L+1)^{2(L+1)} \cdot q^{k(2L+2-\ell)} \cdot \mu^{2L-\ell+3}.$$
Hint: Analyze the probability of both events $\mathscr{E}(\mathbf{y}, S)$ and $\mathscr{E}(\mathbf{z}, T)$ happening together for various intersection sizes $\ell = |S \cap T|$.

4. Argue that C is $\left(\rho, \frac{1-H_q(\rho)}{2\varepsilon}\right)$-list decodable with probability at most $q^{-\Omega_{\rho,\varepsilon}(n)}$.
Hint: Use Chebyschev's inequality.

7.7 Bibliographic Notes

List decoding was defined by Elias [28] and Wozencraft [105].

The result showing that for random error patterns, the list size with high probability is one for the special case of Reed-Solomon codes was shown by McEliece [72]. The result for all codes was proved by Rudra and Uurtamo [86]

In applications of list decoding in complexity theory (see for example [97],[38, Chap. 12]), side information is used crucially to prune the output of a list decoding algorithm to compute a unique answer.

Guruswami [37] showed that the answer to Question 7.3.1 is yes in the sense that there exist linear codes with relative distance δ such that we can find Hamming ball of radius larger than $J_q(\delta)$ with super-polynomially many codewords. This result was proven under a number-theoretic assumption, which was later removed by [47].

(7.24) implies that there exist linear codes with rate $1 - H_q(\rho) - \varepsilon$ that are $\left(\rho, q^{O(1/\varepsilon)}\right)$-list decodable. (This is also true for most linear codes with the appropriate parameters.) However, for a while just for $q = 2$, we knew the *existence* of $\left(\rho, O(1/\varepsilon)\right)$-list decodable codes [41] (though it was not a high probability result). Guruswami, Håstad and Kopparty resolved this "gap" by showing that random linear codes of rate $1 - H_q(\rho) - \varepsilon$ are $(\rho, O(1/\varepsilon))$-list decodable (with high probability) [40].

Chapter 8

What Cannot be Done-II

In this brief interlude of a chapter, we revisit the trade-offs between rate and relative distance for codes. Recall that the best (and only) lower bound on R that we have seen is the GV bound and the best upper bound on R that we have have seen so far is a combination of the Plotkin and Hamming bounds (see Figure 4.5). In this chapter, we will prove the final upper bound on R in this book due to Elias and Bassalygo. Then we will mention the best known upper bound on rate (but without stating or proving it). Finally, we will conclude by summarizing what we have seen so far and laying down the course for the rest of the book.

8.1 Elias-Bassalygo bound

We begin with the statement of a new upper bound on the rate called the Elias-Bassalygo bound.

Theorem 8.1.1 (Elias-Bassalygo bound)**.** *Every q-ary code of rate R, distance δ, and large enough block length, satisfies the following:*

$$R \leq 1 - H_q\left(J_q(\delta)\right) + o(1)$$

See Figure 8.1 for an illustration of the Elias-Bassalygo bound for binary codes. Note that this bound is tighter than all the previous upper bounds on rate that we have seen so far.

The proof of Theorem 8.1.1 uses the following lemma:

Lemma 8.1.2. *Given a q-ary code, $C \subseteq [q]^n$ and $0 \leq e \leq n$, there exists a Hamming ball of radius e with at least $\frac{|C| Vol_q(e,n)}{q^n}$ codewords in it.*

Proof. We will prove the existence of the required Hamming ball by the probabilistic method. Pick a received word $\mathbf{y} \in [q]^n$ at random. It is easy to check that the expected value of $|B(\mathbf{y}, e) \cap C|$ is $\frac{|C| Vol_q(e,n)}{q^n}$. (We have seen this argument earlier in the proof of part (ii) of Theorem 7.4.2.)

This by the probabilistic method implies the existence of a $\mathbf{y} \in [q]^n$ such that

$$|B(\mathbf{y}, e) \cap C| \geq \frac{|C| \, Vol_q(e, n)}{q^n},$$

as desired. □

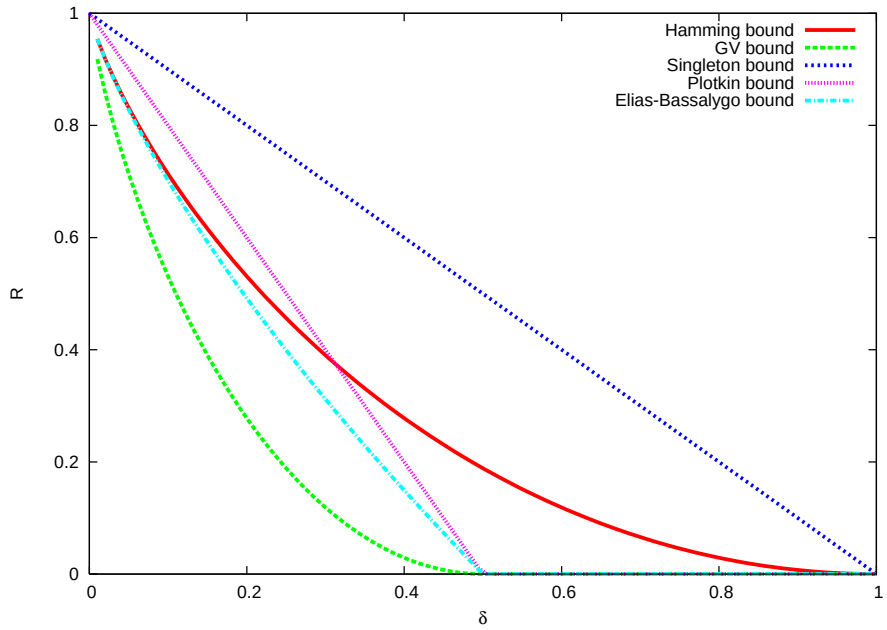


Figure 8.1: Singleton, Hamming, Plotkin, GV and Elias-Bassalygo bounds on rate versus distance for binary codes.

Proof of Theorem 8.1.1. Let $C \subseteq [q]^n$ be any code with relative distance δ. Define $e = nJ_q(\delta) - 1$. By Lemma 8.1.2, there exists a Hamming ball with $\mathscr{B}$ codewords such that the following inequality is true:

$$\mathscr{B} \geq \frac{|C|\, Vol_q(e,n)}{q^n}.$$

By our choice of e and the Johnson bound (Theorem 7.3.1), we have

$$\mathscr{B} \leq qdn.$$

Combining the upper and lower bounds on $\mathscr{B}$ implies the following

$$|C| \leq qnd \cdot \frac{q^n}{Vol_q(e,n)} \leq q^{n\left(1-H_q\left(J_q(\delta)\right)+o(1)\right)},$$

where the second inequality follows from our good old lower bound on the volume of a Hamming ball (Proposition 3.3.1) and the fact that $qdn \leq qn^2 \leq q^{o(n)}$ for large enough n. This implies that the rate R of C satisfies:

$$R \leq 1 - H_q\left(J_q(\delta)\right) + o(1),$$

as desired. □

8.2 The MRRW bound: A better upper bound

The MRRW bound (due to McEliece, Rodemich, Rumsey and Welch) is based on a linear programming approach introduced by Delsarte to bound the rate of a code. The MRRW bound is a better upper bound than the Elias-Bassalygo bound (though we will not state or prove the bound in this book). However, there is a gap between the Gilbert-Varshamov (GV) bound and the MRRW bound. The gap still exists to this day. To give one data point on the gap, consider $\delta = \frac{1}{2} - \varepsilon$ (think of $\varepsilon \to 0$), the GV bound gives a lower bound on R of $\Omega\left(\varepsilon^2\right)$ (see Proposition 3.3.5), while the MRRW bound gives an upper bound on R of $O\left(\varepsilon^2 \log\left(\frac{1}{\varepsilon}\right)\right)$.

8.3 A Breather

Let us now recap the combinatorial results that we have seen so far. Table 8.1 summarizes what we have seen so far for binary codes in Shannon's world and Hamming's world (under both unique and list decoding settings).

Shannon	Hamming	
BSC_p	Unique	List
1-H(p) is capacity	$R \geq 1 - H(\delta)$	$1 - H(p)$ is list decoding capacity
	$R \leq MRRW$	
Explicit codes at capacity?	Explicit Asymptotically good codes?	Explicit codes at capacity?
Efficient decoding algorithm?	Efficient decoding algorithms?	Efficient decoding algorithms?

Table 8.1: High level summary of results seen so far.

For the rest of the section, we remind the reader about the definition of explicit codes (Definition 6.3.1) and strongly explicit codes (Definition 6.3.2).

We begin with BSC_p. We have seen that the capacity of BSC_p is $1 - H(p)$. The most natural open question is to obtain the capacity result but with explicit codes along with efficient decoding (and encoding) algorithms (Question 6.3.1).

Next we consider Hamming's world under unique decoding. For large enough alphabets, we have seen that Reed-Solomon codes (Chapter 5) meet the Singleton bound (Theorem 4.3.1). Further, the Reed-Solomon codes are strongly explicit[1]. The natural question then is

Question 8.3.1. *Can we decode Reed-Solomon codes up to half its distance?*

For smaller alphabets, especially binary codes, as we have seen in the last section, there is a gap between the best known lower and upper bounds on the rate of a code with a given relative

[1] The proof is left as an exercise.

distance. Further, we do not know of an explicit construction of a binary code that lies on the GV bound. These lead to the following questions that are still wide open:

Open Question 8.3.1. *What is the optimal trade-off between R and δ?*

Open Question 8.3.2.

Does there exist an explicit construction of (binary) codes on the GV bound?

If we scale down our ambitions, the following is a natural weaker version of the second question above:

Question 8.3.2. *Do there exist explicit asymptotically good binary codes?*

We also have the following algorithmic counterpart to the above question:

Question 8.3.3. *If one can answer Question 8.3.2, then can we decode such codes efficiently from a non-zero fraction of errors?*

For list decoding, we have seen that the list decoding capacity is $1 - H_q(p)$. The natural open questions are whether we can achieve the capacity with explicit codes (Question 7.4.1) along with efficient list decoding algorithms (Question 7.4.2).

In the remainder of the book, we will focus on the questions mentioned above (and summarized in the last two rows of Table 8.1).

8.4 Bibliographic Notes

The McEliece-Rodemich-Rumsey-Welch (MRRW) bound was introduced in 1977 in the paper [73].

Part III

The Codes

Chapter 9

When Polynomials Save the Day: Polynomial Based Codes

As we saw in Chapter 5, The Reed-Solomon codes give a remarkable family of codes with optimal dimension vs. distance tradeoff. They even match the Singleton bound (recall Theorem 4.3.1), get $k = n - d + 1$ for a code of block length n, distance d and dimension k. However they achieve this remarkable performance only over large alphabets, namely when the alphabet size $q \geq n$. In fact, so far in this book, we have not seen *any explicit* asymptotically good code other than a Reed-Solomon code. This naturally leads to the following question (which is a weaker form for Question 8.3.2):

Question 9.0.1. *Do there exist explicit asymptotically good codes for small alphabets $q \ll n$?*

In this chapter we study an extension of Reed-Solomon codes, called the (generalized) Reed-Muller codes, that lead to codes over smaller alphabets while losing in the dimension-distance tradeoff (but under certain settings do answer Question 9.0.1 in the affirmative).

The main idea is to extend the notion of functions we work with, to multivariate functions. (See Exercise 5.2 for equivalence between certain Reed-Solomon codes and univariate functions.) Just working with bivariate functions (functions on two variables), allows us to get codes of block length $n = q^2$, and more variables can increase the length further for the same alphabet size. We look at functions of *total degree* at most r. Analysis of the dimension of the code reduces to simple combinatorics. Analysis of the distance follows from "polynomial-distance" lemmas, whose use is ubiquitous in algebra, coding theory and computer science, and we describe these in the sections below. We start with the generic construction.

9.1 The generic construction

Recall that for a monomial $\mathbf{X}^{\mathbf{d}} = X_1^{d_1} \cdot X_2^{d_2} \cdots X_m^{d_m}$ its total degree is $d_1 + d_2 + \cdots + d_m$. We next extend this to the definition of the degree of a polynomial:

Definition 9.1.1. The total degree of a polynomial $P(\mathbf{X}) = \sum_{\mathbf{d}} c_{\mathbf{d}} \mathbf{X}^{\mathbf{d}}$ over $\mathbb{F}_q$ (i.e. every $c_{\mathbf{d}} \in \mathbb{F}_q$) is the maximum over $\mathbf{d}$ such that $c_{\mathbf{d}} \neq 0$, of the total degree of $\mathbf{X}^{\mathbf{d}}$. We denote the total degree of P by $\deg(P)$.

For example, the degree of the polynomial $3X^3Y^4 + X^5 + Y^6$ is 7.

In turns out that when talking about Reed-Muller codes, it is convenient to switch back and forth between multivariate functions and multivariate polynomials. We can extend the notion above to functions from $\mathbb{F}_q^m \to \mathbb{F}_q$. For $f : \mathbb{F}_q^m \to \mathbb{F}_q$ let $\deg(f)$ be the minimal degree of a polynomial $P \in \mathbb{F}_q[X_1, \ldots, X_m]$ (where $\mathbb{F}_q[X_1, \ldots, X_m]$ denotes the set of all m-variate polynomials with coefficients from $\mathbb{F}_q$) such that $f(\alpha) = P(\alpha)$ for every $\alpha \in \mathbb{F}_q^m$. Note that since (by Exercise 2.3) for every $a \in \mathbb{F}_q$ we have $a^q - a = 0$, it follows that a minimal degree polynomial does not contain monomials with degree more than $q-1$ any single variable. In what follows,

Definition 9.1.2. We use $\deg_{X_i}(p)$ to denote the degree of polynomial p in variable X_i and $\deg_{X_i}(f)$ to denote the degree of (the minimal polynomial corresponding to) a function f in variable X_i.

For example $\deg_X(3X^3Y^4 + X^5 + Y^6) = 5$ and $\deg_Y(3X^3Y^4 + X^5 + Y^6) = 6$. Further, in this notation we have for every function $f : \mathbb{F}_q^m \to \mathbb{F}_q$, $\deg_{X_i}(f) \leq q-1$ for every $i \in [m]$.

Reed-Muller codes are given by three parameters: a prime power q and positive integers m and r, and consist of the evaluations of m-variate polynomials of degree at most r over all of the domain $\mathbb{F}_q^m$.

Definition 9.1.3 (Reed-Muller Codes)**.** The Reed-Muller code with parameters q, m, r, denoted $\mathrm{RM}(q, m, r)$, is the set of evaluations of all m-variate polynomials in $\mathbb{F}_q[X_1, \ldots, X_m]$ of total degree at most r and individual degree at most $q-1$ over all points in $\mathbb{F}_q^m$. Formally

$$\mathrm{RM}(q, m, r) \stackrel{\text{def}}{=} \left\{ f : \mathbb{F}_q^m \to \mathbb{F}_q \,|\, \deg(f) \leq r \right\}.$$

For example consider the case of $m = q = 2$ and $r = 1$. Note that all bivariate polynomials over $\mathbb{F}_2$ of degree at most 1 are 0, 1, X_1, X_2, $1 + X_1$, $1 + X_2$, $X_1 + X_2$ and $1 + X_1 + X_2$. Thus, we have that (where the evaluation points for (X_1, X_2) are ordered as $(0,0), (0,1), (1,0), (1,1)$):

$$\mathrm{RM}(2,2,1) = \{(0,0,0,0), (1,1,1,1), (0,0,1,1), (0,1,0,1), (1,1,0,0), (1,0,1,0), (0,1,1,0), (1,0,0,1)\}.$$

Also note that $\mathrm{RM}(q, m, 1)$ is almost the Hadamard code (see Exercise 5.6).

The Reed-Muller code with parameters (q, m, r) clearly has alphabet $\mathbb{F}_q$ and block length $n = q^m$. Also it can be verified that $\mathrm{RM}(q, m, r)$ is a linear code (see Exercise 9.1.) This leads to the following question, which will be the primary focus of this chapter:

Question 9.1.1. *What are the dimension and distance of an* $\mathrm{RM}(q,m,r)$ *code?*

The dimension of the code is the number of m-variate monomials of degree at most r, with the condition that degree in each variable is at most $q-1$. No simple closed form expression for this that works for all choices of q, m and r is known, so we will describe the effects only in some cases. The distance analysis of these codes takes a little bit more effort and we will start with two simple settings before describing the general result.

9.2 The low degree case

We start by considering $\mathrm{RM}(q,m,r)$ when $r < q$, i.e., the degree is smaller than the field size. We refer to this setting as the "low-degree" setting.

Dimension. The dimension of $\mathrm{RM}(q,m,r)$ in the low-degree case turns out to have a nice closed form, since we do not have to worry about the constraint that each variable has degree at most $q-1$: this is already imposed by restricting the total degree to at most $r \le q-1$. This leads to a nice expression for the dimension:

Proposition 9.2.1. *The dimension of the Reed Muller code* $\mathrm{RM}(q,m,r)$ *equals* $\binom{m+r}{r}$ *when* $r < q$.

Proof. The dimension equals the size of the set

$$D = \left\{ (d_1,\ldots,d_m) \in \mathbb{Z}^m | d_i \ge 0 \text{ for all } i \in [m], \sum_{i=1}^{m} d_i \le r \right\}, \tag{9.1}$$

since for every $(d_1,\ldots,d_m) \in D$, the monomial $X_1^{d_1}\cdots X_m^{d_m}$ is a monomial of degree at most r and these are all such monomials. The closed form expression for the dimension follows by a simple counting argument. (See Exercise 9.2). □

Distance. Next we turn to the analysis of the distance of the code. To understand the distance we will first state and prove a simple fact about the number of zeroes a multivariate polynomial can have. (We will have three versions of this in this chapter - with the third subsuming the first (Lemma 9.2.2) and second (Lemma 9.3.1), but the first two will be slightly simpler to state and remember.)

Lemma 9.2.2 (Polynomial Zero Lemma (low-degree case))**.** *Let* $f \in \mathbb{F}_q[X_1,\ldots,X_m]$ *be a non-zero polynomial with* $\deg(f) \le r$. *Then the fraction of zeroes of* f *is at most* $\frac{r}{q}$, *i.e.,*

$$\frac{|\{\mathbf{a} \in \mathbb{F}_q^m | f(\mathbf{a}) = 0\}|}{q^m} \le \frac{r}{q}.$$

We make couple of remarks. First note that the above lemma for $m = 1$ is the degree mantra (Proposition 5.2.3). We note that for every $m \geq 1$ the above lemma is tight (see Exercise 9.3). However, there exists polynomials for which the lemma is not tight (see Exercise 9.4).

Proof of Lemma 9.2.2. Note that the lemma statement is equivalent to saying that the probability that $f(\mathbf{a}) = 0$ is at most $\frac{\deg(f)}{q}$ when $\mathbf{a} = (a_1, \ldots, a_m)$ is chosen uniformly at random from $\mathbb{F}_q^m$. We claim that this holds by induction on m.

We will prove the lemma by induction on $m \geq 1$. Note that the base case follows from the degree mantra (Proposition 5.2.3). Now consider the case of $m > 1$ (and we assume that the lemma is true for $m-1$). To apply inductive hypothesis we first write f as a polynomial in X_m with coefficients that are themselves polynomials in $X_1, \ldots, X_{m-1}$. So let

$$f = f_0 X_m^0 + f_1 X_m^1 + \ldots f_t X_m^t,$$

where each $f_i(X_1, \ldots, X_{m-1})$ is a polynomial from $\mathbb{F}_q[X_1, \ldots, X_{m-1}]$ and $\deg(f_i) \leq r - i$. Furthermore let t be the largest index such that f_t is not zero. Now we consider picking $\mathbf{a} \in \mathbb{F}_q^m$ in two steps: We first pick $(a_1, \ldots, a_{m-1})$ uniformly at random from $\mathbb{F}_q^{m-1}$, and then we pick a_m uniformly from $\mathbb{F}_q$. Let

$$f^{(a_1,\ldots,a_{m-1})}(X_m) = f_0(a_1, \ldots, a_{m-1}) X_m^0 + \cdots + \ldots f_t(a_1, \ldots, a_{m-1}) X_m^t.$$

We consider two possible events:

$$\mathscr{E}_1 = \{(a_1, \ldots, a_m) | f_t(a_1, \ldots, a_{m-1}) = 0\}$$

and

$$\mathscr{E}_2 = \{((a_1, \ldots, a_m) | f_t(a_1, \ldots, a_{m-1}) \neq 0 \text{ and } f^{(a_1,\ldots,a_{m-1})}(a_m) = 0\}.$$

By the inductive hypothesis, we have that

$$\Pr[\mathscr{E}_1] \leq \frac{r-t}{q}, \tag{9.2}$$

since $\deg(f_t) \leq r - t$ and $f_t \neq 0$.

For every $(a_1, \ldots, a_{m-1}) \in \mathbb{F}_q^{m-1}$ such that $f_t(a_1, \ldots, a_{m-1}) \neq 0$ we also have that the univariate polynomial $f^{(a_1,\ldots,a_{m-1})}(X_m)$ is non-zero and of degree at most t, and so by the degree mantra it has at most t roots. It follows that for every such $(a_1, \ldots, a_{m-1})$ the probability, over a_m, that $f^{(a_1,\ldots,a_{m-1})}(a_m) = 0$ is at most $\frac{t}{q}$. In turn, it now immediately follows that

$$\Pr[\mathscr{E}_2] \leq \frac{t}{q}. \tag{9.3}$$

Finally, we claim that if neither $\mathscr{E}_1$ nor $\mathscr{E}_2$ occur, then $f(\mathbf{a}) \neq 0$. This is immediate from the definitions of $\mathscr{E}_1$ and $\mathscr{E}_2$, since if $f(a_1, \ldots, a_m) = 0$, it must either be the case that $f_t(a_1, \ldots, a_{m-1}) = 0$ (corresponding to $\mathscr{E}_1$) or it must be that $f_t(a_1, \ldots, a_{m-1}) \neq 0$ and $f^{(a_1,\ldots,a_{m-1})}(a_m) = 0$ (covered by $\mathscr{E}_2$). Note that this implies that $\Pr_{\mathbf{a}}[f(\mathbf{a}) = 0] \leq \Pr[\mathscr{E}_1 \cup \mathscr{E}_2]$. The lemma now follows from the fact that

$$\Pr_{\mathbf{a}}[f(\mathbf{a}) = 0] \leq \Pr[\mathscr{E}_1 \cup \mathscr{E}_2] \leq \Pr[\mathscr{E}_1] + \Pr[\mathscr{E}_2] \leq \frac{r}{q},$$

where the second inequality follows from the union bound (Proposition 3.1.3) and the final inequality follows from (9.2) and (9.3). □

Comparison with other codes

The lemmas above, while quite precise may not be fully transparent in explaining the asymptotics of the performance of the Reed-Muller codes, or contrast them with other codes we have seen. We mention a few basic facts here to get a clearer comparison.

If we set $m = 1$ and $r = k - 1$, then we get the Reed-Solomon codes evaluated on all of F_q (see Chapter 5). If we set $m = k - 1$, $r = 1$ and $q = 2$, then we get family of extended Hadamard codes (extended by including all Hadamard codewords and their complements). For more on this, see Exercise 5.6.

Thus Reed-Muller codes generalize some previously known codes - some with large alphabets and some with small alphabets. Indeed if we wish the alphabet to be small compared to the block length, then we can pick m to be a constant. For instance if we choose $m = 2$, we get codes of length n over an alphabets of size $\sqrt{n}$, while for any choice of relative distance δ, the code has rate $\frac{(1-\delta)^2}{2}$. In general for larger values of m, the code has alphabet size $n^{1/m}$ and rate $\frac{(1-\delta)^m}{m!}$. (See Exercise 9.5.) Thus for small values of m and fixed positive distance $\delta < 1$ there is a rate $R > 0$ such that, by choosing q appropriately large, one get codes on infinitely long block length n and alphabet $n^{1/m}$ with rate R and distance δ, which answers Question 9.0.1 in the affirmative.

This is one of the simplest such families of codes with this feature. We will do better in later in the book (e.g. Chapter 10), and indeed get alphabet size q independent of n with $R > 0$ and $\delta > 0$. But for now this is best we have.

9.3 The case of the binary field

Next we turn to a different extreme of parameter choices for the Reed-Muller codes. Here we fix the alphabet size $q = 2$ and see what varying m and r gets us.

Since we will prove a stronger statement later in Lemma 9.4.1, we only state the distance of the code $\mathrm{RM}(2, m, r)$ below, leaving the proof to Exercise 9.6.

Lemma 9.3.1 (Polynomial distance (binary case))**.** *Let f be a non-zero polynomial from $\mathbb{F}_2[X_1, \ldots, X_m]$ with $\deg_{X_i}(f) \le 1$ for every $i \in [m]$. Then $|\{\mathbf{a} \in \mathbb{F}_2^m | f(\mathbf{a}) \neq 0\}| \ge 2^{m-\deg(f)}$.*

Further, it can be established that the bound in Lemma 9.3.1 is tight (see Exercise 9.7).

The dimension of the code is relatively straightforward to analyze. The dimension is again given by the number of monomials of degree at most r. Since the degree in each variable is either zero or one, this just equals the number of subsets of $[m]$ of size at most r. Thus we have:

Proposition 9.3.2. *For any $r \le m$, the dimension of the Reed-Muller code $\mathrm{RM}(2, m, r)$ is exactly $\sum_{i=0}^{r} \binom{m}{i}$.*

Lemma 9.3.1 and Proposition 9.3.2 imply the following result:

Theorem 9.3.3. *For every $r \le m$, the Reed-Muller code $\mathrm{RM}(2, m, r)$ is a code of block length 2^m, dimension $\sum_{i=0}^{r} \binom{m}{i}$ and distance 2^{m-r}.*

Again, to get a sense of the asymptotics of this code, we can fix $\tau > 0$ and set $r = \tau \cdot m$ and let $m \to \infty$. In this case we get a code of block length n (for infinitely many n) with rate roughly $n^{H(\tau)-1}$ and distance $n^{-\tau}$ (see Exercise 9.8). So both the rate and the distance tend to zero at a rate that is a small polynomial in the block length but the code has a constant sized alphabet. (Note that this implies that we have made some progress towards answering Question 8.3.2.)

9.4 The general case

We now turn to the general case, where q is general and r is allowed to be larger than $q-1$. We will try to analyze the dimension and distance of this code. The distance turns out to still have a clean expression, so we will do that first. The dimension does not have a simple expression describing it exactly, so we will give a few lower bounds that may be generally useful (and are often asymptotically tight).

9.4.1 The general case: Distance

Lemma 9.4.1 (Polynomial distance (general case))**.** *Let f be a non-zero polynomial from $\mathbb{F}_q[X_1 \ldots, X_m]$ with $\deg_{X_i}(f) \le q-1$ for every $i \in [m]$ and $\deg(f) \le r$. Furthermore, let s, t be the unique non-negative integers such that $t \le q-2$ and*

$$s(q-1) + t = r.$$

Then

$$|\{\mathbf{a} \in \mathbb{F}_q^m | f(\mathbf{a}) \neq 0\}| \ge (q-t) \cdot q^{m-s-1} \ge q^{m-\frac{r}{q-1}}.$$

Hence, $\mathrm{RM}(q,m,r)$ *has distance at least* $q^{m-\frac{r}{q-1}}$.

Before proving the lemma we make a few observations: The above lemma clearly generalizes both Lemma 9.2.2 (which corresponds to the case $s = 0$) and Lemma 9.3.1 (where $q = 2$, $s = r-1$ and $t = 1$). In the general case the second lower bound is a little simpler and it shows that the probability that a polynomial is non-zero at a uniformly chosen point in $\mathbb{F}_q^m$ is at least $q^{-r/(q-1)}$. Finally, we note that Lemma 9.4.1 is tight for all settings of parameters (see Exercise 9.9).

Proof of Lemma 9.4.1. The proof is similar to the proof of Lemma 9.2.2 except we take advantage of the fact that the degree in a single variable is at most $q-1$. We also need to prove some simple inequalities.

As in the proof of Lemma 9.2.2 we prove that for a random choice of $\mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{F}_q^m$, the probability that $f(\mathbf{a}) \neq 0$ is at least

$$(q-t) \cdot q^{-(s+1)}. \tag{9.4}$$

Note that in contrast to the proof of Lemma 9.2.2 we focus on the good events — the polynomial being non-zero — rather than on the bad events.

We prove the lemma by induction on m. In the case of $m = 1$ we have by the degree mantra (Proposition 5.2.3) that the probability that $f(a_1) \neq 0$ is at least $\frac{q-r}{q}$. If $r < q-1$ we have $s = 0$ and $t = r$ and so the expression in (9.4) satisfies

$$(q-t) \cdot q^{-1} = \frac{q-r}{q} \leq \Pr[f(a_1) \neq 0].$$

If $r = q-1$ we have $s = 1$ and $t = 0$, but then again we have that (9.4) equals

$$q \cdot q^{-2} = \frac{q-(q-1)}{q} \leq \Pr[f(a_1) \neq 0],$$

where the inequality follows from the degree mantra.

Now we turn to the inductive step. Assume the hypothesis is true for $(m-1)$-variate polynomials and let $f = \sum_{i=0}^{b} f_i X_m^i$ where $f_i \in \mathbb{F}_q[X_1, \ldots, X_{m-1}]$ with $f_b \neq 0$. Note $0 \leq b \leq q-1$ and $\deg(f_b) \leq r-b$. Let $\mathscr{E}$ be the event of interest to us, i.e.,

$$\mathscr{E} = \{(a_1, \ldots, a_m) | f(a_1, \ldots, a_m) \neq 0\}.$$

Let

$$\mathscr{E}_1 = \{(a_1, \ldots, a_{m-1}) | f_b(a_1, \ldots, a_{m-1}) \neq 0\}.$$

We first bound $\Pr[\mathscr{E}|\mathscr{E}_1]$. Fix $a_1, \ldots, a_{m-1}$ such that $f_b(a_1, \ldots, a_{m-1}) \neq 0$ and let

$$P(Z) = \sum_{i=0}^{b} f_i(a_1, \ldots, a_{m-1}) Z^i.$$

Note P is a non-zero polynomial of degree b and we have

$$\Pr[f(a_1, \ldots, a_m) = 0 | a_1, \ldots, a_{m-1}] = \Pr_{a_m}[P(a_m) \neq 0].$$

Since by the degree mantra, a univariate polynomial of degree b has at most b roots, we have

$$\Pr_{a_m}[P(a_m) \neq 0] \geq \frac{q-b}{q}.$$

We conclude

$$\Pr[\mathscr{E}|\mathscr{E}_1] \geq 1 - \frac{b}{q}.$$

Next we will bound $\Pr[\mathscr{E}_1]$. This will allow us to lower bound the probability of $\mathscr{E}$ since

$$\Pr[\mathscr{E}] \geq \Pr[\mathscr{E} \text{ and } \mathscr{E}_1] = \Pr[\mathscr{E}_1] \cdot \Pr[\mathscr{E}|\mathscr{E}_1].$$

Recall that $\deg(f_b) \leq r-b$. Write $r-b = s'(q-1) + t'$ where $s', t' \geq 0$ and $t' \leq q-2$. By induction we have

$$\Pr[\mathscr{E}_1] = \Pr[f_b(a_1, \ldots, a_{m-1}) \neq 0] \geq (q-t') \cdot q^{-(s'+1)}.$$

Putting the two bounds together, we get

$$\Pr[\mathscr{E}] \geq \Pr[\mathscr{E}|\mathscr{E}_1] \cdot \Pr[\mathscr{E}_1] \geq \frac{q-b}{q} \cdot (q-t') \cdot q^{-(s'+1)}.$$

We are now left with a calculation to verify that the bound above is indeed lower bounded by $(q-t) \cdot q^{-(s+1)}$ and we do so in Claim 9.4.2 using the facts that $t, t' \leq q-2$, $b \leq q-1$, $r = s(q-1)+t$, and $r-b = s'(q-1)+t'$. In the claim further below (Claim 9.4.3), we also prove $(q-t) \cdot q^{-(s+1)} \geq q^{-r/(q-1)}$ and this concludes the proof of the lemma. □

Claim 9.4.2. *If q, r, s, t, s', t', b are non-negative integers such that $r = s(q-1)+t$, $r-b = s'(q-1)+t'$, $t, t' \leq q-2$ and $b \leq q-1$ then we have*

$$\frac{q-b}{q} \cdot (q-t') \cdot q^{-(s'+1)} \geq (q-t) \cdot q^{-(s+1)}.$$

Proof. The proof breaks up in to two cases depending on $s-s'$. Note that an equivalent definition of s and s' are that these are the quotients when we divide r and $r-b$ respectively by $q-1$. Since $0 \leq b \leq q-1$, it follows that either $s' = s$ or $s' = s-1$. We consider the two cases separately.

If $s = s'$ we have $t = t'+b$ and then it suffices to show that

$$\frac{q-b}{q} \cdot (q-t') \geq q-(t'+b).$$

In turn this is equivalent to showing

$$(q-b)(q-t') \geq q(q-(t'+b)).$$

But this is immediate since the expression on the left is

$$(q-b)(q-t') = q^2-(b+t')q+bt' = q(q-(b+t'))+bt' \geq q(q-(b+t')),$$

where the final inequality uses $bt' \geq 0$.

If $s = s'+1$ we have a bit more work. Here we have $t+q-1 = t'+b$ and it suffices to show that

$$\frac{q-b}{q} \cdot (q-t') \cdot q \geq (q-t) = (2q-(t'+b+1)).$$

Write $q-b = \alpha$ and $q-t' = \beta$. The expression on the left above simplifies to $\alpha\beta$ and on the right to $\alpha+\beta-1$. Since $b, t' \leq q-1$, we also have $\alpha, \beta \geq 1$. So it suffices to show that $\alpha\beta \geq \alpha+\beta-1$. This is true since $\alpha\beta = \alpha+\alpha(\beta-1)$ and we have $\alpha(\beta-1) \geq \beta-1$ since $\alpha \geq 1$ and $\beta-1 \geq 0$.

We thus conclude that the inequality holds for both $s = s'$ and $s = s'+1$ and this yields the claim. □

Claim 9.4.3. *Let q, r, s, t be non-negative real numbers such that $q \geq 2$, $r = s(q-1)+t$ and $t \leq q-2$. Then*

$$(q-t) \cdot q^{-(s+1)} \geq q^{-r/(q-1)}.$$

We remark that while the inequality is quite useful, the proof below is not particularly insightful. We include it for completeness, but we recommend that the reader skip it unless necessary.

Proof of Claim 9.4.3. We have four parameters in the inequality above. We will simplify it in steps removing parameters one at a time. First we get rid of r by substituting $r = s(q-1)+t$. So it suffices to prove:

$$(q-t)\cdot q^{-(s+1)} \geq q^{-(s(q-1)+t)/(q-1)} = q^{-s}\cdot q^{-t/(q-1)}.$$

We can get rid of q^{-s} from both sides (since the remaining terms are non-negative) and so it suffices to prove:

$$\frac{q-t}{q} \geq q^{-t/(q-1)}.$$

Let $f_q(t) = \frac{t}{q} + q^{-t/(q-1)} - 1$. The inequality above is equivalent to proving $f_q(t) \leq 0$ for $0 \leq t \leq q-2$. We use some basic calculus to prove the above. Note that the first and second derivatives of f_q with respect to t are given by $f_q'(t) = \frac{1}{q} - \frac{\ln q}{q-1} q^{-t/(q-1)}$ and $f_q''(t) = \left(\ln(q)/(q-1)\right)^2 q^{-t/(q-1)}$. In particular the second derivative is always positive which means $f_q(t)$ is maximized at one of the two end points of the interval $t \in [0, q-2]$. We have $f_q(0) = 0 \leq 0$ as desired and so it suffices to prove that

$$f_q(q-2) = q^{-(q-2)/(q-1)} - \frac{2}{q} \leq 0.$$

Multiplying the expression above by q we have that it suffices to show $q^{1/(q-1)} \leq 2$ which in turn is equivalent to proving $q \leq 2^{q-1}$ for every $q \geq 2$. The final inequality follows easily from Bernoulli's inequality (Lemma B.1.4) $1 + kx \leq (1+x)^k$ which holds for every $x \geq -1$ and $k \geq 1$. In our case we substitute $x = 1$ and $k = q-1$ to conclude $q \leq 2^{q-1}$ as desired. □

9.4.2 The general case: Dimension

For integers q, m, r let

$$S_{q,m,r} = \left\{ \mathbf{d} = (d_1, \ldots, d_m) \in \mathbb{Z}^m | 0 \leq d_i \leq q-1 \text{ for all } i \in [m] \text{ and }, \sum_{i=1}^{m} d_i \leq r \right\} \tag{9.5}$$

and let

$$K_{q,m,r} = |S_{q,m,r}|.$$

We start with the following, almost tautological, proposition.

Proposition 9.4.4. *For every prime power q and integers $m \geq 1$ and $r \geq 0$, the dimension of the code* $\mathrm{RM}(q,m,r)$ *is $K_{q,m,r}$.*

Proof. Follows from the fact that for every $\mathbf{d} = (d_1, \ldots, d_m) \in S_{q,m,r}$ the associated monomial $\mathbf{X}^{\mathbf{d}} = X_1^{d_1} \cdots X_m^{d_m}$ is a monomial of degree at most r and individual degree at most $q-1$. Thus these monomials (i.e., their evaluations) form a basis for the Reed-Muller code $\mathrm{RM}(q,m,r)$. (See Exercise 9.10.) □

The definition of $K_{q,m,r}$ does not give a good hint about its growth so below we give a few bounds on $K_{q,m,r}$ that help estimate its growth. Specifically the proposition below gives a lower bound $K^-_{q,m,r}$ and an upper bound $K^+_{q,m,r}$ on $K_{q,m,r}$ that are (1) given by simple expressions and (2) within polynomial factors of each other for every setting of q, m, and r.

Proposition 9.4.5. *For integers $q \geq 2$, $m \geq 1$ and $r \geq 0$, let*

$$K^+_{q,m,r} \triangleq \min\left\{q^m, \binom{m+r}{r}\right\}$$

and let

$$K^-_{q,m,r} \triangleq \begin{cases} \max\left\{q^m/2, q^m - K^+_{q,m,(q-1)m-r}\right\} & \text{if } r \geq (q-1)m/2 \\ \max\left\{\binom{m}{r}, \frac{1}{2}\left(\left\lfloor \frac{2r+m}{m} \right\rfloor\right)^m\right\} & \text{if } r < (q-1)m/2 \end{cases}$$

Then there are universal constants c_1, c_2 ($c_1 < 3.1$ and $c_2 < 8.2$ suffice) such that

$$K^-_{q,m,r} \leq K_{q,m,r} \leq K^+_{q,m,r} \leq c_1 \cdot (K^-_{q,m,r})^{c_2}$$

.

Proof. We tackle the inequalities in order of growing complexity of the proof. In our bounds we use the fact that $K_{q,m,r}$ is monotone non-decreasing in q as well as r (when other parameters are fixed)– see Exercise 9.11.

First we prove $K_{q,m,r} \leq K^+_{q,m,r}$. On the one hand we have

$$K_{q,m,r} \leq K_{q,m,(q-1)m} = q^m,$$

which follows by ignoring the total degree restriction and on the other hand we have

$$K_{q,m,r} \leq K_{r,m,r} = \binom{m+r}{r},$$

whereas here we ignored the individual degree restriction.

Next we show $K^-_{q,m,r} \leq K_{q,m,r}$. First we consider the case $r \geq (q-1)m/2$. Here we argue via symmetry. Consider a map that maps vectors $\mathbf{d} = (d_1, \ldots, d_m) \in \mathbb{Z}^m$ with $0 \leq d_i < q$ to $\overline{\mathbf{d}} = (q-1-d_1, \ldots, q-1-d_m)$. The map $\mathbf{d} \to \overline{\mathbf{d}}$ is a one-to-one map which maps vectors with $\sum_i d_i > r$ to vectors with $\sum_i \overline{d_i} < (q-1)m - r$. In other words either $d \in \{0, \ldots, q-1\}^m$ is in $S_{q,m,r}$ or $\overline{\mathbf{d}} \in S_{q,m,(q-1)m-r}$, thus establishing

$$K_{q,m,r} = q^m - K_{q,m,(q-1)m-r}.$$

Since $r \geq (q-1)m/2$ we have $(q-1)m - r \leq r$ and so

$$K_{q,m,r} \geq K_{q,m,(q-1)m-r},$$

which in turn implies

$$K_{q,m,r} \geq q^m/2.$$

This establishes $K_{q,m,r} \geq K^-_{q,m,r}$ when $r \geq (q-1)m/2$. Next, turning to the case $r < (q-1)m/2$, first let $q' = \left\lfloor \frac{2r+m}{m} \right\rfloor$. We have

$$K_{q,m,r} \geq K_{q',m,r} \geq (q')^m/2$$

since $r \geq (q'-1)m/2$, and this yields

$$K_{q,m,r} \geq (q')^m/2 = \frac{1}{2}\left(\left\lfloor \frac{2r+m}{m} \right\rfloor\right)^m.$$

Finally we also have

$$K_{q,m,r} \geq K_{2,m,r} = \sum_{i=0}^{r} \binom{m}{i} \geq \binom{m}{r},$$

thus establishing $K_{q,m,r} \geq K^-_{q,m,r}$ when $r < (q-1)m/2$.

Finally we turn to the inequalities showing $K^+_{q,m,r} \leq c_1 \cdot (K^-_{q,m,r})^{c_2}$. If $r \geq (q-1)m/2$ we have

$$\frac{q^m}{2} \leq K^-_{q,m,r} \leq K^+_{q,m,r} \leq q^m$$

establishing $K^+_{q,m,r} \leq 2K^-_{q,m,r}$. Next we consider the case $r < m/2$. In this case we have

$$K^-_{q,m,r} \geq \binom{m}{r} \geq (m/r)^r \geq 2^r.$$

On the other hand we also have

$$\binom{m+r}{r} \leq \left(\frac{e(m+r)}{r}\right)^r \leq \left(\frac{e\cdot(3/2)\cdot m}{r}\right)^r = \left(\frac{3e}{2}\right)^r \cdot \left(\frac{m}{r}\right)^r.$$

From $2^r \leq K^-_{q,m,r}$ we get $\left(\frac{3e}{2}\right)^r \leq (K^-_{q,m,r})^{\log_2(3e/2)}$. Combining with $\left(\frac{m}{r}\right)^r \leq K^-_{q,m,r}$ and $K^+_{q,m,r} \leq \binom{m+r}{r}$ we get

$$K^+_{q,m,r} \leq \left(\frac{3e}{2}\right)^r \cdot \left(\frac{m}{r}\right)^r \leq (K^-_{q,m,r})^{1+\log_2(3e/2)}$$

. Finally, we consider the case $m/2 \leq r < (q-1)m/2$. In this range we have

$$\left\lfloor \frac{2r+m}{m} \right\rfloor = 1 + \left\lfloor \frac{2r}{m} \right\rfloor \geq 1 + \frac{r}{m} = \frac{m+r}{m}.$$

Thus

$$K^-_{q,m,r} \geq \frac{1}{2}\left(\left\lfloor \frac{2r+m}{m} \right\rfloor\right)^m \geq \frac{1}{2}\left(\frac{m+r}{m}\right)^m \geq \frac{1}{2}\left(\frac{3}{2}\right)^m.$$

On the other hand we have

$$K^+_{q,m,r} \leq \binom{m+r}{m} \leq \left(\frac{e(m+r)}{m}\right)^m = e^m \cdot \left(\frac{m+r}{m}\right)^m.$$

Again we have $\left(\frac{m+r}{m}\right)^m \leq 2K^-_{q,m,r}$ and $e^m \leq (2K^-_{q,m,r})^{\log_2(3e/2)}$ and so $K^+_{q,m,r} \leq (2K^-_{q,m,r})^{1+\log_2(3e/2)}$. Thus in all cases we have $K^+_{q,m,r} \leq c_1 \cdot (K^-_{q,m,r})^{c_2}$ for $c_2 = 1 + \log_2(3e/2) < 3.1$ and $c_1 = 2^{c_2} < 8.2$, as desired. □

We now give a few examples of codes that can be derived from the bounds above, to illustrate the variety offered by Reed-Muller codes. In each of the cases we set one or more of the parameters among alphabet size, rate, (relative) distance or absolute distance to a constant and explore the behavior in the other parameters. In all cases we use Lemma 9.4.1 to lower bound the distance and Proposition 9.4.5 to lower bound the dimension.

Example 9.4.6 (RM Codes of constant alphabet size and (relative) distance.)**.** *Fix q and $r < q-1$ and consider $m \to \infty$. Then the Reed-Muller codes* $\mathrm{RM}(q,m,r)$ *are $[N,K,D]_q$ codes with block length $N = q^m$, distance $D = \delta \cdot N$ for $\delta = 1 - r/q$, with dimension*

$$K \geq \binom{m}{r} \geq \left(\frac{\log_q N}{r}\right)^r.$$

In other words Reed-Muller codes yield codes of constant alphabet size and relative distance with dimension growing as an arbitrary polynomial in the logarithm of the block length.

Example 9.4.7 (Binary RM Codes of rate close to 1 with constant (absolute) distance.)**.** *Fix $q = 2$ and d and let $m \to \infty$. Then the Reed-Muller codes* $\mathrm{RM}(2,m,m-d)$ *are $[N,K,D]_2$ codes with $N = 2^m$, $D = 2^d$ and*

$$K \geq N - \binom{\log_2 N + d}{d} \geq N - (\log_2 N)^d.$$

(See Exercise 9.12 for bound on K.) Note that the rate $\to 1$ as $N \to \infty$.

Example 9.4.8 (RM codes of constant rate and relative distance over polynomially small alphabets.)**.** *Given any $\varepsilon > 0$ and let $m = \lceil \frac{1}{\varepsilon} \rceil$ and now consider $q \to \infty$ with $r = q/2$. Then the Reed-Muller codes* $\mathrm{RM}(q,m,r)$ *are $[N,K,D]_q$ codes with $N = q^m$, $D = \frac{N}{2}$ and*

$$K \geq \frac{1}{2}\left(\frac{q+m}{m}\right)^m \geq \frac{1}{2m^m} \cdot N.$$

Expressed in terms of N and ε, the codes have length N, dimension $\Omega\left(\varepsilon^{1/\varepsilon}\right) \cdot N$ and relative distance $1/2$ over an alphabet of size N^ε.

Another natural regime is to consider the case of constant rate $1/2$: see Exercise 9.13 for more.

Finally we mention a range of parameters that has been very useful in the theory of computer science. Here the alphabet size is growing with N, but very slowly. But the code has a fixed relative distance and dimension that is polynomially related to the block length.

Example 9.4.9 (RM Codes over polylogarithmic alphabets with polynomial dimension.)**.** *Given $0 < \varepsilon < 1$, let $q \to \infty$ and let $r = q/2$ and $m = q^\varepsilon$. Then the Reed-Muller codes* $\mathrm{RM}(q,m,r)$ *are $[N,K,D]_q$ codes with $N = q^m$, $D = \frac{N}{2}$ and*

$$K \geq \frac{1}{2}\left(\frac{q+m}{m}\right)^m \geq \frac{1}{2}\left(q^{1-\varepsilon}\right)^m = \frac{1}{2} \cdot N^{1-\varepsilon}.$$

Expressed in terms of N and ε, the codes have length N, dimension $\Omega(N^{1-\varepsilon})$ and relative distance $1/2$ over an alphabet of size $(\log N)^{1/\varepsilon}$. (See Exercise 9.14 for claim on the bound on q.)

9.5 Exercises

Exercise 9.1. Argue that any $\mathrm{RM}(q, m, r)$ is a linear code.

Exercise 9.2. Argue that for D as defined in (9.1), we have

$$|D| = \binom{m+r}{r}.$$

Exercise 9.3. Show that Lemma 9.2.2 is tight in the sense that for every prime power q and integers $m \geq 1$ and $1 \leq r \leq q-1$, there exists a polynomial with exactly $r \cdot q^{m-1}$ roots.

Exercise 9.4. Show that Lemma 9.2.2 is not tight for most polynomials. In particular show that for every prime power q and integers $m \geq 1$ and $1 \leq r \leq q-1$, a random polynomial in $\mathbb{F}_q[X_1, \ldots, X_m]$ of degree r has q^{m-1} expected number of roots.

Exercise 9.5. Show that the Reed-Muller codes of Section 9.2 give rise to codes of relative distance δ (for any $0 < \delta < 1$) and block length n such that they have alphabet size of $\sqrt[m]{n}$ and rate $\frac{(1-\delta)^m}{m!}$.

Exercise 9.6. Prove Lemma 9.3.1.

Exercise 9.7. Prove that the lower bound in Lemma 9.3.1 is tight.

Exercise 9.8. Show that there exists a binary RM code with block length n, rate $n^{H(\tau)-1}$ and relative distance $n^{-\tau}$ for any $0 < \tau < 1/2$.

Exercise 9.9. Prove that the (first) lower bound in Lemma 9.4.1 is tight for all settings of the parameters.

Exercise 9.10. Prove that the evaluations of $\mathbf{X}^{\mathbf{d}}$ for every $\mathbf{d} \in S_{q,m,r}$ (as in (9.5)) form a basis for $\mathrm{RM}(q, m, r)$.

Exercise 9.11. Argue that $K_{q,m,r}$ is monotone non-decreasing in q as well as r (when other parameters are fixed).

Exercise 9.12. Argue the claimed bound on K in Example 9.4.7.

Exercise 9.13. Figure out a RM code that has rate $\frac{1}{2}$ and has as large a distance as possible and as small an alphabet as possible.

Exercise 9.14. Prove the claimed bound on q in Example 9.4.9.

Exercise 9.15. In this problem we will talk about the dual of Reed-Muller codes, which turn out to be Reed-Muller codes (with a different degree) themselves. We do so in a sequence of subproblems:

1. Show that for $1 \leq j \leq q-1$

$$\sum_{\alpha \in \mathbb{F}_q} \alpha^j \neq 0$$

   if and only if $j = q-1$.

   Hint: Use Exercise 2.2.

2. Argue that for any $m \geq 1$ and $1 \leq j_1, \ldots, j_m \leq q-1$,

$$\sum_{(c_1,\ldots,c_m)\in\mathbb{F}_q^m} \prod_{i=1}^{m} c_i^{j_1} = 0$$

   if and only if $j_1 = j_2 = \cdots = j_m = q-1$.

3. Using the above or otherwise, show that for any $0 \leq r < (q-1) - s$, we have

$$\mathrm{RM}(q,m,r)^{\perp} = \mathrm{RM}(q,m,m(q-1)-r-1).$$

9.6 Bibliographic Notes

We point out that the original code considered by Reed and Muller is the one in Section 9.3.

Chapter 10

From Large to Small Alphabets: Code Concatenation

Recall Question 8.3.2 that we had asked before: Is there an explicit asymptotically good binary code (that is, rate $R > 0$ and relative distance $\delta > 0$)? In this chapter, we will consider this question when we think of explicit code in the sense of Definition 6.3.1 as well as the stronger notion of a strongly explicit code (Definition 6.3.2).

Let us recall all the (strongly) explicit codes that we have seen so far. (See Table 10.1 for an overview.)

Code	R	δ
Hamming	$1 - O\left(\frac{\log n}{n}\right)$	$O\left(\frac{1}{n}\right)$
Hadamard	$O\left(\frac{\log n}{n}\right)$	$\frac{1}{2}$
Reed-Solomon	$\frac{1}{2}$	$O\left(\frac{1}{\log n}\right)$

Table 10.1: Strongly explicit binary codes that we have seen so far.

Hamming code (Section 2.4), which has rate $R = 1 - O(\log n/n)$ and relative distance $\delta = O(1/n)$ and the Hadamard code (Section 2.7), which has rate $R = O(\log n/n)$ and relative distance $1/2$. Both of these codes have extremely good R or δ at the expense of the other parameter. Next, we consider the Reed-Solomon code (of say $R = 1/2$) as a binary code, which does much better– $\delta = \frac{1}{\log n}$, as we discuss next.

Consider the Reed-Solomon over $\mathbb{F}_{2^s}$ for some large enough s. It is possible to get an $\left[n, \frac{n}{2}, \frac{n}{2}+1\right]_{2^s}$ Reed-Solomon code (i.e. $R = 1/2$). We now consider a Reed-Solomon codeword, where every symbol in $\mathbb{F}_{2^s}$ is represented by an s-bit vector. Now, the "obvious" binary code created by viewing symbols from $\mathbb{F}_{2^s}$ as bit vectors as above is an $\left[ns, \frac{ns}{2}, \frac{n}{2}+1\right]_2$ code[1]. Note that the distance of this code is only $\Theta\left(\frac{N}{\log N}\right)$, where $N = ns$ is the block length of the final binary code. (Recall that $n = 2^s$ and so $N = n \log n$.)

[1]The proof is left as an exercise.

The reason for the (relatively) poor distance is that the bit vectors corresponding to two different symbols in $\mathbb{F}_{2^s}$ may only differ by one bit. Thus, d positions which have different $\mathbb{F}_{2^s}$ symbols might result in a distance of only d as bit vectors.

To fix this problem, we can consider applying a function to the bit-vectors to increase the distance between those bit-vectors that differ in smaller numbers of bits. Note that such a function is simply a code! We define this recursive construction more formally next. This recursive construction is called concatenated codes and will help us construct (strongly) explicit asymptotically good codes.

10.1 Code Concatenation

For $q \geq 2$, $k \geq 1$ and $Q = q^k$, consider two codes which we call *outer* code and *inner* code:

$$C_{\text{out}} : [Q]^K \to [Q]^N,$$

$$C_{\text{in}} : [q]^k \to [q]^n.$$

Note that the alphabet size of C_{out} exactly matches the number of messages for C_{in}. Then given $\mathbf{m} = (m_1, \ldots, m_K) \in [Q]^K$, we have the code $C_{\text{out}} \circ C_{\text{in}} : [q]^{kK} \to [q]^{nN}$ defined as

$$C_{\text{out}} \circ C_{\text{in}}(\mathbf{m}) = (C_{\text{in}}(C_{\text{out}}(\mathbf{m})_1), \ldots, C_{\text{in}}(C_{\text{out}}(\mathbf{m})_N)),$$

where

$$C_{\text{out}}(\mathbf{m}) = (C_{\text{out}}(\mathbf{m})_1, \ldots, C_{\text{out}}(\mathbf{m})_N).$$

This construction is also illustrated in Figure 10.1.

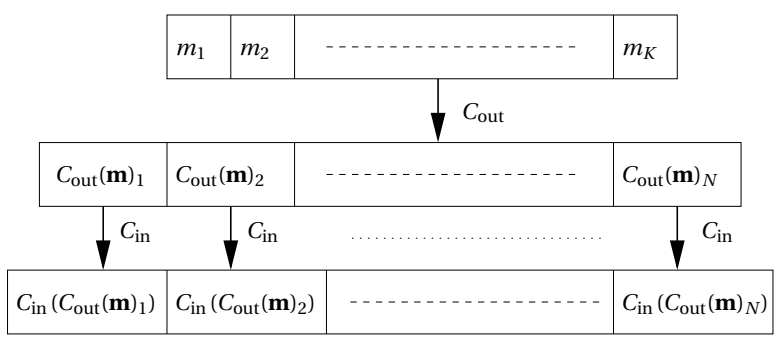


Figure 10.1: Concatenated code $C_{\text{out}} \circ C_{\text{in}}$.

We now look at some properties of a concatenated code.

Theorem 10.1.1. *If C_{out} is an $(N, K, D)_{q^k}$ code and C_{in} is an $(n, k, d)_q$ code, then $C_{\text{out}} \circ C_{\text{in}}$ is an $(nN, kK, dD)_q$ code. In particular, if C_{out} (C_{in} resp.) has rate R (r resp.) and relative distance δ_{out} (δ_{in} resp.) then $C_{\text{out}} \circ C_{\text{in}}$ has rate Rr and relative distance $\delta_{\text{out}} \cdot \delta_{\text{in}}$.*

Proof. The first claim immediately implies the second claim on the rate and relative distance of $C_{\text{out}} \circ C_{\text{in}}$. The claims on the block length, dimension and alphabet of $C_{\text{out}} \circ C_{\text{in}}$ follow from the definition.[2] Next we show that the distance is at least dD. Consider arbitrary $\mathbf{m}_1 \neq \mathbf{m}_2 \in [Q]^K$. Then by the fact that C_{out} has distance D, we have

$$\Delta\left(C_{\text{out}}\left(\mathbf{m}_1\right), C_{\text{out}}\left(\mathbf{m}_2\right)\right) \geq D. \tag{10.1}$$

Thus for each position $1 \leq i \leq N$ that contributes to the distance above, we have

$$\Delta\left(C_{\text{in}}\left(C_{\text{out}}\left(\mathbf{m}_1\right)_i\right), C_{\text{in}}\left(C_{\text{out}}\left(\mathbf{m}_2\right)_i\right)\right) \geq d, \tag{10.2}$$

as C_{in} has distance d. Since there are at least D such positions (from (10.1)), (10.2) implies

$$\Delta\left(C_{\text{out}} \circ C_{\text{in}}\left(\mathbf{m}_1\right), C_{\text{out}} \circ C_{\text{in}}\left(\mathbf{m}_2\right)\right) \geq dD.$$

The proof is complete as the choices of $\mathbf{m}_1$ and $\mathbf{m}_2$ were arbitrary. □

If C_{in} and C_{out} are linear codes, then so is $C_{\text{out}} \circ C_{\text{in}}$, which can be proved for example, by defining a generator matrix for $C_{\text{out}} \circ C_{\text{in}}$ in terms of the generator matrices of C_{in} and C_{out}. The proof is left as an exercise.

10.2 Zyablov Bound

We now instantiate an outer and inner codes in Theorem 10.1.1 to obtain a new lower bound on the rate given a relative distance. We'll initially just state the lower bound (which is called the Zyablov bound) and then we will consider the explicitness of such codes.

We begin with the instantiation of C_{out}. Note that this is a code over a large alphabet and we have seen an optimal code over large enough alphabet: Reed-Solomon codes (Chapter 5). Recall that the Reed-Solomon codes are optimal because they meet the Singleton bound 4.3.1. Hence, let us assume that C_{out} meets the Singleton bound with rate of R, i.e. C_{out} has relative distance $\delta_{\text{out}} > 1 - R$. Note that now we have a chicken and egg problem here. In order for $C_{\text{out}} \circ C_{\text{in}}$ to be an asymptotically good code, C_{in} needs to have rate $r > 0$ and relative distance $\delta_{\text{in}} > 0$ (i.e. C_{in} also needs to be an asymptotically good code). This is precisely the kind of code we are looking for to answer Question 8.3.2! However the saving grace will be that k can be much smaller than the block length of the concatenated code and hence, we can spend "more" time searching for such an inner code.

Suppose C_{in} meets the GV bound (Theorem 4.2.1) with rate of r and thus with relative distance $\delta_{\text{in}} \geq H_q^{-1}(1-r) - \varepsilon$, for some $\varepsilon > 0$. Then by Theorem 10.1.1, $C_{\text{out}} \circ C_{\text{in}}$ has rate of rR and $\delta = (1-R)(H_q^{-1}(1-r) - \varepsilon)$. Expressing R as a function of δ and r, we get the following:

$$R = 1 - \frac{\delta}{H_q^{-1}(1-r) - \varepsilon}.$$

[2] Technically, we need to argue that the q^{kK} messages map to distinct codewords to get the dimension of kK. However, this follows from the fact, which we will prove soon, that $C_{\text{out}} \circ C_{\text{in}}$ has distance $dD \geq 1$, where the inequality follows for $d, D \geq 1$.

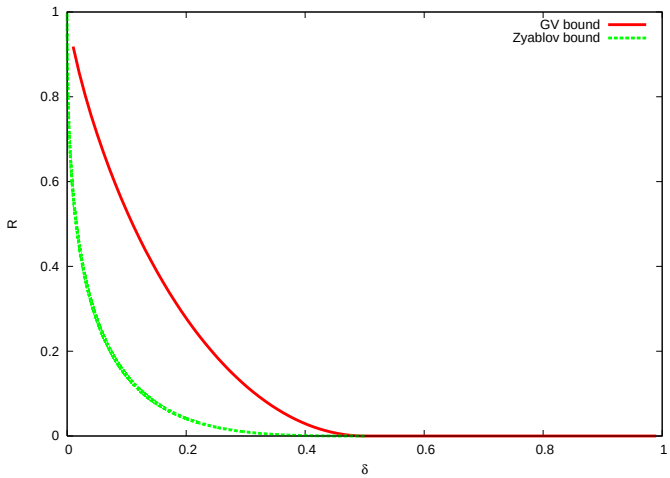


Figure 10.2: The Zyablov bound for binary codes. For comparison, the GV bound is also plotted.

Then optimizing over the choice of r, we get that the rate of the concatenated code satisfies

$$\mathscr{R} \geq \max_{0<r<1-H_q(\delta+\varepsilon)} r\left(1-\frac{\delta}{H_q^{-1}(1-r)-\varepsilon}\right),$$

where the bound of $r < 1 - H_q(\delta+\varepsilon)$ is necessary to ensure that $\mathscr{R} > 0$. This lower bound on the rate is called the Zyablov bound. See Figure 10.2 for a plot of this bound for binary codes.

To get a feel for how the bound behaves, consider the case when $\delta = \frac{1}{2} - \varepsilon$. We claim that the Zybalov bound states that $\mathscr{R} \geq \Omega(\varepsilon^3)$. (Recall that the GV bound for the same δ has a rate of $\Omega(\varepsilon^2)$.) The proof of this claim is left as an exercise.

Note that the Zyablov bound implies that for every $\delta > 0$, there exists a (concatenated) code with rate $R > 0$. However, we already knew about the existence of an asymptotically good code by the GV bound (Theorem 4.2.1). Thus, a natural question to ask is the following:

Question 10.2.1. *Can we construct an explicit code on the Zyablov bound?*

We will focus on linear codes in seeking an answer to the question above because linear codes have polynomial size representation. Let C_{out} be an $[N,K]_Q$ Reed-Solomon code where $N = Q-1$ (evaluation points being $\mathbb{F}_Q^*$ with $Q = q^k$). This implies that $k = \Theta(\log N)$. However we still need an efficient construction of an inner code that lies on the GV bound. We do not expect to construct such a C_{in} in time poly(k) as that would answer Open Question 8.3.2! However, since $k = O(\log N)$, note that an exponential time in k algorithm is still a polynomial (in N) time algorithm.

There are two options for this exponential (in k) time construction algorithm for C_{in}:

- Perform an exhaustive search among all generator matrices for one satisfying the required property for C_{in}. One can do this because the Varshamov bound (Theorem 4.2.1) states that there exists a linear code which lies on the GV bound. This will take $q^{O(kn)}$ time. Using $k = rn$ (or $n = O(k)$), we get $q^{O(kn)} = q^{O(k^2)} = N^{O(\log N)}$, which is upper bounded by $(nN)^{O(\log(nN))}$, a quasi-polynomial time bound.

- The second option is to construct C_{in} in $q^{O(n)}$ time and thus use $(nN)^{O(1)}$ time overall. See Exercise 10.1 for one way to construct codes on the GV bound in time $q^{O(n)}$.

Thus,

Theorem 10.2.1. *We can construct a code that achieves the Zyablov bound in polynomial time.*

In particular, we can construct explicit asymptotically good code in polynomial time, which answers Question 10.2.1 in the affirmative.

A somewhat unsatisfactory aspect of this construction (in the proof of Theorem 10.2.1) is that one needs a brute force search for a suitable inner code (which led to the polynomial construction time). A natural followup question is

Question 10.2.2. *Does there exist a strongly explicit asymptotically good code?*

10.3 Strongly Explicit Construction

We will now consider what is known as the *Justesen code*. The main insight in these codes is that if we are only interested in asymptotically good codes, then the arguments in the previous section would go through even if (i) we pick different inner codes for each of the N outer codeword positions and (ii) most (but not necessarily all) inner code lie on the GV bound. It turns out that constructing an "ensemble" of codes such that most of the them lie on the GV bound is much easier than constructing a single code on the GV bound. For example, the ensemble of all linear codes have this property– this is exactly what Varshamov proved. However, it turns out that we need this ensemble of inner codes to be a smaller one than the set of all linear codes.

Justesen code is concatenated code with multiple, *different* linear inner codes. Specifically, it is composed of an $(N, K, D)_{q^k}$ outer code C_{out} and different inner codes $C_{\text{in}}^i : 1 \le i \le N$. Formally, the concatenation of these codes, denoted by $C_{\text{out}} \circ \left(C_{\text{in}}^1, \ldots, C_{\text{in}}^N\right)$, is defined as follows: given a message $\mathbf{m} \in \left[q^k\right]^K$, let the outer codeword be denoted by $(c_1, \ldots, c_N) \stackrel{\text{def}}{=} C_{\text{out}}(\mathbf{m})$. Then $C_{\text{out}} \circ \left(C_{\text{in}}^1, \ldots, C_{\text{in}}^N\right)(\mathbf{m}) = \left(C_{\text{in}}^1(c_1), C_{\text{in}}^2(c_2), \ldots, C_{\text{in}}^n(c_N)\right)$.

We will need the following result.

Theorem 10.3.1. *Let $\varepsilon > 0$. There exists an ensemble of inner codes $C_{\text{in}}^1, C_{\text{in}}^2, \ldots, C_{\text{in}}^N$ of rate $\frac{1}{2}$, where $N = q^k - 1$, such that for at least $(1-\varepsilon)N$ values of i, C_{in}^i has relative distance $\ge H_q^{-1}\left(\frac{1}{2} - \varepsilon\right)$.*

In fact, this ensemble is the following: for $\alpha \in \mathbb{F}_{q^k}^*$, the inner code $C_{\text{in}}^{\alpha} : \mathbb{F}_q^k \to \mathbb{F}_q^{2k}$ is defined as $C_{\text{in}}^{\alpha}(x) = (x, \alpha x)$. This ensemble is called the *Wozencraft ensemble*. We claim that C_{in}^{α} for every $\alpha \in \mathbb{F}_{q^k}^*$ is linear and is strongly explicit. (The proof if left as an exercise.)

10.3.1 Justesen code

For the Justesen code, the outer code C_{out} is a Reed-Solomon code evaluated over $\mathbb{F}_{q^k}^*$ of rate R, $0 < R < 1$. The outer code C_{out} has relative distance $\delta_{\text{out}} = 1 - R$ and block length of $N = q^k - 1$. The set of inner codes is the Wozencraft ensemble $\{C_{\text{in}}^{\alpha}\}_{\alpha \in \mathbb{F}_{q^k}^*}$ from Theorem 10.3.1. So the Justesen code is the concatenated code $C^* \stackrel{\text{def}}{=} C_{\text{out}} \circ (C_{\text{in}}^1, C_{\text{in}}^2, \ldots, C_{\text{in}}^N)$ with the rate $\frac{R}{2}$. The following proposition estimates the distance of C^*.

Proposition 10.3.2. *Let $\varepsilon > 0$. C^* has relative distance at least $(1 - R - \varepsilon) \cdot H_q^{-1}\left(\frac{1}{2} - \varepsilon\right)$*

Proof. Consider $\mathbf{m}^1 \neq \mathbf{m}^2 \in (\mathbb{F}_{q^k})^K$. By the distance of the outer code $|S| \geq (1 - R)N$, where

$$S = \left\{ i | C_{\text{out}}(\mathbf{m}^1)_i \neq C_{\text{out}}(\mathbf{m}^2)_i \right\}.$$

Call the ith inner code *good* if C_{in}^i has distance at least $d \stackrel{\text{def}}{=} H_q^{-1}\left(\frac{1}{2} - \varepsilon\right) \cdot 2k$. Otherwise, the inner code is considered bad. Note that by Theorem 10.3.1, there are at most εN bad inner codes. Let S_g be the set of all good inner codes in S, while S_b is the set of all bad inner codes in S. Since $S_b \leq \varepsilon N$,

$$|S_g| = |S| - |S_b| \geq (1 - R - \varepsilon)N. \tag{10.3}$$

For each good $i \in S$, by definition we have

$$\Delta\left(C_{\text{in}}^i\left(C_{\text{out}}\left(\mathbf{m}^1\right)_i\right), C_{\text{in}}^i\left(C_{\text{out}}\left(\mathbf{m}^2\right)_i\right)\right) \geq d. \tag{10.4}$$

Finally, from (10.3) and (10.4), we obtain that the distance of C^* is at least

$$(1 - R - \varepsilon) \cdot Nd = (1 - R - \varepsilon) H_q^{-1}\left(\frac{1}{2} - \varepsilon\right) N \cdot 2k,$$

as desired. □

Since the Reed-Solomon codes as well as the Wozencraft ensemble are strongly explicit, the above result implies the following:

Corollary 10.3.3. *The concatenated code C^* from Proposition 10.3.2 is an asymptotically good code and is strongly explicit.*

Thus, we have now satisfactorily answered Question 10.2.2 modulo Theorem 10.3.1, which we prove next.

Proof of Theorem 10.3.1. Fix $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{F}_q^{2k} \setminus \{\mathbf{0}\}$. Note that this implies that $\mathbf{y}_1 = \mathbf{0}$ and $\mathbf{y}_2 = \mathbf{0}$ are not possible. We claim that $\mathbf{y} \in C_{\text{in}}^{\alpha}$ for at most one $\alpha \in \mathbf{F}_{2^k}^*$. The proof is by a simple case analysis. First, note that if $\mathbf{y} \in C_{\text{in}}^{\alpha}$, then it has to be the case that $\mathbf{y}_2 = \alpha \cdot \mathbf{y}_1$.

- Case 1: $\mathbf{y}_1 \neq \mathbf{0}$ and $\mathbf{y}_2 \neq \mathbf{0}$, then $\mathbf{y} \in C_{\text{in}}^{\alpha}$, where $\alpha = \frac{\mathbf{y}_2}{\mathbf{y}_1}$.
- Case 2: $\mathbf{y}_1 \neq \mathbf{0}$ and $\mathbf{y}_2 = \mathbf{0}$, then $\mathbf{y} \notin C_{\text{in}}^{\alpha}$ for every $\alpha \in \mathbb{F}_{2^k}^*$ (as $\alpha \mathbf{y}_1 \neq \mathbf{0}$ since product of two elements in $\mathbb{F}_{2^k}^*$ also belongs to $\mathbb{F}_{2^k}^*$).
- Case 3: $\mathbf{y}_1 = \mathbf{0}$ and $\mathbf{y}_2 \neq \mathbf{0}$, then $\mathbf{y} \notin C_{\text{in}}^{\alpha}$ for every $\alpha \in \mathbb{F}_{2^k}^*$ (as $\alpha \mathbf{y}_1 = \mathbf{0}$).

Now assume that $wt(\mathbf{y}) < H_q^{-1}(1-\varepsilon)n$. Note that if $\mathbf{y} \in C_{\text{in}}^{\alpha}$, then C_{in}^{α} is "bad"(i.e. has relative distance $< H_q^{-1}\left(\frac{1}{2} - \varepsilon\right)$). Since $\mathbf{y} \in C_{\text{in}}^{\alpha}$ for at most one value of α, the total number of bad codes is at most

$$\begin{aligned}
\left|\left\{\mathbf{y} | wt(\mathbf{y}) < H_q^{-1}\left(\frac{1}{2} - \varepsilon\right) \cdot 2k\right\}\right| &\leq Vol_q\left(H_q^{-1}\left(\frac{1}{2} - \varepsilon\right) \cdot 2k, 2k\right) \\
&\leq q^{H_q(H_q^{-1}(\frac{1}{2}-\varepsilon)) \cdot 2k} && (10.5) \\
&= q^{(\frac{1}{2}-\varepsilon) \cdot 2k} \\
&= \frac{q^k}{q^{2\varepsilon k}} \\
&< \varepsilon(q^k - 1) && (10.6) \\
&= \varepsilon N. && (10.7)
\end{aligned}$$

In the above, (10.5) follows from our good old upper bound on the volume of a Hamming ball (Proposition 3.3.1) while (10.6) is true for large enough k. Thus for at least $(1-\varepsilon)N$ values of α, C_{in}^{α} has relative distance at least $H_q^{-1}\left(\frac{1}{2} - \varepsilon\right)$, as desired. □

By concatenating an outer code of distance D and an inner code of distance d, we can obtain a code of distance at least $\geq Dd$ (Theorem 10.1.1). Dd is called the concatenated code's *design distance*. For asymptotically good codes, we have obtained polynomial time construction of such codes (Theorem 10.2.1, as well as strongly explicit construction of such codes (Corollary 10.3.3). Further, since these codes were linear, we also get polynomial time encoding. However, the following natural question about decoding still remains unanswered.

Question 10.3.1. *Can we decode concatenated codes up to half their design distance in polynomial time?*

10.4 Exercises

Exercise 10.1. In Section 4.2.1, we saw that the Gilbert construction can compute an $(n,k)_q$ code in time $q^{O(n)}$. Now the Varshamov construction (Section 4.2.2) is a randomized construction and it is natural to ask how quickly can we compute an $[n,k]_q$ code that meets the GV bound. In this exercise, we will see that this can also be done in $q^{O(n)}$ deterministic time, though the deterministic algorithm is not that straight-forward anymore.

1. (*A warmup*) Argue that Varshamov's proof gives a $q^{O(kn)}$ time algorithm that constructs an $[n,k]_q$ code on the GV bound. (Thus, the goal of this exercise is to "shave" off a factor of k from the exponent.)

2. A $k \times n$ *Toeplitz Matrix* $A = \{A_{i,j}\}_{i=1,\ j=1}^{k\ ,\ n}$ satisfies the property that $A_{i,j} = A_{i-1,j-1}$. In other words, any diagonal has the same value. For example, the following is a 4×6 Toeplitz matrix:
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 1 & 2 & 3 & 4 & 5 \\ 8 & 7 & 1 & 2 & 3 & 4 \\ 9 & 8 & 7 & 1 & 2 & 3 \end{pmatrix}$$
A *random* $k \times n$ Toeplitz matrix $T \in \mathbb{F}_q^{k\times n}$ is chosen by picking the entries in the first row and column uniformly (and independently) at random.

   Prove the following claim: For any non-zero $\mathbf{m} \in \mathbb{F}_q^k$, the vector $\mathbf{m} \cdot T$ is uniformly distributed over $\mathbb{F}_q^n$, that is for every $\mathbf{y} \in \mathbb{F}_q^n$, $\Pr\left[\mathbf{m} \cdot T = \mathbf{y}\right] = q^{-n}$.

   (*Hint:* Write down the expression for the value at each of the n positions in the vector $\mathbf{m} \cdot T$ in terms of the values in the first row and column of T. Think of the values in the first row and column as variables. Then divide these variables into two sets (this "division" will depend on $\mathbf{m}$) say S and $\overline{S}$. Then argue the following: for every fixed $\mathbf{y} \in \mathbb{F}_q^n$ and for every fixed assignment to variables in S, there is a unique assignment to variables in $\overline{S}$ such that $\mathbf{m}T = \mathbf{y}$.)

3. Briefly argue why the claim in part (b) implies that a random code defined by picking its generator matrix as a random Toeplitz matrix with high probability lies on the GV bound.

4. Conclude that an $[n,k]_q$ code on the GV bound can be constructed in time $2^{O(k+n)}$.

10.5 Bibliographic Notes

Code concatenation was first proposed by Forney[30].

Justesen codes were constructed by Justesen [58]. In his paper, Justesen attributes the Wozencraft ensemble to Wozencraft.

Chapter 11

Information Theory Strikes Back: Polar Codes

We begin by recalling Question 13.4.1, which we re-produce for the sake of completeness:

Question 11.0.1. *Can we get to within ε of capacity for* BSC_p *(i.e. rate* $1 - H(p) - \varepsilon$*) via codes with block length and decoding times that are* $\mathrm{poly}(1/\varepsilon)$*?*

In this chapter we introduce *Polar codes*, a class of codes developed from purely information-theoretic insights. We then show how these codes lead to a resolution of Question 11.0.1, namely how to get arbitrarily close to capacity on the binary symmetric channel with block length, and decoding complexity growing polynomially in the inverse of the gap to capacity. This answers in the affirmative one of the most crucial questions in the Shannon setting.

This chapter is organized as follows. We define the precise question, after some simplification, in Section 11.1. In the same section, we discuss why the simplified question solves a much more general problem. We then switch the problem from an error-correction question to a *linear-compression* question in Section 11.2. This switch is very straightforward but very useful in providing insight into the working of Polar codes. In Section 11.3 we introduce the idea of polarization, which provides the essential insight to Polar codes. (We remark that this section onwards is based on notions from Information Theory. The reader unfamiliar with the theory should first consult Appendix E to get familiar with the basic concepts.) In Section 11.4 we then give a complete description of the Polar codes, and describe the encoding and decoding algorithms. In Section 11.5 we then describe the analysis of these codes. We remark that the only complex part of this chapter is this analysis and the construction and algorithms themselves are quite simple (and extremely elegant) modulo this analysis.

11.1 Achieving Gap to Capacity

The goal of this section is to present a simple question that formalizes what it means to achieve capacity with polynomial convergence, and to explain why this is the right question (and how answering this positively leads to much more powerful results by standard methods).

Recall that for $p \in [0,1/2)$ the BSC_p is the channel that takes as input a sequence of bits $\mathbf{X} = (X_1, \ldots, X_n)$ and outputs the sequence $\mathbf{Y} = (Y_1, \ldots, Y_n)$ where for each i, $X_i = Y_i$ with probability $1-p$ and $X_i \neq Y_i$ with probability p; and this happens independently for each $i \in \{1, \ldots, n\}$. We use the notation $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathrm{Bern}(p)^n$ to denote the error pattern, so that $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$.

Our target for this chapter is to prove the following theorem:

Theorem 11.1.1. *For every $p \in [0,1/2)$ there is a polynomial $n_0(\cdot)$ such that for every $\varepsilon > 0$ there exist k, n and an encoder $E : \mathbb{F}_2^k \to \mathbb{F}_2^n$ and decoder $D : \mathbb{F}_2^n \to \mathbb{F}_2^k$ satisfying the following conditions:*

Length and Rate *The codes are short, and the rate is close to capacity, specifically, $1/\varepsilon \leq n \leq n_0(1/\varepsilon)$ and $k \geq (1 - H(p) - \varepsilon) \cdot n$.*

Running times *The encoder and decoder run in time $O(n \log n)$ (where the $O(\cdot)$ notation hides a universal constant independent of p and ε).*

Failure Probability *The probability of incorrect decoding is at most ε. Specifically, for every $\mathbf{m} \in \mathbb{F}_2^k$,*

$$\Pr_{\mathbf{Z} \in \mathrm{Bern}(p)^n}[\mathbf{m} \neq D(E(\mathbf{m}) + \mathbf{Z})] \leq \varepsilon.$$

Theorem 11.1.1 is not the ultimate theorem we may want for dealing with the binary symmetric channel. For starters, it does not guarantee codes of all lengths, but rather only of one fixed length n for any fixed choice of ε. Next, Theorem 11.1.1 only guarantees a small probability of decoding failure, but not one that say goes to zero exponentially fast in the length of the code. The strength of the theorem is (1) its simplicity - it only takes one parameter ε and delivers a good code with rate ε close to capacity and (2) Algorithmic efficiency: the running time of the encoder and decoder is a polynomial in $1/\varepsilon$. It turns out both the weaknesses can be addressed by applying the idea of concatenation of codes (Chapter 10) while preserving the strength. We present the resulting theorem, leaving the proof of this theorem from Theorem 11.1.1 as an exercise. (See Exercise 11.1.)

Theorem 11.1.2. *There exists polynomially growing functions $n_0 : [0,1] \to \mathbb{Z}^+$ and $T : \mathbb{Z}^+ \to \mathbb{Z}^+$ such that for all $p \in [0,1]$, $\varepsilon > 0$ there exists $\delta > 0$, a function $k : \mathbb{Z}^+ \to \mathbb{Z}^+$ and an ensemble of function $E = \{E_n\}_n$ and $D = \{D_n\}_n$ such that for all $n \in \mathbb{Z}^+$ with $n \geq n_0(1/\varepsilon)$ the following hold:*

1. *The codes are ε-close to capacity: Specifically, $k = k(n) \geq (1 - H(p) - \varepsilon)n$, $E_n : \mathbb{F}_2^k \to \mathbb{F}_2^n$ and $D_n : \mathbb{F}_2^n \to \mathbb{F}_2^k$.*

2. *The codes correct p-fraction of errors with all but exponentially small probability: Specifically*

$$\Pr_{\mathbf{Z} \sim \mathrm{Bern}(p)^n, \mathbf{X} \sim U(\mathbb{F}_2^k)}[D_n(E_n(\mathbf{X}) + \mathbf{Z}) \neq \mathbf{X}] \leq \exp(-\delta n).$$

3. *Encoding and Decoding are efficient: Specifically E_n and D_n run in time at most $T(n/\varepsilon)$.*

11.2 Reduction to Linear Compression

In this section we change our problem from that of coding for error-correction to compressing a vector of independent Bernoulli random variables i.e., the error-pattern. (Recall that we encountered compression in Exercise 6.10). We show that if the compression is linear and the decompression algorithm is efficient, then this turns into a linear code with efficient decoding (this is the converse of what we saw in Exercise 6.11). By virtue of being linear the code is also polynomial time encodable, given the generator matrix. We explain this simple connection below.

For $n \geq m$, we say that a pair $(\mathbf{H}, D)$ where $\mathbf{H} \in \mathbb{F}_2^{n\times m}$, and $D : \mathbb{F}_2^m \to \mathbb{F}_2^n$ forms an τ-error *linear compression* scheme for $\mathrm{Bern}(p)^n$ if $\mathbf{H}$ has rank m and

$$\Pr_{\mathbf{Z}\sim \mathrm{Bern}(p)^n}[D(\mathbf{Z}\cdot\mathbf{H}) \neq \mathbf{Z}] \leq \tau.$$

We refer to the ratio $\frac{m}{n}$ as the *(compression) rate* of the scheme (recall Exercise 6.10).

Proposition 11.2.1. *Let $(\mathbf{H}, D)$ be a τ-error linear compression scheme for $\mathrm{Bern}(p)^n$ with $\mathbf{H} \in \mathbb{F}_2^{n\times m}$. Let $k = n - m$ and let $\mathbf{G} \in \mathbb{F}_2^{k\times n}$ and $\mathbf{G}^* : \mathbb{F}_2^{n\times k}$ be full-rank matrices such that $\mathbf{G}\cdot\mathbf{H} = \mathbf{0}$ and $\mathbf{G}\cdot\mathbf{G}^* = \mathbf{I}_k$ (the $k \times k$ identity matrix). Then the encoder $E : \mathbb{F}_2^k \to \mathbb{F}_2^n$ given by*

$$E(X) = \mathbf{X}\cdot\mathbf{G}$$

and the decoder $D' : \mathbb{F}_2^n \to \mathbb{F}_2^k$ given by

$$D'(\mathbf{Y}) = (\mathbf{Y} - D(\mathbf{Y}\cdot\mathbf{H}))\cdot\mathbf{G}^*$$

satisfy for every $\mathbf{m} \in \mathbb{F}_2^k$:

$$\Pr_{\mathbf{Z}\in \mathrm{Bern}(p)^n}[\mathbf{m} \neq D'(E(\mathbf{m}) + \mathbf{Z})] \leq \tau.$$

Remark 11.2.1.

1. Recall that straightforward linear algebra implies the existence of matrices $\mathbf{G}$ and $\mathbf{G}^*$ above (see Exercise 11.2).
2. Note that the complexity of encoding is simply the complexity of multiplying a vector by $\mathbf{G}$. The complexity of decoding is bounded by the complexity of decompression plus the complexity of multiplying by $\mathbf{G}^*$. In particular if all these operations can be carried out in $O(n\log n)$ time then computing E and D' takes $O(n\log n)$ time as well.
3. Note that the above proves the converse of Exercise 6.11. These two results show that (at least for linear codes), channel and source coding are equivalent.

Proof. Suppose $D(\mathbf{Z}\cdot\mathbf{H}) = \mathbf{Z}$. Then we claim that if $\mathbf{Z}$ is the error pattern, then decoding is successful. To see this, note that

$$\begin{aligned}
D'(E(\mathbf{m}) + \mathbf{Z}) &= (E(\mathbf{m}) + \mathbf{Z} - D((E(\mathbf{m}) + \mathbf{Z})\cdot\mathbf{H}))\cdot\mathbf{G}^* && (11.1)\\
&= (E(\mathbf{m}) + \mathbf{Z} - D(\mathbf{Z}\cdot\mathbf{H}))\cdot\mathbf{G}^* && (11.2)\\
&= (E(\mathbf{m}) + \mathbf{Z} - \mathbf{Z}))\cdot\mathbf{G}^* && (11.3)\\
&= E(\mathbf{m})\cdot\mathbf{G}^*\\
&= \mathbf{m}. && (11.4)
\end{aligned}$$

In the above (11.1) follows by definition of D', (11.2) follows from the fact that $E(\mathbf{m}) \cdot \mathbf{H} = \mathbf{m} \cdot \mathbf{G} \cdot \mathbf{H} = \mathbf{0}$, (11.3) follows by assumption that $D(\mathbf{Z} \cdot \mathbf{H}) = \mathbf{Z}$ and (11.4) follows since $E(\mathbf{m}) = \mathbf{m} \cdot \mathbf{G}$ and $\mathbf{G} \cdot \mathbf{G}^* = \mathbf{I}$. Thus the probability of a decoding failure is at most the probability of decompression failure, which by definition is at most τ, as desired. □

Thus our updated quest from now on will be to

Question 11.2.1. *Design a linear compression scheme for* $\mathrm{Bern}(p)^n$ *of rate at most* $H(p) + \varepsilon$.

See Exercise 11.3 on how one can answer Question 11.2.1 with a non-linear compression scheme.

In what follows we will introduce the polarization phenomenon that will lead us to such a compression scheme.

11.3 The Polarization Phenomenon

11.3.1 Information Theory Review

The only information theoretic notions that we need in this chapter are that of Entropy (see Definition E.1.1) and Conditional Entropy (Definition E.2.2). We use the notations $H(X)$ to denote the entropy of a variable X and $H(X|Y)$ to be the conditional entropy, conditioned on Y. The main properties of these notions we will use are the chain rule (see Theorem E.2.2):

$$H(X, Y) = H(Y) + H(X|Y),$$

the fact that conditioning does not increase entropy (see Lemma E.2.4):

$$H(X|Y) \leq H(X).$$

We also use the basic fact that the uniform distribution maximizes entropy (see Lemma E.1.2) and so

$$H(X) \leq \log|\Omega|$$

if Ω denotes the support of X. A final fact that will be useful to keep in mind as we develop the polar codes is that variables with low entropy are essentially determined, and variables with low conditional entropy are predictable. We formalize this (with very loose bounds) below.

Proposition 11.3.1. *Let* $\alpha \geq 0$.

1. *Let* X *be a random variable with* $H(X) \leq \alpha$. *Then there exists* x *such that* $\Pr_X[X \neq x] \leq \alpha$.

2. *Let (X, Y) be jointly distributed variables with $H(X|Y) \le \alpha$. Then the function*

$$A(y) = \operatorname*{argmax}_x \{\Pr[X = x | Y = y]\}$$

satisfies

$$\Pr_{(X,Y)}[X \neq A(Y)] \le \alpha.$$

We defer the proof to Section 11.6.1.

11.3.2 Polarized matrices and decompression

We now return to the task of designing a matrix $\mathbf{H}$ (and corresponding matrices $\mathbf{G}$ and $\mathbf{G}^*$) such that the map $\mathbf{Z} \mapsto \mathbf{Z} \cdot \mathbf{H}$ is a good compression scheme. As such we are seeking rectangular matrices with some extra properties, but in this section we will convert this question into a question about square invertible matrices.

For an invertible matrix $\mathbf{P} \in \mathbb{F}_2^{n \times n}$ we consider its effect on $\mathbf{Z} = (Z_1, \ldots, Z_n) \sim \mathrm{Bern}(p)^n$. Let $\mathbf{W} = (W_1, \ldots, W_n)$ be given by $\mathbf{W} = \mathbf{Z} \cdot \mathbf{P}$. Now consider the challenge of predicting the W_i's as they arrive online. So when attempting to predict W_i we get to see

$$\mathbf{W}_{<i} \triangleq (W_1, \ldots, W_{i-1})$$

and can use this to predict W_i. For arbitrary matrices (in particular the identity matrix), seeing $W_{<i}$ gives us no advantage on predicting W_i. A matrix will be considered polarized if for an appropriately large fraction of i's, W_i is *highly predictable* from $\mathbf{W}_{<i}$.

To formalize the notion of highly predictable we turn to information theory and simply require $H(W_i|\mathbf{W}_{<i}) \le \tau$ for some very small parameter τ of our choice.[1] How many i's should be very predictable? Let

$$S = S_\tau = \{i \in [n] | H(W_i|\mathbf{W}_{<i}) \ge \tau\}$$

denote the set of unpredictable bits. An entropy calculation tells us how small we can hope S would be. Since $\mathbf{P}$ is invertible, we have (see Exercise 11.4):

$$H(\mathbf{W}) = H(\mathbf{Z}) = n \cdot H(p), \tag{11.5}$$

But by the chain rule we have

$$\begin{aligned} H(W) &= \sum_{i=1}^{n} H(W_i|\mathbf{W}_{<i}) \\ &\le \sum_{i \in S} H(W_i|\mathbf{W}_{<i}) + n\tau \qquad (11.6) \\ &\le \sum_{i \in S} H(W_i) + n\tau \qquad (11.7) \\ &\le |S| + n\tau. \qquad (11.8) \end{aligned}$$

[1] Eventually we will set $\tau = o(1/n)$.

In the above, (11.6) follows by using definition of S and that $|\{i \notin S\}| \le n$, (11.7) follows from the fact that conditioning does not increase entropy and (11.8) follows from the fact that uniformity maximizes entropy and so $H(W_i) \le 1$. We thus conclude that $|S| \ge H(p) \cdot n - n\tau \approx H(p) \cdot n$. So the smallest that the set S can be is $H(p) \cdot n$ and we will allow an εn additive slack to get the following definition.

Definition 11.3.1 (Polarizing matrix, unpredictable columns)**.** We say that an invertible matrix $\mathbf{P} \in \mathbb{F}_2^{n \times n}$ is (ε, τ)*-polarizing* for $\mathrm{Bern}(p)^n$ if for $\mathbf{W} = \mathbf{Z} \cdot \mathbf{P}$ and

$$S = S_\tau = \{i \in [n] | H(W_i | \mathbf{W}_{<i}) \ge \tau\}$$

we have $|S| \le (H(p) + \varepsilon)n$. We refer to the set S as the set of *unpredictable* columns of $\mathbf{P}$ (and $\{W_i\}_{1 \in S}$ as the *unpredictable bits* of $\mathbf{W}$).

We next show how to get a compression scheme from a polarizing matrix (without necessarily having an efficient decompressor). The idea is simple: the compressor simply outputs the "unpredictable" bits of $\mathbf{W}$. Let $\mathbf{W}_S = (W_i)_{i \in S}$, the compression of $\mathbf{Z}$ is simply $(\mathbf{Z} \cdot \mathbf{P})_S$. For the sake of completeness, we record this in Algorithm 11.

Algorithm 9 POLAR COMPRESSOR($\mathbf{Z}, S$)

INPUT: String $\mathbf{Z} \in \mathbb{F}_2^n$ and subset $S \subseteq [n]$

OUTPUT: Compressed string $\mathbf{W} \in \mathbb{F}_2^{|S|}$ ▷ Assumes a polarizing matrix $\mathbf{P} \in \mathbb{F}_2^{n \times n}$

1: RETURN $(\mathbf{Z} \cdot \mathbf{P})_S$

Equivalently if we let $\mathbf{P}_S$ denote the $n \times |S|$ matrix whose columns correspond to indices in S, then the compression of $\mathbf{Z}$ is $\mathbf{Z} \cdot \mathbf{P}_S$. So $\mathbf{P}_S$ will be the matrix $\mathbf{H}$ we are seeking. Before turning to the decompression, let us also specify $\mathbf{G}$ and $\mathbf{G}^*$ for this construction above. Indeed Exercise **??** shows that $\mathbf{G} = \left(P^{-1}\right)_{\bar{S}}$ and $\mathbf{G}^* = \mathbf{P}_{\bar{S}}$. In particular, the complexity of multiplying by $\mathbf{P}$ and $\mathbf{P}^{-1}$ dominate the cost of the matrix multiplications needed in the encoding and decoding.

We finally turn to the task of describing the decompressor corresponding to compression with a polarizing matrix $\mathbf{P}$ with unpredictable columns S. The method is a simple iterative one, based on the predictor from Proposition 11.3.1 and is presented in Algorithm 10.

Next we argue that Algorithm 10 has low failure probability:

Lemma 11.3.2. *If P is (ε, τ)-polarizing for $\mathrm{Bern}(p)^n$ with unpredictable columns S, then the successive cancellation decoder has failure probability at most $\tau \cdot n$, i.e.,*

$$\Pr_{\mathbf{Z} \in \mathrm{Bern}(p)^n}[\mathbf{Z} \neq \mathrm{SCD}((\mathbf{Z} \cdot \mathbf{P})_S, \mathbf{P}, S)] \le \tau \cdot n.$$

Thus if $\mathbf{P}$ is $(\varepsilon, \varepsilon/n)$-polarizing for $\mathrm{Bern}(p)^n$ then the failure probability of the successive cancellation decoder is at most ε.

[2]Note that the Successive Cancellation Decompressor, and Decompressors in general are not expected to work correctly on every input. So the INPUT/OUTPUT relations don't fully capture the goals of the algorithm. In such cases, in the rest of the chapter we will include a PERFORMANCE PARAMETER which we wish to minimize that attempts to capture the real goal of the algorithm.

Algorithm 10 Successive Cancellation Decompressor SCD($\mathbf{W}, \mathbf{P}, S$)

INPUT: $S \subseteq [n]$, $\mathbf{W} \in \mathbb{F}_2^S$ and $\mathbf{P} \in \mathbb{F}_2^{n \times n}$
OUTPUT: $\tilde{\mathbf{Z}} \in \mathbb{F}_2^n$ such that $(\tilde{\mathbf{Z}} \cdot \mathbf{P})_S = \mathbf{W}$
PERFORMANCE PARAMETER: [2] $\Pr_{\mathbf{Z} \sim \mathrm{Bern}(p)^n}[\mathrm{SCD}((\mathbf{Z} \cdot \mathbf{P})_S, \mathbf{P}, S) \neq \mathbf{Z}]$ (smaller is better)

1: FOR $i = 1$ to n DO
2: IF $i \in S$ THEN
3: $\tilde{W}_i \leftarrow W_i$
4: ELSE
5: $\tilde{W}_i \leftarrow \mathrm{argmax}_{b \in \mathbb{F}_2} \{\Pr[W_i = b | \mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}]\}$
6: RETURN $\tilde{\mathbf{Z}} \leftarrow \mathbf{P}^{-1} \cdot \tilde{\mathbf{W}}$

Proof. Let $\mathbf{W} = \mathbf{Z} \cdot \mathbf{P}$. Note that by Step 3 in Algorithm 10, we have for every $i \in S$, $\tilde{W}_i = W_i$. For any $i \notin S$, by Proposition 11.3.1 applied with $X = W_i$, $Y = \mathbf{W}_{<i}$, $\alpha = \tau$ we get that

$$\Pr_{\mathbf{Z}}[W_i \neq A^i(\tilde{\mathbf{W}}_{<i})] \leq \tau,$$

where $A^i(\cdot)$ is the corresponding function defined for $i \notin S$. By a union bound, we get that

$$\Pr_{\mathbf{Z}}[\exists i \text{ s.t. } W_i \neq A^i(\tilde{\mathbf{W}}_{<i})] \leq \tau \cdot n.$$

Note that by step 5 in Algorithm 10 and the definition of $A^i(\cdot)$, we have $\tilde{W}_i = A^i(\tilde{\mathbf{W}}_{<i})$. But if $\mathbf{W} \neq \tilde{\mathbf{W}}$ there must exists a least i such that $\tilde{W}_i \neq W_i$. Thus we get $\Pr[\mathbf{W} \neq \tilde{\mathbf{W}}] \leq \tau n$ and so probability that $\mathrm{SCD}(\mathbf{W}_S, \mathbf{P}, S) \neq \mathbf{Z}$ is at most τn. □

To summarize this section we have learned that to prove Theorem 11.1.1 it suffices to answer the following question.

Question 11.3.1. *Find an $(\varepsilon, \varepsilon/n)$-polarizing matrix $\mathbf{P} \in \mathbb{F}_2^{n \times n}$ with n is bounded by a polynomial in $1/\varepsilon$; where* SCD *(potentially a re-implementation of Algorithm 10) as well as multiplication and inversion by $\mathbf{P}$ takes $O(n \log n)$ time.*

Next we describe the idea behind the construction, which will answer the above question.

11.3.3 A polarizing primitive

Thus far in the chapter we have essentially only been looking at the problem from different perspectives, but not yet suggested an idea on how to get the codes (or compression schemes) that we desire (i.e. answer Question 11.3.1. It is this section that will provide the essence of the idea, which is to start with a simple and basic *polarization* step, and then iterate it appropriately

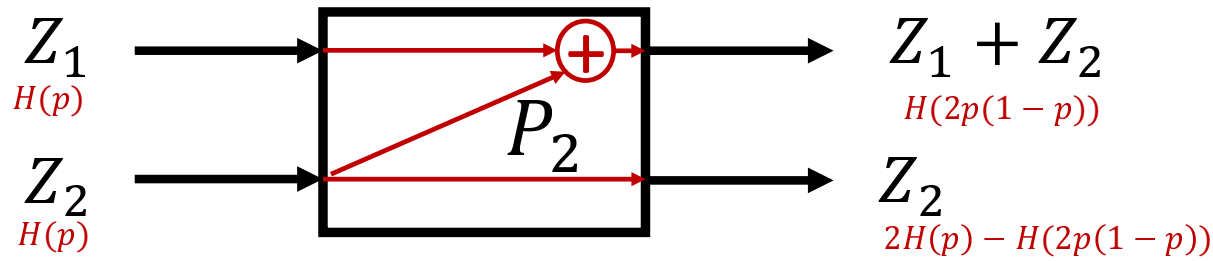


Figure 11.1: The 2×2 Basic Polarizing Transform. Included in brown are the conditional entropies of the variables, conditioned on variables directly above them.

many times to get a highly polarized matrix. Indeed it is this section that will explain the term polarization (and why we use this term to describe the matrices we seek).

Recall that the ultimate goal of polarization is to start with many independent bits $Z_1, \ldots, Z_n$ which are independent and slightly unpredictable (if p is small), and to produce some linear transform that concentrates all the unpredictability into a fewer set of bits. In our basic primitive we will try to do this with two bits. So we have Z_1, Z_2 such that $H(Z_1) = H(Z_2) = p$ and we wish to produce two bits W_1, W_2 such that at least one of these is less predictable than either of Z_i's. Since there are only 4 possible linear combinations of two bits (over $\mathbb{F}_2$) and three are trivial (0, Z_1, and Z_2) we are left with only one candidate function namely $Z_1 + Z_2$. So we will set $W_1 = Z_1 + Z_2$. For W_2 we are left with the trivial functions: 0 carries no information and so is ruled out. Without loss of generality, the only remaining choice is $W_2 = Z_2$. So we look at this transformation: $P_2 : (Z_1, Z_2) \mapsto (W_1, W_2) = (Z_1 + Z_2, Z_2)$. (See Figure 11.1.)

This is an invertible linear transformation given by the matrix

$$\mathbf{P}_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

So (see Exercise 11.5):

$$H(W_1, W_2) = 2H(p). \tag{11.9}$$

But some examination shows that $H(W_1) > H(Z_1), H(Z_2)$. In particular the probability that W_1 is 1 is $2p(1-p) \in (p, 1/2)$ and since $H(\cdot)$ is monotone increasing in this interval (see Exercise 11.6) it follows that $H(W_1) > H(p) = H(Z_1) = H(Z_2)$. Thus W_1 is less predictable than either of the input bits; and now, thanks to the chain rule

$$H(W_2|W_1) = H(W_1, W_2) - H(W_1) = H(Z_1, Z_2) - H(W_1) = H(Z_1) + H(Z_2) - H(W_1) < H(Z_1), H(Z_2).$$

In other words multiplying by $\mathbf{P}_2$ has separated the entropy of two equally entropic bits, into a more entropic bit and a (conditionally) less entropic one. But of course this may be only slight polarization and what we are hoping for is many bits that are almost completely determined by preceding ones.

To get more polarization we apply this 2×2 operation repeatedly. Specifically, let $\mathbf{P}_2(Z_1, Z_2) = (Z_1 + Z_2, Z_2)$. Then we let $\mathbf{P}_4(Z_1, Z_2, Z_3, Z_4) = (W_1, W_2, W_3, W_4) = (\mathbf{P}_2(U_1, U_2), \mathbf{P}_2(U_3, U_4))$ where

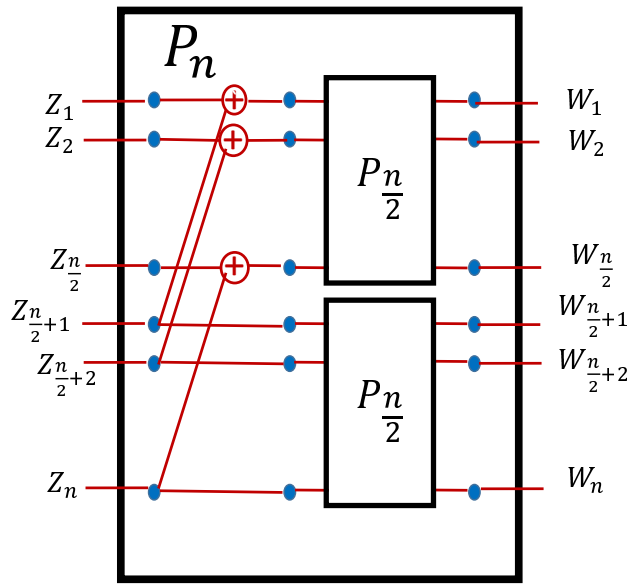


Figure 11.2: The $n \times n$ Basic Polarizing Transform.

$(U_1, U_2) = \mathbf{P}_2(Z_1, Z_3)$ and $(U_3, U_4) = \mathbf{P}_2(Z_2, Z_4)$. Thus the bit $W_1 = Z_1 + Z_2 + Z_3 + Z_4$ has higher entropy than say Z_1 or even $U_1 = Z_1 + Z_3$, whereas $W_4 = Z_4$ conditioned on (W_1, W_2, W_3) can be shown to have much lower entropy than Z_4 (unconditionally) or even $U_4 = Z_4$ conditioned on U_2.

The composition above can be extended easily to n bit inputs, when n is a power of two, to get a linear transform $\mathbf{P}_n$ (See Figure 11.2. We will also give this transformation explicitly in the next section).

It is also clear that some bits will get highly entropic due to these repeated applications of P_2. What is remarkable is that the polarization is nearly "perfect" - most bits W_i have conditional entropy (conditioned on $(W_j)_{j<i}$) close to 1, or close to 0. This leads to the simple construction of the polarizing matrix we will describe in the next section. A striking benefit of this simple construction is multiplying a vector by $\mathbf{P}$ only takes $O(n \log n)$ time, and so does multiplying by the inverse of $\mathbf{P}$. And most importantly a version of SCD (Algorithm 10) is also computable in time $O(n \log n)$ and this leads to an overall compression and decompression algorithms running in time $O(n \log n)$, which we describe in the next section.

We note that one missing element in our description is the challenge of determining the exact set of indices S that include all the high-conditional-entropy bits. We will simply skirt the issue and assume that this set is known for a given matrix $\mathbf{P}$ and given to the compression/decompression algorithms. So this leads to a non-uniform solution to compression and decompression problem, as well as the task of achieving Shannon capacity on the binary symmetric channel. We stress that this is not an inherent problem with the polarization approach. An explicit algorithm to compute S (or a small superset of S) is actually known and we discuss this in Section 11.7.

11.4 Polar codes, Encoder and Decoder

We begin with the description of polar codes along with the statement of the main results in Section 11.4.1. We explicitly state the encoding algorithm and analyze its runtime in Section 11.4.2. We present the decoder as well as its proof of correctness in Section 11.4.3.

11.4.1 The Code and Polarization Claims

We are now ready to describe the code.

Definition 11.4.1 (Basic polarizing matrix)**.** We define the $n \times n$ polarization matrix $\mathbf{P}_n$ recursively for $n = 2, 4, 8, \ldots$, by describing the linear map $P_n : \mathbf{Z} \mapsto \mathbf{Z} \cdot \mathbf{P}_n$. For $n = 2$ and $\mathbf{Z} \in \mathbb{F}_2^2$ we define $P_2(\mathbf{Z}) = (Z_1 + Z_2, Z_2)$.[3] For $n = 2^t$ and $\mathbf{Z} = (\mathbf{U}, \mathbf{V})$ for $\mathbf{U}, \mathbf{V} \in \mathbb{F}_2^{n/2}$ we define

$$P_n(\mathbf{Z}) = (P_{n/2}(\mathbf{U} + \mathbf{V}), P_{n/2}(\mathbf{V})).$$

Exercise 11.7 talks about explicit description of $\mathbf{P}_n$ as well as some extra properties.

In Section 11.5 we show that this matrix polarizes fast as $n \to \infty$. The main theorem we will prove is the following:

Theorem 11.4.1 ((Polynomially) Strong Polarization)**.** *Fix $p \in (0, 1/2)$ and constant c. There exists a polynomial function n_0 such that for every $\varepsilon > 0$ there exists $n = 2^t$ with $1/\varepsilon \le n \le n_0(1/\varepsilon)$ and a set $E \subseteq [n]$ with $|E| \le (\varepsilon/2) \cdot n$ such that for every $i \notin E$, the conditional entropy $H(W_i | \mathbf{W}_{<i}) \notin (n^{-c}, 1 - n^{-c}))$. Furthermore, if we let $S = \{i \in [n] | H(W_i | \mathbf{W}_{<i}) \ge n^{-c}\}$ then $|S| \le (H(p) + \varepsilon) \cdot n$ and the matrix $\mathbf{P}_n$ is $(\varepsilon, 1/n^c)$-polarizing for* $\mathrm{Bern}(p)^n$ *with unpredictable columns S.*

This theorem allows us to choose the threshold on how close to zero the conditional entropy of the polarized bits can be, and this threshold can be the inverse of an arbitrarily high degree polynomial in n. We will prove Theorem 11.4.1 in Section 11.5, which will be quite technical. But with the theorem in hand, it is quite simple to complete the description of the Basic Polar Code along with the associated encoding and decoding algorithms, and to analyze their performance.

Definition 11.4.2 (Basic Polar (Compressing) Code)**.** Given p and $\varepsilon \le 1/4$, let n and $S \subseteq [n]$ be as given by Theorem 11.4.1 for $c = 2$. Then the Basic Polar Code with parameters p and ε maps $\mathbf{Z} \in \mathbb{F}_2^n$ to $P_n(\mathbf{Z})_S$.

Proposition 11.4.2 (Rate of the Basic Polar Code)**.** *For every $p \in (0, 1/2)$ and $\varepsilon > 0$, the rate of the Basic Polar Code with parameters p and ε is at most $H(p) + \varepsilon$.*[4]

11.4.2 Encoding

The description of the map $P_n(\mathbf{Z})$ is already explicitly algorithmic, modulo the computation of the set S. For the sake of concreteness we write the algorithm below in Algorithm 11, assuming S is given as input, and argue its running time.

[3] We will be using $\mathbf{P}_n$ to denote the $n \times n$ matrix and P_n to denote the corresponding linear map that acts on $\mathbf{Z}$.

[4] Recall that we are trying to solve the compression problem now.

Algorithm 11 BASIC POLAR ENCODER($\mathbf{Z}; n, S$)

INPUT: n power of 2, $\mathbf{Z} \in \mathbb{F}_2^n$ and $S \subseteq [n]$
OUTPUT: Compression $\mathbf{W} \in \mathbb{F}_2^S$ of $\mathbf{Z}$ given by $\mathbf{W} = (P_n(\mathbf{Z}))_S$

```
RETURN W = (P(n, Z))_S

function P(n, Z)
    IF n = 1 THEN
        RETURN Z
    ELSE
        Let U = (Z_1, ..., Z_{n/2}) and V = (Z_{n/2+1}, ..., Z_n)
        RETURN (P(n/2, U + V), P(n/2, V))
```

Proposition 11.4.3. *The running time of the* BASIC POLAR ENCODER *algorithm is* $O(n \log n)$.

Proof. If $T(n)$ denotes the running time of the algorithm on inputs of length n, then $T(\cdot)$ satisfies the recurrence $T(n) = 2T(n/2) + O(n)$, which implies the claimed runtime. □

11.4.3 Decoding

Note that a polynomial time algorithm to compute $\Pr\left[W_i = b | \mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}\right]$ given $b \in \mathbb{F}_2$ and $\tilde{\mathbf{W}}_{<i} \in \mathbb{F}_2^{i-1}$ would lead to a polynomial time implementation of the SUCCESSIVE CANCELLATION DECOMPRESSOR (Algorithm 10). However, we will go a step better and exploit the nice structure of the Basic Polarizing Matrix to get an $O(n \log n)$ algorithm without much extra effort. The key insight to this faster decoder is that the decoder works even if $\mathbf{Z} \sim \mathrm{Bern}(p_1) \times \cdots \times \mathrm{Bern}(p_n)$, i.e., even when the bits of $\mathbf{Z}$ are not identically distributed, as long as they are independent. This stronger feature allows for a simple recursive algorithm. Specifically we use the facts that

1. If $Z_1 \sim \mathrm{Bern}(p_1)$ and $Z_2 \sim \mathrm{Bern}(p_2)$ are independent then $Z_1 + Z_2$ is a Bernoulli random variable. Let $b^+(p_1, p_2)$ denote the bias (i.e., probability of being 1) of $Z_1 + Z_2$ (and so $Z_1 + Z_2 \sim \mathrm{Bern}(p^+(p_1, p_2))$ (see Exercise 11.8).

2. If $Z_1 \sim \mathrm{Bern}(p_1)$ and $Z_2 \sim \mathrm{Bern}(p_2)$ are sampled conditioned on $Z_1 + Z_2 = a$ (for some $a \in \mathbb{F}_2$) then Z_2 is still a Bernoulli random variable. Let $b^|(p_1, p_2, a)$ denote the bias of Z_2 conditioned on $Z_1 + Z_2 = a$. (Note that $b^|(p_1, p_2, 0)$ is not necessarily equal to $b^|(p_1, p_2, 1)$.) See Exercise 11.9 for more.

We now use the functions b^+ and $b^|$ defined above to describe our decoding algorithm. We switch our notation slightly to make for a cleaner description. Rather than being given the vector $\mathbf{W}_S \in \mathbb{F}_2^{|S|}$, we assume that our decoder is given as input a vector $\mathbf{W} \in (\mathbb{F}_2 \cup \{?\})^n$ where $W_i =?$ if and only if $i \notin S$.

The main idea behind the decoding algorithm is that to complete $\mathbf{W}$ to a vector in $\mathbb{F}_2^n$, we can first focus on computing

$$\mathbf{W}[1, \ldots, n/2] \stackrel{\text{def}}{=} (W_1, \ldots, W_{n/2}).$$

It will turn out that we can use the fact that $\mathbf{W}[1,\dots,n/2] = P_{n/2}(Z'_1,\dots,Z'_{n/2})$ where[5] $Z'_i \in \mathrm{Bern}(b^+(p,p))$ are independent. This problem can be solved by recursion, and if we compute also the Z'_i's along the way, then we can turn to computing $W[n/2+1,\dots,n] = P_{n/2}(Z_{n/2+1},\dots,Z_n)$ but now Z_i is no longer drawn from $\mathrm{Bern}(p)$. Instead we have $Z_i \sim \mathrm{Bern}(b^{|}(p,p,Z'_{i-n/2}))$ but the Z_i's are still independent. The stronger recursive condition allows us to also solve this problem recursively. Details are given in Algorithm 12.

Algorithm 12 BASIC POLAR DECODER: $\mathrm{BPD}(\mathbf{W};n,p)$

INPUT: n power of 2, $\mathbf{W} \in (\mathbb{F}_2 \cup \{?\})^n$ and $0 \le p < 1/2$
OUTPUT: $\tilde{\mathbf{Z}} \in \mathbb{F}_2^n$ such that for every i either $\mathbf{W}_i = ?$ or $(\tilde{\mathbf{Z}} \cdot \mathbf{P})_i = \mathbf{W}_i$
PERFORMANCE PARAMETER: $\Pr_{\mathbf{Z}\sim\mathrm{Bern}(p)^n}[\mathbf{Z} \neq \mathrm{BPD}((\mathbf{Z}\cdot\mathbf{P}_n)_S;n,p)]$

$(\tilde{\mathbf{Z}},\tilde{\mathbf{W}},\boldsymbol{\rho}) = \mathrm{RPD}(\mathbf{W};n,(p,\cdots,p))$
RETURN $\tilde{\mathbf{Z}}$

function RECURSIVE POLAR DECODER: $\mathrm{RPD}((\mathbf{W};n,(p_1,\dots,p_n)))$
INPUT: $\mathbf{W} \in (\mathbb{F}_2 \cup \{?\})^n$ and $p_1,\dots,p_n \in [0,1]$.
OUTPUT: $\tilde{\mathbf{Z}},\tilde{\mathbf{W}} \in \mathbb{F}_2^n$ with $\tilde{\mathbf{W}} = P_n(\tilde{\mathbf{Z}})$ and $\left(\tilde{\mathbf{W}}\right)_S = \mathbf{W}_S$. $\boldsymbol{\rho} = (\rho_1,\dots,\rho_n) \in [0,1]^n$. ▷ $\tilde{\mathbf{Z}}$ is the main output while $\tilde{\mathbf{W}}$ and $\boldsymbol{\rho}$ will help us reason about correctness.
PERFORMANCE PARAMETER: $\Pr_{\mathbf{Z}\sim\mathrm{Bern}(p_1)\times\cdots\times\mathrm{Bern}(p_n)}[\tilde{\mathbf{Z}} \neq \mathbf{Z}]$ where $\tilde{\mathbf{Z}}$ is the first element of the triple output by $\mathrm{RPD}((\mathbf{Z}\cdot\mathbf{P}_n)_S;n,(p_1,\dots,p_n))$

IF $n = 1$ and $W_1 \in \mathbb{F}_2$ THEN
  RETURN (W_1,W_1,p_1)
ELSE IF $n = 1$ and $W_1 = ?$ THEN
  RETURN $(1,1,p_1)$ if $p_1 \ge 1/2$ and $(0,0,p_1)$ otherwise
ELSE ▷ $n \ge 2$
  $\mathbf{W} = (\mathbf{W}^{(1)},\mathbf{W}^{(2)})$ where $\mathbf{W}^{(1)},\mathbf{W}^{(2)} \in (\mathbb{F}_2 \cup \{?\})^{n/2}$
  FOR $i = 1$ to $n/2$ DO
    let $q_i = b^+(p_i,p_{n/2+i})$
  Let $(\tilde{\mathbf{X}},\tilde{\mathbf{W}}^1,\boldsymbol{\rho}^{(1)}) =$ RECURSIVE POLAR DECODER$(\mathbf{W}^{(1)};n/2,(q_1,\dots,q_{n/2}))$
  FOR $i = 1$ to $n/2$ DO
    let $r_i = b^{|}(p_i,p_{n/2+i},\tilde{X}_i)$
  Let $(\tilde{\mathbf{Y}},\tilde{\mathbf{W}}^2,\boldsymbol{\rho}^{(2)}) =$ RECURSIVE POLAR DECODER$(\mathbf{W}^{(2)};n/2,(r_1,\dots,r_{n/2}))$
  Let $\tilde{\mathbf{Z}} = (\tilde{\mathbf{X}}+\tilde{\mathbf{Y}},\tilde{\mathbf{Y}})$, and let $\tilde{\mathbf{W}} = (\tilde{\mathbf{W}}^1,\tilde{\mathbf{W}}^2)$ and $\boldsymbol{\rho} = (\boldsymbol{\rho}^{(1)},\boldsymbol{\rho}^{(2)})$
  RETURN $(\tilde{\mathbf{Z}},\tilde{\mathbf{W}},\boldsymbol{\rho})$

We assume that p^+ and $p^{|}$ can be computed in constant time, and with this assumption it is straightforward to see that the above algorithm also has a running time of $O(n\log n)$, by using the same recursive analysis that we used in the proof of Proposition 11.4.3. The correctness is a bit more involved and we argue this in the next lemma.

[5] Note that this follows from definition of P_n.

Lemma 11.4.4. *Let* $\mathbf{Z} \sim \mathrm{Bern}(p_1) \times \cdots \times \mathrm{Bern}(p_n)$*, and* $\mathbf{W} = P_n(\mathbf{Z})$*. Further let* $\mathbf{W}' = (W'_1, \ldots, W'_n)$ *be given by* $W'_i = W_i$ *if* $i \in S$ *and* $W'_i =?$ *otherwise. Let* $(\tilde{\mathbf{Z}}, \tilde{\mathbf{W}}, \rho) = \textsc{Recursive Polar Decoder}(\mathbf{W}'; n, (p_1, \ldots, p_n))$*. Then, if* $H(W_i|\mathbf{W}_{<i}) \le \tau$ *for every* $i \notin S$*, we have the following:*

(1) For every i*, it is the case that* $\Pr_{\mathbf{Z}}[W_i = 1|\mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}] = \rho_i$*.*

(2) $\Pr_{\mathbf{Z}}\left[\mathbf{W} \neq \tilde{\mathbf{W}}\right] \le \tau n$*.*

(3) $\Pr_{\mathbf{Z}}\left[\mathbf{Z} \neq \tilde{\mathbf{Z}}\right] \le \tau n$*.*

Note that part (3) of Lemma 11.4.4 proves the same decoding error bound that we proved in Lemma 11.3.2 for the Successive Cancellation Decoder (Algorithm 10).

Proof of Lemma 11.4.4. The main part of the lemma is part (1). Part (2) follows almost immediately from part (1) and Proposition 11.3.1. And part (3) is straightforward. We prove the parts in turn below.

We begin by first arguing that for every n that is a power of two:

$$\tilde{\mathbf{W}} = P_n(\tilde{\mathbf{Z}}). \tag{11.10}$$

For the base case of $n = 1$, lines 5 and 7 show that $W_1 = Z_1$ as desired.[6] By induction (and lines 12 and 15) we have that $\tilde{\mathbf{W}}^1 = P_{n/2}(\tilde{\mathbf{X}})$ and $\tilde{\mathbf{W}}^2 = P_{n/2}(\tilde{\mathbf{Y}})$. Finally, line 16 implies that (the second equality follows from the definition of P_n):

$$P_n(\tilde{\mathbf{Z}}) = P_n((\tilde{\mathbf{X}} + \tilde{\mathbf{Y}}, \tilde{\mathbf{Y}})) = \left(P_{n/2}(\mathbf{X} + \tilde{\mathbf{Y}} + \mathbf{Y}), P_{n/2}(\mathbf{Y})\right) = (\tilde{\mathbf{W}}^1, \tilde{\mathbf{W}}^2) = \tilde{\mathbf{W}},$$

as desired.

Part (1) follows by induction on n. If $n = 1$ (where say we call $\mathrm{RPD}((W_1), 1, (p''_1))$) then the claim follows since here we have $\rho_1 = p''_1$, $W_1 = Z_1$ and $Z_1 \sim \mathrm{Bern}(p''_1)$. For larger values of n, we consider two cases (below we have $\mathbf{Z} = (\mathbf{U}, \mathbf{V})$).

If $i \le n/2$ then we have that $W_i = P_{n/2}(\mathbf{U} + \mathbf{V})_i$ (via definition of $\mathbf{P}_n$). Furthermore $(\mathbf{U} + \mathbf{V}) \sim \mathrm{Bern}(q_1) \times \cdots \times \mathrm{Bern}(q_{n/2})$. Thus by the inductive claim, the recursive call $\textsc{Recursive Polar Decoder}(\mathbf{W}'[1, \ldots, n/2]; n/2, (q_1, \ldots, q_{n/2})$ satisfies

$$\begin{aligned}
\rho_i &= \rho_i^{(1)} \\
&= \Pr_{(\mathbf{U}+\mathbf{V}) \sim \mathrm{Bern}(q_1) \times \cdots \times \mathrm{Bern}(q_{n/2})}\left[W_i = 1|\mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}\right] \\
&= \Pr_{\mathbf{Z} \sim \mathrm{Bern}(p_1) \times \cdots \times \mathrm{Bern}(p_n)}\left[W_i = 1|\mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}\right]
\end{aligned}$$

as desired. (In the above, the first equality follows from the algorithm definition and the third equality follows from the definition of q_i.)

Now if $i > n/2$, note that the condition $\mathbf{W}[1, \ldots, n/2] = \tilde{\mathbf{W}}[1, \ldots, n/2]$ is equivalent to the condition that $\mathbf{U} + \mathbf{V} = \tilde{\mathbf{X}}$ (since $\mathbf{W}[1, \ldots, n/2] = P_{n/2}(\mathbf{U} + \mathbf{V})$ and $\tilde{\mathbf{X}} = P_{n/2}^{-1}(\tilde{\mathbf{W}}[1, \ldots, n/2])$, where the

[6] Note that when defining P_n, the base case was $n = 2$ but note that if we started with $n = 1$ and define $P_1(Z) = W$, then the resulting definition is the same as on the we saw in Definition 11.4.1.

latter equality follows from (11.10) and the fact that the map $P_{n/2}$ is invertible). And now from definition of $\mathbf{P}_n$, we have $\mathbf{W}[n/2+1,\ldots,n] = P_{n/2}(\mathbf{V})$. Conditioning on $\mathbf{U}+\mathbf{V}=\tilde{\mathbf{X}}$ implies that $\mathbf{V} \sim \mathrm{Bern}(r_1)\times\cdots\times\mathrm{Bern}(r_{n/2})$ — this is exactly how r_i's were defined. Thus we have

$$\begin{aligned}
&\Pr_{\mathbf{Z}\sim\mathrm{Bern}(p_1)\times\cdots\times\mathrm{Bern}(p_n)}[W_i = 1|\mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}] \\
&= \Pr_{\mathbf{V}\sim\mathrm{Bern}(r_1)\times\mathrm{Bern}(r_{n/2})}\left[P_{n/2}(\mathbf{V})_{i-n/2} = 1|P_{n/2}(\mathbf{V})[1,\ldots,i-n/2-1] = \tilde{\mathbf{W}}[n/2+1,\ldots,i-1] \text{ and } \mathbf{U}+\mathbf{V}=\tilde{\mathbf{X}}\right].
\end{aligned}$$

By induction again on the recursive call to RECURSIVE POLAR DECODER($\mathbf{W}'[n/2+1,\ldots,n]; n/2, (r_1,\ldots,r_{n/2})$) we have the final quantity above equals $\rho^{(2)}_{i-n/2} = \rho_i$ (where the last equality follows from the algorithm definition). This concludes the proof of part (1).

Part (2) follows from Proposition 11.3.1. We first note that if $i \in S$, then by line 5 of the algorithm we have $W_i = \tilde{W}_i$. Now assume $i \notin S$. We claim that $\tilde{W}_i = 1$ if and only if $\rho_i = \Pr_{\mathbf{Z}}[W_i = 1|\mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}] \geq 1/2$. To see this note that $\rho_i = p'_i$ when RECURSIVE POLAR DECODER is called on the input $(W'_i, 1, p'_i)$. Then the claim follows from line 7 of the algorithm. Thus $\tilde{W}_i = \mathrm{argmax}_{b\in\mathbb{F}_2}\{\{\Pr[W_i = b|\mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}]$. By Proposition 11.3.1, applied to the variables $X = W_i$ and $Y = \mathbf{W}_{<i}$ with $\alpha = \tau$ we get

$$\Pr[W_i \neq \tilde{W}_i|\mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}] \leq \tau.$$

By a union bound over i, we thus have

$$\Pr[\exists i \in [n] \text{ s.t. } W_i \neq \tilde{W}_i|\mathbf{W}_{<i} = \tilde{\mathbf{W}}_{<i}] \leq \tau n.$$

But if $\mathbf{W} \neq \tilde{\mathbf{W}}$ there must exist a least index i such that $W_i \neq \tilde{W}_i$ and so we have $\Pr[\mathbf{W} \neq \tilde{\mathbf{W}}] \leq \tau n$ concluding proof of part (2).

For part (3), note that by (11.10), if $\tilde{\mathbf{W}} = \mathbf{W}$ (which holds with probability at least $1-\tau n$ from part (2)), we have $\tilde{\mathbf{Z}} = P_n^{-1}(\tilde{\mathbf{W}}) = P_n^{-1}(\mathbf{W}) = \mathbf{Z}$ as desired. □

The correctness of BASIC POLAR DECODER (Algorithm 12) follows immediately and we state this explicitly below (see Exercise 11.10).

Corollary 11.4.5. *For every input $n = 2^t$, $p \in [0,1/2)$ and $\mathbf{W} \in (\mathbb{F}_2 \cup \{?\})^n$, the* BASIC POLAR DECODER *(Algorithm 12) runs in time $O(n\log n)$ and computes an output $\tilde{\mathbf{Z}} \in \mathbb{F}_2^n$ such that $P_n(\tilde{\mathbf{Z}})_i = \mathbf{W}_i$ holds for every i for which $\mathbf{W}_i \neq ?$. Furthermore if (1) $\mathbf{Z} \sim \mathrm{Bern}(p)^n$, (2) $\mathbf{W}_S = P_n(\mathbf{Z})_S$ and (3) $H(W_i|\mathbf{W}_{<i}) \leq \tau$ for every $i \notin S$ then $\Pr_{\mathbf{Z}}[\tilde{\mathbf{Z}} \neq \mathbf{Z}] \leq \tau n$.*

To summarize the claims of this section, Theorem 11.4.1 guarantees the existence of a polarizing matrix as desired to satisfy the information-theoretic conditions of Question 11.3.1. And Proposition 11.4.3 and Corollary 11.4.5 ensure that the encoding and descoding times are $O(n\log n)$. This allows us to complete the proof of Theorem 11.1.1 (modulo the proof of Theorem 11.4.1 — which will be proved in the next section).

Proof of Theorem 11.1.1. Recall from Proposition 11.2.1 and Question 11.3.1 that it suffices to find, given $p \in [0,1/2)$ and $\varepsilon > 0$, an $(\varepsilon, \varepsilon/n)$-polarizing matrix $\mathbf{P} \in \mathbb{F}_2^{n\times n}$ with n bounded by a polynomial in $1/\varepsilon$; such that multiplication by $\mathbf{P}$ and decompression take $O(n\log n)$ time.

Theorem 11.4.1 applied with parameters p, ε and $c = 2$ yields n and the matrix $\mathbf{P} = \mathbf{P}_n \in \mathbb{F}_2^{n\times n}$ that is $(\varepsilon, 1/n^2)$-polarizing. Since Theorem 11.4.1 also guarantees $\varepsilon \geq 1/n$, we have that $\mathbf{P}_n$ is $(\varepsilon, \varepsilon/n)$-polarizing. Furthermore we have that there exists a set $S \subseteq [n]$ with $|S| \leq (H(p)+\varepsilon)n$ such that $H(\mathbf{W}_{\bar{S}}|\mathbf{W}_S) \leq \varepsilon$ when $\mathbf{W} = \mathbf{Z}\cdot\mathbf{P}$ and $\mathbf{Z} \sim \mathrm{Bern}(p)^n$. Given such a set S we have, by Proposition 11.4.3, that the time to compress $\mathbf{Z} \in \mathbb{F}_2^n$ to $(\mathbf{Z}\cdot\mathbf{P}_n)_S$ is $O(n\log n)$. Finally Corollary 11.4.5 asserts that a decompression $\tilde{\mathbf{Z}}$ can be computed given $(\mathbf{Z}\cdot\mathbf{P}_n)_S$ in time $O(n\log n)$ and $\tilde{\mathbf{Z}}$ equals $\mathbf{Z}$ with probability at least $1-\varepsilon$ thereby completing the proof of Theorem 11.1.1. □

11.5 Analysis: Speed of Polarization

We now turn to the most crucial aspect of polarization - the fact that it happens, and it is fast enough to deliver polynomial convergence to capacity. In this section we first give an overview of how we will think about polarization. We will then analyze the convergence by first exploring what happens in a single polarization step (i.e., the action of P_2) and then showing how the local effects aggregate after $\log_2 n$ steps of polarization. This will lead us to the proof of Theorem 11.4.1.

11.5.1 Overview of Analysis

We start by setting up some more notation. Recall $n = 2^t$. In this section it will be convenient for us to give names to intermediate variables $\{Z_i^{(j)}\}_{(i\in[n], 0\leq j\leq t)}$ that are computed during the computation of $P_n(Z_1,\ldots,Z_n)$. Let $Z_i^{(0)} = Z_i$. For $1 \leq j \leq t$, let

$$(Z_i^{(j)}, Z_{i+2^{t-j}}^{(j)}) = P_2(Z_i^{(j-1)}, Z_{i+2^{t-j}}^{(j-1)}) = (Z_i^{(j-1)} + Z_{i+2^{t-j}}^{(j-1)}, Z_{i+2^{t-j}}^{(j-1)}). \tag{11.11}$$

for every $i \in [n]$ such that i and $i+2^{t-j}$ are in the same "*block at the jth*" level. i.e., $\lceil i/2^{t-j+1}\rceil = \lceil (i+2^{t-j})/2^{t-j+1}\rceil$. (Alternatively one could say i and $i+2^{t-j}$ as in the same *dyadic interval* of size 2^{t-j+1}.) Further,

Definition 11.5.1. We will say that a pair (i, i') are *jth-level siblings* if they are in the same block at the jth level and $i' = i + 2^{t-j}$.

Note that if one unravels the recursion in (11.11), then if $\mathbb{Z}_i, Z_{i'}$ are on the LHS, then (i, i') are siblings at the jth level.

We now claim that (see Exercise 11.11):

$$P_n(\mathbf{Z}) = (Z_1^{(t)}, \ldots, Z_n^{(t)}). \tag{11.12}$$

Figure 11.3 illustrates the block structure of P_n and the notion of jth level blocks and siblings at the jth level.

In what follows we will pick a random $i \in [n]$ and analyze the conditional entropy of $Z_i^{(j)}|\mathbf{Z}_{<i}^{(j)}$ as j progresses from 0 to t (we independently pick i for different values of j). Indeed let

$$X_j = H(Z_i^{(j)}|\mathbf{Z}_{<i}^{(j)}).$$

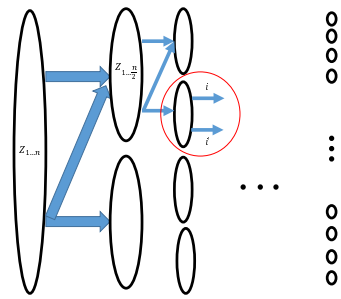


Figure 11.3: Block structure of the Basic Polarizing Transform. Circled are a block at the 2nd level and two 2nd level siblings.

Clearly $X_0 = H(p)$ since $Z_i^{(0)} = Z_i$ is independent of $\mathbf{Z}_{<i}$ and is distributed according to Bern(p). In what follows we will show that for every constant c if t is a sufficient large then with high probability *over the choice of* i, $X_t \notin (n^{-c}, 1-n^{-c})$. To track the evolution of X_j as j increases, we will first try to analyze *local polarization* which will study how X_j compares with X_{j-1}. Definition 11.5.2 below captures the desired effect of a local step, and the following lemma asserts that the operator P_2 does indeed satisfy the conditions of local polarization.

Definition 11.5.2 (Local Polarization)**.** A sequence of random variables $X_0, \ldots, X_j, \ldots$, with $X_j \in [0,1]$ is *locally polarizing* if the following conditions hold:

(1) **(Unbiased):** For every j, and $a \in [0,1]$ we have $\mathbb{E}[X_{j+1} | X_j = a] = a$.

(2) **(Variance in the middle):** For every $\tau > 0$, there exists $\theta = \theta(\tau) > 0$ such that for all j, we have: If $X_j \in (\tau, 1-\tau)$ then $|X_{j+1} - X_j| \geq \theta$.

(3) **(Suction at the ends):** For every $c < \infty$, there exists $\tau = \tau(c) > 0$ such that (i) if $X_j \leq \tau$ then $\Pr[X_{j+1} \leq X_j / c] \geq 1/2$; and similarly (ii) if $1 - X_j \leq \tau$ then $\Pr[(1 - X_{j+1} \leq (1 - X_j)/c] \geq 1/2$.

We further say a sequence is *simple* if for every sequence $a_0, \ldots, a_j$, conditioned on $X_0 = a_0, \ldots, X_j = a_j$, there are two values a^+ and $a^|$ such that X_{j+1} takes value a^+ with probability $1/2$ and $a^|$ with probability $1/2$.

Lemma 11.5.1 (Local Polarization)**.** *The sequence $X_0, \ldots, X_j, \ldots$, with $X_j = H(Z_i^{(j)} | \mathbf{Z}_{<i}^{(j)})$ where i is drawn uniformly from $[n]$ is a simple and locally polarizing sequence.*

We will prove Lemma 11.5.1 in Section 11.5.2 but use it below. But first let us see what it does and fails to do. While local polarization prevents the conditional entropies from staying static, it doesn't assert that eventually all conditional entropies will be close to 0 or close to 1, the kind of strong polarization that we desire. The following definition captures our desire from a strong polarizing process, and the lemma afterwards asserts that local polarization does imply strong polarization.

Definition 11.5.3 ((Polynomially) Strong Polarization)**.** A sequence of random variables $X_0, X_1, \ldots, X_t, \ldots$ with $X_j \in [0,1]$ *strongly polarizes* if for all $\gamma > 0$ there exist $\alpha < 1$ and $\beta < \infty$ such that for all t we have

$$\Pr[X_t \in (\gamma^t, 1-\gamma^t)] \le \beta \cdot \alpha^t.$$

Lemma 11.5.2 (Local vs. Global Polarization)**.** *If a sequence $X_0, \ldots, X_t, \ldots$, with $X_t \in [0,1]$ is simple and locally polarizing, then it is also strongly polarizing.*

Armed with Lemmas 11.5.1 and 11.5.2, proving Theorem 11.4.1 is just a matter of setting parameters.

Proof of Theorem 11.4.1. We assume without loss of generality that $c \ge 1$. (Proving the theorem for larger c implies it also for smaller values of c.) Given p and $c \ge 1$, let $\gamma = 2^{-c}$. Let $\beta < \infty$ and $\alpha < 1$ be the constants given by the definition of strong polarization (Definition 11.5.3) for this choice of γ. We prove the theorem for $n_0(x) = \max\{8x, 2(2\beta x)^{\lceil \log(1/\alpha) \rceil}\}$. Note that n_0 is a polynomial in x. Given $\varepsilon > 0$, let $t = \max\left\{\lfloor \log(8/\varepsilon) \rfloor, \left\lceil \frac{\log(2\beta/\varepsilon)}{\log(1/\alpha)} \right\rceil\right\}$ so that

$$n = 2^t \le \max\left\{8/\varepsilon, 2 \cdot (2\beta/\varepsilon)^{1/\log(1/\alpha)}\right\} = n_0(1/\varepsilon).$$

Note that the choice of t gives $\beta \cdot \alpha^t \le \varepsilon/2$ and $\gamma^t = 2^{-ct} = n^{-c}$. We also have $n \ge 4/\varepsilon$ and thus $2n^{-c} \le 2n^{-1} \le \varepsilon/2$. We show that for this choice and t and n, the polarizing transform P_n has the desired properties — namely that the set S of variables of noticeably large conditional entropy is of small size.

We first show that the set of variables with intermediate conditional entropies is small. Let us recall some notations from above, specifically (11.11). Let $(Z_i^{(j)})$ denote the intermediate results of the computation $\mathbf{W} = P_n(\mathbf{Z}) = (Z_1^{(t)}, \ldots, Z_n^{(t)})$, and let $X_j = H(Z_i^{(j)} | \mathbf{Z}_{<i}^{(j)})$ for a uniformly random choice of $i \in [n]$. By Lemmas 11.5.1 and 11.5.2 we have that the sequence $X_0, \ldots, X_t, \ldots$ is strongly polarizing. By the definition of strong polarization, we have that

$$\begin{aligned}
\Pr_{i \in [n]}\left[H(W_i|\mathbf{W}_{<i}) \in (n^{-c}, 1-n^{-c})\right] &= \Pr_i\left[H(Z_i^{(t)}|\mathbf{Z}_{<i}^{(t)}) \in (\gamma^t, 1-\gamma^t)\right] \\
&= \Pr\left[X_t \in (\gamma^t, 1-\gamma^t)\right] \\
&\le \beta\alpha^t \\
&\le \varepsilon/2.
\end{aligned}$$

Thus we have that the set $E = \{i \in [n] | H(W_i|\mathbf{W}_{<i}) \in (n^{-c}, 1-n^{-c})\}$ satisfies $|E| \le \varepsilon n/2$.

Finally, we argue the "Further" part. Indeed, we have

$$nH(p) = \sum_{i \in [n]} H(W_i|\mathbf{W}_{<i}) \ge \sum_{i \in S \setminus E} H(W_i|\mathbf{W}_{<i}) \ge (|S| - |E|)(1-n^{-c}),$$

where the first equality follows from the chain rule and the last inequality follows from definitions of S and E. Re-arranging one gets that

$$|S| \le \frac{nH(p)}{1-n^{-c}} + \varepsilon n/2 \le nH(p)(1+2n^{-c}) + \varepsilon n/2 \le nH(p) + \varepsilon n.$$

□

It remains to prove Lemmas 11.5.1 and 11.5.2 which we prove in the rest of this chapter.

11.5.2 Local Polarization

To understand how X_j compares with X_{j-1}, we start with some basic observations about these variables, or more importantly the variables $Z_i^{(j)}$ and $Z_i^{(j+1)}$ (recall (11.11)). Let i and i' be jth level siblings, so that $(Z_i^{(j)}, Z_{i'}^{(j)}) = P_2(Z_i^{(j-1)}, Z_{i'}^{(j-1)})$. Our goal is to compare the pairs of conditional entropies $(H(Z_i^{(j)}|\mathbf{Z}_{<i}^{(j)}), H(Z_{i'}^{(j)}|\mathbf{Z}_{<i'}^{(j)}))$ with $(H(Z_i^{(j-1)}|\mathbf{Z}_{<i}^{(j-1)}), H(Z_{i'}^{(j-1)}|\mathbf{Z}_{<i'}^{(j-1)}))$. The collection of variables involved and conditioning seem messy, so let us look at the structure of P_n more carefully to simplify the above. We do this by noticing the $Z_i^{(j)}$ is really independent of most $Z_{i'}^{(j)}$ (at least for small values of j) and in particular the set of variables that $Z_i^{(j-1)}$ and $Z_{i'}^{(j-1)}$ depend on are disjoint. Furthermore these two variables, and the sets of variables that they depend on are identically distributed. Next, we present the details.

We begin with a useful notation. Given $i \in [n = 2^t]$ and $0 \le j \le t$, let

$$S_{i,j} \stackrel{\text{def}}{=} \left\{k \in [n] | i \equiv k \mod 2^{t-j}\right\}.$$

Note that the $\ell_1, \ell_2 \in S_{i,j}$ need not be siblings at the jth level.

Proposition 11.5.3. *For every $1 \le j \le t$ and jth level siblings i and i' with $i < i'$ the following hold:*

(1) $S_{i,j} = S_{i',j} = S_{i,j-1} \cup S_{i',j-1}$.

(2) $\left\{Z_k^{(j)} | k \in S_{i,j}\right\}$ *is independent of* $\left\{Z_k^{(j)} | k \notin S_{i,j}\right\}$.

(3)

$$\begin{aligned} H(Z_i^{(j-1)}|\mathbf{Z}_{<i}^{(j-1)}) &= H\left(Z_i^{(j-1)}|\mathbf{Z}_{\{k \in S_{i,j-1}, k<i\}}^{(j-1)}\right) \\ &= H\left(Z_i^{(j-1)}|\mathbf{Z}_{\{k \in S_{i,j}, k<i\}}^{(j-1)}\right). \end{aligned}$$

(4)

$$\begin{aligned} H(Z_{i'}^{(j-1)}|\mathbf{Z}_{<i'}^{(j-1)}) &= H\left(Z_{i'}^{(j-1)}|\mathbf{Z}_{\left\{k \in S_{i',j-1}, k<i'\right\}}^{(j-1)}\right) \\ &= H\left(Z_{i'}^{(j-1)}|\mathbf{Z}_{\{k \in S_{i,j}, k<i\}}^{(j-1)}\right). \end{aligned}$$

(5)

$$\begin{aligned} H(Z_i^{(j)}|\mathbf{Z}_{<i}^{(j)}) &= H\left(Z_i^{(j)}|\mathbf{Z}_{\{k \in S_{i,j}, k<i\}}^{(j)}\right) \\ &= H\left(Z_i^{(j)}|\mathbf{Z}_{\{k \in S_{i,j}, k<i\}}^{(j-1)}\right). \end{aligned}$$

(6)

$$H(Z_{i'}^{(j)}|\mathbf{Z}_{<i'}^{(j)}) = H\left(Z_{i'}^{(j)}|\left\{\mathbf{Z}_i^{(j)}\right\} \cup \mathbf{Z}_{\{k\in S_{i,j},k<i\}}^{(j)}\right)$$
$$= H\left(Z_{i'}^{(j)}|\left\{Z_i^{(j)}\right\} \cup \mathbf{Z}_{\{k\in S_{i,j},k<i\}}^{(j-1)}\right).$$

Proof. Part (1) follows from the definition of $S_{i,j}$ and the definition of siblings. Indeed, since $i' = i + 2^{t-j}$, we have $i \equiv i' \mod 2^{t-j}$, which implies the first equality. The second equality follows from the observations that $k_1 \equiv k_2 \mod 2^{t-j+1}$ implies $k_1 \equiv k_2 \mod 2^{t-j}$ (this in turn implies $S_{i,j-1}, S_{i',j-1} \subseteq S_{i,j}$) and that if $k \equiv i \mod 2^{t-j+1}$ then $k \not\equiv i' \mod 2^{t-j}$ and vice versa (which in turn implies that $S_{i,j-1}$ and $S_{i',j-1}$ are disjoint. Part (2) follows from the fact that (see Exercise 11.12):

Lemma 11.5.4. *For every i, the set $\left\{Z_k^{(j)}|k \in S_{i,j}\right\}$ is determined completely by $\left\{Z_k|k \in S_{i,j}\right\}$,*

and the Z_k's are all independent. The first equality in part (3) follows immediately from part (2), and the second uses part (1) and the fact that $Z_i^{(j-1)}$ is independent of $\left\{Z_k|k \in S_{i',j-1}, k < i\right\}$ (the latter claim follows from the fact that $S_{i,j-1}$ and $S_{i',j-1}$ as disjoint, as argued in the proof of part (1) above). The first equality in part (4) is similar, whereas the second uses the additional fact that $S_{i',j-1}$ contains no elements between i and i'. Indeed the latter observation implies that $H\left(Z_{i'}^{(j-1)}|\mathbf{Z}_{\left\{k\in S_{i',j-1},k<i'\right\}}^{(j-1)}\right) = H\left(Z_{i'}^{(j-1)}|\mathbf{Z}_{\left\{k\in S_{i',j-1},k\le i\right\}}^{(j-1)}\right)$. But by part (2), $Z_{i'}^{(j-1)}$ is independent of $Z_i^{(j-1)}$ and hence we have $H\left(Z_{i'}^{(j-1)}|\mathbf{Z}_{\left\{k\in S_{i',j-1},k<i'\right\}}^{(j-1)}\right) = H\left(Z_{i'}^{(j-1)}|\mathbf{Z}_{\left\{k\in S_{i',j-1},k<i\right\}}^{(j-1)}\right)$. The second equality in (4) then follows from parts (1) and (2). The first equalities in parts (5) and (6) are similar to the first equality in part (3) with part (6) using the fact that $\left\{k \in S_{i',j}|k < i'\right\} = \{i\} \cup \left\{k \in S_{i,j}|k < i\right\}$. The second equality follows from the fact that (see Exercise 11.13):

Lemma 11.5.5. *There is a one-to-one map from the variables $\mathbf{Z}_{\{k\in S_{i,j},k<i\}}^{(j-1)}$ to the variables $\mathbf{Z}_{\{k\in S_{i,j},k<i\}}^{(j)}$,*

and so conditioning on one set is equivalent to conditioning on the other. □

To summarize the effect of Proposition 11.5.3 above, let us name the random variables $U = Z_i^{(j-1)}$ and $V = Z_{i'}^{(j-1)}$ and further let $A = \left\{Z_k^{(j-1)}|k \in S_{i,j-1}, k < i\right\}$ and $B = \left\{Z_k^{(j-1)}|k \in S_{i',j-1}, k < i'\right\}$. By the proposition above, the conditional entropies of interest (i.e., those of i and i') at the $(j-1)$th stage are $H(U|A)$ and $H(V|B)$. On the other hand the conditional entropies of interest one stage later (i.e., at the jth stage) are $H(U+V|A,B)$ and $H(V|A,B,U)$. (Here we use that fact that $P_2(U,V) = (U+V,V)$.) By part (2) of Proposition 11.5.3 we also have that (U,A) and (V,B) are independent of each other. Finally, by examination we also have that (see Exercise 11.14)

Lemma 11.5.6. *(U,A) and (V,B) are identically distributed.*

So our concern turns to understanding the local polarization of two independent and identically distributed bits. If one could ignore the conditioning then this is just a problem about two bits (U, V) and their polarization when transformed to $(U+V, V)$.

In the following lemma, we show how in the absence of conditioning these variables show local polarization effects. (For our application it will be useful for us to allow the variables to be not identically distributed, though still independent.) Suppose $H(U) = H(p_1)$ and $H(V) = H(p_2)$, then notice that $H(U+V) = H(p_1 \circ p_2)$ where

$$p_1 \circ p_2 \stackrel{\text{def}}{=} p_1(1-p_2) + p_2(1-p_1).$$

In the following lemma we show how $H(p_1 \circ p_2)$ relates to $H(p_1)$ and $H(p_2)$.

Lemma 11.5.7. *Let $p_1, p_2 \in [0, 1/2]$ with $p_1 < p_2$ and $\tau \in (0, 1/2)$. Then we have:*

(1) $H(p_1 \circ p_2) \geq H(p_2)$.

(2) There exists $\theta = \theta(\tau) > 0$ such that if $H(p_1), H(p_2) \in (\tau, 1-\tau)$ then

$$H(p_1 \circ p_2) - H(p_2) \geq \theta.$$

(3) If $H(p_1), H(p_2) \leq \tau$ then

$$H(p_1 \circ p_2) \geq (1 - 9/\log(1/\tau))(H(p_1) + H(p_2)).$$

In particular, for every $c < \infty$, if $\tau \leq 2^{-9c}$ then

$$H(p_1) + H(p_2) - H(p_1 \circ p_2) \leq (H(p_1) + H(p_2))/c.$$

(4) If $H(p_1), H(p_2) \geq 1-\tau$ and $\tau \leq 1 - H(1/4)$ then

$$H(p_1 \circ p_2) \geq 1 - 20\tau(1 - H(p_2)).$$

In particular, for every $c' < \infty$, if $\tau < 1/(20c')$ then

$$1 - H(p_1 \circ p_2) \leq (1 - H(p_2))/c'.$$

We defer the proof of the above lemma to Section 11.6.2.

The lemma above essentially proves that $H(U+V)$ satisfies the requirements for local polarization relative to $H(U)$ and $H(V)$, but we still need to deal with the conditioning with respect to A and B. We do this below using some careful applications of Markov's inequality.

Lemma 11.5.8. *If (U, A) and (V, B) are identical and independent random variables with U, V being elements of $\mathbb{F}_2$ with $H(U|A) = H(V|B) = H(p)$, then the following hold:*

(1) For every $\tau > 0$ there exists $\theta > 0$ such that if $H(p) \in (\tau, 1-\tau)$ then

$$H(U+V|A,B) \geq H(p) + \theta.$$

(2) *For every $c < \infty$ there exists $\tau > 0$ such that if $H(p) \le \tau$ then*

$$H(U+V|A,B) \ge (2-1/c)H(p),$$

and if $H(p) \ge 1-\tau$ then

$$H(U+V|A,B) \ge 1-1/c(1-H(p)).$$

Proof. Let $p_a = \Pr[U=1|A=a]$ so that $H(p) = H(U|A) = \mathbb{E}_A[H(p_A)]$. Similarly let $q_b = \Pr[V = 1|B=b]$. In what follows we consider what happens when A and B are chosen at random. If $H(p_A)$ and $H(q_B)$ are close to their expectations, then the required polarization comes from Lemma 11.5.7. But if $H(p_A)$ or $H(q_B)$ can deviate significantly from their expectation, then polarization happens simply due to the fact that one of them is much larger than the other and $H(U+V|A=a,B=b) \ge \max\{H(p_a), H(q_b)\}$. The details are worked out below.

We start with part (1). Let $\theta(\cdot)$ be the function from part (2) of Lemma 11.5.7 so that if $H(p_1), H(p_2) \in (\rho, 1-\rho)$ then $H(p_1 \circ p_2) - H(p_2) \ge \theta(\rho)$. Given $\tau > 0$ let $\theta_1 = \theta(\tau/2)$. We prove this part for $\theta = \min\left\{\frac{\theta_1}{9}, \frac{\tau^2}{36}\right\} > 0$.

Let

$$r_1 = \Pr_A[H(p_A) \le \tau/2],$$
$$r_2 = \Pr_A[H(p_A) \in (\tau/2, 1-\tau/2)],$$

and

$$r_3 = \Pr_A[H(p_A) \ge 1-\tau/2].$$

(Note that since (U,A) and (V,B) are identically and independently distributed if one replaces p_A by q_B in the above equalities, then the equalities still remain valid. We will be implicitly using this for the rest of the proof.) Since $r_1 + r_2 + r_3 = 1$, at least one of them must be greater than or equal to $1/3$. Suppose $r_2 \ge 1/3$, then we have with probability at least $1/9$, both $H(p_A) \in (\tau/2, 1-\tau/2)$ and $H(q_B) \in (\tau/2, 1-\tau/2)$. Let a, b be such that $H(p_a), H(q_b) \in (\tau/2, 1-\tau/2)$. Then, since $U+V \sim \mathrm{Bern}(p_a \circ p_b)$, by Lemma 11.5.7 part (2),

$$H(U+V|A=a,B=b) \ge H(p_a) + \theta_1.$$

And by Lemma 11.5.7 part (1), we have for all a, b,

$$H(U+V|A=a,B=b) \ge H(p_a).$$

Putting it together, we have

$$H(U+V|A,B) \ge \mathbb{E}_A[p_A] + \frac{1}{9}\cdot\theta_1 = H(p) + \frac{\theta_1}{9}.$$

Next we consider the case where $r_3 \ge 1/3$. Now consider the probability that $\Pr_A[H(p_A) \le 1-\tau]$. Notice that

$$1-\tau \ge H(p) \ge (1 - r_3 - \Pr_A[H(p_A) \le 1-\tau]) \cdot (1-\tau) + r_3 \cdot (1-\tau/2).$$

Rearranging we conclude

$$\Pr_A[H(p_A) \le (1-\tau)] \ge \frac{r_3\tau}{2(1-\tau)} \ge \frac{\tau}{6}.$$

Thus with probability at least $\tau/18$ we have A such that $H(p_A) \le (1-\tau)$ and B such that $H(q_B) \ge 1-\tau/2$. Let a, b be such that $H(p_a) \le 1-\tau$ and $H(q_b) \ge 1-\tau/2$. We have (from part (1) of Lemma 11.5.7)

$$H(U+V|A=a, B=a) \ge H(q_b) \ge H(p_a) + \frac{\tau}{2}.$$

We conclude that in this case

$$H(U+V|A,B) \ge \mathbb{E}_A[H(p_A)] + \frac{\tau^2}{36} = H(p) + \frac{\tau^2}{36}.$$

The case $r_1 \ge 1/3$ is similar and also yields

$$H(U+V|A,B) \ge H(p) + \tau^2/36.$$

Thus in all cases we have

$$H(U+V|A,B) \ge H(p) + \theta,$$

which completes the proof of part (1).

We now turn to part (2). We only prove the case where $H(p) \le \tau$. The case where $H(p) \ge 1-\tau$ is similar and we omit that part (see Exercise 11.15). Given $c < \infty$, let $\tau' = \tau(4c)$ be the constant from part (3) of Lemma 11.5.7 for constant $4c$, so that if $H(p_1), H(p_2) \le \tau'$ then

$$H(p_1 \circ p_2) \ge \left(1 - \frac{1}{4c}\right) \cdot \big(H(p_1) + H(p_2)\big).$$

Now let $\tau = \frac{\tau'}{2c}$ and $H(p) \le \tau$. Define

$$\alpha = \Pr_A\left[H(p_A) \ge \tau'\right].$$

By Markov's inequality (Lemma 3.1.4) we have $\alpha \le 1/(2c)$. Let

$$\gamma = \mathbb{E}_A\left[H(p_A)|H(p_A) \ge \tau'\right].$$

and

$$\delta = \mathbb{E}_A\left[H(p_A)|H(p_A) < \tau'\right].$$

We have

$$H(p) = \gamma\alpha + \delta(1-\alpha). \tag{11.13}$$

We divide our analysis into four cases depending on whether $H(p_A) \ge \tau'$ or not, and whether $H(q_B) \ge \tau'$ or not. Let S_{11} denote the event that $H(p_A) \ge \tau'$ and $H(q_B) \ge \tau'$ and S_{00} denotes the event that $H(p_A) < \tau'$ and $H(q_B) < \tau'$. Define S_{10} and S_{01} similarly.

We start with the case of S_{10}. Let a, b be such that $H(p_a) \geq \tau'$ and $H(q_b) < \tau'$. We have by part (1) of Lemma 11.5.7, $H(U+V|A=a, B=b) \geq H(U|A=a) = H(p_a)$ (and similarly $H(U+V|A=a, B=b) \geq H(q_b)$. Thus taking expectation after conditioning on $(A,B) \in S_{10}$ we have

$$\mathbb{E}_{(A,B)|(A,B)\in S_{10}}[H(U+V|A,B)] \geq \mathbb{E}\left[H(p_A)|H(p_A) \geq \tau'\right] = \gamma.$$

Similarly we have

$$\mathbb{E}_{(A,B)|(A,B)\in S_{01}}[H(U+V|A,B)] \geq \gamma$$

as well as

$$\mathbb{E}_{(A,B)|(A,B)\in S_{11}}[H(U+V|A,B)] \geq \gamma.$$

Note that $S_{11} \cup S_{10} \cup S_{01}$ happen with probability $2\alpha - \alpha^2$. Now we turn to S_{00}. Let a, b be such that $H(p_a), H(q_b) < \tau'$. By Lemma 11.5.7 part (3) we have $H(U+V|A=a, B=b) \geq (1 - 1/(4c))\left(H(p_a) + H(q_b)\right)$. Taking expectations conditioned on $(A,B) \in S_{00}$ we get

$$\begin{aligned}\mathbb{E}_{(A,B)|(A,B)\in S_{00}}[H(U+V|A,B)] &\geq \left(1 - \frac{1}{4c}\right) \cdot \left(\mathbb{E}_A\left[H(p_A)|H(p_A) < \tau'\right] + \mathbb{E}_B\left[H(q_B)|H(q_B) < \tau'\right]\right) \\ &= \left(1 - \frac{1}{4c}\right) \cdot 2\delta.\end{aligned}$$

Note finally that S_{00} happens with probability $(1-\alpha)^2$. Combining the four cases we have

$$\begin{aligned}H(U+A|A,B) &\geq (2\alpha - \alpha^2)\gamma + (1-\alpha)^2\left(1 - \frac{1}{4c}\right)(2\delta) \\ &= 2\alpha\gamma + (1-\alpha)2\delta - \alpha^2\gamma - (\alpha)(1-\alpha)\delta - \frac{1}{4c} \cdot (1-\alpha)^2 2\delta \\ &= 2H(p) - \alpha \cdot H(p) - \frac{1}{2c} \cdot (1-\alpha)((1-\alpha)\delta).\end{aligned}$$

In the above, the last equality follows from (11.13). Part (2) now follows by using $(1-\alpha)\delta \leq H(p)$ (which in turn follows from (11.13)) and $\alpha \leq 1/(2c)$. □

We are now ready to prove the local polarization lemma.

Proof of Lemma 11.5.1. Recall that $X_j = \mathbb{E}_{I\sim[n]}[Z_I^{(j)}]$. Let $X_j = H(p)$. Note that conditioned on the value of X_j, for any $(j+1)$-level siblings $i < i'$, I is equally likely to equal i or i'. Conditioning on $I \in \{i, i'\}$, with probability $1/2$, $I = i$ and with probability $1/2$ $I = i'$. Let $U = Z_i^{(j)}$, $V = Z_{i'}^{(j)}$, $A = \left\{Z_k^{(j)} | k < i, k \in S_{i,j}\right\}$ and $B = \left\{Z_k^{(j)} | k < i', k \in S_{i',j}\right\}$ then if $I = i$, $X_j = H(U|A)$ (this follows from Lemma 11.5.3 part (3)) and if $I = i'$ then $X_j = H(V|B)$ (this follows from Lemma 11.5.3 part (4)). Furthermore if $I = i$ then $X_{j+1} = H(U+V|A,B)$ (this follows from (11.11) and parts (1) and (5) from Lemma 11.5.3) and if $I = i'$ then $X_{j+1} = H(V|A,B,U)$ (this follows from (11.11), parts (1) and (6) from Lemma 11.5.3 and the fact that $V|U$ and $V|U+V$ have the same distribution). With this setup, we are now ready to prove that the sequence $X_0, X_1, \ldots,$ satisfy the conditions of local polarization, and furthermore are simple.

We argue the conditions hold for each conditioning $I \in \{i, i'\}$ and so hold without the conditioning (the latter holds because the pairs $\{i, i'\}$ make a disjoint cover of $[n]$ and hence for a random $I \sim [n]$ is equally likely to fall in one of these pairs). The condition $\mathbb{E}[X_{j+1}|X_j = a] = a$ follows from the fact that there is a bijection from (U, V) to $(U + V, V)$, and so

$$H(U + V|A, B) + H(V|A, B, U) = H(U|A) + H(V|B).$$

Indeed, note that $2a$ is the RHS and $2X_{j+1}$ is the LHS of the above equality.

Now note that (see Exercise 11.16):

Lemma 11.5.9. *(U, A) and (V, B) are independently and identically distributed.*

The variance in the middle condition follows from Lemma 11.5.8 part (1) and the suction at the ends condition follows from Lemma 11.5.8 part (2). Finally simplicity follows from the fact that with probability $1/2$, $X_j = H(U + V|A, B)$ and with probability $1/2$, $X_j = H(V|A, B, U)$. □

11.5.3 Local vs. Global Polarization

Finally we prove Lemma 11.5.2 which shows that simple local polarization implies strong polarization. We prove this part in two phases. First, we show that in the first $t/2$ steps, the sequence shows moderate polarization — namely, with all but exponentially small probability $X_{t/2}$ is an inverse exponential in t, but with a small constant base (so $X_t \notin \left(\alpha_1^t, 1 - \alpha_1^t\right)$ for some $\alpha_1 < 1$, but α_1 is close to 1). Next we show that conditioned on this moderate polarization, the sequence gets highly polarized (so $X_t \notin \left(\gamma^t, 1 - \gamma^t\right)$ for any $\gamma > 0$ of our choice), again with an exponentially small failure probability. We start with part (1).

In what follows, let $\gamma > 0$ be given and let

$$c = \max\left\{4, \frac{\gamma^8}{16}\right\}.$$

Let $\tau = \tau(c)$ be given by condition (3) of the definition of Local Polarization (Definition 11.5.2) and

$$\theta = \min\left\{1 - \frac{1}{c}, \theta(\tau)\right\}$$

where $\theta(\tau)$ is the constant given by condition (2) of the same definition.

We start with the first phase. We consider a potential function

$$\phi_j \stackrel{\text{def}}{=} \min\left\{\sqrt{X_j}, \sqrt{1 - X_j}\right\}.$$

We first notice that ϕ_j is expected to drop by a constant factor in each step of polarization.

Lemma 11.5.10.

$$\mathbb{E}\left[\phi_{j+1}|\phi_j = a\right] \leq \left(1 - \frac{\theta^2}{16}\right) \cdot a.$$

Proof. Without loss of generality assume $X_j \le 1/2$ (see Exercise 11.17) and so $a = \phi_j = \sqrt{X_j}$ and so $X_j = a^2$. Using the simplicity of the sequence $X_0, \ldots$ as well as the fact that $\mathbb{E}\left[X_{j+1} | X_j = a\right] = a$, we have that there exists δ such that $X_{j+1} = a^2 + \delta$ with probability $1/2$ and $a^2 - \delta$ with probability $1/2$. Furthermore, if $X_j \le \tau$, by the unbiasedness and suctions at the ends conditions, we have $\delta \ge (1 - 1/c)a^2$ and if $X_j > \tau$ by the variance in the middle condition, we have $\delta \ge \theta(\tau) \ge \theta(\tau)a^2$. Thus in either case we have

$$\delta \ge \theta a^2. \tag{11.14}$$

We now bound $\mathbb{E}\left[\phi_{j+1}\right]$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\phi_{j+1}\right] &\le \mathbb{E}\left[\sqrt{X_{j+1}}\right] \\
&= \frac{1}{2}\sqrt{a^2+\delta} + 1/2\sqrt{a^2-\delta} \\
&= \frac{a}{2}\left(\sqrt{1+\frac{\delta}{a^2}} + \sqrt{1-\frac{\delta}{a^2}}\right) \\
&\le \frac{a}{2}\left(1 + \frac{\delta}{2a^2} - \frac{\delta^2}{16a^4} + 1 - \frac{\delta}{2a^2} - \frac{\delta^2}{16a^4}\right) \\
&= a\left(1 - \frac{\delta^2}{16a^4}\right) \\
&\le a\left(1 - \frac{\theta^2}{16}\right).
\end{aligned}$$

In the above, the first inequality follows from Lemma B.1.5 while the second inequality follows from (11.14). □

Lemma 11.5.11 (Weakly Polynomial Polarization)**.** *There exists $\alpha_1 < 1$ such that for all even t, we have*

$$\Pr\left[X_{t/2} \in (\alpha_1^t, 1 - \alpha_1^t)\right] \le \alpha_1^t.$$

Proof. We first prove by induction on j that

$$\mathbb{E}\left[\phi_j\right] \le \left(1 - \frac{\theta^2}{16}\right)^j.$$

This is certainly true for $j = 0$ since $\phi_0 \le 1$. For higher j, by Lemma 11.5.10 we have

$$\mathbb{E}\left[\phi_j\right] \le \left(1 - \frac{\theta^2}{16}\right)\mathbb{E}\left[\phi_{j-1}\right] \le \left(1 - \frac{\theta^2}{16}\right)^j$$

as claimed. Let

$$\beta = \sqrt{1 - \frac{\theta^2}{16}}$$

and $\alpha_1 = \sqrt{\beta}$ (note that $\alpha_1 < 1$). By our claim, we have $\mathbb{E}\left[\phi_{t/2}\right] \le \beta^t$. By Markov's inequality (Lemma 3.1.4), we now get that

$$\Pr\left[\phi_{t/2} \ge \alpha_1^t\right] \le \frac{\beta^t}{\alpha_1^t} = \alpha_1^t.$$

Finally we note that if $\phi_{t/2} \le \alpha_1^t$ then $X_{t/2} \notin (\alpha_1^{2t}, 1-\alpha_1^{2t})$ and so in particular $X_{t/2} \notin (\alpha_1^t, 1-\alpha_1^t)$. We conclude that the probability that $X_{t/2} \in (\alpha_1^t, 1-\alpha_1^t)$ is at most α_1^t, yielding the lemma. □

We now turn to phase two of the polarization. Here we use the fact that if $X_{t/2}$ is much smaller than τ, then X_j is very unlikely to become larger than τ for any $t/2 \le j \le t$. Furthermore if it does not ever become larger than τ then X_t is very likely to be close to its expected value (which grows like γ^t). The following lemmas provide the details. In the following recall that $\tau = \tau(c)$ where $c \ge 4$.

Lemma 11.5.12. *For all $\lambda > 0$, if $X_0 \le \lambda$, then the probability there exists $j > 0$ such that $X_j \ge \tau$ is at most λ/τ. Similarly if $X_0 \ge 1-\lambda$, then the probability there exists $j > 0$ such that $X_j \le 1-\tau$ is at most λ/τ.*

The lemma above is a special case of Doob's inequality for martingales. We give the (simple) proof below.

Proof. We give the proof for the case $X_0 \le \lambda$. The case $X_0 \ge 1-\lambda$ is symmetrical (see Exercise 11.18). Notice that we wish to show that for every integer $T > 0$

$$\Pr\left[\max_{0\le t\le T}\{X_t\} \ge \tau\right] \le \lambda/\tau.$$

Let us create a related sequence of variables Y_i as follows. Let $Y_0 = X_0$ and for $i \ge 1$, if $Y_{i-1} < \tau$ then $Y_i = X_i$, else $Y_i = Y_{i-1}$. Note that for every i and a, by the simplicity of X_i's, we have $\mathbb{E}[Y_i | Y_{i-1} = a] = a$. Note further that $\max_{0\le t\le T}\{X_t\} \ge \tau$ if and only if $Y_T \ge \tau$. Thus

$$\Pr\left[\max_{0\le t\le T}\{X_t\} \ge \tau\right] = \Pr[Y_T \ge \tau] \le \frac{\mathbb{E}[Y_T]}{\tau},$$

where the final inequality is Markov's inequality (Lemma 3.1.4). But

$$\mathbb{E}[Y_T] = \mathbb{E}[Y_{T-1}] = \cdots = \mathbb{E}[Y_0] = \mathbb{E}[X_0] \le \lambda$$

and this yields the lemma. □

Lemma 11.5.13 (Weak to Strong Polarization)**.** *There exists $\alpha_2 < 1$ such that for every $\lambda > 0$ if $X_0 \notin (\lambda, 1-\lambda)$, then the probability that $X_{t/2} \in (\gamma^t, 1-\gamma^t)$ is at most $\lambda/\tau + \alpha_2^t$.*

Proof. Again we consider the case $X_0 \le \lambda$ with the other case being symmetrical (see Exercise 11.19).

Let $Z_i = 1$ if $X_i < X_{i-1}$ and 0 otherwise. For simple sequences, notice that Z_i's are independent and 1 with probability 1/2. Let $Z = \sum_{i=1}^{t/2} Z_i$. We consider two possible "error" events. $\mathscr{E}_1$ is

the event that there exists $1 \le j \le t/2$ such that $X_j \ge \tau$, and $\mathscr{E}_2$ is the event that $Z \le t/8$. Note that by Lemma 11.5.12, $\mathscr{E}_1$ happens with probability at most λ/τ and (by the Chernoff bounds–Theorem 3.1.6) $\mathscr{E}_2$ happens with probability at most α_2^t for some $\alpha_2 < 1$. Now, if event $\mathscr{E}_1$ does not occur, then

$$X_{t/2} \le 2^{t/2} \cdot c^{-Z} X_0 \le 2^{t/2} c^{-Z}.$$

The first inequality follows from the subsequent argument. Using simplicity, we have with probability $1/2$, $X_1 \le (1/c)X_0 \le X_0/4$ (because of the suction at the ends condition) and with probability $1/2$ $X_1 \le 2X_0$ (this follows the bound in the other case and the unbiasedness of the X_is). Further if $\mathscr{E}_2$ also does not occur we have

$$X_{t/2} \le 2^{t/2} \cdot c^{-t/8} \le \gamma^t$$

by the choice of $c = 1/(2/\gamma^2)^4$. □

Proof of Lemma 11.5.2. Recall that we wish to show, given $\gamma > 0$, that there exists $\alpha < 1$ and $\beta < \infty$ such that for all t we have

$$\Pr\left[X_t \in \left(\gamma^t, 1-\gamma^t\right)\right] \le \beta \cdot \alpha^t.$$

Let $\alpha_1 < 1$ be the constant from Lemma 11.5.11. Let $\alpha_2 < 1$ be the constant from Lemma 11.5.13. We prove this for $\alpha = \max\{\alpha_1, \alpha_2\} < 1$ and $\beta = 2 + 1/\tau < \infty$.

Let $\mathscr{E}$ be the event that $X_{t/2} \in (\alpha_1^t, 1-\alpha_1^t)$. By Lemma 11.5.11 we have that $\Pr[\mathscr{E}] \le \alpha_1^t$. Now conditioned of $\mathscr{E}$ not occurring, using Lemma 11.5.13 with $\lambda = \alpha_1^t$, we have $\Pr\left[X_t \in (\gamma^t, 1-\gamma^t)\right] \le \alpha_1^t/\tau + \alpha_2^t$. Thus putting the two together we get

$$\Pr\left[X_t \in (\gamma^t, 1-\gamma^t)\right] \le \alpha_1^t + \frac{\alpha_1^t}{\tau} + \alpha_2^t \le \alpha^t + \frac{\alpha^t}{\tau} + \alpha^t = \beta \cdot \alpha^t,$$

as desired. □

11.6 Entropic Calculations

In this section, we present the omitted proofs on various properties of entropy and probability that mostly need some calculations.

11.6.1 Proof of Proposition 11.3.1

We begin with the first part, which is a straightforward calculation expanding the definition. For any i in the support of X, let p_i denote $\Pr_X[X = i]$ and let $x = \operatorname{argmax}_i\{p_i\}$ be the value maximizing this probability. Let $p_x = 1-\gamma$. We wish to show that $\gamma \le \alpha$. We now perform some crude calculations that lead us to this bound.

If $\gamma \le 1/2$ we have

$$\alpha \ge H(X)$$

$$
\begin{aligned}
&= \sum_i p_i \log \frac{1}{p_i} \\
&\geq \sum_{i \neq x} p_i \log \frac{1}{p_i} && (11.15) \\
&\geq \sum_{i \neq x} p_i \log \frac{1}{\sum_{j \neq x} p_j} && (11.16) \\
&= \left(\sum_{i \neq x} p_i \right) \cdot \log \left(\frac{1}{\sum_{j \neq x} p_j} \right) \\
&= \gamma \cdot \log 1/\gamma \\
&\geq \gamma, && (11.17)
\end{aligned}
$$

as desired. In the above, (11.15) follows since all summands are non-negative, (11.16) follows since for every $i \neq x$, $p_i \leq \sum_{j \neq x} p_j$ and (11.17) follows since $\gamma \leq 1/2$ and so $\log 1/\gamma \geq 1$.

Now if $\gamma > 1/2$ we have a much simpler case since now we have

$$
\begin{aligned}
\alpha &\geq H(X) \\
&= \sum_i p_i \log \frac{1}{p_i} \\
&\geq \sum_i p_i \log \frac{1}{p_x} && (11.18) \\
&= \log \frac{1}{p_x} && (11.19) \\
&= \log \frac{1}{1-\gamma} \\
&\geq 1. && (11.20)
\end{aligned}
$$

(In the above, (11.18) follows since $p_i \leq p_x$, (11.19) follows since $\sum_i p_i = 1$ and (11.20) follows from the assumption that $\gamma \geq 1/2$.) But γ is always at most 1 so in this case also we have $\alpha \geq 1 \geq \gamma$ as desired.

We now consider the second part, which follows from the previous part via a simple averaging argument. Given y and i, let $p_{i,y} = \Pr_X[X = i | Y = y]$ and let $x_y = \operatorname{argmax}_i \{p_{i,y}\}$ be the value maximizing this probability. Let $\gamma_y = 1 - p_{x_y,y}$ and note that $\gamma = \mathbb{E}_Y[\gamma_Y]$. Let $\alpha_y = H(X|Y = y)$ and note again we have $\alpha = \mathbb{E}_Y[\alpha_Y]$. By the first part, we have for every y, $\gamma_y \leq \alpha_y$ and so it follows that

$$\gamma = \mathbb{E}_Y[\gamma_Y] \leq \mathbb{E}_Y[\alpha_Y] = \alpha.$$

11.6.2 Proof of Lemma 11.5.7

The lemma follows in a relatively straightforward manner with parts (3) and (4) using Lemma B.2.3.

Part (1) is immediate from the monotonicity of the entropy function in the interval $[0, 1/2]$ (see Exercise 11.6). For $0 \leq p_1, p_2 \leq 1/2$ we have $p_2 \leq p_1 \circ p_2 \leq 1/2$ and so (see Exercise 11.20)

$$H(p_1 \circ p_2) \geq H(p_2). \tag{11.21}$$

Next we turn to part (2). Let $H^{-1}(x) = p$ such that $0 \le p \le 1/2$ such that $H(p) = x$. Note H^{-1} is well defined and satisfies $H^{-1}(x) > 0$ if $x > 0$ and $H^{-1}(x) < 1/2$ if $x < 1$. Let

$$\alpha = \alpha(\tau) = H^{-1}(\tau)(1 - 2H^{-1}(1-\tau))$$

and

$$\beta = \beta(\tau) = 2H^{-1}(1-\tau)(1 - H^{-1}(1-\tau))$$

and

$$\gamma = \gamma(\tau) = \log((1-\beta)/\beta).$$

Note that $\alpha > 0$ and $\beta < 1/2$ and so $\gamma > 0$. We prove that $H(p_1 \circ p_2) - H(p_2) \ge \alpha \cdot \gamma$, and this will yield part (2) for $\theta = \theta(\tau) = \alpha \cdot \gamma > 0$.

First note that since $H(p_1), H(p_2) \in (\tau, 1-\tau)$, we have $p_1, p_2 \in (H^{-1}(\tau), H^{-1}(1-\tau))$. Thus

$$p_1 \circ p_2 - p_2 = p_2(1-2p_1) + p_1 - p_2 = p_1(1-2p_2) \ge H^{-1}(\tau)(1 - 2H^{-1}(1-\tau)) = \alpha.$$

Next we consider $\min_{p_2 \le q \le p_1 \circ p_2}\{H'(q)\}$. Note that by Exercise 11.21 $H'(q) = \log((1-q)/q)$ and this is minimized when q is maximum. The maximum value of $q = p_1 \circ p_2$ is in turn maximized by using the maximum values of $p_1, p_2 = H^{-1}(1-\tau)$. Thus we have that $\min_{p_2 \le q \le p_1 \circ p_2}\{H'(q)\} \ge H'(\beta) = \gamma$. By elementary calculus we now conclude that

$$H(p_1 \circ p_2) - h(p_2) \ge (p_1 \circ p_2 - p_2) \cdot \min_{p_2 \le q \le p_1 \circ p_2}\{H'(q)\} \ge \alpha \cdot \gamma = \theta.$$

This concludes the proof of part (2).

Next we move to part (3). For this we first describe some useful bounds on $H(p)$. On the one hand we have $H(p) \ge p \log 1/p$. For $p \le 1/2$ we also have $-(1-p)\log(1-p) \le (1/\ln 2)(1-p)(p + p^2) \le (1/\ln 2)p \le 2p$. And so we have $H(p) \le p(2 + \log 1/p)$.

Summarizing, we have for $p \le 1/2$,

$$p \log(1/p) \le H(p) \le p \log(1/p) + 2p. \tag{11.22}$$

We now consider $H(p_1) + H(p_2) - H(p_1 \circ p_2)$. We have

$$\begin{aligned}
&H(p_1) + H(p_2) - H(p_1 \circ p_2) \\
&\quad \le p_1(\log(1/p_1) + 2) + p_2(\log(1/p_2) + 2) - (p_1 \circ p_2)\log(1/(p_1 \circ p_2)) && (11.23)\\
&\quad \le p_1(\log(1/p_1) + 2) + p_2(\log(1/p_2) + 2) - (p_1 + p_2 - 2p_1p_2)\log(1/(2p_2)) && (11.24)\\
&\quad = p_1 \log(2p_2/p_1) + p_2 \log(2p_2/p_2) + 2p_1p_2 \log(1/(2p_2)) + 2(p_1 + p_2) \\
&\quad \le p_1 \log(p_2/p_1) + 2p_1p_2 \log(1/(p_2)) + 6p_2 && (11.25)\\
&\quad \le 2p_1 H(p_2) + 7p_2 && (11.26)\\
&\quad \le 2p_1 H(p_2) + 7H(p_2)/\log(1/p_2) && (11.27)\\
&\quad \le 9H(p_2)/\log(1/\tau). && (11.28)
\end{aligned}$$

In the above, (11.23) follows from (11.22), (11.24) follows from the fact that $p_1 \circ p_2 \le p_1 + p_2 \le 2p_2$, (11.25) follows from the fact that $3(p_1 + p_2) \le 6p_2$, (11.26) follows from the fact that $p_1 \log(p_2/p_1) \le$

p_2, (11.27) follows from (11.22) and (11.28) follows from the subsequent argument. Indeed by definition of τ and (11.22) we have $p_2 \log(1/p_2) \le \tau$. Using the fact that $p_2 \le 1/2$, this implies that $p_2 \le \tau$, which in turn implies $\log(1/p_2) \ge \log(1/\tau)$. Similarly, we have $p_1 \log(1/p_1) \le \tau$, which again with $p_1 \le 1/2$, we have $p_1 \le \tau \le 1/\log(1/\tau)$ (where the second equality uses the fact that $\tau < 1/2$). This concludes the proof of part (3).

Finally we turn to part (4). Here we let $H(p_i) = 1 - y_i$ and $p_i = 1/2 - x_i$ for $i \in \{1,2\}$. Since $\tau \le 1 - H(1/4)$ and $H(p_i) \ge 1 - \tau$, we have $x_i \le 1/4$. By Lemma B.2.4, we have $1 - 5x^2 \le H(1/2 - x) \le 1 - x^2$ for $x \le 1/4$. Returning to our setup if $1 - \tau \ge H(1/4)$ and $1 - y_i = H(p_i) \ge 1 - \tau$, and $p_i = 1/2 - x_i$, then $1 - x_i^2 \ge 1 - y_i$, so

$$x_i \le \sqrt{y_i}. \tag{11.29}$$

Furthermore, $p_1 \circ p_2 = 1/2 - 2x_1x_2$ and

$$\begin{aligned} H(p_1 \circ p_2) &= H(1/2 - 2x_1x_2) \\ &\ge 1 - 20(x_1x_2)^2 \\ &\ge 1 - 20y_1y_2 \\ &= 1 - 20(1 - H(p_1))(1 - H(p_2)) \\ &\ge 1 - 20\tau(1 - H(p_2)), \end{aligned}$$

where the second inequality follows from (11.29). This yields part (4).

11.7 Summary and additional information

In this chapter we showed how a very simple phenomenon leads to a very effective coding and decoding mechanism. Even the idea of reducing error-correction to compression is novel, though perhaps here the novelty is in the realization that this can idea can be put to good use. The idea of using polarization to create a compression scheme, as well as the exact procedure to create polarization are both radically novel, and remarkably effective.

Our description of this compression mechanism is nearly complete. The one omission is that we do not show which columns of the matrix $\mathbf{P}_n$ should be used to produce the compressed output — we only showed that a small subset exists. The reader should know that this aspect can also be achieved effectively, and this was first shown by Tal and Vardy [98], and adapted to the case of strong polarization by Guruswami and Xia. Specifically there is a polynomial time algorithm that given p, ε and c outputs $n \le \text{poly}(1/\varepsilon)$, $P_n \in \mathbb{F}_2^{n\times n}$ and a set $S \subseteq [n]$ such that P_n is (ε, n^{-c})-polarizing for $\text{Bern}(p)^n$ with unpredictable columns S, and $|S| \le (H(p) + \varepsilon)n$. The details are not very hard given the work so far, but still out of scope of this chapter.

Our analysis of local polarization differs from the literature in the absence of the use of "Mrs. Gerber's Lemma" due to Wyner and Ziv, which is a convexity claim that provides a convenient way to deal with conditional entropies (essentially implying that the conditioning can be ignored). In particular, it yields the following statement whose proof can be found as Lemma 2.2 in [21].

Lemma 11.7.1. *If (U, A) and (V, B) are independent and U, V are binary valued random variables with $H(U|A) = H(p)$ and $H(V|B) = H(q)$, then $H(U + V|A, B) \geq H(p(1-q) + q(1-p))$.*

The proof of the lemma uses the convexity of the function $H(a \circ H^{-1}(x))$ which turns out to have a short, but delicate and technical proof which led us to omit it here. This lemma would be a much cleaner bridge between the unconditioned polarization statement (Lemma 11.5.7) and its conditional variant (Lemma 11.5.8). Unfortunately Lemma 11.7.1 is known to be true only in the binary case whereas our proof method is applicable to larger alphabets (as shown by Guruswami and Velingker [49]).

11.8 Exercises

Exercise 11.1. Prove Theorem 11.1.2 (assuming Theorem 11.1.1).

Exercise 11.2. Argue that the matrices $\mathbf{G}$ and $\mathbf{G}^*$ in Proposition 11.2.1 exist.

Exercise 11.3. Show that there exists a non-linear comrepssion scheme for $\text{Bern}(p)^n$ of rate at most $H(p) + \varepsilon$.

Exercise 11.4. Prove (11.5).

Exercise 11.5. Prove (11.9).

Exercise 11.6. Show that $H(p)$ is monotonically increasing for $0 \leq p \leq \frac{1}{2}$.

Exercise 11.7. Give an explicit description of the polarizing matrix $\mathbf{P}_n$ such that $P_n(\mathbf{Z}) = \mathbf{Z} \cdot \mathbf{P}_n$. Further, prove that $\mathbf{P}_n$ is its own inverse.

Exercise 11.8. Show that

$$p^+(p_1, p_2) = p_1(1 - p_2) + (1 - p_1)p_2.$$

Exercise 11.9. Show that

$$p^|(p_1, p_2, 0) = p_1 p_2 / (p_1 p_2 + (1 - p_1)(1 - p_2))$$

and

$$p^|(p_1, p_2, 1) = (1 - p_1)p_2 / ((1 - p_1)p_2 + p_1(1 - p_2)).$$

Exercise 11.10. Prove Corollary 11.4.5.

Exercise 11.11. Prove (11.12).

Exercise 11.12. Prove Lemma 11.5.4.

Exercise 11.13. Prove Lemma 11.5.5.

Exercise 11.14. Prove Lemma 11.5.6.

Exercise 11.15. Prove part (2) of Lemma 11.5.8 for the case $H(p) \geq 1 - \tau$.

Exercise 11.16. Prove Lemma 11.5.9.

Exercise 11.17. Prove Lemma 11.5.10 for the case $X_j > 1/2$.

Exercise 11.18. Prove Lemma 11.5.12 when $X_0 \geq 1 - \lambda$.

Exercise 11.19. Prove Lemma 11.5.13 when $X_0 > \lambda$.

Exercise 11.20. Prove (11.21).

Exercise 11.21.

$$H'(q) = \log((1-q)/q).$$

11.9 Bibliographic Notes

Polar codes were invented in the remarkable paper by Arıkan [4] where he showed that they achieve capacity in the limit of large block lengths $n \to \infty$ with $O(n \log n)$ encoding time and $O(n \log n)$ decoding complexity via the successive cancellation decoder. In particular, Arıkan proved that the transform $P_2^{\otimes t}$ is polarizing in the limit of $t \to \infty$, in the sense that for any fixed $\gamma > 0$, the fraction of indices for which $H(\mathbf{W}_i \mid \mathbf{W}_{<i}) \in (\gamma, 1-\gamma)$, where $\mathbf{W} = P_2^{\otimes t}\mathbf{Z}$, is vanishing for large t. In fact, Arıkan showed that one could take $\gamma = \gamma(t) = 2^{-5t/4}$, which led to an upper bound of $n \cdot \gamma = O(1/n^{1/4})$ (block) decoding error probability for the successive cancellation decoder. Soon afterwards, Arıkan and Teletar proved that one can take $\gamma < 2^{-\Omega(2^{\beta t})}$ for any $\beta < 1/2$, which led to improved decoding error probability of $2^{-n^{-\beta}}$ as a function of the block length $n = 2^t$. The fall-off of the parameter γ in n was referred to as the "rate" of (limiting) polarization.

These works considered the basic 2×2 transform P_2 and binary codes. More general transforms, and non-binary codes, were considered later in [88, 64, 74]. These results showed that limiting polarization is universal, as long as some minimal conditions are met by the basic matrix being tensored.

The 2012 survey by Şaşoğlu is an excellent and highly recommended resource for some of the early works on polarization and polar codes [21]. Polar codes were widely described as the first constructive capacity achieving codes. Further, polarization was also found to be a versatile technique to asymptotically resolve several other fundamental problems in information theory such as lossless and lossy source coding problem, coding for broadcast, multiple access, and wiretap channels, etc.

However, none of these works yield effective finite length bounds on the block length n needed to achieve rates within ε of capacity, i.e., a rate at least $1 - h(p) - \varepsilon$ for the binary symmetric channel with crossover probability p. Without this it was not clear in what theoretical sense polar codes are better than say Forney's construction, which can also get within any desired $\varepsilon > 0$ of capacity, but have complexity growing exponentially in $1/\varepsilon^2$ due to the need for inner codes of length $1/\varepsilon^2$ that are decoded by brute-force.

A finite length analysis of polar codes, and *strong* polarization where the probability of not polarizing falls off exponentially in t, and thus is polynomially small in the block length $n = 2^t$, was established in independent works by Guruswami and Xia [50] and Hassani, Alishahi, and Urbanke [54]. The latter tracked channel "Bhattacharyya parameters" whereas the former tracked conditional entropies (as in the present chapter) which are a bit cleaner to deal with as they form a martingale. This form of fast polarization made polar codes the first, and so far only

known, family with block length and complexity scaling polynomially in $1/\varepsilon$ where ε is the gap to capacity,

This analyis of strong polarization in the above works applied only to the 2×2 transofrm and binary case. The strong polarization of the basic 2×2 transform was also established for all prime alphabets in [49], leading to the first construction of codes achieving the symmetric capacity of all discrete memoryless channels (for prime alphabets) with polynomial complexity in the gap to capacity. However, these analyses relied on rather specific inequalities (which were in particular somewhat painful to establish for the non-binary case) and it was not clear what exactly made them tick.

The recent work of the authors and Błasiok and Nakkiran [5] gave a modular and conceptually clear analysis of strong polarization by abstracting the properties needed from each local step to conclude fast global polarization. This made the demands on the local evolution of the conditional entropies rather minimal and qualitative, and enabled showing strong polarization and polynomially fast convergence to capacity for the *entire* class of polar codes, not just the binary 2×2 case. We followed this approach in this chapter, and in particular borrowed the concepts of variance in the middle and suction at the ends for local polarization from this work. However, we restrict attention to the basic 2×2 transform, and the binary symmetric channel, and gave elementary self-contained proofs of the necessary entropic inequalities needed to establish the properties required of the local polarization step.

Another difference in our presentation is that we described the successive cancellation decoder for the polarizing transform $P_2^{\otimes t}$, which leads to clean recursive description based on a more general primitive of decoding copies of independent but not necessary identical random variables. In contrast, in many works, including Arıkan's original paper [4], the decoding is described for the transform followed by the bit reversal permutation. The polarization property of the bit reversed transform is, however, notationally simpler to establish. Nevertheless, the transform $P_2^{\otimes t}$ commutes with the bit reversal permutation, so both the transforms, with or without bit reversal, end up having *identical* polarization properties.

Part IV

The Algorithms

Chapter 12

Decoding Concatenated Codes

In this chapter we study Question 10.3.1. Recall that the concatenated code $C_{\text{out}} \circ C_{\text{in}}$ consists of an outer $[N, K, D]_{Q=q^k}$ code C_{out} and an inner $[n, k, d]_q$ code C_{in}, where $Q = O(N)$. (Figure 12.1 illustrates the encoding function.) Then $C_{\text{out}} \circ C_{\text{in}}$ has design distance Dd and Question 10.3.1 asks if we can decode concatenated codes up to half the design distance (say for concatenated codes that we saw in Section 10.2 that lie on the Zyablov bound). In this chapter, we begin with a very natural unique decoding algorithm that can correct up to $Dd/4$ errors. Then we will consider a more sophisticated algorithm that will allow us to answer Question 10.3.1 in the affirmative.

12.1 A Natural Decoding Algorithm

We begin with a natural decoding algorithm for concatenated codes that “reverses" the encoding process (as illustrated in Figure 12.1). In particular, the algorithm first decodes the inner code and then decodes the outer code.

For the time being let us assume that we have a polynomial time unique decoding algorithm $D_{C_{\text{out}}} : [q^k]^N \to [q^k]^K$ for the outer code that can correct up to $D/2$ errors.

This leaves us with the task of coming up with a polynomial time decoding algorithm for the inner code. Our task of coming up with such a decoder is made easier by the fact that the running time needs to be polynomial in the *final* block length. This in turn implies that we would be fine if we pick a decoding algorithm that runs in singly exponential time in the inner block length as long as the inner block length is logarithmic in the outer code block length. (Recall that we put this fact to good use in Section 10.2 when we constructed explicit codes on the Zyablov bound.) Note that the latter is what we have assumed so far and thus, we can use the Maximum Likelihood Decoder (or MLD) (e.g. its implementation in Algorithm 2, which we will refer to as $D_{C_{\text{in}}}$). Algorithm 13 formalizes this algorithm.

It is easy to check that each step of Algorithm 13 can be implemented in polynomial time. In particular,

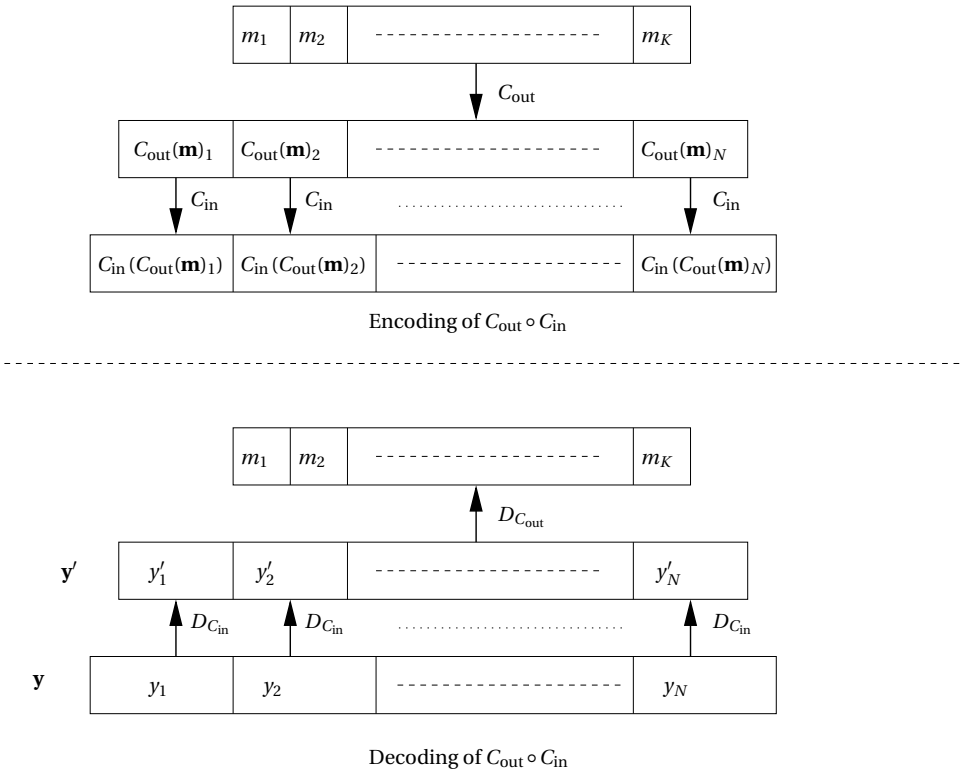


Figure 12.1: Encoding and Decoding of the concatenated code $C_{\text{out}} \circ C_{\text{in}}$. $D_{C_{\text{out}}}$ is a unique decoding algorithm for C_{out} and $D_{C_{\text{in}}}$ is a unique decoding algorithm for the inner code (e.g. MLD).

Algorithm 13 Natural Decoder for $C_{\text{out}} \circ C_{\text{in}}$

INPUT: Received word $\mathbf{y} = (y_1, \cdots, y_N) \in [q^n]^N$
OUTPUT: Message $\mathbf{m}' \in [q^k]^K$

1: $\mathbf{y}' \leftarrow (y'_1, \cdots, y'_N) \in [q^k]^N$ where

$$C_{\text{in}}(y'_i) = D_{C_{\text{in}}}(y_i) \quad 1 \le i \le N.$$

2: $\mathbf{m}' \leftarrow D_{C_{\text{out}}}(\mathbf{y}')$
3: RETURN $\mathbf{m}'$

1. The time complexity of Step 1 is $O(nq^k)$, which for our choice of $k = O(\log N)$ (and constant rate) for the inner code, is $(nN)^{O(1)}$ time.

2. Step 2 needs polynomial time by our assumption that the unique decoding algorithm $D_{C_{\text{out}}}$ takes $N^{O(1)}$ time.

Next, we analyze the error-correction capabilities of Algorithm 13:

Proposition 12.1.1. *Algorithm 13 can correct $< \frac{Dd}{4}$ many errors.*

Proof. Let $\mathbf{m}$ be the (unique) message such that $\Delta\left(C_{\text{out}} \circ C_{\text{in}}(\mathbf{m}), \mathbf{y}\right) < \frac{Dd}{4}$.

We begin the proof by defining a bad event as follows. We say a *bad event* has occurred (at position $1 \le i \le N$) if $y_i \neq C_{\text{in}}(C_{\text{out}}(\mathbf{m})_i)$. More precisely, define the set of all bad events to be

$$\mathscr{B} = \left\{i | y_i \neq C_{\text{in}}(C_{\text{out}}(\mathbf{m})_i)\right\}.$$

Note that if $|\mathscr{B}| < \frac{D}{2}$, then the decoder in Step 2 will output the message $\mathbf{m}$. Thus, to complete the proof, we only need to show that $|\mathscr{B}| < D/2$. To do this, we will define a superset $\mathscr{B}' \supseteq \mathscr{B}$ and then argue that $|\mathscr{B}'| < D/2$, which would complete the proof.

Note that if $\Delta\left(y_i, C_{\text{in}}(C_{\text{out}}(\mathbf{m})_i)\right) < \frac{d}{2}$ then $i \notin \mathscr{B}$ (by the proof of Proposition 1.4.1)– though the other direction does not hold. We define $\mathscr{B}'$ to be the set of indices where $i \in \mathscr{B}'$ if and only if

$$\Delta\left(y_i, C_{\text{in}}(C_{\text{out}}(\mathbf{m})_i)\right) \ge \frac{d}{2}.$$

Note that $\mathscr{B} \subseteq \mathscr{B}'$.

Now by definition, note that the total number of errors is at least $|\mathscr{B}'| \cdot \frac{d}{2}$. Thus, if $|\mathscr{B}'| \ge \frac{D}{2}$, then the total number of errors is at least $\frac{D}{2} \cdot \frac{d}{2} = \frac{Dd}{4}$, which is a contradiction. Thus, $|\mathscr{B}'| < \frac{D}{2}$, which completes the proof. □

Note that Algorithm 13 (as well the proof of Proposition 12.1.1) can be easily adapted to work for the case where the inner codes are different, e.g. Justesen codes (Section 10.3).

Thus, Proposition 12.1.1 and Theorem 12.3.3 imply that

Theorem 12.1.2. *There exist an explicit linear code on the Zyablov bound that can be decoded up to a fourth of the Zyablov bound in polynomial time.*

This of course is predicated on the fact that we need a polynomial time unique decoder for the outer code. Note that Theorem 12.1.2 implies the existence of an explicit asymptotically good code that can be decoded from a constant fraction of errors.

We now state two obvious open questions. The first is to get rid of the assumption on the existence of $D_{C_{\text{out}}}$:

Question 12.1.1. *Does there exist a polynomial time unique decoding algorithm for outer codes, e.g. for Reed-Solomon codes?*

Next, note that Proposition 12.1.1 does not quite answer Question 10.3.1. We move to answering this latter question next.

12.2 Decoding From Errors and Erasures

Now we digress a bit from answering Question 10.3.1 and talk about decoding Reed-Solomon codes. For the rest of the chapter, we will assume the following result.

Theorem 12.2.1. *An $[N,K]_q$ Reed-Solomon code can be corrected from e errors (or s erasures) as long as $e < \frac{N-K+1}{2}$ (or $s < N-K+1$) in $O(N^3)$ time.*

We defer the proof of the result on decoding from errors to Chapter 15 and leave the proof of the erasure decoder as an exercise. Next, we show that we can get the best of both worlds by correcting errors and erasures simultaneously:

Theorem 12.2.2. *An $[N,K]_q$ Reed-Solomon code can be corrected from e errors and s erasures in $O(N^3)$ time as long as*

$$2e + s < N - K + 1. \tag{12.1}$$

Proof. Given a received word $\mathbf{y} \in (\mathbb{F}_q \cup \{?\})^N$ with s erasures and e errors, let $\mathbf{y}'$ be the sub-vector with no erasures. This implies that $\mathbf{y}' \in \mathbb{F}_q^{N-s}$ is a valid received word for an $[N-s,K]_q$ Reed-Solomon code. (Note that this new Reed-Solomon code has evaluation points that corresponding to evaluation points of the original code, in the positions where an erasure did not occur.) Now run the error decoder algorithm from Theorem 12.2.1 on $\mathbf{y}'$. It can correct $\mathbf{y}'$ as long as

$$e < \frac{(N-s)-K+1}{2}.$$

This condition is implied by (12.1). Thus, we have proved one can correct e errors under (12.1). Now we have to prove that one can correct the s erasures under (12.1). Let $\mathbf{z}'$ be the output after correcting e errors. Now we extend $\mathbf{z}'$ to $\mathbf{z} \in (\mathbb{F}_q \cup \{?\})^N$ in the natural way. Finally, run the erasure decoding algorithm from Theorem 12.2.1 on $\mathbf{z}$. This works as long as $s < (N-K+1)$, which in turn is true by (12.1).

The time complexity of the above algorithm is $O(N^3)$ as both the algorithms from Theorem 12.2.1 can be implemented in cubic time. □

Next, we will use the above errors and erasure decoding algorithm to design decoding algorithms for certain concatenated codes that can be decoded up to half their design distance (i.e. up to $Dd/2$).

12.3 Generalized Minimum Distance Decoding

Recall the natural decoding algorithm for concatenated codes from Algorithm 13. In particular, we performed MLD on the inner code and then fed the resulting vector to a unique decoding algorithm for the outer code. A drawback of this algorithm is that it does not take into account the information that MLD provides. For example, it does not distinguish between the situations where a given inner code's received word has a Hamming distance of one vs where the received word has a Hamming distance of (almost) half the inner code distance from the closest codeword. It seems natural to make use of this information. Next, we study an algorithm called the Generalized Minimum Distance (or GMD) decoder, which precisely exploits this extra information.

In the rest of the section, we will assume C_{out} to be an $[N, K, D]_{q^k}$ code that can be decoded (by $D_{C_{\text{out}}}$) from e errors and s erasures in polynomial time as long as $2e + s < D$. Further, let C_{in} be an $[n, k, d]_q$ code with $k = O(\log N)$ which has a unique decoder $D_{C_{\text{in}}}$ (which we will assume is the MLD implementation from Algorithm 2).

We will in fact look at three versions of the GMD decoding algorithm. The first two will be randomized algorithms while the last will be a deterministic algorithm. We will begin with the first randomized version, which will present most of the ideas in the final algorithm.

12.3.1 GMD algorithm- I

Before we state the algorithm, let us look at two special cases of the problem to build some intuition.

Consider the received word $\mathbf{y} = (y_1, \ldots, y_N) \in [q^n]^N$ with the following special property: for every i such that $1 \le i \le N$, either $y_i = y_i'$ or $\Delta(y_i, y_i') \ge d/2$, where $y_i' = MLD_{C_{\text{in}}}(y_i)$. Now we claim that if $\Delta(\mathbf{y}, C_{\text{out}} \circ C_{\text{in}}) < dD/2$, then there are $< D$ positions in $\mathbf{y}$ such that $\Delta(y_i, C_{\text{in}}(y_i')) \ge d/2$ (we call such a position *bad*). This is because, for every bad position i, by the definition of y_i', $\Delta(y_i, C_{\text{in}}) \ge d/2$. Now if there are $\ge D$ bad positions, this implies that $\Delta(\mathbf{y}, C_{\text{out}} \circ C_{\text{in}}) \ge dD/2$, which is a contradiction. Now note that we can decode $\mathbf{y}$ by just declaring an erasure at every bad position and running the erasure decoding algorithm for C_{out} on the resulting vector.

Now consider the received word $\mathbf{y} = (y_1, \ldots, y_N)$ with the special property: for every i such that $i \in [N]$, $y_i \in C_{\text{in}}$. In other words, if there is an error at position $i \in [N]$, then a valid codeword in C_{in} gets mapped to another valid codeword $y_i \in C_{\text{in}}$. Note that this implies that a position with error has at least d errors. By a counting argument similar to the ones used in the previous paragraph, we have that there can be $< D/2$ such error positions. Note that we can now decode $\mathbf{y}$ by essentially running a unique decoder for C_{out} on $\mathbf{y}$ (or more precisely on $(x_1, \ldots, x_N)$, where $y_i = C_{\text{in}}(x_i)$).

Algorithm 14 generalizes these observations to decode arbitrary received words. In particular, it smoothly "interpolates" between the two extreme scenarios considered above.

Note that if $\mathbf{y}$ satisfies one of the two extreme scenarios considered earlier, then Algorithm 14 works exactly the same as discussed above.

By our choice of $D_{C_{\text{out}}}$ and $D_{C_{\text{in}}}$, it is easy to see that Algorithm 14 runs in polynomial time (in the final block length). More importantly, we will show that the final (deterministic) version

Algorithm 14 Generalized Minimum Decoder (ver 1)

INPUT: Received word $\mathbf{y} = (y_1, \cdots, y_N) \in [q^n]^N$
OUTPUT: Message $\mathbf{m}' \in [q^k]^K$

1: FOR $1 \le i \le N$ DO
2: $\quad y'_i \leftarrow D_{C_{\text{in}}}(y_i)$.
3: $\quad w_i \leftarrow \min\left(\Delta(y'_i, y_i), \frac{d}{2}\right)$.
4: $\quad$ With probability $\frac{2w_i}{d}$, set $y''_i \leftarrow ?$, otherwise set $y''_i \leftarrow x$, where $y'_i = C_{\text{in}}(x)$.
5: $\mathbf{m}' \leftarrow D_{C_{\text{out}}}(\mathbf{y}'')$, where $\mathbf{y}'' = (y''_1, \ldots, y''_N)$.
6: RETURN $\mathbf{m}'$

of Algorithm 14 can do unique decoding of $C_{\text{out}} \circ C_{\text{in}}$ up to half of its design distance.

As a first step, we will show that in expectation, Algorithm 14 works.

Lemma 12.3.1. *Let* $\mathbf{y}$ *be a received word such that there exists a codeword* $C_{\text{out}} \circ C_{\text{in}}(\mathbf{m}) = (c_1, \ldots, c_N) \in [q^n]^N$ *such that* $\Delta(C_{\text{out}} \circ C_{\text{in}}(\mathbf{m}), \mathbf{y}) < \frac{Dd}{2}$. *Further, if* $\mathbf{y}''$ *has* e' *errors and* s' *erasures (when compared with* $C_{\text{out}} \circ C_{\text{in}}(\mathbf{m})$*), then*

$$\mathbb{E}\left[2e' + s'\right] < D.$$

Note that if $2e' + s' < D$, then by Theorem 12.2.2, Algorithm 14 will output $\mathbf{m}$. The lemma above says that in expectation, this is indeed the case.

Proof of Lemma 12.3.1. For every $1 \le i \le N$, define $e_i = \Delta(y_i, c_i)$. Note that this implies that

$$\sum_{i=1}^{N} e_i < \frac{Dd}{2}. \tag{12.2}$$

Next for every $1 \le i \le N$, we define two indicator variables:

$$X_i^? = \mathbb{1}_{y''_i = ?},$$

and

$$X_i^e = \mathbb{1}_{C_{\text{in}}(y''_i) \ne c_i \text{ and } y''_i \ne ?}.$$

We claim that we are done if we can show that for every $1 \le i \le N$:

$$\mathbb{E}\left[2X_i^e + X_i^?\right] \le \frac{2e_i}{d}. \tag{12.3}$$

Indeed, by definition we have: $e' = \sum_i X_i^e$ and $s' = \sum_i X_i^?$. Further, by the linearity of expectation (Proposition 3.1.2), we get

$$\mathbb{E}\left[2e' + s'\right] \le \frac{2}{d}\sum_i e_i < D,$$

where the inequality follows from (12.2).

To complete the proof, we will prove (12.3) by a case analysis. Towards this end, fix an arbitrary $1 \le i \le N$.

Case 1: ($c_i = y'_i$) First, we note that if $y''_i \neq ?$ then since $c_i = y'_i$, we have $X_i^e = 0$. This along with the fact that $\Pr[y''_i = ?] = \frac{2w_i}{d}$ implies

$$\mathbb{E}\left[X_i^?\right] = \Pr[X_i^? = 1] = \frac{2w_i}{d},$$

and

$$\mathbb{E}\left[X_i^e\right] = \Pr[X_i^e = 1] = 0.$$

Further, by definition we have

$$w_i = \min\left(\Delta(y'_i, y_i), \frac{d}{2}\right) \le \Delta(y'_i, y_i) = \Delta(c_i, y_i) = e_i.$$

The three relations above prove (12.3) for this case.

Case 2: ($c_i \neq y'_i$) As in the previous case, we still have

$$\mathbb{E}\left[X_i^?\right] = \frac{2w_i}{d}.$$

Now in this case, if an erasure is not declared at position i, then $X_i^e = 1$. Thus, we have

$$\mathbb{E}\left[X_i^e\right] = \Pr[X_i^e = 1] = 1 - \frac{2w_i}{d}.$$

Next, we claim that as $c_i \neq y'_i$,

$$e_i + w_i \ge d, \tag{12.4}$$

which implies

$$\mathbb{E}\left[2X_i^e + X_i^?\right] = 2 - \frac{2w_i}{d} \le \frac{2e_i}{d},$$

as desired.

To complete the proof, we show (12.4) via yet another case analysis.

Case 2.1: ($w_i = \Delta(y'_i, y_i) < d/2$) By definition of e_i, we have

$$e_i + w_i = \Delta(y_i, c_i) + \Delta(y'_i, y_i) \ge \Delta(c_i, y'_i) \ge d,$$

where the first inequality follows from the triangle inequality and the second inequality follows from the fact that C_{in} has distance d.

Case 2.2: ($w_i = \frac{d}{2} \le \Delta(y'_i, y_i)$) As y'_i is obtained from MLD, we have

$$\Delta(y'_i, y_i) \le \Delta(c_i, y_i).$$

This along with the assumption on $\Delta(y'_i, y_i)$, we get

$$e_i = \Delta(c_i, y_i) \ge \Delta(y'_i, y_i) \ge \frac{d}{2}.$$

This in turn implies that

$$e_i + w_i \ge d,$$

as desired. □

12.3.2 GMD Algorithm- II

Note that Step 4 in Algorithm 14 uses “fresh" randomness for each i. Next we look at another randomized version of the GMD algorithm that uses the *same* randomness for every i. In particular, consider Algorithm 15.

Algorithm 15 Generalized Minimum Decoder (ver 2)

INPUT: Received word $\mathbf{y} = (y_1, \cdots, y_N) \in [q^n]^N$
OUTPUT: Message $\mathbf{m}' \in [q^k]^K$

1: Pick $\theta \in [0,1]$ uniformly at random.
2: FOR $1 \le i \le N$ DO
3: $\quad y'_i \leftarrow D_{C_{\text{in}}}(y_i)$.
4: $\quad w_i \leftarrow \min\left(\Delta(y'_i, y_i), \frac{d}{2}\right)$.
5: $\quad$ If $\theta < \frac{2w_i}{d}$, set $y''_i \leftarrow ?$, otherwise set $y''_i \leftarrow x$, where $y'_i = C_{\text{in}}(x)$.
6: $\mathbf{m}' \leftarrow D_{C_{\text{out}}}(\mathbf{y}'')$, where $\mathbf{y}'' = (y''_1, \ldots, y''_N)$.
7: RETURN $\mathbf{m}'$

We note that in the proof of Lemma 12.3.1, we only use the randomness to show that

$$\Pr\left[y''_i = ?\right] = \frac{2w_i}{d}.$$

In Algorithm 15, we note that

$$\Pr\left[y''_i = ?\right] = \Pr\left[\theta \in \left[0, \frac{2w_i}{d}\right)\right] = \frac{2w_i}{d},$$

as before (the last equality follows from our choice of θ). One can verify that the proof of Lemma 12.3.1 can be used to show the following lemma:

Lemma 12.3.2. *Let* $\mathbf{y}$ *be a received word such that there exists a codeword* $C_{\text{out}} \circ C_{\text{in}}(\mathbf{m}) = (c_1, \ldots, c_N) \in [q^n]^N$ *such that* $\Delta(C_{\text{out}} \circ C_{\text{in}}(\mathbf{m}), \mathbf{y}) < \frac{Dd}{2}$. *Further, if* $\mathbf{y}''$ *has* e' *errors and* s' *erasures (when compared with* $C_{\text{out}} \circ C_{\text{in}}(\mathbf{m})$*), then*

$$\mathbb{E}_\theta\left[2e' + s'\right] < D.$$

Next, we will see that Algorithm 15 can be easily “derandomized."

12.3.3 Derandomized GMD algorithm

Lemma 12.3.2 along with the probabilistic method shows that there exists a value $\theta^* \in [0,1]$ such that Algorithm 15 works correctly even if we fix θ to be θ^* in Step 1. Obviously we can obtain such a θ^* by doing an exhaustive search for θ. Unfortunately, there are uncountable choices of θ because $\theta \in [0,1]$. However, this problem can be taken care of by the following discretization trick.

Define $Q = \{0,1\} \cup \{\frac{2w_1}{d}, \cdots, \frac{2w_N}{d}\}$. Then because for each i, $w_i = \min(\Delta(y'_i, y_i), d/2)$, we have

$$Q = \{0,1\} \cup \{q_1, \cdots, q_m\}$$

where $q_1 < q_2 < \cdots < q_m$ for some $m \leq \left\lfloor \frac{d}{2} \right\rfloor$. Notice that for every $\theta \in [q_i, q_{i+1})$, just before Step 6, Algorithm 15 computes the same $\mathbf{y}''$. (See Figure 12.2 for an illustration as to why this is the case.)

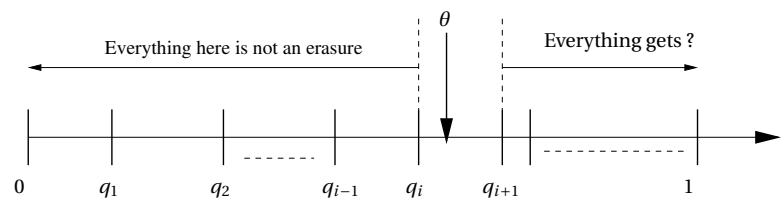


Figure 12.2: All values of $\theta \in [q_i, q_{i+1})$ lead to the same outcome

Thus, we need to cycle through all possible values of $\theta \in Q$, leading to Algorithm 16.

Algorithm 16 Deterministic Generalized Minimum Decoder‘

INPUT: Received word $\mathbf{y} = (y_1, \cdots, y_N) \in [q^n]^N$
OUTPUT: Message $\mathbf{m}' \in [q^k]^K$

1: $Q \leftarrow \{\frac{2w_1}{d}, \cdots, \frac{2w_N}{d}\} \cup \{0,1\}$.
2: FOR $\theta \in Q$ DO
3: FOR $1 \leq i \leq N$ DO
4: $y'_i \leftarrow D_{C_{\text{in}}}(y_i)$.
5: $w_i \leftarrow \min\left(\Delta(y'_i, y_i), \frac{d}{2}\right)$.
6: If $\theta < \frac{2w_i}{d}$, set $y''_i \leftarrow ?$, otherwise set $y''_i \leftarrow x$, where $y'_i = C_{\text{in}}(x)$.
7: $\mathbf{m}'_\theta \leftarrow D_{C_{\text{out}}}(\mathbf{y}'')$, where $\mathbf{y}'' = (y''_1, \ldots, y''_N)$.
8: RETURN $\mathbf{m}'_{\theta^*}$ for $\theta^* = \arg\min_{\theta \in Q} \Delta\left(C_{\text{out}} \circ C_{\text{in}}\left(\mathbf{m}'_\theta\right), \mathbf{y}\right)$

Note that Algorithm 16 is Algorithm 15 repeated $|Q|$ times. Since $|Q|$ is $O(n)$, this implies that Algorithm 16 runs in polynomial time. This along with Theorem 10.2.1 implies that

Theorem 12.3.3. *For every constant rate, there exists an explicit linear binary code on the Zyablov bound. Further, the code can be decoded up to half of the Zyablov bound in polynomial time.*

Note that the above answers Question 10.3.1 in the affirmative.

12.4 Bibliographic Notes

Forney in 1966 designed the Generalized Minimum Distance (or GMD) decoding [31].

Chapter 13

Efficiently Achieving the Capacity of the $\mathrm{BSC_p}$

Table 13.1 summarizes the main results we have seen so far for (binary codes).

	Shannon	Hamming	
		Unique Decoding	List Decoding
Capacity	$1-H(p)$ (Thm 6.3.1)	$\geq$ GV (Thm 4.2.1) $\leq$ MRRW (Sec 8.2)	$1-H(p)$ (Thm 7.4.1)
Explicit Codes	?	Zyablov bound (Thm 10.2.1)	?
Efficient Algorithms	?	$\frac{1}{2}\cdot$ Zyablov bound (Thm 12.3.3)	?

Table 13.1: An overview of the results seen so far

In this chapter, we will tackle the open questions in the first column of Table 13.1. Recall that there exist linear codes of rate $1-H(p)-\varepsilon$ such that decoding error probability is not more than $2^{-\delta n}$, $\delta = \Theta(\varepsilon^2)$ on the BSC_p (Theorem 6.3.1 and Exercise 6.3). This led to Question 6.3.1, which asks if we can achieve the BSC_p capacity with explicit codes and efficient decoding algorithms?

13.1 Achieving capacity of BSC_p

We will answer Question 6.3.1 in the affirmative by using concatenated codes. The main intuition in using concatenated codes is the following. As in the case of construction of codes on the Zyablov bound, we will pick the inner code to have the property that we are after: i.e. a code that achieves the BSC_p capacity. (We will again exploit the fact that since the block length of the inner code is small, we can construct such a code in a brute-force manner.) However, unlike the case of the Zyablov bound construction, we do not know of an explicit code that is optimal over say the $q\mathrm{SC}_p$ channel. The main observation here is that the fact that the BSC_p noise is memory-less can be exploited to pick the outer code that can correct from some small but constant fraction of *worst-case* errors.

Before delving into the details, we present the main ideas. We will use an outer code C_{out} that has rate close to 1 and can correct from some fixed constant (say γ) fraction of worst-case errors. We pick an inner code C_{in} that achieves the BSC_p capacity with parameters as guaranteed by Theorem 6.3.1. Since the outer code has rate almost 1, the concatenated code can be made to have the required rate (since the final rate is the product of the rates of C_{out} and C_{in}). For decoding, we use the natural decoding algorithm for concatenated codes from Algorithm 13. Assume that each of the inner decoders has a decoding error probability of (about) γ. Then the intermediate received word $\mathbf{y}'$ has an expected γ fraction of errors (with respect to the outer codeword of the transmitted message), though we might not have control over where the errors occur. However, we picked C_{out} so that it can correct up to γ fraction of worst-case errors. This shows that everything works in expectation. To make everything work with high probability (i.e. achieve exponentially small overall decoding error probability), we make use of the fact that since the noise in BSC_p is independent, the decoding error probabilities of each of the inner decodings is independent and thus, by the Chernoff bound (Theorem 3.1.6), with all but an exponentially small probability $\mathbf{y}'$ has $\Theta(\gamma)$ fraction of errors, which we correct with the worst-case error decoder for C_{out}. See Figure 13.1 for an illustration of the main ideas. Next, we present the details.

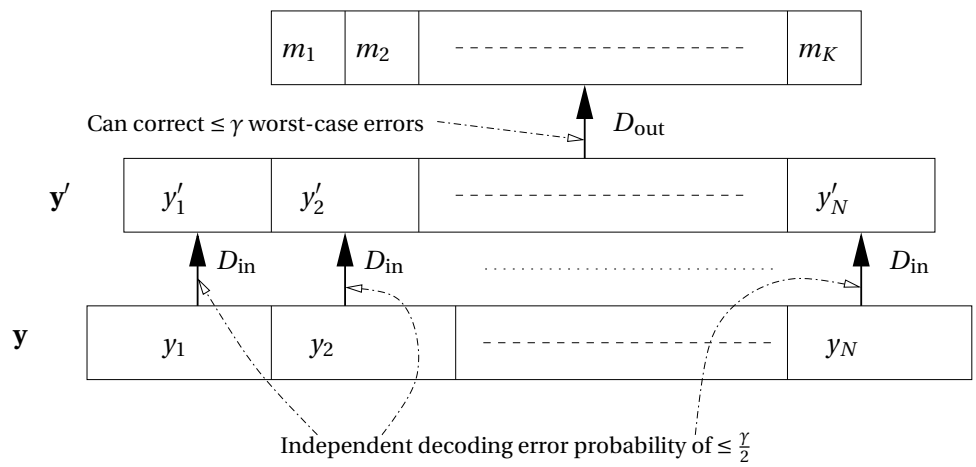


Figure 13.1: Efficiently achieving capacity of BSC_p.

We answer Question 6.3.1 in the affirmative by using a concatenated code $C_{\text{out}} \circ C_{\text{in}}$ with the following properties (where $\gamma > 0$ is a parameter that depends only on ε and will be fixed later on):

(i) C_{out}: The outer code is a linear $[N, K]_{2^k}$ code with rate $R \geq 1 - \frac{\varepsilon}{2}$, where $k = O(\log N)$. Further, the outer code has a unique decoding algorithm D_{out} that can correct at most γ fraction of worst-case errors in time $T_{\text{out}}(N)$.

(ii) C_{in}: The inner code is a linear binary $[n, k]_2$ code with a rate of $r \geq 1 - H(p) - \varepsilon/2$. Further, there is a decoding algorithm D_{in} (which returns the transmitted codeword) that runs in time $T_{\text{in}}(k)$ and has decoding error probability no more than $\frac{\gamma}{2}$ over BSC_p.

Table 13.2 summarizes the different parameters of C_{out} and C_{in}.

	Dimension	Block length	q	Rate	Decoder	Decoding time	Decoding guarantee
C_{out}	K	N	2^k	$1-\frac{\varepsilon}{2}$	D_{out}	$T_{\text{out}}(N)$	$\leq \gamma$ fraction of worst-case errors
C_{in}	$k \leq O(\log N)$	n	2	$1-H(p)-\frac{\varepsilon}{2}$	D_{in}	$T_{\text{in}}(k)$	$\leq \frac{\gamma}{2}$ decoding error probability over BSC_p

Table 13.2: Summary of properties of C_{out} and C_{in}

Suppose $C^* = C_{\text{out}} \circ C_{\text{in}}$. Then, it is easy to check that

$$R(C^*) = R \cdot r \geq \left(1-\frac{\varepsilon}{2}\right) \cdot \left(1-H(p)-\frac{\varepsilon}{2}\right) \geq 1-H(p)-\varepsilon,$$

as desired.

For the rest of the chapter, we will assume that p is an absolute constant. Note that this implies that $k = \Theta(n)$ and thus, we will use k and n interchangeably in our asymptotic bounds. Finally, we will use $\mathcal{N} = nN$ to denote the block length of C^*.

The decoding algorithm for C^* that we will use is Algorithm 13, which for concreteness we reproduce as Algorithm 17.

Algorithm 17 Decoder for efficiently achieving BSC_p capacity

INPUT: Received word $\mathbf{y} = (y_1, \cdots, y_N) \in [q^n]^N$
OUTPUT: Message $\mathbf{m}' \in [q^k]^K$

1: $\mathbf{y}' \leftarrow (y'_1, \cdots, y'_N) \in [q^k]^N$ where

$$C_{\text{in}}(y'_i) = D_{\text{in}}(y_i) \quad 1 \leq i \leq N.$$

2: $\mathbf{m}' \leftarrow D_{\text{out}}(\mathbf{y}')$
3: RETURN $\mathbf{m}'$

Note that encoding C^* takes time

$$O(N^2k^2) + O(Nkn) \leq O(N^2n^2) = O(\mathcal{N}^2),$$

as both the outer and inner codes are linear[1]. Further, the decoding by Algorithm 17 takes time

$$N \cdot T_{\text{in}}(k) + T_{\text{out}}(N) \leq \text{poly}(N),$$

[1]Note that encoding the outer code takes $O(N^2)$ operations over $\mathbb{F}_{q^k}$. The term $O(N^2k^2)$ then follows from the fact that each operation over $\mathbb{F}_{q^k}$ can be implemented with $O(k^2)$ operations over $\mathbb{F}_q$.

where the inequality is true as long as

$$T_{\text{out}}(N) = N^{O(1)} \text{ and } T_{\text{in}}(k) = 2^{O(k)}. \tag{13.1}$$

Next, we will show that decoding via Algorithm 17 leads to an exponentially small decoding error probability over BSC_p. Further, we will use constructions that we have already seen in this book to instantiate C_{out} and C_{in} with the required properties.

13.2 Decoding Error Probability

We begin by analyzing Algorithm 17.

By the properties of D_{in}, for any fixed i, there is an error at y'_i with probability $\leq \frac{\gamma}{2}$. Each such error is independent, since errors in BSC_p itself are independent by definition. Because of this, and by linearity of expectation, the expected number of errors in $\mathbf{y}'$ is $\leq \frac{\gamma N}{2}$.

Taken together, those two facts allow us to conclude that, by the (multiplicative) Chernoff bound (Theorem 3.1.6), the probability that the total number of errors will be more than γN is at most $e^{-\frac{\gamma N}{6}}$. Since the decoder D_{out} fails only when there are more than γN errors, this is also the final decoding error probability. Expressed in asymptotic terms, the error probability is $2^{-\Omega(\frac{\gamma \mathcal{N}}{n})}$.

13.3 The Inner Code

We find C_{in} with the required properties by an exhaustive search among linear codes of dimension k with block length n that achieve the BSC_p capacity by Shannon's theorem (Theorem 6.3.1). Recall that for such codes with rate $1 - H(p) - \frac{\varepsilon}{2}$, the MLD has a decoding error probability of $2^{-\Theta(\varepsilon^2 n)}$ (Exercise 6.3). Thus, if k is at least $\Omega\left(\frac{\log(\frac{1}{\gamma})}{\varepsilon^2}\right)$, Exercise 6.3 implies the existence of a linear code with decoding error probability at most $\frac{\gamma}{2}$ (which is what we need). Thus, with the restriction on k from the outer code, we have the following restriction on k:

$$\Omega\left(\frac{\log(\frac{1}{\gamma})}{\varepsilon^2}\right) \leq k \leq O\left(\log N\right).$$

Note, however, that since the proof of Theorem 6.3.1 uses MLD on the inner code and Algorithm 2 is the only known implementation of MLD, we have $T_{\text{in}} = 2^{O(k)}$ (which is what we needed in (13.1). The construction time is even worse. There are $2^{O(kn)}$ generator matrices; for each of these, we must check the error rate for each of 2^k possible transmitted codewords, and for each codeword, computing the decoding error probability requires time $2^{O(n)}$.[2] Thus, the construction time for C_{in} is $2^{O(n^2)}$.

[2] To see why the latter claim is true, note that there are 2^n possible received words and given any one of these received words, one can determine (i) if the MLD produces a decoding error in time $2^{O(k)}$ and (ii) the probability that the received word can be realized, given the transmitted codeword in polynomial time.

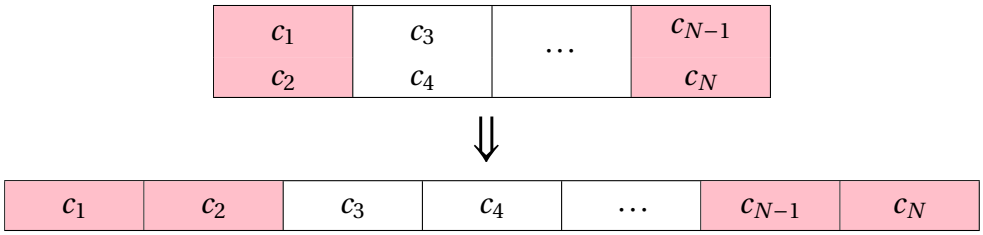


Figure 13.2: Error Correction cannot decrease during "folding." The example has $k = 2$ and a pink cell implies an error.

13.4 The Outer Code

We need an outer code with the required properties. There are several ways to do this.

One option is to set C_{out} to be a Reed-Solomon code over $\mathbb{F}_{2^k}$ with $k = \Theta(\log N)$ and rate $1-\frac{\varepsilon}{2}$. Then the decoding algorithm D_{out}, could be the error decoding algorithm from Theorem 12.2.2. Note that for this D_{out} we can set $\gamma = \frac{\varepsilon}{4}$ and the decoding time is $T_{\text{out}}(N) = O(N^3)$.

Till now everything looks on track. However, the problem is the construction time for C_{in}, which as we saw earlier is $2^{O(n^2)}$. Our choice of n implies that the construction time is $2^{O(\log^2 N)} \leq N^{O(\log N)}$, which of course is not polynomial time. Thus, the trick is to find a C_{out} defined over a smaller alphabet (certainly no larger than $2^{O(\sqrt{\log N})}$). This is what we do next.

13.4.1 Using a binary code as the outer code

The main observation is that we can also use an outer code which is some explicit binary linear code (call it C') that lies on the Zyablov bound and can be corrected from errors up to half its design distance. We have seen that such a code can be constructed in polynomial time (Theorem 12.3.3).

Note that even though C' is a binary code, we can think of C' as a code over $\mathbb{F}_{2^k}$ in the obvious way: every k consecutive bits are considered to be an element in $\mathbb{F}_{2^k}$ (say via a linear map). Note that the rate of the code does not change. Further, any decoder for C' that corrects bit errors can be used to correct errors over $\mathbb{F}_{2^k}$. In particular, if the algorithm can correct β fraction of bit errors, then it can correct that same fraction of errors over $\mathbb{F}_{2^k}$. To see this, think of the received word as $\mathbf{y} \in (\mathbb{F}_{2^k})^{N'/k}$, where N' is the block length of C' (as a binary code), which is at a fractional Hamming distance at most ρ away from $\mathbf{c} \in (\mathbb{F}_{2^k})^{N'/k}$. Here $\mathbf{c}$ is what once gets by "folding" consecutive k bits into one symbol in some codeword $\mathbf{c}' \in C'$. Now consider $\mathbf{y}' \in \mathbb{F}_2^{N'}$, which is just "unfolded" version of $\mathbf{y}$. Now note that each symbol in $\mathbf{y}$ that is in error (w.r.t. $\mathbf{c}$) leads to at most k bit errors in $\mathbf{y}'$ (w.r.t. $\mathbf{c}'$). Thus, in the unfolded version, the total number of errors is at most

$$k \cdot \rho \cdot \frac{N'}{k} = \rho \cdot N'.$$

(See Figure 13.2 for an illustration for $k = 2$.) Thus to decode $\mathbf{y}$, one can just "unfold" $\mathbf{y}$ to $\mathbf{y}'$ and use the decoding algorithm for C' (which can handle ρ fraction of errors) on $\mathbf{y}'$.

We will pick C_{out} to be C' when considered over $\mathbb{F}_{2^k}$, where we choose

$$k = \Theta\left(\frac{\log(\frac{1}{\gamma})}{\varepsilon^2}\right).$$

Further, D_{out} is the GMD decoding algorithm (Algorithm 16) for C'.

Now, to complete the specification of C^*, we relate γ to ε. The Zyablov bound gives $\delta_{\text{out}} = (1-R)H^{-1}(1-r)$, where R and r are the rates of the outer and inners codes for C'. Now we can set $1-R = 2\sqrt{\gamma}$ (which implies that $R = 1-2\sqrt{\gamma}$) and $H^{-1}(1-r) = \sqrt{\gamma}$, which implies that r is[3] $1-O\left(\sqrt{\gamma}\log\frac{1}{\gamma}\right)$. Since we picked D_{out} to be the GMD decoding algorithm, it can correct $\frac{\delta_{\text{out}}}{2} = \gamma$ fraction of errors in polynomial time, as desired.

The overall rate of C_{out} is simply $R \cdot r = \left(1-2\sqrt{\gamma}\right)\cdot\left(1-O\left(\sqrt{\gamma}\log\frac{1}{\gamma}\right)\right)$. This simplifies to $1-O\left(\sqrt{\gamma}\log\left(\frac{1}{\gamma}\right)\right)$. Recall that we need this to be at least $1-\frac{\varepsilon}{2}$. Thus, we would be done here if we could show that ε is $\Omega\left(\sqrt{\gamma}\log\frac{1}{\gamma}\right)$, which would follow by setting

$$\gamma = \varepsilon^3.$$

13.4.2 Wrapping Up

We now recall the construction, encoding and decoding time complexity for our construction of C^*. The construction time for C_{in} is $2^{O(n^2)}$, which substituting for n, is $2^{O\left(\frac{1}{\varepsilon^4}\log^2\left(\frac{1}{\varepsilon}\right)\right)}$. The construction time for C_{out}, meanwhile, is only $\text{poly}(N)$. Thus, our overall, construction time is $\text{poly}(\mathcal{N}) + 2^{O\left(\frac{1}{\varepsilon^4}\log^2\left(\frac{1}{\varepsilon}\right)\right)}$.

As we have seen in Section 13.1, the encoding time for this code is $O(\mathcal{N}^2)$, and the decoding time is $N^{O(1)} + N\cdot 2^{O(n)} = \text{poly}(\mathcal{N}) + \mathcal{N}\cdot 2^{O\left(\frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right)\right)}$. We also have shown that the decoding error probability is exponentially small: $2^{-\Omega(\frac{\gamma\mathcal{N}}{n})} = 2^{-\Omega(\varepsilon^6\mathcal{N})}$. Thus, we have proved the following result:

Theorem 13.4.1. *For every constant p and $0 < \varepsilon < 1-H(p)$, there exists a linear code C^* of block length $\mathcal{N}$ and rate at least $1-H(p)-\varepsilon$, such that*

(a) C^* can be constructed in time $\text{poly}(\mathcal{N}) + 2^{O(\varepsilon^{-5})}$;

(b) C^* can be encoded in time $O(\mathcal{N}^2)$; and

(c) There exists a $\text{poly}(\mathcal{N}) + \mathcal{N}\cdot 2^{O(\varepsilon^{-5})}$ time decoding algorithm that has an error probability of at most $2^{-\Omega(\varepsilon^6\mathcal{N})}$ over the BSC_p.

[3] Note that $r = 1-H(\sqrt{\gamma}) = 1+\sqrt{\gamma}\log\sqrt{\gamma} + (1-\sqrt{\gamma})\log(1-\sqrt{\gamma})$. Noting that $\log(1-\sqrt{\gamma}) = -\sqrt{\gamma}-\Theta(\gamma)$, we can deduce that $r = 1-O(\sqrt{\gamma}\log(1/\gamma))$.

Thus, we have answered in the affirmative Question 6.3.1, which was the central open question from Shannon's work. However, there is a still somewhat unsatisfactory aspect of the result above. In particular, the exponential dependence on $1/\varepsilon$ in the decoding time complexity is not nice. This leads to the following question:

Question 13.4.1. *Can we bring the high dependence on ε down to* $\mathrm{poly}\left(\frac{1}{\varepsilon}\right)$ *in the decoding time complexity?*

13.5 Discussion and Bibliographic Notes

Forney answered Question 6.3.1 in the affirmative by using concatenated codes. (As was mentioned earlier, this was Forney's motivation for inventing code concatenation: the implication for the rate vs. distance question was studied by Zyablov later on.)

We now discuss Question 13.4.1. For the binary erasure channel, the decoding time complexity can be brought down to $\mathcal{N} \cdot \mathrm{poly}(\frac{1}{\varepsilon})$ using LDPC codes, specifically a class known as Tornado codes developed by Luby et al. [71]. The question for binary symmetric channels, however, is still open. Recently there have been some exciting progress on this front by the construction of the so-called Polar codes.

We conclude by noting an improvement to Theorem 13.4.1. We begin with a theorem due to Spielman:

Theorem 13.5.1 ([94])**.** *For every small enough $\beta > 0$, there exists an explicit C_{out} of rate $\frac{1}{1+\beta}$ and block length N, which can correct $\Omega\left(\frac{\beta^2}{(\log \frac{1}{\beta})^2}\right)$ errors, and has $O(N)$ encoding and decoding.*

Clearly, in terms of time complexity, this is superior to the previous option in Section 13.4.1. Such codes are called "Expander codes." One can essentially do the same calculations as in Section 13.4.1 with $\gamma = \Theta\left(\frac{\varepsilon^2}{\log^2(1/\varepsilon)}\right)$.[4] However, we obtain an encoding and decoding time of $\mathcal{N} \cdot 2^{\mathrm{poly}(\frac{1}{\varepsilon})}$. Thus, even though we obtain an improvement in the time complexities as compared to Theorem 13.4.1, this does not answer Question 13.4.1.

[4] This is because we need $1/(1+\beta) = 1 - \varepsilon/2$, which implies that $\beta = \Theta(\varepsilon)$.

Chapter 14

Decoding Reed-Muller Codes

In this chapter we describe decoding algorithms for the Reed-Muller codes, introduced in Chapter 9. Recall that these are the codes obtained by evaluations of multivariate polynomials over all possible assignments to the variables. We will see several decoding algorithms for these codes, ranging from simplistic ones that correct a constant fraction of the minimum distance (with the constant depending on q), to algorithms based on more sophisticated concepts that correct up to half the minimum distance.

To elaborate on the above, recall that the Reed-Muller code with parameters q, m, r is the set of functions

$$\mathrm{RM}(q,m,r) \stackrel{\text{def}}{=} \left\{ f : \mathbb{F}_q^m \to \mathbb{F}_q \,|\, \deg(f) \le r \right\}.$$

The minimum distance distance of the code is

$$\Delta_{q,m,r} \stackrel{\text{def}}{=} (q-t) \cdot q^{m-s-1},$$

where s, t satisfy $r = s(q-1) + t$ and $0 \le t \le q-2$ (recall Lemma 9.4.1). We will first describe an algorithm to correct $\varepsilon \cdot \Delta_{q,m,r}$ for some constant $\varepsilon > 0$ that depends only on q. Later we will give algorithms that correct $\left\lfloor \frac{\Delta_{q,m,r}-1}{2} \right\rfloor$ errors.

14.1 A natural decoding algorithm

The main insight behind our first decoding algorithm is the simple fact that the degree of polynomials does not increase on *affine substitutions*. Let us introduce this notion and then explain why this might be useful in building decoding algorithms.

Definition 14.1.1. A one-dimensional s-variate *affine form* $a \in \mathbb{F}_q[Z_1, \ldots, Z_s]$ is a polynomial of the form $a(Z_1, \ldots, Z_s) = \sum_{i=1}^{s} \alpha_i Z_i + \alpha_0$. In other words an affine form is a polynomial of degree at most 1. An m-dimensional s-variate *affine form* $A = \langle a_1, \ldots, a_m \rangle$ is simply an m-tuple of one-dimensional affine forms.

For example $A_0 = \langle Z_1 + Z_2, Z_1, Z_2 \rangle$ is a 3-dimensional 2-variate affine form over $\mathbb{F}_2$.

Definition 14.1.2. Given an m-variate polynomial $P \in \mathbb{F}_q[X_1, \ldots, X_m]$ and an m-dimensional s-variate affine form $A = \langle a_1, \ldots, a_m \rangle \in (\mathbb{F}_q[Z])^m$ where $Z = (Z_1, \ldots, Z_s)$, the affine substitution of A into P is given by the polynomial $P \circ A \in \mathbb{F}_q[Z]$ given by $(P \circ A)(Z) = P(a_1(Z), \ldots, a_m(Z))$.

Let A_0 be the affine form as above and let $P_0(X_1, X_2, X_3) = X_1X_2 + X_1X_2X_3 + X_3$ over $\mathbb{F}_2$. Then we have

$$(P_0 \circ A_0)(Z_1, Z_2) = (Z_1 + Z_2)Z_1 + (Z_1 + Z_2)Z_1Z_2 + Z_2 = Z_1^2 + Z_1Z_2 + Z_1^2Z_2 + Z_1Z_2^2 + Z_2 = Z_1 + Z_1Z_2 + Z_2,$$

where the last equality follows since we are working over $\mathbb{F}_2$.

Notice that the notion of affine substitutions extends to functions naturally, viewing both f and A as functions (given by the evaluations of corresponding polynomials) in the definition above.

Affine substitutions have nice algebraic, geometric, and probabilistic properties and these combine to give us the decoding algorithm of this section. We introduce these properties in order.

Proposition 14.1.1 (Degree of Affine Substitutions)**.** *Affine substitutions do not increase the degree of a polynomial. Specifically, if A is an affine form, then for every polynomial P, we have* $\deg(P \circ A) \leq \deg(P)$.

Proof. The proof is straightforward. First note that for any single monomial $M = \prod_{i=1}^m X_i^{r_i}$ the affine substitution $M \circ A = \prod_{i=1}^m a_i(Z)^{r_i}$ has degree at most $\deg(M)$. Next note that if we write a general polynomial as a sum of monomials, say $P = \sum_M c_M \cdot M$, then the affine substitution is additive and so $P \circ A = \sum_M c_M(M \circ A)$. The proposition now follows from the fact that

$$\deg(P \circ A) = \deg(\sum_M c_M(M \circ A)) \leq \max_M\{\deg(M \circ A)\} \leq \max_M \deg(M)\} = \deg(P).$$

□

We remark that the bound above can be tight (see Exercise 14.1) and that the result above generalizes to the case when we replace each term by a degree d-polynomial instead of a degree 1-polynomial (see Exercise 14.2).

Next we turn to the geometric aspects of affine substitutions. These aspects will be essential for some intuition, though we will rarely invoke them formally.

One way to view affine substitutions into functions is that we are viewing the restriction of a function on a small subset of the domain. For example, when $s = 1$, then an affine substitution A into a function f, restricts the domain of the function to the set $\{A(z) | z \in \mathbb{F}_q\}$ where $A(z)$ is of the form $\mathbf{a}z + \mathbf{b}$ for some $\mathbf{a}, \mathbf{b} \in \mathbb{F}_q^m$. This set forms a *line* in $\mathbb{F}_q^m$ with slope $\mathbf{a}$ and passing through the point $\mathbf{b}$. When s becomes larger, we look at higher dimensional (affine) subspaces such as planes ($s = 2$) and cubes ($s = 3$). While lines, planes and cubes are not exactly the same as in the Euclidean space they satisfy many similar properties and this will be used to drive some of the probabilistic thinking below.

In what follows we will be looking restrictions of two functions f and g on small-dimensional affine subspaces. On these subspaces we would like to argue that f and g disagree roughly as

often as they do on $\mathbb{F}_q^m$. To do so, we use the fact that "random" affine substitutions sample uniformly from $\mathbb{F}_q^m$. We formalize this below.

Consider a uniform choice of an affine form $A(\mathbf{z}) = \mathbf{M}\mathbf{z} + \mathbf{b}$, i.e., where $\mathbf{M} \in \mathbb{F}_q^{m\times s}$ and $\mathbf{b} \in \mathbb{F}_q^m$ are chosen uniformly and independently from their respective domains. (Note that this allows M to be of less than full rank with positive probability, and we will allow this to keep calculations simple and clean. We do warn the reader that this can lead to degenerate lines and subspaces - e.g., when M is the zero matrix then these subspaces contain only one point.)

Proposition 14.1.2. *(1) Fix $\mathbf{z} \in \mathbb{F}_q^s$. Then, for a uniformly random A, the point $A(\mathbf{z})$ is distributed uniformly in $\mathbb{F}_q^m$.*

(2) Fix $\mathbf{z} \in \mathbb{F}_q^s \setminus \{0\}$ and $\mathbf{x} \in \mathbb{F}_q^m$. and let A be chosen uniformly subject to the condition $A(\mathbf{0}) = \mathbf{x}$. Then the point $A(\mathbf{z})$ is distributed uniformly in $\mathbb{F}_q^m$. Consequently, for every pair of functions $f, g : \mathbb{F}_q^m \to \mathbb{F}_q$, we have

$$\Pr_A\left[f \circ A(\mathbf{z}) \neq g \circ A(\mathbf{z})\right] = \delta(f, g).$$

Proof. Let $A(\mathbf{z}) = \mathbf{M}\mathbf{z} + \mathbf{b}$ where $\mathbf{M} \in \mathbb{F}_q^{m\times s}$ and $\mathbf{b} \in \mathbb{F}_q^m$ are chosen uniformly and independently.

For part (1), we use the fact that for every fixed $\mathbf{M}$ and $\mathbf{z}$, $\mathbf{M}\mathbf{z} + \mathbf{b}$ is uniform over $\mathbb{F}_q^m$ when $\mathbf{b}$ is uniform. In particular, for every $\mathbf{y} \in \mathbb{F}_q^m$ we have

$$\Pr_{\mathbf{b}}\left[\mathbf{M}\mathbf{z} + \mathbf{b} = \mathbf{y}\right] = \Pr_{\mathbf{b}}\left[\mathbf{b} = \mathbf{y} - \mathbf{M}\mathbf{z}\right] = q^{-m}.$$

Since this holds for every $\mathbf{M}$, it follows that

$$\Pr_{\mathbf{M},\mathbf{b}}\left[\mathbf{M}\mathbf{z} + \mathbf{b} = \mathbf{y}\right] = q^{-m}$$

and so we conclude that $A(\mathbf{z}) = \mathbf{M}\mathbf{z} + \mathbf{y}$ is uniformly distributed over $\mathbb{F}_q^m$.

For part (2), note that the condition $A(\mathbf{0}) = \mathbf{x}$ implies $\mathbf{b} = \mathbf{x}$. So, for fixed $\mathbf{y} \in \mathbb{F}_q^m$, we have

$$\Pr_{\mathbf{M},\mathbf{b}}\left[A(\mathbf{z}) = \mathbf{y} \mid A(\mathbf{0}) = \mathbf{x}\right] = \Pr_{\mathbf{M}}\left[\mathbf{M}\mathbf{z} + \mathbf{x} = \mathbf{y}\right].$$

Now let $\mathbf{z} = (z_1, \ldots, z_s)$ and denote the columns of $\mathbf{M}$ by $M_1, \ldots, M_s$ so that

$$\mathbf{M}\mathbf{z} = z_1 M_1 + \cdots + z_s M_s.$$

Since $\mathbf{z} \neq \mathbf{0}$ we must have some $z_i \neq 0$ and let i be the largest such index. We note that for every choice of $M_1, \ldots, M_{i-1}, M_{i+1}, \ldots, M_s$, the probability, over the choice of M_i that $\mathbf{M}\mathbf{z}+\mathbf{x} = \mathbf{y}$ is q^{-m}, since this happens if and only if $M_i = z_i^{-1}(y - (z_1 M_1 + \cdots z_{i-1} M_{i-1} + x)$, and this event happens with probability q^{-m}. Averaging over the choices of the remaining columns of $\mathbf{M}$, we still have

$$\Pr_{\mathbf{M},\mathbf{b}}\left[A(\mathbf{z}) = \mathbf{y} \mid A(\mathbf{0}) = \mathbf{x}\right] = \Pr_{\mathbf{M}}\left[\mathbf{M}\mathbf{z} + \mathbf{x} = \mathbf{y}\right] = q^{-m},$$

thus establishing that $A(\mathbf{z})$ is distributed uniformly over $\mathbb{F}_q^m$ even when conditioned on $A(\mathbf{0}) = \mathbf{x}$.

Finally to see the final implication of part (2), fix functions $f, g : \mathbb{F}_q^m$ and let

$$E = \left\{ \mathbf{y} \in \mathbb{F}_q^m | f(\mathbf{y}) \neq g(\mathbf{y}) \right\},$$

so that $\delta(f, g) = \Pr_{\mathbf{y}}\left[\mathbf{y} \in E\right]$. We have

$$\Pr_A\left[f \circ A(\mathbf{z}) \neq g \circ A(\mathbf{z})\right] = \Pr_A\left[A(\mathbf{z}) \in E\right],$$

but since $A(\mathbf{z})$ is uniform in $\mathbb{F}_q^m$ even given $A(\mathbf{0}) = \mathbf{x}$, we have

$$\Pr_A\left[A(\mathbf{z}) \in E\right] = \Pr_{\mathbf{y}}\left[\mathbf{y} \in E\right] = \delta(f, g),$$

as claimed. □

14.1.1 The Actual Algorithm

Now we explain why affine substitutions might help in decoding the Reed-Muller code. Recall that the decoding problem for the Reed-Muller codes is the following:

- **Input**: Parameters q, m, r and e (bound on number of errors) and a function $f : \mathbb{F}_q^m \to \mathbb{F}_q$.
- **Output**: Polynomial $P \in \mathbb{F}_q[X_1, \ldots, X_m]$ with $\deg(P) \leq r$ such that $|\{\mathbf{x} \in \mathbb{F}_q^m | f(\mathbf{x}) \neq P(\mathbf{x})\}| \leq e$.

One way to recover the desired polynomial P is to output its value at every given point $\mathbf{x} \in \mathbb{F}_q^m$. (This works provided the polynomial P to be output is uniquely determined by f and the number of errors, and that is the setting we will be working with.) In what follows we will do exactly this. The main idea behind the algorithm of this section is the following: We will pick an affine form A such that $A(\mathbf{0}) = \mathbf{x}$ and attempt to recover the polynomial $P \circ A$. Evaluating $P \circ A(\mathbf{0})$ gives us $P(\mathbf{x})$ and so this suffices, but why is the task of computing $P \circ A$ any easier? Suppose we use an s-variate form A for small s. Then the function $P \circ A$ is given by q^s values with $s < m$ this can be a much smaller sized function and so brute force methods would work faster. But an even more useful observation is that if A is chosen at random such that $A(\mathbf{0}) = \mathbf{x}$, then most of the q^s points (in fact all but one) are random points and so unlikely to be erroneous. In particular for any fixed non-zero $\mathbf{z}$, we would have with high probability $f \circ A(\mathbf{z}) = P \circ A(\mathbf{z})$, where the probability is over the choice of A, assuming the number of errors is small. Since $q^s < q^m$ one can apply a union bound over the roughly q^s choices of z to (hopefully) establish that all the points $\mathbf{z} \in \mathbb{F}_q^s \setminus \{\mathbf{0}\}$ are not errors, and if this happens a further hope would be that $P \circ A$ is uniquely determined by its values on $\mathbb{F}_q^s \setminus \{\mathbf{0}\}$. The two hopes are in tension with each other — the former needs small values of s and the latter needs s to be large; and so we pick an intermediate s, specifically $s = \left\lceil \frac{r+1}{q-1} \right\rceil$, where both conditions are realized and this yields the algorithm below. We describe the algorithm first and then explain this choice of parameters later.

Algorithm 18 SIMPLE REED-MULLER DECODER

INPUT: $r < m$, $0 < e \leq \frac{1}{3} \cdot q^{m - \lceil (r+1)/(q-1) \rceil}$, and function $f : \mathbb{F}_q^m \to \mathbb{F}_q$.

OUTPUT: Polynomial $P \in \mathbb{F}_q[X_1, \ldots, X_m]$ with $\deg(P) \leq r$ such that $|\{\mathbf{x} \in \mathbb{F}_q^m | f(\mathbf{x}) \neq P(\mathbf{x})\}| \leq e$, if such a polynomial exists and NULL otherwise

FOR $\mathbf{x} \in \mathbb{F}_q^m$ DO
  $g(\mathbf{x})$ =LOCAL-DECODE-RM-SIMPLE$(\mathbf{x}, f)$.
RETURN INTERPOLATE$\left(q, m, g, r, \mathbb{F}_q^m\right)$

procedure LOCAL-DECODE-RM-SIMPLE$(\mathbf{x}, f)$
  Repeat LOCAL-DECODE-RM-SIMPLE-ITER$(\mathbf{x}, f)$ $O(m \log q)$ times and return most frequent answer.

procedure LOCAL-DECODE-RM-SIMPLE-ITER$(\mathbf{x}, f)$
  Let $s \leftarrow \left\lceil \frac{r+1}{q-1} \right\rceil$
  Select an m-dimensional s-variate affine form A uniformly conditioned on $A(\mathbf{0}) = \mathbf{x}$.
  $g \leftarrow$ INTERPOLATE$\left(q, s, f \circ A, r, F_q^s \setminus \{\mathbf{0}\}\right)$
  IF g is NULL THEN
    $g \leftarrow 0$
  RETURN $g(\mathbf{0})$.

procedure INTERPOLATE(q, m, f, r, S) ▷ Returns a polynomial $P \in \mathbb{F}_q(Z_1, \ldots, Z_s)$ such that $\deg(P) \leq r$ and $P(\mathbf{x}) = f(\mathbf{x})$ for every $\mathbf{x} \in S$ and return NULL is no such P exists. ▷ See comments in Section 14.1.2 for more on how this algorithm can be implemented.

The detailed algorithm is given as Algorithm 18. Roughly the algorithm contains two loops. The outer loop enumerates $\mathbf{x} \in \mathbb{F}_q^n$ and invokes a subroutine LOCAL-DECODE-RM-SIMPLE that determines $P(\mathbf{x})$ correctly with very high probability. This subroutine creates an inner loop which invokes a less accurate subroutine LOCAL-DECODE-RM-SIMPLE-ITER, which computes $P(\mathbf{x})$ correctly with probability $\frac{2}{3}$, many times and reports the most commonly occurring answer. The crux of the algorithm is thus LOCAL-DECODE-RM-SIMPLE-ITER. This algorithm picks a random affine form A such that $A(\mathbf{0}) = \mathbf{x}$ and assumes that $f \circ A(\mathbf{z}) = P \circ A(\mathbf{z})$ for every non-zero $\mathbf{z}$. Based on this assumption it interpolates the polynomial $P \circ A$ and returns $P \circ A(\mathbf{0}) = P(A(\mathbf{0})) = P(\mathbf{x})$.

The crux of the analysis is to show that the assumption holds with probability at least $\frac{2}{3}$ over the random choice of A provided the number of errors is small. We will undertake this analysis next.

14.1.2 Analysis of the simple decoding algorithm

We first show that each invocation of LOCAL-DECODE-RM-SIMPLE-ITER succeeds with high probability:

Lemma 14.1.3. *Let $P \in \mathbb{F}_q[X_1, \ldots, X_m]$ be a polynomial of degree at most r and let $f : \mathbb{F}_q^m \to \mathbb{F}_q$ be such that*

$$e = |\{\mathbf{x} \in \mathbb{F}_q^m | f(\mathbf{x}) \neq P(\mathbf{x})\}| \leq \frac{1}{3} \cdot q^{m - \lceil (r+1)/(q-1) \rceil}.$$

Then, for every $\mathbf{x} \in \mathbb{F}_q^m$, the probability that LOCAL-DECODE-RM-SIMPLE-ITER$(f, \mathbf{x})$ *returns $P(\mathbf{x})$ is at least $2/3$.*

Proof. Recall $s = \left\lceil \frac{r+1}{q-1} \right\rceil$ and so $e \leq \frac{q^{m-s}}{3}$. We use this condition in the analysis below.

Fix $\mathbf{z} \in \mathbb{F}_q^s \setminus \{\mathbf{0}\}$. Since A was picked conditioned on $A(\mathbf{0}) = \mathbf{x}$, by part (2) of Proposition 14.1.2 we have that $A(\mathbf{z})$ is a uniformly random element of $\mathbb{F}_q^m$ (and in particular this is independent of $\mathbf{x}$). So the probability that $f(A(\mathbf{z})) \neq P(A(\mathbf{z}))$ is exactly $\frac{e}{q^m}$. Taking the union bound over all $\mathbf{z} \in \mathbb{F}_q^s \setminus \{\mathbf{0}\}$ we get that

$$\Pr_A \left[\exists \mathbf{z} \in \mathbb{F}_q^s \setminus \{\mathbf{0}\} \text{ s.t. } f(A(\mathbf{z})) \neq P(A(\mathbf{z})) \right] \leq (q^s - 1) \cdot \frac{e}{q^m} \leq \frac{e}{q^{m-s}} \leq \frac{1}{3}. \tag{14.1}$$

So, with probability at least $\frac{2}{3}$, we have that $f \circ A(\mathbf{z}) = P \circ A(\mathbf{z})$ for every $\mathbf{z} \in \mathbb{F}_q^s \setminus \{\mathbf{0}\}$. We argue below that if this holds, then LOCAL-DECODE-RM-SIMPLE-ITER$(f, \mathbf{x})$ returns $P(\mathbf{x})$ and this proves the lemma.

Since $P \circ A$ is a polynomial in $\mathbb{F}_q[Z_1, \ldots, Z_s]$ of degree at most r that agrees with $f \circ A$ on every $\mathbf{z} \in \mathbb{F}_q^s \setminus \{\mathbf{0}\}$, we have that the there exists at least one polynomial satisfying the condition of the final interpolation step in LOCAL-DECODE-RM-SIMPLE-ITER$(f, \mathbf{x})$. It suffices to show that this polynomial is unique, but this follows from Exercise 14.4, which asserts that $\delta(P \circ A, h) \geq \frac{2}{q^s}$ for every polynomial $h \in \mathbb{F}_q[Z_1, \ldots, Z_s]$ of degree at most r, provided $r < (q-1)s$. (Note that our choice of $s = \left\lceil \frac{r+1}{q-1} \right\rceil$ ensures this.) In particular this implies that every pair of polynomials disagree on at least two points in $\mathbb{F}_q^s$ and so on at least one point in $\mathbb{F}_q^s \setminus \{\mathbf{0}\}$. Thus $P \circ A$ is

the unique polynomial that fits the condition of the interpolation step in LOCAL-DECODE-RM-SIMPLE-ITER$(f, \mathbf{x})$ and so this subroutine returns $P(\mathbf{x})$ with probability at least $\frac{2}{3}$. □

We note that one can push the $\frac{1}{3}$ fraction of errors to $\frac{1}{2} - \gamma$ for any $0 < \gamma < 1/2$ with a success probability of $\frac{1}{2} + \gamma$ (see Exercise 14.5).

With Lemma 14.1.3 in hand, some routine analysis suffices to show the correctness of the Simple Reed-Muller Decoder (Algorithm 18) and we do so in the theorem below.

Theorem 14.1.4. *The Simple Reed-Muller Decoder (Algorithm 18) is a correct (randomized) polynomial in n time algorithm decoding the Reed-Muller code* $\mathrm{RM}(q, m, r)$ *from* $e \le \frac{1}{3} \cdot q^{m - \lceil (r+1)/(q-1) \rceil}$ *errors.*

Proof. Fix $P \in \mathbb{F}_q[X_1, \ldots, X_m]$ be a polynomial of degree at most r and $f : \mathbb{F}_q^m \to \mathbb{F}_q$ such that

$$e = |\{\mathbf{x} \in \mathbb{F}_q^m | f(\mathbf{x}) \neq P(\mathbf{x})\}| \le \frac{1}{3} \cdot q^{m - \lceil (r+1)/(q-1) \rceil}.$$

Further fix $x \in q^m$. Lemma 14.1.3 asserts that a call to LOCAL-DECODE-RM-SIMPLE-ITER$(f, \mathbf{x})$ returns $P(\mathbf{x})$ with probability at least $\frac{2}{3}$. By an application of the Chernoff bounds (in particular, see Exercise 3.3), the majority of the $O(m \log q)$ calls to LOCAL-DECODE-RM-SIMPLE-ITER(f, x) is $P(\mathbf{x})$ except with probability $\exp(-m \log q)$ and by setting the constant in the $O(\cdot)$ appropriately, we an ensure this probability is at most $\frac{q^{-m}}{3}$. We thus conclude that for every fixed $\mathbf{x} \in \mathbb{F}_q^m$ the probability that LOCAL-DECODE-RM-SIMPLE$(f, \mathbf{x})$ does not return $P(\mathbf{x})$ is at most $\frac{q^{-m}}{3}$. By the union bound, we could that the probability that there exists $\mathbf{x} \in \mathbb{F}_q^m$ such that LOCAL-DECODE-RM-SIMPLE$(f, \mathbf{x}) \neq P(\mathbf{x})$ is at most $\frac{1}{3}$. Thus with probability at least $\frac{2}{3}$ the algorithm computes $P(\mathbf{x})$ correctly for every $\mathbf{x}$ and thus the interpolation returns P with probability at least $\frac{2}{3}$.

The running time of the algorithm is easy to establish. Let $T_{\mathrm{int}}(n)$ denote the time it takes to interpolate to find the coefficients of a polynomial P given its n evaluations. It is well-known that T_{int} is a polynomial with near linear running time. (See Remarks on Interpolation below.) We have that the LOCAL-DECODE-RM-SIMPLE-ITER takes time at most $T_{\mathrm{int}}(q^s)$ per invocation, and thus LOCAL-DECODE-RM-SIMPLE takes $O(m \cdot T_{\mathrm{int}}(q^s) \cdot \log q)$ steps per invocation. Since LOCAL-DECODE-RM-SIMPLE is executed q^m times by the overall algorithm, the overall running time is bounded by $T_{\mathrm{int}}(q^m) + O(m \cdot q^m \cdot T_{\mathrm{int}}(q^s) \log q)$. Expressed in terms of $n = q^m$ and $e_{\max} = q^{m-s}/3$ and crudely bounding interpolation cost by a cubic function, this translates into a running time of $O(n^3) + O\left(\frac{n^4}{e_{\max}^3} \log n\right)$. □

Remarks on Interpolation. As mentioned in the proof above, the running time of the algorithm depends on the running time of the two interpolation steps in the algorithm DECODE-RM-SIMPLE. To get polynomial time algorithms for either step, it suffices to note that interpolation is just solving a system of linear equations and thus can always be solved in cubic time by Gaussian elimination (see Exercise 14.6). To make the steps more efficient, one can use the structure of polynomial interpolation to get some speedups for the first interpolation step (see Section 14.5). For the second, since we are only interested in evaluating $P \circ A(0)$, interpolation

is a bit of overkill. It turns out one can explicitly determine the exact linear form which describes $P \circ A(0)$ in terms of $\{P \circ A(\mathbf{z}) | \mathbf{z} \in \mathbb{F}_q^s \setminus \{\mathbf{0}\}\}$ and this turns out to be extremely simple: In fact $P \circ A(\mathbf{0}) = -\sum_{\mathbf{z} \in \mathbb{F}_q^s \setminus \{\mathbf{0}\}} P \circ A(\mathbf{z})$ (see Exercise 14.7).

Remark on Fraction of Errors corrected. The number of errors corrected by the Simple Reed-Muller Decoder, $\frac{1}{3} \cdot q^{m - \lceil (r+1)/(q-1) \rceil}$, is complex and requires some explanation. It turns out that this quantity is closely related to the distance of the code. For $s = \left\lceil \frac{r+1}{q-1} \right\rceil$ if we now let t be such that $r = s(q-1) - t$ (note that this is different from the way we did this splitting in Lemma 9.4.1), then from Lemma 9.4.1 we have that the distance of the code $\mathrm{RM}(q, m, r)$ is $(t+1)q^{m-s}$ where $1 \le t \le q-1$ (see Exercise 14.8). So in particular the distance of the code is between q^{m-s} and q^{m-s+1}. In contrast, our algorithm corrects $\frac{q^{m-s}}{3}$ errors, which is at least a $\frac{1}{3q}$-fraction of the distance of the code. Ideally we would like algorithms correcting up to $\frac{1}{2}$ as many errors as the distance, and this algorithm falls short by a "constant" factor, if q is a constant. In the rest of the chapter we will try to achieve this factor of $\frac{1}{2}$.

14.2 Majority Logic Decoding

The algorithm of the previous section corrects errors up to a constant fraction of the distance (with the constant depending on q, but not on m or r) but is not the best one could hope for. In this section we develop an algorithm that corrects the optimal number of errors over $\mathbb{F}_2$. The main idea is to continue to explore the function over "affine subspaces" but now the substitutions will be much simpler. Specifically they will be of the form $x_i = b_i$ for many different choices of i where $b_i \in \mathbb{F}_2$. This will leave us with a function on the unset variables and while we won't be able to determine the function completely on the remaining variables, we will be able to determine some coefficients and this will allow us to make progress.

The main idea driving this algorithm is the following proposition about degree r polynomials.

Proposition 14.2.1. *Let $P \in \mathbb{F}_2[X_1, \ldots, X_m]$ be of degree r and let $C \in \mathbb{F}_2$ be the coefficient of the monomial $\prod_{i=1}^r X_i$ in P. Then, for every $\mathbf{b} \in \mathbb{F}_2^{m-r}$, it is the case that $\sum_{\mathbf{a} \in \mathbb{F}_2^r} P(\mathbf{a}, \mathbf{b}) = C$.*

Proof. Let $P_{\mathbf{b}}(X_1, \ldots, X_r) = P(X_1, \ldots, X_r, \mathbf{b})$, i.e., $P_{\mathbf{b}}$ is P restricted to the subspace $X_i = b_i$ for $r < i \le m$. Note that the coefficient of the monomial $X_1 \cdots X_r$ in $P_{\mathbf{b}}$ remains C, since all other monomial now have degree strictly less than r after the substitutions $X_i = b_i$. (Note that we used the fact that P has degree at most r to make this assertion.) So we can write $P_{\mathbf{b}} = C \cdot X_1 \cdots X_r + g$ where $\deg(g) < r$. We wish to show that

$$\sum_{\mathbf{a} \in \mathbb{F}_2^r} P_{\mathbf{b}}(\mathbf{a}) = \sum_{\mathbf{a} \in \mathbb{F}_2^r} C \cdot \left(\prod_{j=1}^r a_j \right) + \sum_{\mathbf{a} \in \mathbb{F}_2^r} g(\mathbf{a}) = C.$$

We first note that the first summation is trivially C since all terms except when $a_1 = \cdots = a_r = 1$ evaluate to zero and the term corresponding to $a_1 = \cdots = a_r = 1$ evaluates to C. The proposition

now follows from Exercise 14.7 which asserts that for every polynomial g of degree less than r, the summation $\sum_{\mathbf{a}\in\mathbb{F}_2^r} g(\mathbf{a}) = 0$. $\square$

As such the proposition above only seems useful in the error-free setting — after all, it assumes P is given correctly everywhere. But it extends to the setting of a small number of *errors* immediately. Note that if a function f disagrees with polynomial P on at most e points in $\mathbb{F}_2^m$ then there are at most e choices of $\mathbf{b} \in \mathbb{F}_2^{m-r}$ for which $\sum_{\mathbf{a}\in\mathbb{F}_2^r} f(\mathbf{a},\mathbf{b})$ does not equal C. In particular, if $e < 2^{m-r}/2$ then the *majority* of choices of $\mathbf{b}$ lead to the correct value of C. (And remarkably, for the class of degree r polynomials, this is exactly one less than half the distance of the associated code.) Of course, the monomial $\prod_{i=1}^r X_i$ is not (very) special. The same reasoning allows us to compute the coefficient of any monomial of degree r. For example $\text{majority}_{\mathbf{b}\in\mathbb{F}_2^{m-r}}\{\sum_{\mathbf{a}\in\mathbb{F}_2^r} f(\mathbf{a},\mathbf{b})\}$ gives the coefficient of $X_1\cdots X_r$, and $\text{majority}_{\mathbf{b}\in\mathbb{F}_2^{m-r}}\{\sum_{\mathbf{a}\in\mathbb{F}_2^r} f(\mathbf{b},\mathbf{a})\}$ (note the exchange of $\mathbf{a}$ and $\mathbf{b}$) gives the coefficient of $X_{m-r+1}\cdots X_m$. (See Exercise 14.9.) With appropriate notation for substituting $\mathbf{a}$ and $\mathbf{b}$ into the right places, we can calculate any other monomial of degree r as well. And then downward induction on r allows us to compute coefficients of lower degree monomials. This leads us to the algorithm described next. For the sake of completeness we also give the full analysis afterwards.

14.2.1 The Majority Logic Decoding Algorithm

We start with some notation that will help us describe the algorithm more precisely.

Definition 14.2.1. For $S \subseteq [m]$ we let X_S denote the monomial $\prod_{i\in S} X_i$.

Definition 14.2.2. For $S \subseteq [m]$ with $|S| = t$ and vectors $\mathbf{a} \in \mathbb{F}_2^t$ and $\mathbf{b} \in \mathbb{F}_2^{m-t}$, let $(S \leftarrow \mathbf{a}, \bar{S} \leftarrow \mathbf{b})$ denote the vector whose coordinates in S are given by $\mathbf{a}$ and coordinates in $\bar{S}$ are given by $\mathbf{b}$.

Definition 14.2.3. For $S \subseteq [m]$ with $|S| = t$ and vectors $\mathbf{a} \in \mathbb{F}_2^t$ and $\mathbf{b} \in \mathbb{F}_2^{m-t}$, let $f(S \leftarrow \mathbf{a}, \bar{S} \leftarrow \mathbf{b})$ denote the evaluation of f on $(S \leftarrow \mathbf{a}, \bar{S} \leftarrow \mathbf{b})$. In other words, let $S = \{i_1, \ldots, i_t\}$ with $i_k < i_{k+1}$ and let $\bar{S} = \{j_1, \ldots, j_{m-t}\}$ with $j_k < j_{k+1}$. Then $f(S \leftarrow \mathbf{a}, \bar{S} \leftarrow \mathbf{b}) = f(\mathbf{c})$ where $\mathbf{c} \in \mathbb{F}_2^m$ is the vector such that $c_{i_k} = a_k$ and $c_{j_\ell} = b_\ell$.

The majority logic decoder details are presented in Algorithm 19.

14.2.2 The analysis

We next argue that the algorithm MAJORITY LOGIC DECODER (Algorithm 19) correct up to half the errors for the $\text{RM}(2, m, r)$ code.

Lemma 14.2.2. *On input $f : \mathbb{F}_2^m \to \mathbb{F}_2$ that disagrees with a polynomial $Q \in \mathbb{F}_2[X_1, \ldots, X_m]$ of degree at most r on at most $e < \frac{1}{2}\cdot 2^{m-r}$ points, the algorithm* MAJORITY LOGIC DECODER(f) *correctly outputs Q.*

Proof. Let $Q(X) = \sum_{S\subseteq[m]} C'_S X_S$ and let $Q_t(X) = \sum_{S\subseteq[m]:|S|\geq t} C'_S X_S$. We argue by downward induction on t (from $r+1$ down to 0) that $Q_t = P_t$ where P_t's are the polynomials computed by

Algorithm 19 Majority Logic Decoder

INPUT: $r < m$, $0 \le e < \frac{1}{2}2^{m-r}$, and function $f : \mathbb{F}_2^m \to \mathbb{F}_2$.
OUTPUT: Output P such $\deg(P) \le r$ and $|\{x \in \mathbb{F}_2^m | P(x) \ne f(x)\}| \le e$.

```
P_{r+1} ← 0
FOR t = r downto 0 DO
    f_t ← f − P_{t+1}
    FOR every S ⊆ [m] with |S| = t DO
        FOR every b ∈ F_2^{m−t} DO
            C_{S,b} ← Σ_{a∈F_2^t} f_t(S ← a, S̄ ← b).
        C_S ← majority_b {C_{S,b}}
    P_t ← P_{t+1} + Σ_{S⊆[m],|S|=t} C_S X_S
RETURN P_0
```

our algorithm. The base case is obvious since $P_{r+1} = Q_{r+1} = 0$. Assume now that $P_{t+1} = Q_{t+1}$ and so we have that $f_t = f - P_{t+1}$ disagrees with $Q - Q_{t+1}$ on at most e points. (See Exercise 14.10.) We now argue that for every subset $S \subseteq [m]$ with $|S| = t$, $C_S = C'_S$. Fix such a set S. For $\mathbf{b} \in \mathbb{F}_2^{m-t}$, we refer to the partial assignment $\bar{S} \leftarrow \mathbf{b}$ as a "subcube" corresponding to the points $\{(S \leftarrow \mathbf{a}, \bar{S} \leftarrow \mathbf{b}) | \mathbf{a} \in \mathbb{F}_2^t\}$. We say that a subcube $\bar{S} \leftarrow \mathbf{b}$ is in error if there exists a such that $f_t(S \leftarrow \mathbf{a}, \bar{S} \leftarrow \mathbf{b}) \ne (Q - Q_{t+1})(S \leftarrow \mathbf{a}, \bar{S} \leftarrow \mathbf{b})$. By Proposition 14.2.1, we have that if a subcube is not in error then $C_{S,\mathbf{b}} = C'_S$, since C'_S is the coefficient of X_S in the polynomial $(Q - Q_{t+1})$ (whose degree is at most t). Furthermore at most e subcubes can be in error (see Exercise 14.11). Finally, since the total number of subcubes is $2^{m-t} \ge 2^{m-r} > 2e$ we thus have that a majority of subcubes are not in error and so $C_S = \text{majority}_{\mathbf{b}}\{C_{S,\mathbf{b}}\} = C'_S$.

Thus we have for every S with $|S| = t$ that $C_S = C'_S$ and so $Q_t = P_t$, giving the inductive step. So we have for every t, $P_t = Q_t$ and in particular $P_0 = Q_0 = Q$ as desired.

□

The running time of the algorithm is easy to see as being at most $n = 2^m$ times the number of coefficients of a degree r polynomial, which is $\sum_{i=0}^{r} \binom{m}{i} \le n$ in the binary case. Thus $O(n^2)$ is a crude upper bound on the running time of this algorithm. Note that this algorithm corrects up to exactly $\left\lfloor \frac{d-1}{2} \right\rfloor$ errors where $d = 2^{m-r}$ is the minimum distance of $\mathrm{RM}(2, m, r)$. (Since d is even, this quantity equals $\frac{d}{2} - 1$.) We thus have the following theorem.

Theorem 14.2.3. *For every $0 \le r < m$, The Majority Logic Decoder (Algorithm 19), corrects up to $\frac{d}{2} - 1$ errors in the Reed-Muller code, $\mathrm{RM}(2, m, r)$, in $O(n^2)$ time, where $n = 2^m$ is the block length of the code and $d = 2^{m-r}$ is its minimum distance.*

14.3 Decoding by reduction to Reed-Solomon decoding

The algorithms described so far were based on very basic ideas, but they have their limitations. The SIMPLE REED-MULLER DECODER (Algorithm 18) fails to correct errors up to half the min-

imum distance. And the majority logic algorithm (Algorithm 19) seems to work only over $\mathbb{F}_2$ (where the monomial structure is especially simple). The final algorithm we give uses a slightly more sophisticated algebraic idea, but then ends up yielding an almost 'trivial' reduction to Reed-Solomon decoding. (It is trivial in the sense that the reduction algorithm almost does no work.) The resulting reduction can use any algorithm for Reed-Solomon decoding including any of the ones from Chapter 15.

The crux of the reduction is a natural bijection between the vector space $\mathbb{F}_q^m$ and the field $\mathbb{F}_{q^m}$. This bijection converts the space of functions $\{f|\mathbb{F}_q^m \to \mathbb{F}_q\}$ to the space of functions $\{f : \mathbb{F}_{q^m} \to \mathbb{F}_q\} \subseteq \{f : \mathbb{F}_{q^m} \to \mathbb{F}_{q^m}\}$. Algorithmically, it is important that the bijection only acts on the domain and so almost no work is needed to convert a function $g \in \{f|\mathbb{F}_q^m \to \mathbb{F}_q\}$ to its image $G \in \{f : \mathbb{F}_{q^m} \to \mathbb{F}_{q^m}\}$ under the bijection. Thus Reed-Muller codes get transformed to a subcode of some Reed-Solomon code, and corrupted Reed-Muller codewords get mapped to corrupted Reed-Solomon codewords. Now comes the algebraic part: namely, analyzing how good is the distance of the so-obtained Reed-Solomon code, or equivalently upper bounding the degree of the polynomials G obtained by applying the bijection to $g \in \mathrm{RM}(q, m, r)$. It turns out the bijection preserves the distance exactly and so algorithms correcting the Reed-Solomon code up to half its distance does the same for the Reed-Muller code.

In the rest of this section we first describe the 'nice' bijection Φ from $\mathbb{F}_{q^m} \to \mathbb{F}_q^m$ and introduce a parameter called the *extension degree* that captures how good the bijection is. Then, we analyze the extension degree of the bijection map and show that it ends up mapping Reed-Muller codes to Reed-Solomon codes of the same distance, and thus an algorithm to decode Reed-Solomon codes with errors up to half the distance of the code also yield algorithms to decode Reed-Muller code with errors up to half the distance of the code.

14.3.1 A bijection from $\mathbb{F}_q^m$ vs. $\mathbb{F}_{q^m}$

The bijections we will eventually work with in this section will be *linear-bijections*. We introduce this concept first.

Definition 14.3.1. A function $\Phi : \mathbb{F}_{q^m} \to \mathbb{F}_q^m$ is said to be an $\mathbb{F}_q$-linear bijection if

1. Φ is a bijection, i.e., $\Phi(u) = \Phi(v) \Rightarrow u = v$ for every $u, v \in \mathbb{F}_{q^m}$.
2. Φ is $\mathbb{F}_q$-linear, i.e., for every $\alpha, \beta \in \mathbb{F}_q$ and $u, v \in \mathbb{F}_{q^m}$ it is the case that $\Phi(\alpha u + \beta v) = \alpha\Phi(u) + \beta\Phi(v)$.

(Above and throughout this section it will be useful to remember that $\mathbb{F}_q \subseteq \mathbb{F}_{q^m}$ and so operations such as αu for $\alpha \in \mathbb{F}_q$ and $u \in \mathbb{F}_{q^m}$ are well-defined.)

Note that a linear bijection $\Phi : \mathbb{F}_{q^m} \to \mathbb{F}_q^m$ can be viewed as m functions $(\Phi_1, \ldots, \Phi_m)$ with $\Phi_i : \mathbb{F}_{q^m} \to \mathbb{F}_q$ so that $\Phi(u) = (\Phi_1(u), \ldots, \Phi_m(u))$. Furthermore each Φ_i is a linear function from $\mathbb{F}_{q^m} \to \mathbb{F}_q$ and since $\mathbb{F}_q \subseteq \mathbb{F}_{q^m}$, Φ_i can be viewed as a polynomial in $\mathbb{F}_{q^m}[Z]$ (see Exercise 14.12). Our proposition below recalls the basic properties of such *linearized polynomials*.

Proposition 14.3.1. *There exists an $\mathbb{F}_q$-linear bijection from $\mathbb{F}_{q^m}$ to $\mathbb{F}_q^m$. If $\Phi = (\Phi_1, \ldots, \Phi_m) : \mathbb{F}_{q^m} \to \mathbb{F}_q^m$ is an $\mathbb{F}_q$-linear bijection then each $\Phi_i(Z)$ is a trace function. Specifically there exist*

$\lambda_1 \ldots, \lambda_m \in \mathbb{F}_{q^m}$, *which are linearly independent over* $\mathbb{F}_q$, *such that* $\Phi_i(Z) = \mathrm{Tr}(\lambda_i Z) = \sum_{j=0}^{m-1} \lambda_i^{q^j} Z^{q^j}$. *In particular* $\deg(\Phi_i) = q^{m-1}$ *and* Φ *is a* linearized polynomial *(i.e., only non-zero coefficients are for monomials of the form* Z^{q^j}*).*

Proof. Given a bijection Φ the fact that it is a Trace function follows from Proposition D.5.17, which implies its degree and linearized nature (see Exercise 14.13). All that remains to show is that a linear bijection exists. We claim that if $\lambda_1, \ldots, \lambda_m \in \mathbb{F}_{q^m}$ are $\mathbb{F}_q$-linearly independent then $\Phi = (\Phi_1, \ldots, \Phi_m)$, with $\Phi_i(Z) = \mathrm{Tr}(\lambda_i Z)$, is a $\mathbb{F}_q$ linear bijection. Linearity follows from the linearity of Trace so all we need to verify is that this is a bijection. And since the domain and range of Φ have the same cardinality, it suffices to show that Φ is surjective. Consider the set

$$S = \{(\Phi(u) | u \in \mathbb{F}_{q^m}\} \subseteq \mathbb{F}_q^m.$$

By the linearity of Φ, S is a subspace of $\mathbb{F}_q^m$. If $S \neq \mathbb{F}_q^m$ (i.e., if Φ is not surjective) we must have that elements of S satisfy some non-trivial constraint, i.e., there exists $(\alpha_1, \ldots, \alpha_m) \in \mathbb{F}_q^m \setminus \{\mathbf{0}\}$ such that $\sum_{i=1}^m \alpha_i \beta_i = 0$ for every $(\beta_1, \ldots, \beta_m) \in S$ (see Exercise 14.14). But now consider

$$\sum_i \alpha_i \Phi_i(Z) = \sum_i \alpha_i \mathrm{Tr}(\lambda_i Z) = \mathrm{Tr}\left(\left(\sum_i \alpha_i \lambda_i\right) \cdot Z\right), \tag{14.2}$$

where the last equality follows from the fact that Tr is a linear map (see Proposition D.5.17). On the one hand (see Exercise 14.15) this is a non-zero polynomial in Z of degree at most q^{m-1}, while on the other hand it evaluates to zero on every $u \in \mathbb{F}_{q^m}$, which contradicts the degree mantra. We conclude Φ is surjective, and hence it is an $\mathbb{F}_q$-linear bijection. □

Our goal is to use a linear bijection from $\Phi : \mathbb{F}_{q^m} \to \mathbb{F}_q^m$ (any such bijection will do for us) to convert functions whose domain is $\mathbb{F}_q^m$ (which is the case for codewords of the Reed-Muller code) to functions whose domain is $\mathbb{F}_{q^m}$. Specifically, given $f : \mathbb{F}_q^m \to \mathbb{F}_q$ and $\Phi : \mathbb{F}_{q^m} \to \mathbb{F}_q^m$, let $f \circ \Phi : \mathbb{F}_{q^m} \to \mathbb{F}_q^m$ be the function $(f \circ \Phi)(u) = f(\Phi(u))$.

Key to the utility of this transformation is the analysis of how this blows up the degree of the underlying polynomials. Recall that for a function $F : \mathbb{F}_{q^m} \to \mathbb{F}_{q^m}$, its degree is defined to be the degree of the (unique) polynomial $P \in \mathbb{F}_{q^m}[Z]$ with $\deg(P) < q^m$ such that $P(u) = F(u)$ for every $u \in \mathbb{F}_{q^m}$. Our main challenge henceforth is to understand the question: If $f : \mathbb{F}_q^m \to \mathbb{F}_q$ is a degree r polynomial, how large can the degree of $f \circ \Phi$ be? We do not answer this question right away, but define the parameter quantifying this effect next, and then design and analyze a Reed-Muller decoding algorithm in terms of this parameter.

Definition 14.3.2. For prime power q and non-negative integers m and r, let the *extension degree* of r over $\mathbb{F}_{q^m}$, denoted $R_{q,m}(r)$, be the maximum degree of $p \circ \Phi$ over all choices of $p \in \mathbb{F}_q[X_1, \ldots, X_m]$ (or the associated function $p : \mathbb{F}_q^m \to \mathbb{F}_q$) of degree at most r and over all $\mathbb{F}_q$-linear bijections $\Phi : \mathbb{F}_{q^m} \to \mathbb{F}_q$.

Our algorithm and its analysis are quite natural modulo the analysis of $R_{q,m}(r)$ and we describe them below.

Algorithm 20 REED-SOLOMON-BASED DECODER

INPUT: q, $r < m(q-1)$, $0 \le e < (q^m - R_{q,m}(r))/2$, and function $f : \mathbb{F}_q^m \to \mathbb{F}_q$.
OUTPUT: Output p such $\deg(p) \le r$ and $|\{x \in \mathbb{F}_q^m | P(x) \neq f(x)\}| \le e$.

Let $\Phi : \mathbb{F}_{q^m} \to \mathbb{F}_q^m$ be an $\mathbb{F}_q$-linear bijection
FOR $u \in \mathbb{F}_{q^m}$ DO
  $F(u) \leftarrow f(\Phi(u))$
Let P be the output of decoding F using Reed-Solomon codes by Algorithm 21 ▷ With inputs $k = R_{q,m}(r) + 1$, $n = q^m$ and n pairs $(u, F(u))$ for every $u \in \mathbb{F}_{q^m}$.
FOR $u \in \mathbb{F}_q^m$ DO
  $p(u) \leftarrow P(\Phi^{-1}(u))$
RETURN p

We note that the decoder here outputs a polynomial $p \in \mathbb{F}_q[X_1, \ldots, X_m]$ in terms of its function representation. If a coefficient representation is desired one can use some interpolation algorithm to convert it. Other than such interpolation, most of the transformation above is mostly syntactic, since a normal representation of $\mathbb{F}_{q^m}$ is already in the form of vectors in $\mathbb{F}_q^m$ via some $\mathbb{F}_q$-linear bijection. The only real work is in the call to the Reed-Solomon decoder.

Below we show that the algorithm above is correct for $e < (q^m - R_{q,m}(r))/2$. The crux of the analysis is in showing that this quantity is actually half the distance of the Reed-Muller code, and we defer that analysis to the next section.

Proposition 14.3.2. *Let $f : \mathbb{F}_q^m \to \mathbb{F}_q$ be any function and let $g : \mathbb{F}_q^m \to \mathbb{F}_q$ be a degree r polynomial such that $|\{u \in \mathbb{F}_q^m | f(u) \neq g(u)\}| \le e < (q^m - R_{q,m}(r))/2$. Then* REED-SOLOMON-BASED DECODER(f) *returns g*

Proof. Let $G = g \circ \Phi$. Then we have that $\deg(G) \le R_{q,m}(r)$ and we have that $\{v \in \mathbb{F}_{q^m} | F(v) \neq G(v)\}| \le e < (q^m - R_{q,m}(r))/2$. Since the distance of the Reed-Solomon code with $N = q^m$ and $K = R_{m,q}(r) + 1$ is $N - K + 1 = q^m - R_{q,m}(r)$ and e is less than half the distance, we have that G is the unique polynomial with this property, and so the Reed-Solomon decoder must return $P = G$. We conclude that $p = P \circ \Phi^{-1} = G \circ \Phi^{-1} = g$, as desired. □

14.3.2 Analysis of Extension Degree

We start with a simple warmup result that already leads to an optimal algorithm for decoding when $r < q$.

Proposition 14.3.3. $R_{q,m}(r) \le r \cdot q^{m-1}$.

Proof. The proof following immediately from the definition and the fact that linear functions are polynomials of degree q^{m-1}. Specifically, let $p \in \mathbb{F}_q[X_1, \ldots, X_m]$ be of degree at most r and let $\Phi = (\Phi_1, \ldots, \Phi_m)$ be an $\mathbb{F}_q$-linear bijection. Then since each Φ_i is a polynomial of degree q^{m-1} we have that $p(\Phi_1(Z), \ldots, \Phi_m(Z))$ is a polynomial of degree at most $r \cdot q^{m-1}$. Finally since the reduction modulo $(Z^{q^m} - Z)$ does not increase the degree we have $p \circ \Phi$ is a polynomial of degree at most $r \cdot q^{m-1}$, as desried. □

Corollary 14.3.4. *If $r < q$ then* REED-SOLOMON-BASED DECODER *decodes* $\mathrm{RM}(q,m,r)$ *up to half its minimum distance.*

Proof. By Lemma 9.2.2 we have that the distance of the Reed-Muller code $\mathrm{RM}(q,m,r)$ is $(q-r)\cdot q^{m-1}$. From Proposition 14.3.3 we have that REED-SOLOMON-BASED DECODER decodes provided $e < (q^m - R_{m,q}(r))/2 = (q^m - r\cdot q^{m-1})/2 = (q-r)\cdot q^{m-1}/2$, i.e., up to half the minimum distance. □

Finally we turn to the general case (i.e. $r \geq q$). For this part, the crude bound that the degree of $f\circ\Phi$ is at most $q^{m-1}\cdot\deg(f)$ is no longer good enough. This bound is larger than q^m, whereas every function has degree at most q^m-1. To get the 'right' degree bound on the degree of $f\circ\Phi$, we now need to use the fact that we can reduce any polynomial modulo $(Z^{q^m}-Z)$ and this leaves the underlying function on $\mathbb{F}_{q^m}$ unchanged. Thus from this point on we will try to understand the degree of $f\circ\Phi \pmod{Z^{q^m}-Z}$. Using this reduction properly we will eventually be able to get the correct bound on the degree of $f\circ\Phi$. We state the bound next and then work our way towards proving it.

Lemma 14.3.5. *Let $r = s(q-1)+t$ where $0 \leq t < q-1$. Then $R_{q,m}(r) = q^m - (q-t)q^{m-s-1}$.*

We first state the immediate consequence of Lemma 14.3.5 to the error-correction bound of the Reed-Solomon-based decoder.

Theorem 14.3.6. *For every prime power q, integers $m \geq 1$ and $0 \leq r < m(q-1)$, the* REED-SOLOMON-BASED DECODER *decodes* $\mathrm{RM}(q,m,r)$ *up to half its minimum distance.*

Proof. Let $r = s(q-1)+t$ with $0 \leq t < q-1$. By the polynomial distance lemma (Lemma 9.4.1) we have that the distance of the Reed-Muller code $\mathrm{RM}(q,m,r)$ is $(q-t)\cdot q^{m-s-1}$. Combining Proposition 14.3.2 with Lemma 14.3.5 we have that REED-SOLOMON-BASED DECODER decodes provided $e < (q^m - R_{m,q}(r))/2 = (q-t)\cdot q^{m-s-1}/2$, i.e., up to half the minimum distance of $\mathrm{RM}(q,r,m)$. □

The proof of Lemma 14.3.5 is somewhat involved and needs some new notions. We introduce these notions next.

Definition 14.3.3 (q-degree)**.** For integer d, let $d_0, d_1, d_2, \ldots$ denote its expansion in base q, i.e., $d = \sum_{i=0}^{\infty} d_i q^i$ and $0 \leq d_i < q$ for every i. For a monomial Z^d, define its q-degree, denoted $\deg_q(Z^d)$, to be the quantity $\sum_{i=0}^{\infty} d_i$. For a polynomial $p(Z) = \sum_d c_d Z^d$, define its q-degree, denoted $\deg_q(p)$, to be $\max_{d|c_d\neq 0}\{\deg_q(Z^d)\}$.

For example $\deg_2(X^8+X^3+X+1) = \deg_2(X^3) = 2$.

We describe some basic properties of q-degree below. Informally, the proposition below proves (in parts (1)-(3)) that the q-degree behaves just like the regular degree when it comes to addition, multiplication and reduction modulo $Z^{q^m}-Z$. Note that while parts (1) and (2) are natural, part (3) is already special in that it only holds for reduction modulo some special polynomials. Finally part (4) of the proposition allows us to related the q-degree of a polynomial with its actual degree and this will come in useful when we try to bound the degree of $f\circ\Phi(\pmod{Z)^{q^m}-Z}$.

Proposition 14.3.7. *For every $\alpha, \beta \in \mathbb{F}_{q^m}$ and $P, P_1, P_2 \in \mathbb{F}_{q^m}[Z]$ we have:*

(1) $\deg_q(\alpha P_1 + \beta P_2) \le \max\{\deg_q(P_1), \deg_q(P_2)\}$.

(2) $\deg_q(P_1 \cdot P_2) \le \deg_q(P_1) + \deg_q(P_2)$.

(3) $\deg_q(P \pmod{Z^{q^m} - Z}) \le \deg_q(P)$. *(Note that here the total degree* $\deg(P)$ *can be strictly greater than* q^m*.)*

(4) $\deg(P) < q^m$ *and* $\deg_q(P) = s(q-1) + t$ *for* $0 \le t < q-1$ *implies*

$$\deg(P) \le q^m - (q-t)q^{m-s-1}.$$

Proof. Part (1) is immediate from the definition since the monomials of $\alpha P_1 + \beta P_2$ is in the union of the monomials of P_1 and P_2. Next, note that due to part (1), it suffices to prove parts (2)-(4) for monomials.

For part (2) for monomials, we wish to show that that $\deg_q(Z^d \cdot Z^e) \le \deg_q(Z^d) + \deg_q(Z^e)$. Let $d = \sum_i d_i q^i$, $e = \sum_i e_i q^i$ and let $f = d + e = \sum_i f_i q^i$. Then it can be verified that for every i we have $\sum_{j=0}^{i} f_j \le \sum_{j=0}^{i}(d_j + e_j)$ (see Exercise 14.16) and this eventually implies

$$\deg_q(Z^{d+e}) = \sum_j f_j \le \sum_j (d_j + e_j) = \deg_q(Z^d) + \deg_q(Z^e). \tag{14.3}$$

For part (3), note that since $Z^{q^i} = \left(\left(Z^{q^{\lfloor i/m \rfloor}}\right)^m\right)^{i \mod m}$,

$$Z^{q^i} \pmod{Z^{q^m} - Z} = Z^{q^{(i \mod m)}}.$$

So

$$Z^{d_i q^i} \mod (Z^{q^m} - Z) = Z^{d_i q^{(i \mod m)}}. \tag{14.4}$$

We conclude that for $d = \sum_i d_i q^i$ with $0 \le d_i < q$ we have:

$$\begin{aligned}
\deg_q\left(Z^d \mod \left(Z^{q^m} - Z\right)\right) &= \deg_q\left(Z^{\sum_i d_i q^i} \mod \left(Z^{q^m} - Z\right)\right) \\
&\le \deg_q\left(Z^{\sum_i d_i q^{i \mod m}}\right) \\
&\le \sum_i \deg_q\left(Z^{d_i q^{i \mod m}}\right) \\
&= \sum_i d_i \\
&= \deg_q\left(Z^d\right),
\end{aligned}$$

as desried. In the above, the first inequality follows from Exercise 14.17 and the second inequality follows from part (2) while the final two equalities follows from definition of $\deg_q(\cdot)$.

Finally we turn to part (4), which compares the (actual) degree of a polynomial to its q-degree. Again it suffices to prove for monomials. Let $d < q^m$ be given by $d = \sum_{i=0}^{m-1} d_i q^i$. Subject

to the condition $\sum_i d_i \le s(q-1)+t$, we note that d is maximized when $d_{m-1} = \cdots = d_{m-s} = (q-1)$ and $d_{m-s-1} = t$ (see Exercise 14.18) in which case we get $d + (q-t)q^{m-s-1} = q^m$, or $d = q^m - (q-t)q^{m-s-1}$. □

Our next lemma relates the degree of a multivariate function $f : \mathbb{F}_q^m \to \mathbb{F}_q$ to the q-degree of $f \circ \Phi$ for a linear bijection Φ.

Lemma 14.3.8. *For every polynomial $p \in \mathbb{F}_q[X_1,\ldots,X_m]$ and every $\mathbb{F}_q$-linear bijection, we have* $\deg_q(p \circ \Phi) \le \deg(p)$.

Proof. By Proposition 14.3.7, part (1), it suffices to prove the lemma for the special case of monomials. Fix a monomial $M(X_1,\ldots,X_m) = X_1^{r_1}\cdots X_m^{r_m}$ with $\sum_j r_j = r$. Also fix an $\mathbb{F}_q$-linear bijection $\Phi = (\Phi_1,\ldots,\Phi_m)$. Let $\tilde{M}$ denote the univariate polynomial $M \circ \Phi \pmod{Z^{q^m} - Z}$. Note that $M \circ \Phi(Z) = \prod_{i=1}^m \Phi_i(Z)^{r_i}$. And note further that $\deg_q(\Phi_i(Z)) = 1$ for every i. By Proposition 14.3.7 part (2), we now conclude that $\deg_q\left(\prod_{i=1}^m \Phi_i(Z)^{r_i}\right) \le \sum_{i=1}^m r_i = r$. Finally by Proposition 14.3.7 part (3) we have that

$$\deg_q\left(\tilde{M}\right) = \deg_q\left(M \circ \Phi \pmod{Z^{q^m} - Z}\right) \le \deg_q(M \circ \Phi) \le r,$$

as desried. □

We are now ready to prove Lemma 14.3.5 which asserts that $R_{q,m}\left(s(q-1)+t\right) = q^m - (q-t)q^{m-s-1}$.

Proof of Lemma 14.3.5. We focus on proving $R_{q,m}(s(q-1)+t) \le q^m - (q-t)q^{m-s-1}$. The other direction follows from Exercise 14.19. Fix a polynomial $p \in \mathbb{F}_q[X_1,\ldots,X_m]$ of degree at most $r = s(q-1)+t$ and consider the function $p \circ \Phi : \mathbb{F}_{q^m} \to \mathbb{F}_{q^m}$. This corresponds to the polynomial $\tilde{p}(Z) = p(\Phi_1(Z),\ldots,\Phi_m(Z)) \mod \left(Z^{q^m} - Z\right)$. By Lemma 14.3.8 we have that $\deg_q(\tilde{p}) \le r$. And by construction $\deg(\tilde{p}) < q^m$. So by Proposition 14.3.7, part (4), we have that $\deg(\tilde{p}) \le q^m - (q-t)q^{m-s-1}$, yielding the lemma.

□

14.4 Exercises

Exercise 14.1. Show that if $q > \deg(f)$, then for any polynomial $P \in \mathbb{F}_q[X_1,\ldots,X_m]$, there exists an m-dimensional affine form A such that $\deg(P \circ A) = \deg(P)$.

Exercise 14.2. An s-variate d-form is a polynomial $a(Z_1,\ldots,Z_m)$ of degree at most d. Note that $d = 1$ gives Definition 14.1.1. Similar to Definition 14.1.1 one can define an m-dimensional s-variate d-form $\langle a_1,\ldots,a_m\rangle$. Finally, given such a d-form analogous to Definition 14.1.2, for an m-variate polynomial P one can define $P \circ A$.

Prove that

$$\deg(P \circ A) \le d \cdot \deg(P).$$

Exercise 14.3. Prove that for every pair $\mathbf{z}_1 \ne \mathbf{z}_2 \in \mathbb{F}_q^s$, for a random affine form A, $A(\mathbf{z}_1)$ and $A(\mathbf{z}_2)$ are distributed uniformly and independently over $\mathbb{F}_q^m$.

Exercise 14.4. Show that any two distinct polynomials in $\mathbb{F}_q[Z_1,\ldots,Z_s]$ of degree at most $r < q(s-1)$ differ in at least two positions $\mathbf{x} \in \mathbb{F}_q^s$.

Hint: Use the polynomial distance lemma (Lemma 9.4.1).

Exercise 14.5. Let $P \in \mathbb{F}_q[X_1,\ldots,X_m]$ be a polynomial of degree at most r, $0 < \gamma < \frac{1}{2}$ and let $f : \mathbb{F}_q^m \to \mathbb{F}_q$ be such that

$$e = |\{\mathbf{x} \in \mathbb{F}_q^m \mid f(\mathbf{x}) \neq P(\mathbf{x})\}| \leq \left(\frac{1}{2} - \gamma\right) \cdot q^{m - \lceil (r+1)/(q-1) \rceil}.$$

Then, for every $\mathbf{x} \in \mathbb{F}_q^m$, the probability that LOCAL-DECODE-RM-SIMPLE-ITER$(f, \mathbf{x})$ returns $P(\mathbf{x})$ is at least $\frac{1}{2} + \gamma$.

Exercise 14.6. Let $g : \mathbb{F}_q^m \to \mathbb{F}_q$ and integer $0 \leq r < m(q-1) - 1$ be given. Then one can in $O\left(q^{3m}\right)$ operations compute a polynomial $P \in \mathbb{F}_q[X_1,\ldots,X_m]$ of degree at most r such that for every $\mathbf{x} \in \mathbb{F}_q^m$, $P(\mathbf{x}) = g(\mathbf{x})$ (if such a polynomial exists).

Hint: Use Exercise 2.6.

Exercise 14.7. If $P : \mathbb{F}_q^s \to \mathbb{F}_q$ is a polynomial of degree $r < s(q-1)$ then

$$P(\mathbf{0}) = -\sum_{\mathbf{z} \in \mathbb{F}_q^s \setminus \{\mathbf{0}\}} P(\mathbf{z}).$$

Hint: Use Exercise 9.15.

Exercise 14.8. Let $r = s(q-1) - t$. Then $\mathrm{RM}(q, s, r)$ has distance at least $(t+1)q^{m-s}$.

Exercise 14.9. Let $f : \mathbb{F}_2^m \to \mathbb{F}_2$ differ from a degree r polynomial P (of degree r) in $< 2^{m-r-1}$ locations. Show that $\mathrm{majority}_{\mathbf{b} \in \mathbb{F}_2^{m-r}}\{\sum_{\mathbf{a} \in \mathbb{F}_2^r} f(\mathbf{a}, \mathbf{b})\}$ gives the coefficient of $X_1 \cdots X_r$ in $P(X_1,\ldots,X_m)$ and $\mathrm{majority}_{\mathbf{b} \in \mathbb{F}_2^{m-r}}\{\sum_{\mathbf{a} \in \mathbb{F}_2^r} f(\mathbf{b}, \mathbf{a})\}$ (note the exchange of $\mathbf{a}$ and $\mathbf{b}$) gives the coefficient of $X_{m-r+1} \cdots X_m$ in $P(X_1,\ldots,X_m)$.

Exercise 14.10. Let $f, g : \mathbb{F}_q^m \to \mathbb{F}_q$ disagree in e positions $\mathbf{x} \in \mathbb{F}_q^m$. Then for any function $h : \mathbb{F}_q^m \to \mathbb{F}_q$, the functions $f - h$ and $g - h$ also disagree in e positions.

Exercise 14.11. Let $f, g : \mathbb{F}_q^m \to \mathbb{F}_q$ disagree in e positions $\mathbf{x} \in \mathbb{F}_q^m$. Fix a subset $S \subseteq [m]$ of size t. Argue that there are at most e values of $\mathbf{b} \in \mathbb{F}_q^{m-t}$ for which there is an $\mathbf{a} \in \mathbb{F}_q^t$ such that $f(S \leftarrow \mathbf{a}, \bar{S} \leftarrow \mathbf{b}) \neq g(S \leftarrow \mathbf{a}, \bar{S} \leftarrow \mathbf{b})$.

Exercise 14.12. Let $\Phi : \mathbb{F}_{q^m} \to \mathbb{F}_q^m$ be a linear bijection. How that

1. Φ can be viewed as m functions $(\Phi_1,\ldots,\Phi_m)$ with $\Phi_i : \mathbb{F}_{q^m} \to \mathbb{F}_q$ so that $\Phi(u) = (\Phi_1(u),\ldots,\Phi_m(u))$.

2. Furthermore each Φ_i is a linear function from $\mathbb{F}_{q^m} \to \mathbb{F}_q$ and each Φ_i can be viewed as a polynomial in $\mathbb{F}_{q^m}[Z]$.

Exercise 14.13. Show that a linear-bijection is a linearized polynomial of degree q^{m-1}.

Exercise 14.14. Argue that if a linear subspace $S \subset \mathbb{F}_q^m$ of dimension $< m$, then there $(\alpha_1,\ldots,\alpha_m) \in \mathbb{F}_q^m \setminus \{\mathbf{0}\}$ such that $\sum_{i=1}^m \alpha_i \beta_i = 0$ for every $(\beta_1,\ldots,\beta_m) \in S$.

Exercise 14.15. Argue that the polynomial $\mathrm{Tr}((\sum_i \alpha_i \lambda_i) \cdot Z)$ from (14.2) is

1. Non-zero and has degree at most q^{m-1}; and
2. Evaluates to zero on all $u \in \mathbb{F}_{q^m}$.

Exercise 14.16. Let $(d_0, d_1, \ldots)$ and $(e_0, e_1, \ldots)$ be d and e in base $q \geq 2$. Let $f = d + e = \sum_i f_i q^i$. Then show that for every i we have $\sum_{j=0}^{i} f_j \leq \sum_{j=0}^{i} (d_j + e_j)$.
Hint: Use induction.

Exercise 14.17. Show that

$$\deg_q \left(Z^{\sum_i d_i q^i} \mod \left(Z^{q^m} - Z \right) \right) \leq \deg_q \left(Z^{\sum_i d_i q^i \mod m} \right).$$

Exercise 14.18. Show that subject to the condition $\sum_{i=0}^{m-1} d_i \leq s(q-1) + t$, that $d = \sum_{i=0}^{m-1} d_i q^i$ is maximized when $d_{m-1} = \cdots = d_{m-s} = (q-1)$ and $d_{m-s-1} = t$.

Exercise 14.19. $T \subseteq \mathbb{F}_q$ be a set of size t. Consider the polynomial

$$p(X_1, \ldots, X_m) = (X_1^{q-1} - 1) \cdots (X_s^{q-1} - 1) \cdot \prod_{a \in T} (X_{s+1} - a).$$

Note that $\deg(p) = s(q-1) + t$.

1. Prove that for every linear bijection Φ, we have $\deg(p \circ \Phi) \geq q^m - (q-t) q^{m-s-1}$.
   Hint: How many zeroes does p have? What does this say about the degree of $p \circ \Phi$?
2. Conclude that $R_{q,m}(r) \geq q^m - (q-t) q^{m-s-1}$, where $r = s(q-1) + t$ where $0 \leq t < q-1$.

14.5 Bibliographic Notes

We start with the history of the Majority Logic Decoder, Algorithm 19. This algorithm is the first non-trivial algorithm in coding theory, dating back to the paper of Reed [83] from 1954. Indeed the codes were proposed by Muller [75] and the reason Reed's name is associated with the codes is due to the decoding algorithm. Despite their "in hindsight" simplicity, the algorithm is particularly clever and subtle. One proof of the subtlety is in the fact that the algorithm doesn't seem to extend immediately to non-binary settings — indeed even extending to the ternary case ends up involving some careful reasoning (and is settled in the work of Kim and Kopparty [61]).

See [103, Chapter 10] for details on near-linear time polynomial interpolation algorithm.

The Simple Reed-Muller Decoder, Algorithm 18, originates from the work of Beaver and Feigenbaum [7] and Lipton [70]. (We note that neither paper mentions the codes by name, and only consider the case $r < q$.) Subsequent works by Gemmell et al. [32] and Gemmell and Sudan [34] extended these algorithms to correct a number of errors *close to* (but not matching) half the distance of the code. These algorithms also played a central role in the theory of "locally decodable codes" that we will describe later in Chapter **??**.

Finally, the Reed-Solomon-based Decoder, Algorithm 20, is due to Pellikaan and Wu [78], who gave this algorithm in the context of "list-decoding".

Chapter 15

Efficient Decoding of Reed-Solomon Codes

So far in this book, we have shown how to efficiently decode explicit codes up to half of the Zyablov bound (Theorem 12.3.3) and how to efficiently achieve the capacity of the BSC_p (Theorem 13.4.1). The proofs of both of these results assumed that one can efficiently do unique decoding for Reed-Solomon codes up to half its distance (Theorem 12.2.1). In this chapter, we present such a unique decoding algorithm. Then we will generalize the algorithm to a list decoding algorithm that efficiently achieves the Johnson bound (Theorem 7.3.1).

15.1 Unique decoding of Reed-Solomon codes

Consider the $[n, k, d = n-k+1]_q$ Reed-Solomon code with evaluation points $(\alpha_1, \cdots, \alpha_n)$. (Recall Definition 5.2.1.) Our goal is to describe an algorithm that corrects up to $e < \frac{n-k+1}{2}$ errors in polynomial time. Let $\mathbf{y} = \left(y_1, \cdots, y_n\right) \in \mathbb{F}_q^n$ be the received word. We will now do a syntactic shift that will help us better visualize all the decoding algorithms in this chapter better. In particular, we will also think of $\mathbf{y}$ as the set of ordered pairs $\{(\alpha_1, y_1), (\alpha_2, y_2), \ldots, (\alpha_n, y_n)\}$, that is, we think of $\mathbf{y}$ as a collection of "points" in "2-D space." See Figure 15.1 for an illustration. From now on, we will switch back and forth between our usual vector interpretation of $\mathbf{y}$ and this new geometric notation.

Further, let us assume that there exists a polynomial $P(X)$ of degree at most $k-1$ such that $\Delta\left(\mathbf{y}, (P(\alpha_i))_{i=1}^n\right) \le e$. (Note that if such a $P(X)$ exists then it is unique.) See Figure 15.2 for an illustration.

We will use reverse engineering to design a unique decoding algorithm for Reed-Solomon codes. We will assume that we somehow know $P(X)$ and then prove some identities involving (the coefficients of) $P(X)$. Then to design the algorithm we will just use the identities and try to solve for $P(X)$. Towards this end, let us assume that we also magically got our hands on a polynomial $E(X)$ such that

$$E(\alpha_i) = 0 \text{ if and only if } y_i \neq P(\alpha_i).$$

$E(X)$ is called an *error-locator polynomial*. We remark that there exists such a polynomial of

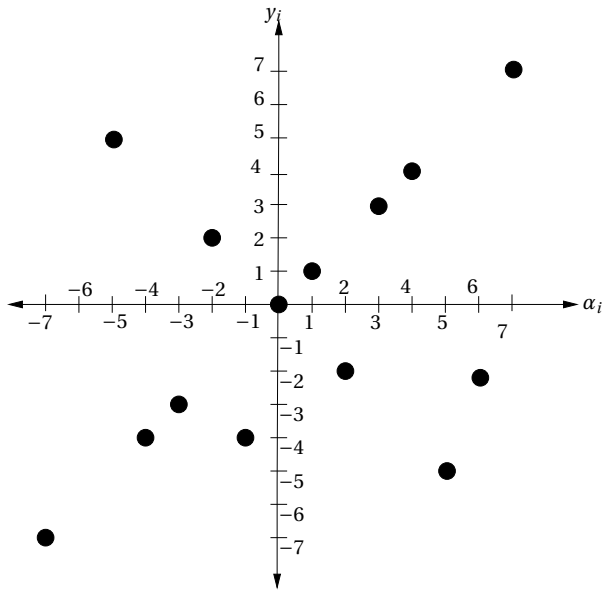


Figure 15.1: An illustration of a received word for a $[14,2]$ Reed-Solomon code (we have implicitly embedded the field $\mathbb{F}_q$ in the set $\{-7,\ldots,7\}$). The evaluations points are $(-7,-5,-4,-3,-2,-1,0,1,2,3,4,5,6,7)$ and the received word is $(-7,5,-4,-3,2,-4,0,1,-2,3,4,-5,-2,7)$.

degree at most e. In particular, consider the polynomial:

$$E(X) = \prod_{i:y_i \neq P(\alpha_i)} (X - \alpha_i).$$

See Figure 15.3 for an illustration of the $E(X)$ corresponding to the received word in Figure 15.1.

Now we claim that for every $1 \leq i \leq n$,

$$y_i E(\alpha_i) = P(\alpha_i) E(\alpha_i). \tag{15.1}$$

To see why (15.1) is true, note that if $y_i \neq P(\alpha_i)$, then both sides of (15.1) are 0 (as $E(\alpha_i) = 0$). On the other hand, if $y_i = P(\alpha_i)$, then (15.1) is obviously true.

All the discussion above does not seem to have made any progress as both $E(X)$ and $P(X)$ are unknown. Indeed, the task of the decoding algorithm is to find $P(X)$! Further, if $E(X)$ is known then one can easily compute $P(X)$ from $\mathbf{y}$ (the proof is left as an exercise). However, note that we can now try and do reverse engineering. If we think of coefficients of $P(X)$ (of which there are k) and the coefficients of $E(X)$ (of which there are $e+1$) as variables, then we have n equations from (15.1) in $e+k+1$ variables. From our bound on e, this implies we have more equations than variables. Thus, if we could solve for these unknowns, we would be done.

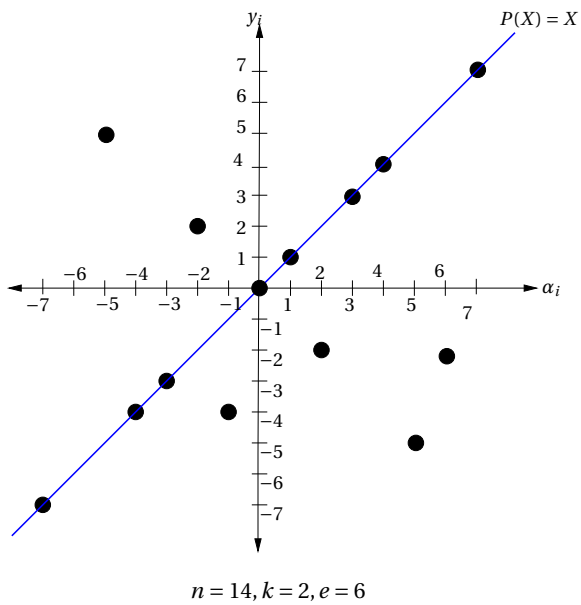


Figure 15.2: An illustration of the closest codeword $P(X) = X$ for the received word from Figure 15.1. Note that we are considering polynomials of degree 1, which are "lines."

However, there is a catch– these n equations are quadratic equations, which in general are NP-hard to solve. However, note that for our choice of e, we have $e + k - 1 \ll n$. Next, we will exploit this with a trick that is sometimes referred to as linearization. The idea is to introduce new variables so that we can convert the quadratic equations into linear equations. Care must be taken so that the number of variables after this linearization step is still smaller than the (now linear) n equations. Now we are in familiar territory as we know how to solve linear equations over a field (e.g. by Gaussian elimination). (See section 15.4 for some more discussion on the hardness of solving quadratic equations and the linearization technique.)

To perform linearization, define $N(X) \stackrel{\text{def}}{=} P(X) \cdot E(X)$. Note that $N(X)$ is a polynomial of degree less than or equal to $e + k - 1$. Further, if we can find $N(X)$ and $E(X)$, then we are done. This is because we can compute $P(X)$ as follows:

$$P(X) = \frac{N(X)}{E(X)}.$$

The main idea in the Welch-Berlekamp algorithm is to "forget" what $N(X)$ and $E(X)$ are meant to be (other than the fact that they are degree bounded polynomials).

15.1.1 Welch-Berlekamp Algorithm

Algorithm 21 formally states the Welch-Berlekamp algorithm.

As we have mentioned earlier, computing $E(X)$ is as hard as finding the solution polynomial $P(X)$. Also in some cases, finding the polynomial $N(X)$ is as hard as finding $E(X)$. E.g., given

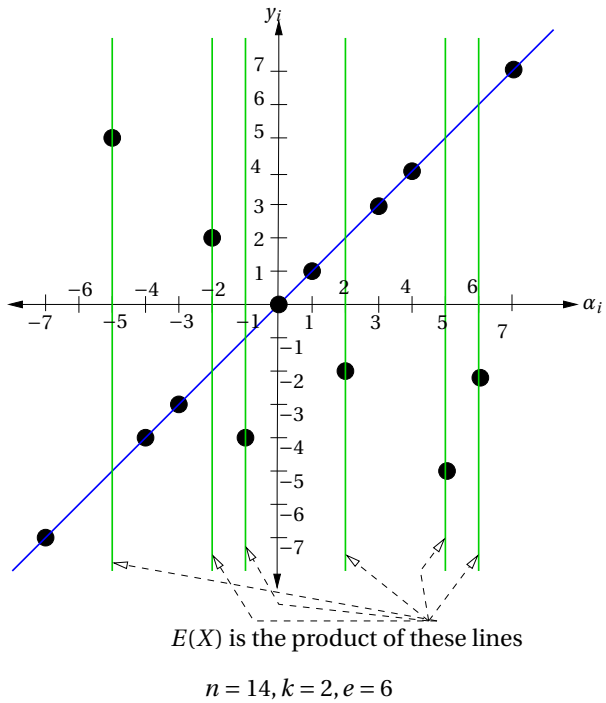


Figure 15.3: An illustration of the the error locator polynomial $E(X) = (X+5)(X+2)(X+1)(X-2)(X-5)(X-6)$ for the received word from Figure 15.1. Note that $E(X)$ is the product of the green lines.

$N(X)$ and $\mathbf{y}$ (such that $y_i \neq 0$ for $1 \leq i \leq n$) one can find the error locations by checking positions where $N(\alpha_i) = 0$. While each of the polynomials $E(X)$, $N(X)$ is hard to find individually, the main insight in the Welch-Berlekamp algorithm is that computing them together is easier.

Next we analyze the correctness and run time of Algorithm 21.

Correctness of Algorithm 21. Note that if Algorithm 21 does not output fail, then the algorithm produces a correct output. Thus, to prove the correctness of Algorithm 21, we just need the following result.

Theorem 15.1.1. *If $(P(\alpha_i))_{i=1}^n$ is transmitted (where $P(X)$ is a polynomial of degree at most $k-1$) and at most $e < \frac{n-k+1}{2}$ errors occur (i.e. $\Delta(\mathbf{y}, (P(\alpha_i))_{i=1}^n) \leq e$), then the Welch-Berlekamp algorithm outputs $P(X)$.*

The proof of the theorem above follows from the subsequent claims.

Claim 15.1.2. *There exist a pair of polynomials $E^*(X)$ and $N^*(X)$ that satisfy* Step 1 *such that $\frac{N^*(X)}{E^*(X)} = P(X)$.*

Note that now it suffices to argue that $\frac{N_1(X)}{E_1(X)} = \frac{N_2(X)}{E_2(X)}$ for any pair of solutions $((N_1(X), E_1(X))$ and $(N_2(X), E_2(X))$ that satisfy Step 1, since Claim 15.1.2 above can then be used to see that ratio must be $P(X)$. Indeed, we will show this to be the case:

Algorithm 21 Welch-Berlekamp Algorithm

INPUT: $n \geq k \geq 1$, $0 < e < \frac{n-k+1}{2}$ and n pairs $\{(\alpha_i, y_i)\}_{i=1}^n$ with α_i distinct
OUTPUT: Polynomial $P(X)$ of degree at most $k-1$ or fail.

1: Compute a non-zero polynomial $E(X)$ of degree exactly e, and a polynomial $N(X)$ of degree at most $e+k-1$ such that
$$y_i E(\alpha_i) = N(\alpha_i) \quad 1 \leq i \leq n. \tag{15.2}$$
2: IF $E(X)$ and $N(X)$ as above do not exist or $E(X)$ does not divide $N(X)$ THEN
3: RETURN fail
4: $P(X) \leftarrow \frac{N(X)}{E(X)}$.
5: IF $\Delta(\mathbf{y}, (P(\alpha_i))_{i=1}^n) > e$ THEN
6: RETURN fail
7: ELSE
8: RETURN $P(X)$

Claim 15.1.3. *If any two distinct solutions* $(E_1(X), N_1(X)) \neq (E_2(X), N_2(X))$ *satisfy* Step 1, *then they will satisfy*
$$\frac{N_1(X)}{E_1(X)} = \frac{N_2(X)}{E_2(X)}.$$

Proof of Claim 15.1.2. We just take $E^*(X)$ to be the error-locating polynomial for $P(X)$ and let $N^*(X) = P(X)E^*(X)$ where $\deg(N^*(X)) \leq \deg(P(X)) + \deg(E^*(X)) \leq e+k-1$. In particular, define $E^*(X)$ as the following polynomial of degree exactly e:
$$E^*(X) = X^{e-\Delta\left(\mathbf{y},(P(\alpha_i))_{i=1}^n\right)} \prod_{1 \leq i \leq n \mid y_i \neq P(\alpha_i)} (X - \alpha_i). \tag{15.3}$$

By definition, $E^*(X)$ is a non-zero polynomial of degree e with the following property:
$$E^*(\alpha_i) = 0 \quad \text{iff} \quad y_i \neq P(\alpha_i).$$

We now argue that $E^*(X)$ and $N^*(X)$ satisfy (15.2). Note that if $E^*(\alpha_i) = 0$, then $N^*(\alpha_i) = P(\alpha_i)E^*(\alpha_i) = y_i E^*(\alpha_i) = 0$. When $E^*(\alpha_i) \neq 0$, we know $P(\alpha_i) = y_i$ and so we still have $P(\alpha_i)E^*(\alpha_i) = y_i E^*(\alpha_i)$, as desired. □

Proof of Claim 15.1.3. Note that the degrees of the polynomials $N_1(X)E_2(X)$ and $N_2(X)E_1(X)$ are at most $2e+k-1$. Let us define polynomial $R(X)$ with degree at most $2e+k-1$ as follows:
$$R(X) = N_1(X)E_2(X) - N_2(X)E_1(X). \tag{15.4}$$

Furthermore, from Step 1 we have, for every $i \in [n]$,
$$N_1(\alpha_i) = y_i E_1(\alpha_i) \quad \text{and} \quad N_2(\alpha_i) = y_i E_2(\alpha_i). \tag{15.5}$$

Substituting (15.5) into (15.4) we get for $1 \le i \le n$:

$$\begin{aligned} R(\alpha_i) &= (y_i E_1(\alpha_i))E_2(\alpha_i) - (y_i E_2(\alpha_i))E_1(\alpha_i) \\ &= 0. \end{aligned}$$

The polynomial $R(X)$ has n roots and

$$\begin{aligned} \deg(R(X)) &\le e + k - 1 + e \\ &= 2e + k - 1 \\ &< n, \end{aligned}$$

Where the last inequality follows from the upper bound on e. Since $\deg(R(X)) < n$, by the degree mantra (Proposition 5.2.3) we have $R(X) \equiv 0$. This implies that $N_1(X)E_2(X) \equiv N_2(X)E_1(X)$. Note that as $E_1(X) \neq 0$ and $E_2(X) \neq 0$, this implies that $\frac{N_1(X)}{E_1(X)} = \frac{N_2(X)}{E_2(X)}$, as desired. □

Implementation of Algorithm 21. In Step 1, $N(X)$ has $e+k$ unknowns and $E(X)$ has $e+1$ unknowns. For each $1 \le i \le n$, the constraint in (15.2) is a linear equation in these unknowns. Thus, we have a system of n linear equations in $2e+k+1 < n+2$ unknowns. By claim 15.1.2, this system of equations have a solution. The only extra requirement is that the degree of the polynomial $E(X)$ should be exactly e. We have already shown $E(X)$ in equation (15.3) to satisfy this requirement. So we add a constraint that the coefficient of X^e in $E(X)$ is 1. Therefore, we have $n+1$ linear equation in at most $n+1$ variables, which we can solve in time $O(n^3)$, e.g. by Gaussian elimination.

Finally, note that Step 4 can be implemented in time $O(n^3)$ by "long division." Thus, we have proved

Theorem 15.1.4. *For any $[n,k]_q$ Reed-Solomon code, unique decoding can be done in $O(n^3)$ time up to $\frac{d-1}{2} = \frac{n-k}{2}$ number of errors.*

Recall that the above is a restatement of the error decoding part of Theorem 12.2.1. Thus, this fills in the final missing piece from the proofs of Theorem 12.3.3 (decoding certain concatenated codes up to half of their design distance) and Theorem 13.4.1 (efficiently achieving the BSC_p capacity).

15.2 List Decoding Reed-Solomon Codes

Recall Question 7.4.3, which asks if there is an efficient list decoding algorithm for a code of rate $R > 0$ that can correct $1 - \sqrt{R}$ fraction of errors. Note that in the question above, explicitness is not an issue as e.g., a Reed-Solomon code of rate R by the Johnson bound is $(1-\sqrt{R}, O(n^2))$-list decodable (Theorem 7.3.1).

We will study an efficient list decoding algorithm for Reed-Solomon codes that can correct up to $1 - \sqrt{R}$ fraction of errors. To this end, we will present a sequence of algorithms for (list) decoding Reed-Solomon codes that ultimately will answer Question 7.4.3.

Before we talk about the algorithms, we restate the (list) decoding problem for Reed-Solomon codes. Consider any $[n,k]_q$ Reed-Solomon code that has the evaluation set $\{\alpha_1,\ldots,\alpha_m\}$. Below is a formal definition of the decoding problem for Reed-Solomon codes:

- **Input**: Received word $\mathbf{y} = (y_1,\ldots,y_n)$, $y_i \in \mathbb{F}_q$ and error parameter $e = n - t$.
- **Output**: All polynomials $P(X) \in \mathbb{F}_q[X]$ of degree at most $k-1$ such that $P(\alpha_i) = y_i$ for at least t values of i.

Our main goal of course is to make t as small as possible.

We begin with the unique decoding regime, where $t > \dfrac{n+k}{2}$. We looked at the Welch-Berlekamp algorithm in Algorithm 21, which we restate below in a slightly different form (that will be useful in developing the subsequent list decoding algorithms).

- **Step 1:** Find polynomials $N(X)$ of degree $k+e-1$, and $E(X)$ of degree e such that

$$N(\alpha_i) = y_i E(\alpha_i), \text{ for every } 1 \le i \le n$$

- **Step 2:** If $Y - P(X)$ divides $Q(X,Y) = YE(X) - N(X)$, then output $P(X)$ (assuming $\Delta(\mathbf{y}, (P(\alpha_i))_{i=1}^n) \le e$).

Note that $Y - P(X)$ divides $Q(X,Y)$ in **Step 2** above if and only if $P(X) = \frac{N(X)}{E(X)}$, which is exactly what Step 4 does in Algorithm 21.

15.2.1 Structure of list decoding algorithms for Reed-Solomon

Note that the Welch-Berlekamp Algorithm has the following general structure:

- **Step 1:** (Interpolation Step) Find non-zero $Q(X,Y)$ such that $Q(\alpha_i, y_i) = 0, 1 \le i \le n$.
- **Step 2:** (Root Finding Step) If $Y - P(X)$ is a factor of $Q(X,Y)$, then output $P(X)$ (assuming it is close enough to the received word).

In particular, in the Welch-Berlekamp algorithm we have $Q(X,Y) = YE(X) - N(X)$ and hence, **Step 2** is easy to implement.

All the list decoding algorithms that we will consider in this chapter will have the above two-step structure. The algorithms will differ in how exactly **Step 1** is implemented. Before we move on to the details, we make one observation that will effectively "take care of" **Step 2** for us. Note that **Step 2** can be implemented if one can factorize the bivariate polynomial $Q(X,Y)$ (and then only retain the linear factors of the form $Y - P(X)$). Fortunately, it is known that factoring bivariate polynomials can be done in polynomial time (see e.g. [59]). We will not prove this result in the book but will use this fact as a given.

Finally, to ensure the correctness of the two-step algorithm above for Reed-Solomon codes, we will need to ensure the following:

- **Step 1** requires solving for the co-efficients of $Q(X,Y)$. This can be done as long as the number of coefficients is greater than the the number of constraints. (The proof of this fact is left as an exercise.) Also note that this argument is a departure from the corresponding argument for the Welch-Berlekamp algorithm (where the number of coefficients is upper bounded by the number of constraints).

- In **Step 2**, to ensure that for every polynomial $P(X)$ that needs to be output $Y - P(X)$ divides $Q(X,Y)$, we will add restrictions on $Q(X,Y)$. For example, for the Welch-Berlekamp algorithm, the constraint is that $Q(X,Y)$ has to be of the form $YE(X) - N(X)$, where $E(X)$ and $N(X)$ are non-zero polynomials of degree e and at most $e+k-1$ respectively.

Next, we present the first instantiation of the algorithm structure above, which leads us to our first list decoding algorithm for Reed-Solomon codes.

15.2.2 Algorithm 1

The main insight in the list decoding algorithm that we will see is that if we carefully control the degree of the polynomial $Q(X,Y)$, then we can satisfy the required conditions that will allow us to make sure **Step 1** succeeds. Then we will see that the degree restrictions, along with the degree mantra (Proposition 5.2.3) will allow us to show **Step 2** succeeds too. The catch is in defining the correct notion of degree of a polynomial. We do that next.

First, we recall the definition of maximum degree of a variable.

Definition 15.2.1. $\deg_X(Q)$ is the maximum degree of X in $Q(X,Y)$. Similarly, $\deg_Y(Q)$ is the maximum degree of Y in $Q(X,Y)$

For example, for $Q(X,Y) = X^2Y^3 + X^4Y^2$ $\deg_X(Q) = 4$ and $\deg_Y(Q) = 3$. Given $\deg_X(Q) = a$ and $\deg_Y(Q) = b$, we can write

$$Q(X,Y) = \sum_{\substack{0 \le i \le a, \\ 0 \le j \le b}} c_{ij} X^i Y^j,$$

where the coefficients $c_{ij} \in \mathbb{F}_q$. Note that the number of coefficients is equal to $(a+1)(b+1)$.

The main idea in the first list decoding algorithm for Reed-Solomon code is to place bounds on $\deg_X(Q)$ and $\deg_Y(Q)$ for **Step 1**. The bounds are chosen so that there are enough variables to guarantee the existence of a $Q(X,Y)$ with the required properties. We will then use these bound along with the degree mantra (Proposition 5.2.3) to argue that **Step 2** works. Algorithm 22 presents the details. Note that the algorithm generalizes the Welch-Berlekamp algorithm (and follows the two step skeleton outlined above).

Correctness of Algorithm 22. To ensure the correctness of Step 1, we will need to ensure that the number of coefficients for $Q(X,Y)$ (which is $(\ell+1)(n/\ell+1)$) is larger than the number of constraints in (15.6 (which is n). Indeed,

$$(\ell+1)\cdot\left(\frac{n}{\ell}+1\right) > \ell\cdot\frac{n}{\ell} = n.$$

Algorithm 22 The First List Decoding Algorithm for Reed-Solomon Codes

INPUT: $n \geq k \geq 1$, $\ell \geq 1$, $e = n - t$ and n pairs $\{(\alpha_i, y_i)\}_{i=1}^n$
OUTPUT: (Possibly empty) list of polynomials $P(X)$ of degree at most $k-1$

1: Find a non-zero $Q(X,Y)$ with $\deg_X(Q) \leq \ell, \deg_Y(Q) \leq \dfrac{n}{\ell}$ such that

$$Q(\alpha_i, y_i) = 0, 1 \leq i \leq n. \tag{15.6}$$

2: $\mathbb{L} \leftarrow \emptyset$
3: FOR every factor $Y - P(X)$ of $Q(X,Y)$ DO
4: IF $\Delta(\mathbf{y}, (P(\alpha_i))_{i=1}^n) \leq e$ and $\deg(P) \leq k-1$ THEN
5: Add $P(X)$ to $\mathbb{L}$.
6: RETURN $\mathbb{L}$

To argue that the final $\mathbb{L}$ in Step 6 contains all the polynomials $P(X)$ that need to be output. In other words, we need to show that if $P(X)$ of degree $\leq k-1$ agrees with Y in at least t positions, then $Y - P(X)$ divides $Q(X,Y)$. Towards this end, we define

$$R(X) \stackrel{\text{def}}{=} Q(X, P(X)).$$

Note that $Y - P(X)$ divides $Q(X,Y)$ if and only if $R(X) \equiv 0$. Thus, we need to show $R(X) \equiv 0$. For the sake of contradiction, assume that $R(X) \not\equiv 0$. Note that

$$\begin{aligned} \deg(R) &\leq \deg_X(Q) + \deg(P) \cdot \deg_Y(Q) && (15.7) \\ &\leq \ell + \frac{n(k-1)}{\ell}. && (15.8) \end{aligned}$$

On the other hand, if $P(\alpha_i) = y_i$ then (15.6) implies that

$$Q(\alpha_i, y_i) = Q(\alpha_i, P(\alpha_i)) = 0.$$

Thus, α_i is a root of $R(X)$. In other words R has at least t roots. Note that the degree mantra (Proposition 5.2.3) this will lead to a contradiction if $t > \deg(R)$, which will be true if

$$t > \ell + \frac{n(k-1)}{\ell}.$$

If we pick $\ell = \sqrt{n(k-1)}$, we will have $t > 2\sqrt{n(k-1)}$. Thus, we have shown

Theorem 15.2.1. *Algorithm 22 can list decode Reed-Solomon codes of rate R from $1 - 2\sqrt{R}$ fraction of errors. Further, the algorithm can be implemented in polynomial time.*

The claim on the efficient run time follows as Step 1 can be implemented by Gaussian elimination and for Step 3, all the factors of $Q(X,Y)$ (and in particular all linear factors of the form $Y - P(X)$) can be computed using e.g. the algorithm from [59].

The bound $1 - 2\sqrt{R}$ is better than the unique decoding bound of $\frac{1-R}{2}$ for $R < 0.07$. This is still far from the $1 - \sqrt{R}$ fraction of errors guaranteed by the Johnson bound. See Figure 15.2.2 for an illustration.

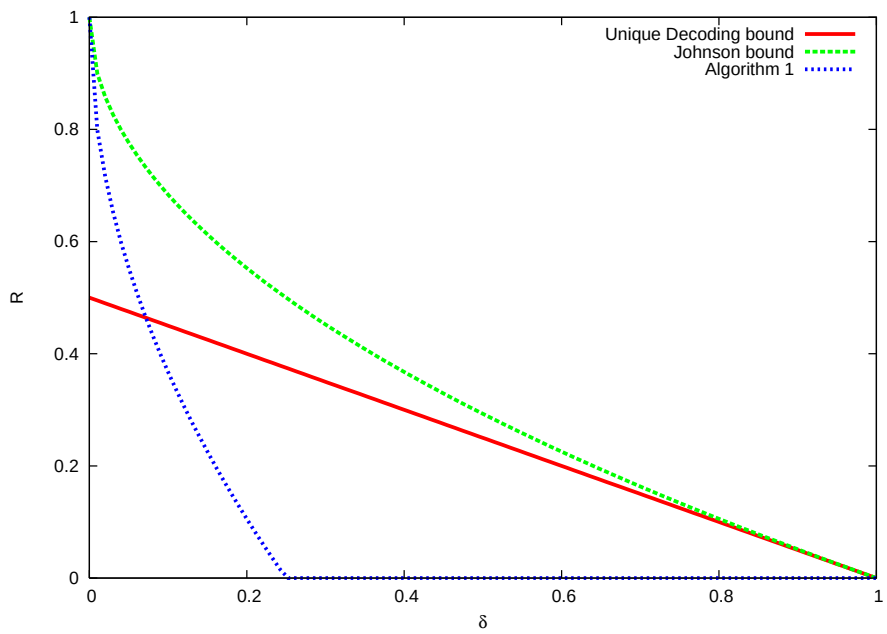


Figure 15.4: The tradeoff between rate R and the fraction of errors that can be corrected by Algorithm 22.

15.2.3 Algorithm 2

To motivate the next algorithm, recall that in Algorithm 22, in order to prove that the root finding step (Steps 3-6 in Algorithm 22) works, we defined a polynomial $R(X) \stackrel{\text{def}}{=} Q(X, P(X))$. In particular, this implied that $\deg(R) \leq \deg_X(Q) + (k-1) \cdot \deg_Y(Q)$ (and we had to select $t > \deg_X(Q) + (k-1) \cdot \deg_Y(Q)$). One shortcoming of this approach is that the maximum degree of X and Y might not occur in the same term. For example, in the polynomial $X^2Y^3 + X^4Y^2$, the maximum X and Y degrees do not occur in the same monomial. The main insight in the new algorithm is to use a more "balanced" notion of degree of $Q(X, Y)$:

Definition 15.2.2. The $(1, w)$ weighted degree of the monomial X^iY^j is $i + wj$. Further, the $(1, w)$-weighted degree of $Q(X, Y)$ (or just its $(1, w)$ degree) is the maximum $(1, w)$ weighted degree of its monomials.

For example, the $(1,2)$-degree of the polynomial $XY^3 + X^4Y$ is $\max(1+3\cdot 2, 4+2\cdot 1) = 7$. Also note that the $(1,1)$-degree of a bivariate polynomial $Q(X, Y)$ is its total degree (or the "usual" definition of degree of a bivariate polynomial). Finally, we will use the following simple lemma (whose proof we leave as an exercise):

Lemma 15.2.2. *Let $Q(X, Y)$ be a bivariate polynomial of $(1, w)$ degree D. Let $P(X)$ be a polynomial such that $\deg(P) \leq w$. Then we have*

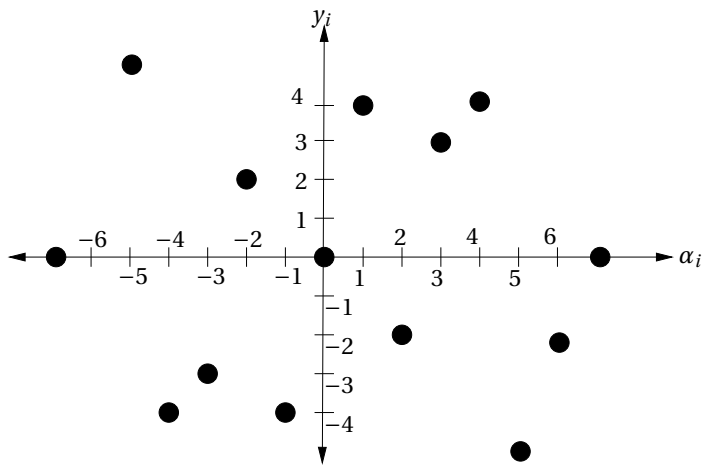
$$\deg(Q(X, P(X))) \leq D.$$

Note that a bivariate polynomial $Q(X,Y)$ of $(1,w)$ degree at most D can be represented as follows:

$$Q(X,Y) \stackrel{\text{def}}{=} \sum_{\substack{i+wj \le D \\ i,j \ge 0}} c_{i,j} X^i Y^j,$$

where $c_{i,j} \in \mathbb{F}_q$.

The new algorithm is basically the same as Algorithm 22, except that in the interpolation step, we compute a bivariate polynomial of bounded $(1,k-1)$ degree. Before we state the precise algorithm, we will present the algorithm via an example. Consider the received word in Figure 15.5.



$n = 14, k = 2, e = 9$

Figure 15.5: An illustration of a received word for the $[14,2]$ Reed-Solomon code from Figure 15.1 (where again we have implicitly embedded the field $\mathbb{F}_q$ in the set $\{-7,\ldots,7\}$). Here we have considered $e = 9$ errors which is more than what Algorithm 21 can handle. In this case, we are looking for lines that pass through at least 5 points.

Now we want to interpolate a bivariate polynomial $Q(X,Y)$ with a $(1,1)$ degree of 4 that "passes" through all the 2-D points corresponding to the received word from Figure 15.5. Figure 15.6 shows such an example.

Finally, we want to factorize all the linear factors $Y - P(X)$ of the $Q(X,Y)$ from Figure 15.6. Figure 15.7 shows the two polynomials X and $-X$ such that $Y - X$ and $Y + X$ are factors of $Q(X,Y)$ from Figure 15.6.

We now precisely state the new list decoding algorithm in Algorithm 23.

Proof of Correctness of Algorithm 23. As in the case of Algorithm 22, to prove the correctness of Algorithm 23, we need to do the following:

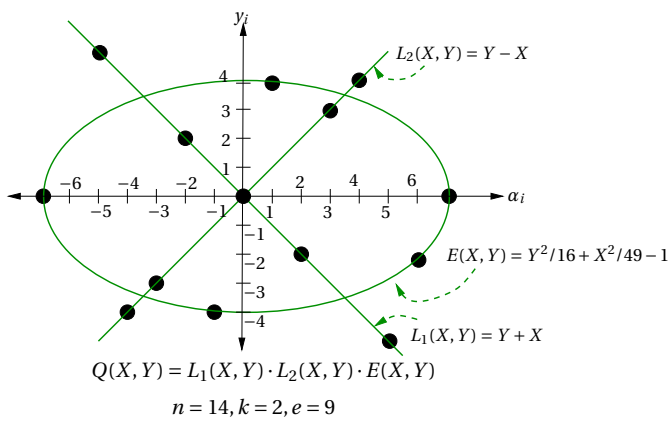


Figure 15.6: An interpolating polynomial $Q(X,Y)$ for the received word in Figure 15.5.

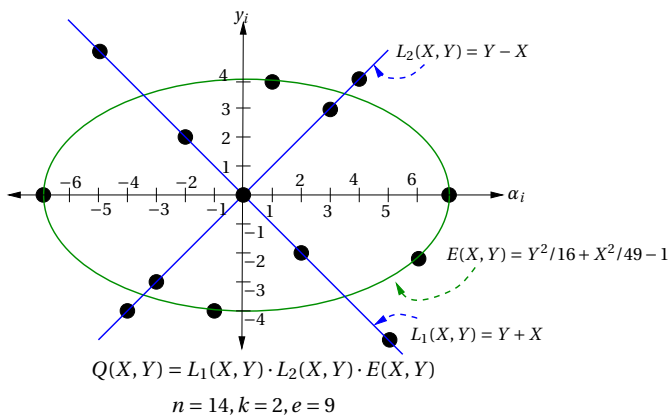


Figure 15.7: The two polynomials that need to be output are shown in blue.

- (Interpolation Step) Ensure that the number of coefficients of $Q(X,Y)$ is strictly greater than n.
- (Root Finding Step) Let $R(X) \stackrel{\text{def}}{=} Q(X, P(X))$. We want to show that if $P(\alpha_i) \geq y_i$ for at least t values of i, then $R(X) \equiv 0$.

To begin with, we argue why we can prove the correctness of the root finding step. Note that since $Q(X,Y)$ has $(1, k-1)$ degree at most D, Lemma 15.2.2 implies that

$$\deg(R) \leq D.$$

Then using the same argument as we used for the correctness of the root finding step of Algorithm 22, we can ensure $R(X) \equiv 0$ if we pick

$$t > D.$$

Algorithm 23 The Second List Decoding Algorithm for Reed-Solomon Codes

INPUT: $n \geq k \geq 1$, $D \geq 1$, $e = n - t$ and n pairs $\{(\alpha_i, y_i)\}_{i=1}^n$
OUTPUT: (Possibly empty) list of polynomials $P(X)$ of degree at most $k-1$

1: Find a non-zero $Q(X,Y)$ with $(1, k-1)$ degree at most D, such that

$$Q(\alpha_i, y_i) = 0, 1 \leq i \leq n. \tag{15.9}$$

2: $\mathbb{L} \leftarrow \emptyset$
3: FOR every factor $Y - P(X)$ of $Q(X,Y)$ DO
4: IF $\Delta(\mathbf{y}, (P(\alpha_i))_{i=1}^n) \leq e$ and $\deg(P) \leq k-1$ THEN
5: Add $P(X)$ to $\mathbb{L}$.
6: RETURN $\mathbb{L}$

Thus, we would like to pick D to be as small as possible. On the other hand, Step 1 will need D to be large enough (so that the number of variables is more than the number of constraints in (15.9). Towards that end, let the number of coefficients of $Q(X,Y)$ be

$$\mathcal{N} = \left|\left\{(i,j) | i + (k-1)j \leq D, i, j \in \mathbb{Z}^+\right\}\right|$$

To bound $\mathcal{N}$, we first note that in the definition above, $j \leq \left\lfloor \frac{D}{k-1} \right\rfloor$. (For notational convenience, define $\ell = \left\lfloor \frac{D}{k-1} \right\rfloor$.) Consider the following sequence of relationships

$$\begin{aligned}
\mathcal{N} &= \sum_{j=1}^{\ell} \sum_{i=0}^{D-(k-1)j} 1 \\
&= \sum_{j=0}^{\ell} (D - (k-1)j + 1) \\
&= \sum_{j=0}^{\ell} (D+1) - (k-1) \sum_{j=0}^{\ell} j \\
&= (D+1)(\ell+1) - \frac{(k-1)\ell(\ell+1)}{2} \\
&= \frac{\ell+1}{2}(2D + 2 - (k-1)\ell) \\
&\geq \left(\frac{\ell+1}{2}\right)(D+2) && (15.10) \\
&\geq \frac{D(D+2)}{2(k-1)}. && (15.11)
\end{aligned}$$

In the above, (15.10) follows from the fact that $\ell \leq \frac{D}{k-1}$ and (15.11) follows from the fact that $\frac{D}{k-1} - 1 \leq \ell$.

Thus, the interpolation step succeeds (i.e. there exists a non-zero $Q(X,Y)$ with the required properties) if

$$\frac{D(D+2)}{2(k-1)} > n.$$

The choice

$$D = \left\lceil \sqrt{2(k-1)n} \right\rceil$$

suffices by the following argument:

$$\frac{D(D+2)}{2(k-1)} > \frac{D^2}{2(k-1)} \geq \frac{2(k-1)n}{2(k-1)} = n.$$

Thus for the root finding step to work, we need $t > \lceil \sqrt{2(k-1)n} \rceil$, which implies the following result:

Theorem 15.2.3. *Algorithm* 2 *can list decode Reed-Solomon codes of rate R from up to* $1-\sqrt{2R}$ *fraction of errors. Further, the algorithm runs in polynomial time.*

Algorithm 2 runs in polynomial time as Step 1 can be implemented using Gaussian elimination (and the fact that the number of coefficients is $O(n)$) while the root finding step can be implemented by any polynomial time algorithm to factorize bivariate polynomials. Further, we note that $1-\sqrt{2R}$ beats the unique decoding bound of $(1-R)/2$ for $R < 1/3$. See Figure 15.2.3 for an illustration.

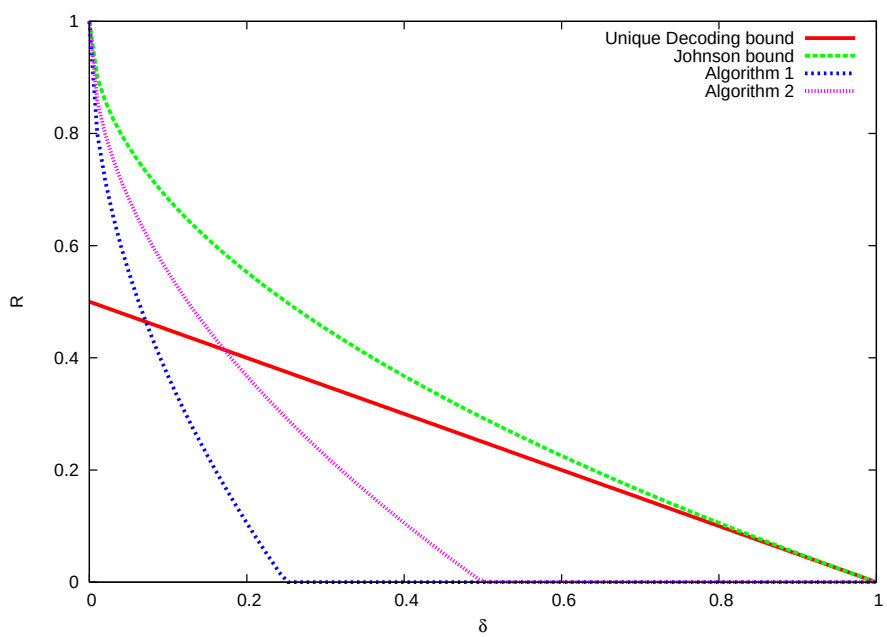


Figure 15.8: The tradeoff between rate R and the fraction of errors that can be corrected by Algorithm 22 and Algorithm 23.

15.2.4 Algorithm 3

Finally, we present the list decoding algorithm for Reed-Solomon codes, which can correct $1-\sqrt{R}$ fraction of errors. The main idea is to add more restrictions on $Q(X,Y)$ (in addition to its $(1,k-1)$-degree being at most D). This change will have the following implications:

- The number of constraints will increase but the number of coefficients will remain the same. This seems to be bad as this results in an increase in D (which in turn would result in an increase in t).

- However, this change will also increases the number of roots of $R(X)$ and this gain in the number of roots more than compensates for the increase in D.

In particular, the constraint is as follows. For some integer parameter $r \geq 1$, we will insist on $Q(X,Y)$ having r roots at $(\alpha_i, y_i), 1 \leq i \leq n$.

To motivate the definition of multiplicity of a root of a bivariate polynomial, let us consider the following simplified examples. In Figure 15.9 the curve $Q(X,Y) = Y - X$ passes through the

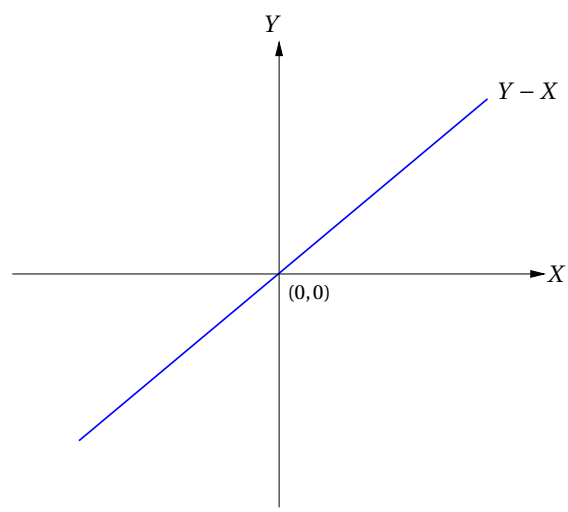


Figure 15.9: Multiplicity of 1

origin once and has no term of degree 0.

In Figure 15.10, the curve $Q(X,Y) = (Y-X)(Y+X)$ passes though the origin twice and has no term with degree at most 1.

In Figure 15.11, the curve $Q(X,Y) = (Y-X)(Y+X)(Y-2X)$ passes through the origin thrice and has no term with degree at most 2. More generally, if r lines pass through the origin, then note that the curve corresponding to their product has no term with degree at most $r-1$. This leads to the following more general definition:

Definition 15.2.3. $Q(X,Y)$ has r roots at $(0,0)$ if $Q(X,Y)$ doesn't have any monomial with degree at most $r-1$.

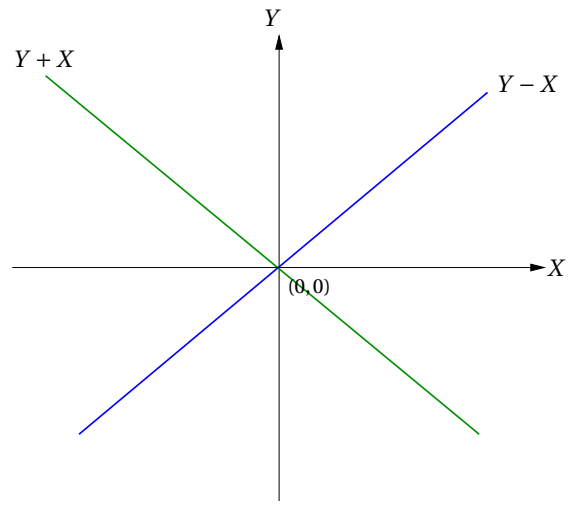


Figure 15.10: Multiplicity of 2

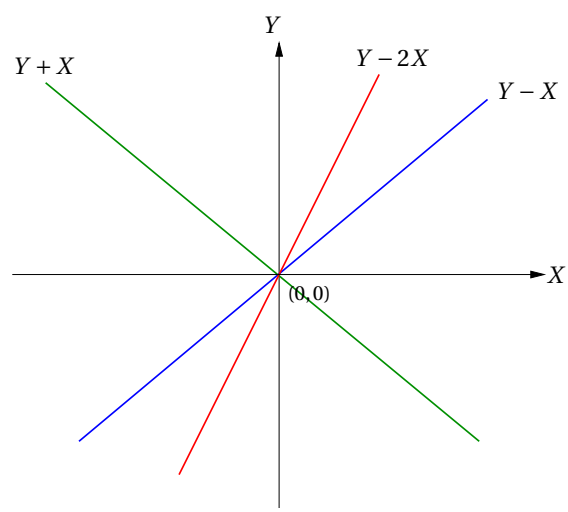


Figure 15.11: Multiplicity of 3

The definition of a root with multiplicity r at a more general point follows from a simple translation:

Definition 15.2.4. $Q(X,Y)$ has r roots at (α,β) if $Q_{\alpha,\beta}(X,Y) \stackrel{\text{def}}{=} Q(x+\alpha, y+\beta)$ has r roots at $(0,0)$.

Before we state the precise algorithm, we will present the algorithm with an example. Consider the received word in Figure 15.12.

Now we want to interpolate a bivariate polynomial $Q(X,Y)$ with $(1,1)$ degree 5 that "passes twice" through all the 2-D points corresponding to the received word from Figure 15.12. Figure 15.13 shows such an example.

Finally, we want to factorize all the linear factors $Y - P(X)$ of the $Q(X,Y)$ from Figure 15.13. Figure 15.14 shows the five polynomials of degree one are factors of $Q(X,Y)$ from Figure 15.13.

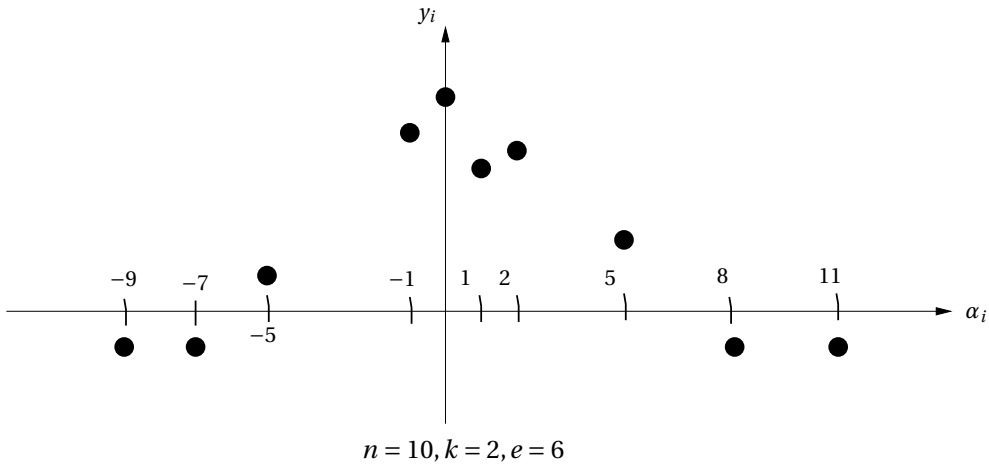


Figure 15.12: An illustration of a received word for the [10,2] Reed-Solomon code (where we have implicitly embedded the field $\mathbb{F}_q$ in the set $\{-9,\ldots,11\}$). Here we have considered $e = 6$ errors which is more than what Algorithm 23 can decode. In this case, we are looking for lines that pass through at least 4 points.

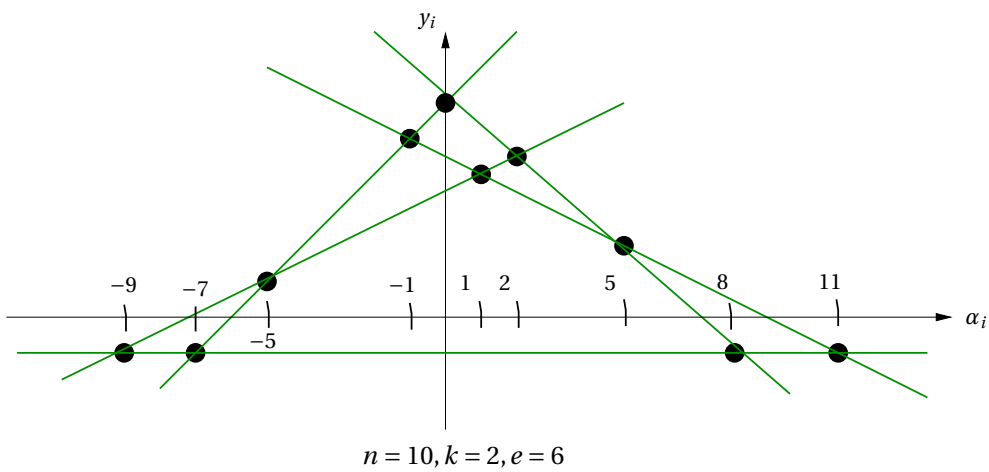
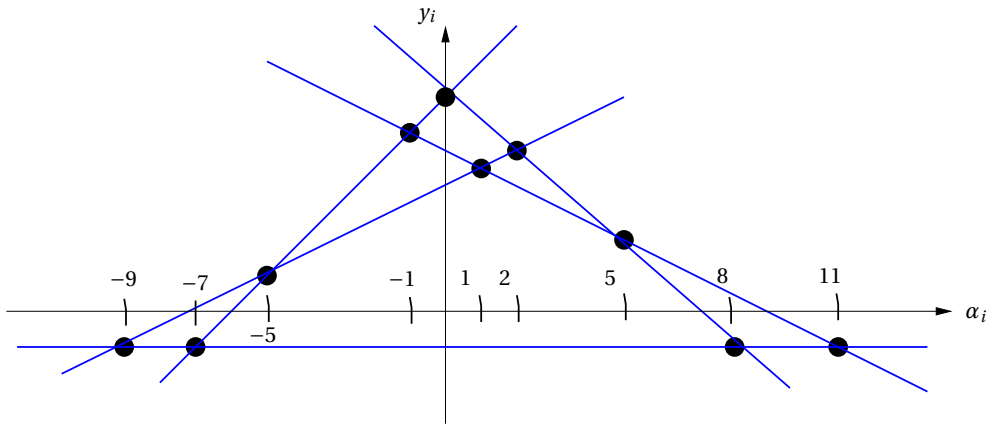


Figure 15.13: An interpolating polynomial $Q(X,Y)$ for the received word in Figure 15.12.

(In fact, $Q(X,Y)$ exactly decomposes into the five lines.)
Algorithm 24 formally states the algorithm.

Correctness of Algorithm 24. To prove the correctness of Algorithm 24, we will need the following two lemmas (we defer the proofs of the lemmas above to Section 15.2.4):

Lemma 15.2.4. *The constraints in (15.12) imply* $\binom{r+1}{2}$ *constraints for each* i *on the coefficients of* $Q(X,Y)$.



$n = 10, k = 2, e = 6$

Figure 15.14: The five polynomials that need to be output are shown in blue.

Algorithm 24 The Third List Decoding Algorithm for Reed-Solomon Codes

INPUT: $n \geq k \geq 1$, $D \geq 1$, $r \geq 1$, $e = n - t$ and n pairs $\{(\alpha_i, y_i)\}_{i=1}^n$
OUTPUT: (Possibly empty) list of polynomials $P(X)$ of degree at most $k-1$

1: Find a non-zero $Q(X,Y)$ with $(1, k-1)$ degree at most D, such that

$$Q(\alpha_i, y_i) = 0, \text{ with multiplicity } r \text{ for every } 1 \leq i \leq n. \tag{15.12}$$

2: $\mathbb{L} \leftarrow \emptyset$
3: FOR every factor $Y - P(X)$ of $Q(X,Y)$ DO
4: IF $\Delta(\mathbf{y}, (P(\alpha_i))_{i=1}^n) \leq e$ and $\deg(P) \leq k-1$ THEN
5: Add $P(X)$ to $\mathbb{L}$.
6: RETURN $\mathbb{L}$

Lemma 15.2.5. *$R(X) \stackrel{\text{def}}{=} Q(X, P(X))$ has r roots for every i such that $P(\alpha_i) = y_i$. In other words, $(X - \alpha_i)^r$ divides $R(X)$.*

Using arguments similar to those used for proving the correctness of Algorithm 23, to argue the correctness of the interpolations step we will need

$$\frac{D(D+2)}{2(k-1)} > n\binom{r+1}{2},$$

where the LHS is an upper bound on the number of coefficients of $Q(X,Y)$ as before from (15.11) and the RHS follows from Lemma 15.2.4. We note that the choice

$$D = \left\lceil \sqrt{(k-1)nr(r-1)} \right\rceil$$

works. Thus, we have shown the correctness of Step 1.

For the correctness of the root finding step, we need to show that the number of roots of $R(X)$ (which by Lemma 15.2.5 is at least rt) is strictly bigger than the degree of $R(X)$, which from Lemma 15.2.2 is D. That is we would be fine we if have,

$$tr > D,$$

which is the same as

$$t > \frac{D}{r},$$

which in turn will follow if we pick

$$t = \left\lceil \sqrt{(k-1)n\left(1-\frac{1}{r}\right)} \right\rceil.$$

If we pick $r = 2(k-1)n$, then we will need

$$t > \left\lceil \sqrt{(k-1)n - \frac{1}{2}} \right\rceil > \left\lceil \sqrt{(k-1)n} \right\rceil,$$

where the last inequality follows because of the fact that t is an integer. Thus, we have shown

Theorem 15.2.6. *Algorithm 24 can list decode Reed-Solomon codes of rate R from up to $1-\sqrt{R}$ fraction of errors. Further, the algorithm runs in polynomial time.*

The claim on the run time follows from the same argument that was used to argue the polynomial running time of Algorithm 23. Thus, Theorem 15.2.6 shows that Reed-Solomon codes can be efficiently decoded up to the Johnson bound. For an illustration of fraction of errors correctable by the three list decoding algorithms we have seen, see Figure 15.2.3.

A natural question to ask is if Reed-Solomon codes of rate R can be list decoded beyond $1-\sqrt{R}$ fraction of errors. The answer is still not known:

Open Question 15.2.1. *Given a Reed-Solomon code of rate R, can it be efficiently list decoded beyond $1-\sqrt{R}$ fraction of errors?*

Recall that to complete the proof of Theorem 15.2.6, we still need to prove Lemmas 15.2.4 and 15.2.5, which we do next.

Proof of key lemmas

Proof of Lemma 15.2.4. Let

$$Q(X,Y) = \sum_{\substack{i,j \\ i+(k-1)j \le D}} c_{i,j} X^i Y^j$$

and

$$Q_{\alpha,\beta}(X,Y) = Q(X+\alpha, Y+\beta) = \sum_{i,j} c_{i,j}^{\alpha,\beta} X^i Y^j.$$

We will show that

(i) $c_{i,j}^{\alpha,\beta}$ are homogeneous linear combinations of $c_{i,j}$'s.

(ii) If $Q_{\alpha,\beta}(X,Y)$ has no monomial with degree $< r$, then that implies $\binom{r+1}{2}$ constraints on $c_{i,j}^{\alpha,\beta}$'s.

Note that (i) and (ii) prove the lemma. To prove (i), note that by the definition:

$$Q_{\alpha,\beta}(X,Y) = \sum_{i,j} c_{i,j}^{\alpha,\beta} X^i Y^j \tag{15.13}$$

$$= \sum_{\substack{i',j' \\ i'+(k-1)j' \le D}} c_{i',j'} (X+\alpha)^{i'} (Y+\beta)^{j'} \tag{15.14}$$

Note that, if $i > i'$ or $j > j'$, then $c_{i,j}^{\alpha,\beta}$ doesn't depend on $c^{i',j'}$. By comparing coefficients of $X^i Y^j$ from (15.13) and (15.14), we obtain

$$c_{i,j}^{\alpha,\beta} = \sum_{\substack{i'>i \\ j'>j}} c_{i',j'} \binom{i'}{i} \binom{j'}{j} \alpha^i \beta^j,$$

which proves (i). To prove (ii), recall that by definition $Q_{\alpha,\beta}(X,Y)$ has no monomial of degree $< r$. In other words, we need to have constraints $c_{i,j}^{\alpha,\beta} = 0$ if $i+j \le r-1$. The number of such constraints is

$$|\{(i,j) | i+j \le r-1,\ i,j \in \mathbb{Z}^{\ge 0}\}| = \binom{r+1}{2},$$

where the equality follows from the following argument. Note that for every fixed value of $0 \le j \le r-1$, i can take $r-j$ values. Thus, we have that the number of constraints is

$$\sum_{j=0}^{r-1} r-j = \sum_{\ell=1}^{r} \ell = \binom{r+1}{2},$$

as desired. □

We now re-state Lemma 15.2.5 more precisely and then prove it.

Lemma 15.2.7. *Let $Q(X,Y)$ be computed by Step 1 in Algorithm 24. Let $P(X)$ be a polynomial of degree $\le k-1$, such that $P(\alpha_i) = y_i$ for at least $t > \frac{D}{r}$ many values of i, then $Y - P(X)$ divides $Q(X,Y)$.*

Proof. Define

$$R(X) \stackrel{\text{def}}{=} Q(X, P(X)).$$

As usual, to prove the lemma, we will show that $R(X) \equiv 0$. To do this, we will use the following claim.

Claim 15.2.8. *If $P(\alpha_i) = y_i$, then $(X - \alpha_i)^r$ divides $R(X)$, that is α_i is a root of $R(X)$ with multiplicity r.*

Note that by definition of $Q(X, Y)$ and $P(X)$, $R(X)$ has degree $\leq D$. Assuming the above claim is correct, $R(X)$ has at least $t \cdot r$ roots. Therefore, by the degree mantra (Proposition 5.2.3), $R(X)$ is a zero polynomial as $t \cdot r > D$. We will now prove Claim 15.2.8. Define

$$P_{\alpha_i, y_i}(X) \stackrel{\text{def}}{=} P(X + \alpha_i) - y_i, \tag{15.15}$$

and

$$\begin{aligned} R_{\alpha_i, y_i}(X) &\stackrel{\text{def}}{=} R(X + \alpha_i) && (15.16)\\ &= Q(X + \alpha_i, P(X + \alpha_i)) && (15.17)\\ &= Q(X + \alpha_i, P_{\alpha_i, y_i}(X) + y_i) && (15.18)\\ &= Q_{\alpha_i, y_i}(X, P_{\alpha_i, y_i}(X)), && (15.19) \end{aligned}$$

where (15.17), (15.18) and (15.19) follow from the definitions of $R(X)$, $P_{\alpha_i, y_i}(X)$ and $Q_{\alpha_i, y_i}(X, Y)$ respectively.

By (15.16) if $R_{\alpha_i, y_i}(0) = 0$, then $R(\alpha_i) = 0$. So, if X divides $R_{\alpha_i, y_i}(X)$, then $X - \alpha_i$ divides $R(X)$. (This follows from a similar argument that we used to prove Proposition 5.2.3.) Similarly, if X^r divides $R_{\alpha_i, y_i}(X)$, then $(X - \alpha_i)^r$ divides $R(X)$. Thus, to prove the lemma, we will show that X^r divides $R_{\alpha_i, y_i}(X)$. Since $P(\alpha_i) = y_i$ when α_i agrees with y_i, we have $P_{\alpha_i, y_i}(0) = 0$. Therefore, X is a root of $P_{\alpha_i, y_i}(X)$, that is, $P_{\alpha_i, y_i}(X) = X \cdot g(X)$ for some polynomial $g(X)$ of degree at most $k - 1$. We can rewrite

$$R_{\alpha_i, y_i}(X) = \sum_{i', j'} c_{i', j'}^{\alpha_i, y_i} X^{i'} (P_{\alpha_i, y_i}(X))^{j'} = \sum_{i', j'} c_{i', j'}^{\alpha_i, y_i} X^{i'} (X g(X))^{j'}.$$

Now for every i', j' such that $c_{i', j'}^{\alpha_i, y_i} \neq 0$, we have $i' + j' \geq r$ as $Q_{\alpha_i, y_i}(X, Y)$ has no monomial of degree $< r$. Thus X^r divides $R_{\alpha_i, y_i}(X)$, since $R_{\alpha_i, y_i}(x)$ has no non-zero monomial X^ℓ for any $\ell < r$. □

15.3 Extensions

We now make some observations about Algorithm 24. In particular, the list decoding algorithm is general enough to solve more general problems than just list decoding. In this section, we present an overview of these extensions.

Recall that the constraint (15.12) states that $Q(X,Y)$ has $r \geq 0$ roots at (α_i, y_i), $1 \leq i \leq n$. However, our analysis did not explicitly use the fact that the multiplicity is same for every i. In particular, given non-zero integer multiplicities $w_i \geq 0$, $1 \leq i \leq n$, Algorithm 24 can be generalized to output all polynomials $P(X)$ of degree at most $k-1$, such that

$$\sum_{i:P(\alpha_i)=y_i} w_i > \sqrt{(k-1)n\sum_{i=0}^{n}\binom{w_i+1}{2}}.$$

(We leave the proof as an exercise.) Note that till now we have seen the special case $w_i = r$, $1 \leq i \leq n$.

Further, we claim that the α_i's need not be distinct for the all of the previous arguments to go through. In particular, one can generalize Algorithm 24 even further to prove the following (the proof is left as an exercise):

Theorem 15.3.1. *Given integer weights $w_{i,\alpha}$ for every $1 \leq i \leq n$ and $\alpha \in \mathbb{F}$, in polynomial time one can output all $P(X)$ of degree at most $k-1$ such that*

$$\sum_i w_{i,P(\alpha_i)} > \sqrt{(k-1)n\sum_{i=0}^{n}\sum_{\alpha\in\mathbb{F}}\binom{w_{i,\alpha}+1}{2}}.$$

This will be useful to solve the following generalization of list decoding called soft decoding.

Definition 15.3.1. Under soft decoding problem, the decoder is given as input a set of non-negative weights $w_{i,d}(1 \leq i \leq n, \alpha \in \mathbb{F}_q)$ and a threshold $W \geq 0$. The soft decoder needs to output all codewords $(c_1, c_2, \ldots, c_n)$ in q-ary code of block length n that satisfy:

$$\sum_{i=1}^{n} w_{i,c_i} \geq W.$$

Note that Theorem 15.3.1 solve the soft decoding problem with

$$W = \sqrt{(k-1)n\sum_{i=0}^{n}\sum_{\alpha\in\mathbb{F}}\binom{w_{i,\alpha}+1}{2}}.$$

Consider the following special case of soft decoding where $w_{i,y_i} = 1$ and $w_{i,\alpha} = 0$ for $\alpha \in \mathbb{F} \setminus \{y_i\}$ $(1 \leq i \leq n)$. Note that this is exactly the list decoding problem with the received word $(y_1, \ldots, y_n)$. Thus, list decoding is indeed a special case of soft decoding. Soft decoding has practical applications in settings where the channel is analog. In such a situation, the "quantizer" might not be able to pinpoint a received symbol y_i with 100% accuracy. Instead, it can use the weight $w_{i,\alpha}$ to denote its confidence level that ith received symbol was α.

Finally, we consider a special case of soft called list recovery, which has applications in designing list decoding algorithms for concatenated codes.

Definition 15.3.2 (List Recovery)**.** Given $S_i \subseteq \mathbb{F}_q$, $1 \le i \le n$ where $|S_i| \le \ell$, output all $[n,k]_q$ codewords $(c_1,\ldots,c_n)$ such that $c_i \in S_i$ for at least t values of i. If for every valid input the number of such codewords is at most L, then the corresponding code is called $(1-t/n,\ell,L)$-list recoverable.

We leave the proof that list decoding is a special case of soft decoding as an exercise. Finally, we claim that Theorem 15.3.1 implies the following result for list recovery (the proof is left as an exercise):

Theorem 15.3.2. *Given $t > \sqrt{(k-1)\ell n}$, the list recovery problem with agreement parameter t for $[n,k]_q$ Reed-Solomon codes can be solved in polynomial time.*

15.4 Bibliographic Notes

In 1960, before polynomial time complexity was regarded as an acceptable notion of efficiency, Peterson designed an $O(N^3)$ time algorithm for the unique decoding of Reed-Solomon codes [80]. This algorithm was the first efficient algorithm for unique decoding of Reed-Solomon codes. The Berlekamp-Massey algorithm, which used shift registers for multiplication, was even more efficient, achieving a computational complexity of $O\left(N^2\right)$. Currently, an even more efficient algorithm, with a computational complexity of $O\left(N\text{poly}(\log N)\right)$, is known [82].

The Welch-Berlekamp algorithm, covered under US Patent [104], has a running time complexity of $O\left(N^3\right)$. We will follow a description of the Welch-Berlekamp algorithm provided by Gemmell and Sudan in [33].

Håstad, Philips and Safra showed that solving a system of quadratic equations (even those without any square terms like we have in (15.1)) over any field $\mathbb{F}_q$ is NP-hard [55]. (In fact, it is even hard to approximately solve this problem: i.e. where one tries to compute an assignment that satisfies as many equations as possible.) Linearization is a trick that has been used many times in theoretical computer science and cryptography. See this blog post by Dick Lipton for more on this.

Algorithm 23 is due to Sudan [96] and Algorithm 24 is due to Guruswami and Sudan [48]. Near-linear time implementations of these list decoding algorithms are also known [2].

It is natural to ask whether Theorem 15.3.2 is tight for list recovery, i.e. generalize Open Question 15.2.1 to list recovery. It was shown by Guruswami and Rudra that Theorem 15.3.2 is indeed the best possible list recovery result for Reed-Solomon codes [43]. Thus, any algorithm that answers Open Question 15.2.1 in some sense has to exploit the fact that in the list decoding problem, the α_i's are distinct. Recently it was shows by Rudra and Wootters that at least combinatorially, Reed-Solomon codes (with random evaluation points) are list decodable beyond the Johnson bound [87]. On the flip side, there are limits known on list decoding Reed-Solomon codes (both unconditional ones due to Ben-Sasson et al. [8] as well as those based on hardness of computing discrete log in the worst-case due to Cheng and Wan [16]) but none of them are close to the Johnson bound (especially for constant rate Reed-Solomon codes).

Chapter 16

Efficiently Achieving List Decoding Capacity

In the previous chapters, we have seen these results related to list decoding:

- Reed-Solomon codes of rate $R > 0$ can be list-decoded in polynomial time from $1 - \sqrt{R}$ errors (Theorem 15.2.6). This is the best algorithmic list decoding result we have seen so far.
- There exist codes of rate $R > 0$ that are $\left(1 - R - \varepsilon, O\left(\frac{1}{\varepsilon}\right)\right)$-list decodable for $q \geq 2^{\Omega(\frac{1}{\varepsilon})}$ (and in particular for $q = \text{poly}(n)$) (Theorem 7.4.1 and Proposition 3.3.2). This of course is the best possible combinatorial result.

Note that there is a gap between the algorithmic result and the best possible combinatorial result. This leads to the following natural question:

Question 16.0.1. *Are there explicit codes of rate $R > 0$ that can be list-decoded in polynomial time from $1 - R - \varepsilon$ fraction of errors for $q \leq poly(n)$?*

In this chapter, we will answer Question 16.0.1 in the affirmative.

16.1 Folded Reed-Solomon Codes

We will now introduce a new type of code called the Folded Reed-Solomon codes. These codes are constructed by combining every m consecutive symbols of a regular Reed-Solomon code into one symbol from a larger alphabet. Note that we have already seen such a folding trick when we instantiated the outer code in the concatenated code that allowed us to efficiently achieve the BSC_p capacity (Section 13.4.1). For a Reed-Solomon code that maps $\mathbb{F}_q^k \to \mathbb{F}_q^n$, the corresponding Folded Reed-Solomon code will map $\mathbb{F}_q^k \to \mathbb{F}_{q^m}^{n/m}$. We will analyze Folded Reed-Solomon codes that are derived from Reed-Solomon codes with evaluation $\{1, \gamma, \gamma^2, \gamma^3, \ldots, \gamma^{n-1}\}$,

$f(1)$	$f(\gamma)$	$f(\gamma^2)$	$f(\gamma^3)$	$\cdots$	$f(\gamma^{n-2})$	$f(\gamma^{n-1})$

Figure 16.1: Encoding $f(X)$ of degree $\leq k-1$ and coefficients in $\mathbb{F}_q$ corresponding to the symbols in the message.

where γ is the generator of $\mathbb{F}_q^*$ and $n \leq q-1$. Note that in the Reed-Solomon code, a message is encoded as in Figure 16.1.

For $m = 2$, the conversion from Reed-Solomon to Folded Reed-Solomon can be visualized as in Figure 16.2 (where we assume n is even).

$f(1)$	$f(\gamma)$	$f(\gamma^2)$	$f(\gamma^3)$	$\cdots$	$f(\gamma^{n-2})$	$f(\gamma^{n-1})$

$$\Downarrow$$

$f(1)$ $f(\gamma)$	$f(\gamma^2)$ $f(\gamma^3)$	$\cdots$	$f(\gamma^{n-2})$ $f(\gamma^{n-1})$

Figure 16.2: Folded Reed-Solomon code for $m = 2$

For general $m \geq 1$, this transformation will be as in Figure 16.3 (where we assume that m divides n).

$f(1)$	$f(\gamma)$	$f(\gamma^2)$	$f(\gamma^3)$	$\cdots$	$f(\gamma^{n-2})$	$f(\gamma^{n-1})$

$$\Downarrow$$

$f(1)$ $f(\gamma)$ $\vdots$ $f(\gamma^{m-1})$	$f(\gamma^m)$ $f(\gamma^{m+1})$ $\vdots$ $f(\gamma^{2m-1})$	$f(\gamma^{2m})$ $f(\gamma^{2m+1})$ $\vdots$ $f(\gamma^{3m-1})$	$\cdots$	$f(\gamma^{n-m})$ $f(\gamma^{n-m+1})$ $\vdots$ $f(\gamma^{n-1})$

Figure 16.3: Folded Reed-Solomon code for general $m \geq 1$

More formally, here is the definition of folded Reed-Solomon codes:

Definition 16.1.1 (Folded Reed-Solomon Code)**.** The m-folded version of an $[n,k]_q$ Reed-Solomon code C (with evaluation points $\{1,\gamma,\ldots,\gamma^{n-1}\}$), call it C', is a code of block length $N = n/m$ over $\mathbb{F}_{q^m}$, where $n \leq q-1$. The encoding of a message $f(X)$, a polynomial over $\mathbb{F}_q$ of degree at most $k-1$, has as its j'th symbol, for $0 \leq j < n/m$, the m-tuple $\left(f\left(\gamma^{jm}\right), f\left(\gamma^{jm+1}\right), \cdots, f\left(\gamma^{jm+m-1}\right)\right)$.

In other words, the codewords of C' are in one-one correspondence with those of the Reed-Solomon code C and are obtained by bundling together consecutive m-tuple of symbols in codewords of C.

16.1.1 The Intuition Behind Folded Reed-Solomon Codes

We first make the simple observation that the folding trick above cannot *decrease* the list decodability of the code. (We have already seen this argument earlier in Section 13.4.1.)

Claim 16.1.1. *If the Reed-Solomon code can be list-decoded from ρ fraction of errors, then the corresponding folded Reed-Solomon code with folding parameter m can also be list-decoded from ρ fraction of errors.*

Proof. The idea is simple: If the Reed-Solomon code can be list decoded from ρ fraction of errors (by say an algorithm $\mathscr{A}$), the Folded Reed-Solomon code can be list decoded by "unfolding" the received word and then running $\mathscr{A}$ on the unfolded received word and returning the resulting set of messages. Algorithm 25 has a more precise statement.

Algorithm 25 Decoding Folded Reed-Solomon Codes by Unfolding

INPUT: $\mathbf{y} = ((y_{1,1},\ldots,y_{1,m}),\ldots,(y_{n/m,1},\ldots,y_{n/m,m})) \in \mathbb{F}_{q^m}^{n/m}$

OUTPUT: A list of messages in $\mathbb{F}_q^k$

1: $\mathbf{y}' \leftarrow (y_{1,1},\ldots,y_{1,m},\ldots,y_{n/m,1},\ldots,y_{n/m,m}) \in \mathbb{F}_q^n$.

2: RETURN $\mathscr{A}(\mathbf{y}')$

The reason why Algorithm 25 works is simple. Let $\mathbf{m} \in \mathbb{F}_q^k$ be a message. Let RS($\mathbf{m}$) and FRS($\mathbf{m}$) be the corresponding Reed-Solomon and folded Reed-Solomon codewords. Now for every $i \in [n/m]$, if $\mathrm{FRS}(\mathbf{m})_i \neq (y_{i,1},\ldots,y_{i,n/m})$ then in the worst-case for every $j \in [n/m]$, $\mathrm{RS}(\mathbf{m})_{(i-1)n/m+j} \neq y_{i,j}$: i.e. one symbol disagreement over $\mathbb{F}_{q^m}$ can lead to at most m disagreements over $\mathbb{F}_q$. See Figure 16.4 for an illustration.

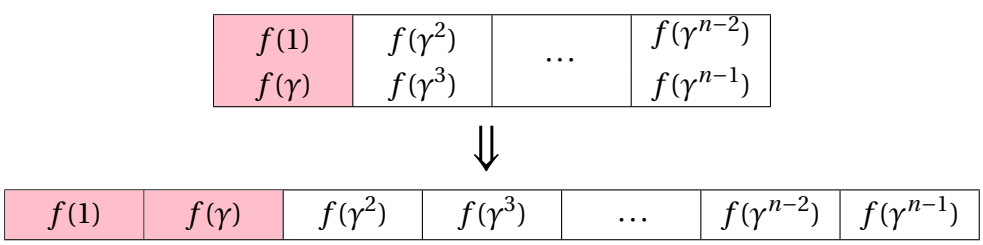


Figure 16.4: Error pattern after unfolding. A pink cell means an error: for the Reed-Solomon code it is for RS($\mathbf{m}$) with $\mathbf{y}'$ and for folded Reed-Solomon code it is for FRS($\mathbf{m}$) with $\mathbf{y}$

This implies that for any $\mathbf{m} \in \mathbb{F}_q^k$ if $\Delta(\mathbf{y}, \mathrm{FRS}(\mathbf{m})) \leq \rho \cdot \frac{n}{m}$, then $\Delta(\mathbf{y}', \mathrm{RS}(\mathbf{m})) \leq m \cdot \rho \cdot \frac{n}{m} = \rho \cdot n$, which by the properties of algorithm $\mathscr{A}$ implies that Step 2 will output $\mathbf{m}$, as desired. □

The intuition for a *strict* improvement by using Folded Reed-Solomon codes is that if the fraction of errors due to folding increases beyond what it can list-decode from, that error pattern does not need to be handled and can be ignored. For example, suppose a Reed-Solomon code that can be list-decoded from up to $\frac{1}{2}$ fraction of errors is folded into a Folded Reed-Solomon code with $m = 2$. Now consider the error pattern in Figure 16.5.

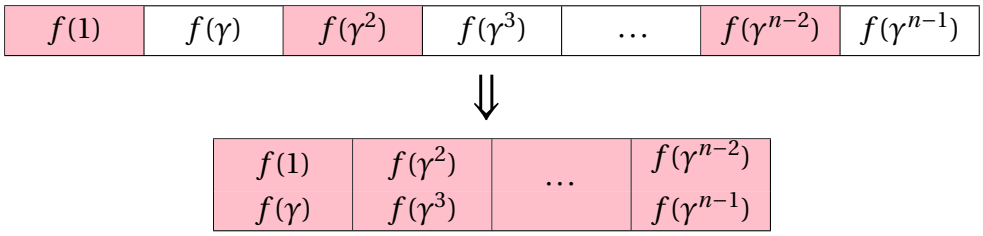


Figure 16.5: An error pattern after folding. The pink cells denotes the location of errors

The error pattern for Reed-Solomon code has $\frac{1}{2}$ fraction of errors, so any list decoding algorithm must be able to list-decode from this error pattern. However, for the Folded Reed-Solomon code the error pattern has 1 fraction of errors which is too high for the code to list-decode from. Thus, this "folded" error pattern case can be discarded from the ones that a list decoding algorithm for folded Reed-Solomon code needs to consider. This is of course one example– however, it turns out that this folding operation actually rules out *a lot* of error patterns that a list decoding algorithm for folded Reed-Solomon code needs to handle (even beyond the current best $1 - \sqrt{R}$ bound for Reed-Solomon codes). Put another way, an algorithm for folded Reed-Solomon codes has to solve the list decoding problem for the Reed-Solomon codes where the error patterns are "bunched" together (technically they're called *bursty* errors). Of course, converting this intuition into a theorem takes more work and is the subject of this chapter.

Wait a second... The above argument has a potential hole– what if we take the argument to the extreme and "cheat" by setting $m = n$ where *any* error pattern for the Reed-Solomon code will result in an error pattern with 100% errors for the Folded Reed-Solomon code: thus, we will only need to solve the problem of error detection for Reed-Solomon codes (which we can easily solve for any linear code and in particular for Reed-Solomon codes)? It is a valid concern but we will "close the loophole" by only using a constant m as the folding parameter. This will still keep q to be polynomially large in n and thus, we would still be on track to answer Question 16.0.1. Further, if we insist on smaller list size (e.g. one independent of n), then we can use code concatenation to achieve capacity achieving results for codes over smaller alphabets. (See Section 16.4 for more.)

General Codes. We would like to point out that the folding argument used above is not specific to Reed-Solomon codes. In particular, the argument for the reduction in the number of error patterns holds for any code. In fact, one can prove that for general random codes, with high

probability, folding does strictly improve the list decoding capabilities of the original code. (The proof is left as an exercise.)

16.2 List Decoding Folded Reed-Solomon Codes: I

We begin with an algorithm for list decoding folded Reed-Solomon codes that works with agreement $t \sim mRN$. Note that this is a factor m larger than the RN agreement we ultimately want. In the next section, we will see how to knock off the factor of m.

Before we state the algorithm, we formally (re)state the problem we want to solve:

- **Input**: An agreement parameter $0 \le t \le N$ and the received word:

$$\mathbf{y} = \begin{pmatrix} y_0 & y_m & & y_{n-m} \\ \vdots & \vdots & \cdots & \vdots \\ y_{m-1} & y_{2m-1} & & y_{n-1} \end{pmatrix} \in \mathbb{F}_q^{m \times N}, \quad N = \frac{n}{m}$$

- **Output**: Return all polynomials $f(X) \in \mathbb{F}_q[X]$ of degree at most $k-1$ such that for at least t values of $0 \le i < N$

$$\begin{pmatrix} f\left(\gamma^{mi}\right) \\ \vdots \\ f\left(\gamma^{m(i+1)-1}\right) \end{pmatrix} = \begin{pmatrix} y_{mi} \\ \vdots \\ y_{m(i+1)-1} \end{pmatrix} \tag{16.1}$$

The algorithm that we will study is a generalization of the Welch-Berlekamp algorithm (Algorithm 21). However unlike the previous list decoding algorithms for Reed-Solomon codes (Algorithms 22, 23 and 24), this new algorithm has more similarities with the Welch-Berlekamp algorithm. In particular, for $m = 1$, the new algorithm is *exactly* the Welch-Berlekamp algorithm. Here are the new ideas in the algorithm for the two-step framework that we have seen in the previous chapter:

- **Step 1**: We interpolate using $(m+1)$-variate polynomial, $Q(X, Y_1, \ldots, Y_m)$, where degree of each variable Y_i is exactly one. In particular, for $m = 1$, this interpolation polynomial is exactly the one used in the Welch-Berlekamp algorithm.

- **Step 2**: As we have done so far, in this step, we output all "roots" of Q. Two remarks are in order. First, unlike Algorithms 22, 23 and 24, the roots $f(X)$ are no longer simpler linear factors $Y - f(X)$, so one cannot use a factorization algorithm to factorize $Q(X, Y_1, \ldots, Y_m)$. Second, the new insight in this algorithm, is to show that all the roots form an (affine) subspace,[1] which we can use to compute the roots.

[1] An affine subspace of $\mathbb{F}_q^k$ is a set $\{\mathbf{v} + \mathbf{u} | \mathbf{u} \in S\}$, where $S \subseteq \mathbb{F}_q^k$ is a linear subspace and $\mathbf{v} \in \mathbb{F}_q^k$.

Algorithm 26 has the details.

Algorithm 26 The First List Decoding Algorithm for Folded Reed-Solomon Codes

INPUT: An agreement parameter $0 \le t \le N$, parameter $D \ge 1$ and the received word:

$$\mathbf{y} = \begin{pmatrix} y_0 & y_m & & y_{n-m} \\ \vdots & \vdots & \cdots & \vdots \\ y_{m-1} & y_{2m-1} & & y_{n-1} \end{pmatrix} \in \mathbb{F}_q^{m \times N}, \quad N = \frac{n}{m}$$

OUTPUT: All polynomials $f(X) \in \mathbb{F}_q[X]$ of degree at most $k-1$ such that for at least t values of $0 \le i < N$

$$\begin{pmatrix} f\left(\gamma^{mi}\right) \\ \vdots \\ f\left(\gamma^{m(i+1)-1}\right) \end{pmatrix} = \begin{pmatrix} y_{mi} \\ \vdots \\ y_{m(i+1)-1} \end{pmatrix} \tag{16.2}$$

1: Compute a non-zero $Q(X, Y_1, \ldots, Y_m)$ where

$$Q(X, Y_1, \ldots, Y_m) = A_0(X) + A_1(X)Y_1 + A_2(X)Y_2 + \cdots + A_m(X)Y_m$$

with $\deg(A_0) \le D + k - 1$ and $\deg(A_j) \le D$ for $1 \le j \le m$ such that

$$Q(\gamma^{mi}, y_{mi}, \cdots, y_{m(i+1)-1}) = 0, \quad \forall 0 \le i < N \tag{16.3}$$

2: $\mathbb{L} \leftarrow \emptyset$
3: FOR every $f(X) \in \mathbb{F}_q[X]$ such that $Q(X, f(X), f(\gamma X), f(\gamma^2 X), \ldots, f(\gamma^{m-1} X)) = 0$ DO
4: IF $\deg(f) \le k-1$ and $f(X)$ satisfies (16.2) for at least t values of i THEN
5: Add $f(X)$ to $\mathbb{L}$.
6: RETURN $\mathbb{L}$

Correctness of Algorithm 26. In this section, we will only concentrate on the correctness of the algorithm and analyze its error correction capabilities. We will defer the analysis of the algorithm (and in particular, proving a bound on the number of polynomials that are output by Step 6) till the next section.

We first begin with the claim that there always exists a non-zero choice for Q in Step 1 using the same arguments that we have used to prove the correctness of Algorithms 23 and 24:

Claim 16.2.1. *If $(m+1)(D+1)+k-1 > N$, then there exists a non-zero $Q(X, Y_1, \ldots Y_m)$ that satisfies the required properties of Step* 1.

Proof. As in the proof of correctness of Algorithms 22, 23 and 24, we will think of the constraints in (16.3) as linear equations. The variables are the coefficients of $A_i(X)$ for $0 \le i \le m$. With the

stipulated degree constraints on the $A_i(X)$'s, note that the number of variables participating in (16.3) is

$$D + k + m(D+1) = (m+1)(D+1) + k - 1.$$

The number of equations is N. Thus, the condition in the claim implies that we have strictly more variables then equations and thus, there exists a non-zero Q with the required properties. □

Next, we argue that the root finding step works (again using an argument very similar to those that we have seen for Algorithms 22, 23 and 24):

Claim 16.2.2. *If $t > D + k - 1$, then all polynomial $f(X) \in \mathbb{F}_q[X]$ of degree at most $k-1$ that agree with the received word in at least t positions is returned by Step* 6.

Proof. Define the univariate polynomial

$$R(X) = Q\left(X, f(X), f\left(\gamma X\right), \ldots\ldots f\left(\gamma^{m-1} X\right)\right).$$

Note that due to the degree constraints on the $A_i(X)$'s and $f(X)$, we have

$$\deg(R) \leq D + k - 1,$$

since $\deg(f(\gamma^i X)) = \deg(f(X))$. On the other hand, for every $0 \leq i < N$ where (16.1) is satisfied we have

$$R\left(\gamma^{mi}\right) = Q\left(\gamma^{mi}, y_{mi}, \ldots, y_{m(i+1)-1}\right) = 0,$$

where the first equality follows from (16.1), while the second equality follows from (16.3). Thus $R(X)$ has at least t roots. Thus, the condition in the claim implies that $R(X)$ has more roots then its degree and thus, by the degree mantra (Proposition 5.2.3) $R(X) \equiv 0$, as desired. □

Note that Claims 16.2.1 and 16.2.2 prove the correctness of the algorithm. Next we analyze the fraction of errors the algorithm can correct. Note that the condition in Claim 16.2.1 is satisfied if we pick

$$D = \left\lfloor \frac{N-k+1}{m+1} \right\rfloor.$$

This in turn implies that the condition in Claim 16.2.2 is satisfied if

$$t > \frac{N-k+1}{m+1} + k - 1 = \frac{N + m(k-1)}{m+1}.$$

Thus, the above would be satisfied if

$$t \geq \frac{N}{m+1} + \frac{mk}{m+1} = N\left(\frac{1}{m+1} + mR\left(\frac{m}{m+1}\right)\right),$$

where the equality follows from the fact that $k = mRN$.

Note that when $m = 1$, the above bound exactly recovers the bound for the Welch-Berlekamp algorithm (Theorem 15.1.4). Thus, we have shown that

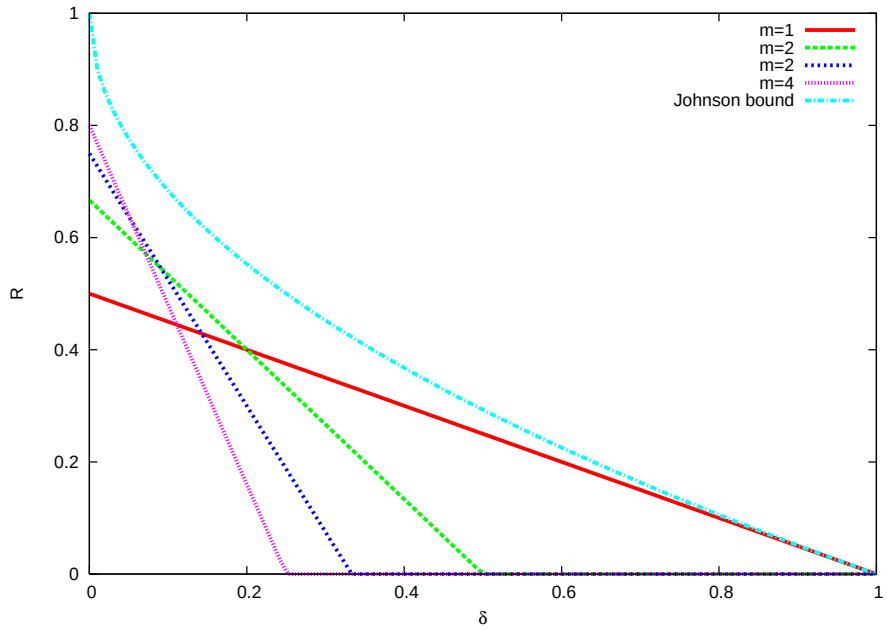


Figure 16.6: The tradeoff between rate R and the fraction of errors that can be corrected by Algorithm 26 for folding parameter $m = 1,2,3$ and 4. The Johnson bound is also plotted for comparison. Also note that the bound for $m = 1$ is the Unique decoding bound achieved by Algorithm 21.

Theorem 16.2.3. *Algorithm 26 can list decode folded Reed-Solomon code with folding parameter $m \geq 1$ and rate R up to $\frac{m}{m+1}(1 - mR)$ fraction of errors.*

See Figure 16.2 for an illustration of the tradeoff for $m = 1,2,3,4$.

Note that if we can replace the mR factor in the bound from Theorem 16.2.3 by just R then we can approach the list decoding capacity bound of $1 - R$. (In particular, we would be able to correct $1 - R - \varepsilon$ fraction of errors if we pick $m = O(1/\varepsilon)$.) Further, we need to analyze the number of polynomials output by the root finding step of the algorithm (and then analyze the runtime of the algorithm). In the next section, we show how we can "knock-off" the extra factor m (and we will also bound the list size).

16.3 List Decoding Folded Reed-Solomon Codes: II

In this section, we will present the final version of the algorithm that will allow us to answer Question 16.0.1 in the affirmative. We start off with the new idea that allows us to knock off the factor of m. (It would be helpful to keep the proof of Claim 16.2.2 in mind.)

To illustrate the idea let us consider the folding parameter to be $m = 3$. Let $f(X)$ be a polynomial of degree at most $k - 1$ that needs to be output and let $0 \leq i < N$ be a position where it agrees with the received word. (See Figure 16.7 for an illustration.)

$f(\gamma^{3i})$	$f(\gamma^{3i+1})$	$f(\gamma^{3i+2})$
y_{3i}	y_{3i+1}	y_{3i+2}

Figure 16.7: An agreement in position i.

The idea is to "exploit" this agreement over *one* $\mathbb{F}_q^3$ symbol and convert it into *two* agreements over $\mathbb{F}_{q^2}$. (See Figure 16.8 for an illustration.)

$f(\gamma^{3i})$	$f(\gamma^{3i+1})$
y_{3i}	y_{3i+1}

$f(\gamma^{3i+1})$	$f(\gamma^{3i+2})$
y_{3i+1}	y_{3i+2}

Figure 16.8: More agreement with a sliding window of size 2.

Thus, in the proof of Claim 16.2.2, for each agreement we can now get *two* roots for the polynomial $R(X)$. In general for an agreement over one $\mathbb{F}_{q^m}$ symbols translates into $m-s+1$ agreement over $\mathbb{F}_q^s$ for any $1 \le s \le m$ (by "sliding a window" of size s over the m symbols from $\mathbb{F}_q$). Thus, in this new idea the agreement is $m-s+1$ times more than before which leads to the mR term in Theorem 16.2.3 going down to $\frac{mR}{m-s+1}$. Then making s smaller than m but still large enough we can get down the relative agreement to $R+\varepsilon$, as desired. There is another change that needs to be done to make the argument go through: the interpolation polynomial Q now has to be $(s+1)$-variate instead of the earlier $(m+1)$-variate polynomial. Algorithm 27 has the details.

Correctness of Algorithm 27. Next, we analyze the correctness of Algorithm 27 as well as compute its list decoding error bound. We begin with the result showing that there exists a Q with the required properties for Step 1.

Lemma 16.3.1. *If $D \ge \left\lfloor \frac{N(m-s+1)-k+1}{s+1} \right\rfloor$, then there exists a non-zero polynomial $Q(X, Y_1, \ldots, Y_s)$ that satisfies* Step 1 *of the above algorithm.*

Proof. Let us consider all coefficients of all polynomials A_i as variables. Then the number of variables will be

$$D + k + s(D+1) = (s+1)(D+1) + k - 1.$$

On the other hand, the number of constraints in (16.5), i.e. the number of equations when all coefficients of all polynomials A_i are considered variables) will be $N(m-s+1)$.

Note that if we have more variables than equations, then there exists a non-zero Q that satisfies the required properties of Step 1. Thus, we would be done if we have:

$$(s+1)(D+1) + k - 1 > N(m-s+1),$$

which is equivalent to:

$$D > \frac{N(m-s+1)-k+1}{s+1} - 1.$$

Algorithm 27 The Second List Decoding Algorithm for Folded Reed-Solomon Codes

INPUT: An agreement parameter $0 \le t \le N$, parameter $D \ge 1$ and the received word:

$$\mathbf{y} = \begin{pmatrix} y_0 & y_m & & y_{n-m} \\ \vdots & \vdots & \cdots & \vdots \\ y_{m-1} & y_{2m-1} & & y_{n-1} \end{pmatrix} \in \mathbb{F}_q^{m \times N}, \quad N = \frac{n}{m}$$

OUTPUT: All polynomials $f(X) \in \mathbb{F}_q[X]$ of degree at most $k-1$ such that for at least t values of $0 \le i < N$

$$\begin{pmatrix} f\left(\gamma^{mi}\right) \\ \vdots \\ f\left(\gamma^{m(i+1)-1}\right) \end{pmatrix} = \begin{pmatrix} y_{mi} \\ \vdots \\ y_{m(i+1)-1} \end{pmatrix} \tag{16.4}$$

1: Compute non-zero polynomial $Q(X, Y_1, .., Y_s)$ as follows:

$$Q(X, Y_1, .., Y_s) = A_0(X) + A_1(X)Y_1 + A_2(X)Y_2 + .. + A_s(X)Y_s,$$

with $\deg[A_0] \le D + k - 1$ and $\deg[A_i] \le D$ for every $1 \le i \le s$ such that for all $0 \le i < N$ and $0 \le j \le m - s$, we have

$$Q(\gamma^{im+j}, y_{im+j}, ..., y_{im+j+s-1}) = 0. \tag{16.5}$$

2: $\mathbb{L} \leftarrow \emptyset$

3: FOR every $f(X) \in \mathbb{F}_q[X]$ such that

$$Q\left(X, f(X), f\left(\gamma X\right), f\left(\gamma^2 X\right), \ldots, f\left(\gamma^{s-1} X\right)\right) \equiv 0 \tag{16.6}$$

DO

4: IF $\deg(f) \le k-1$ and $f(X)$ satisfies (16.2) for at least t values of i THEN

5: Add $f(X)$ to $\mathbb{L}$.

6: RETURN $\mathbb{L}$

The choice of D in the statement of the claim satisfies the condition above, which complete the proof. □

Next we argue that the root finding step works.

Lemma 16.3.2. *If $t > \frac{D+k-1}{m-s+1}$, then every polynomial $f(X)$ that needs to be output satisfies (16.6).*

Proof. Consider the polynomial $R(X) = Q\left(X, f(X), f\left(\gamma X\right), ..., f\left(\gamma^{s-1} X\right)\right)$. Because the degree of $f\left(\gamma^{\ell} X\right)$ (for every $0 \le \ell \le s-1$) is at most $k-1$,

$$\deg(R) \le D + k - 1. \tag{16.7}$$

Let f(X) be one of the polynomials of degree at most $k-1$ that needs to be output, and $f(X)$ agrees with the received word at column i for some $0 \le i < N$, that is:

$$\begin{pmatrix} f\left(\gamma^{mi}\right) \\ f\left(\gamma^{mi+1}\right) \\ \cdot \\ \cdot \\ \cdot \\ f\left(\gamma^{m(i+1)-1}\right) \end{pmatrix} = \begin{pmatrix} y_{mi} \\ y_{mi+1} \\ \cdot \\ \cdot \\ \cdot \\ y_{m(i+1)-1} \end{pmatrix},$$

then for all $0 \le j \le m-s$, we have:

$$\begin{aligned} R\left(\gamma^{mi+j}\right) &= Q\left(\gamma^{mi+j}, f\left(\gamma^{mi+j}\right), f\left(\gamma^{mi+1+j}\right), \ldots, f\left(\gamma^{mi+s-1+j}\right)\right) \\ &= Q\left(\gamma^{mi+j}, y_{mi+j}, y_{mi+1+j}, \ldots, y_{mi+s-1+j}\right) = 0. \end{aligned}$$

In the above, the first equality follows as $f(X)$ agrees with $\mathbf{y}$ in column i while the second equality follows from (16.5). Thus, the number of roots of $R(X)$ is at least

$$t(m-s+1) > D+k-1 \ge \deg(R),$$

where the first inequality follows from the assumption in the claim and the second inequality follows from (16.7). Hence, by the degree mantra $R(X) \equiv 0$, which shows that $f(X)$ satisfies (16.6), as desired. □

16.3.1 Error Correction Capability

Now we analyze the the fraction of errors the algorithm above can handle. (We will come back to the thorny issue of proving a bound on the output list size for the root finding step in Section 16.3.2.)

The argument for the fraction of errors follows the by now standard route. To satisfy the constraint in Lemma 16.3.1, we pick

$$D = \left\lfloor \frac{N(m-s+1)-k+1}{s+1} \right\rfloor.$$

This along with the constraint in Lemma 16.3.2, implies that the algorithm works as long as

$$t > \left\lfloor \frac{D+k-1}{m-s+1} \right\rfloor.$$

The above is satisfied if we choose

$$t > \frac{\frac{N(m-s+1)-k+1}{s+1} + k - 1}{m-s+1} = \frac{N(m-s+1)-k+1+(k-1)(s+1)}{(m-s+1)(s+1)} = \frac{N(m-s+1)+s(k-1)}{(s+1)(m-s+1)}.$$

Thus, we would be fine if we pick

$$t > \frac{N}{s+1} + \frac{s}{s+1} \cdot \frac{k}{m-s+1} = N\left(\frac{1}{s+1} + \left(\frac{s}{s+1}\right)\left(\frac{m}{m-s+1}\right) \cdot R\right),$$

where the equality follows from the fact that $k = mRN$. This implies the following result:

Theorem 16.3.3. *Algorithm 27 can list decode folded Reed-Solomon code with folding parameter $m \geq 1$ and rate R up to $\frac{s}{s+1}(1 - mR/(m-s+1))$ fraction of errors.*

See Figure 16.3.1 for an illustration of the bound above.

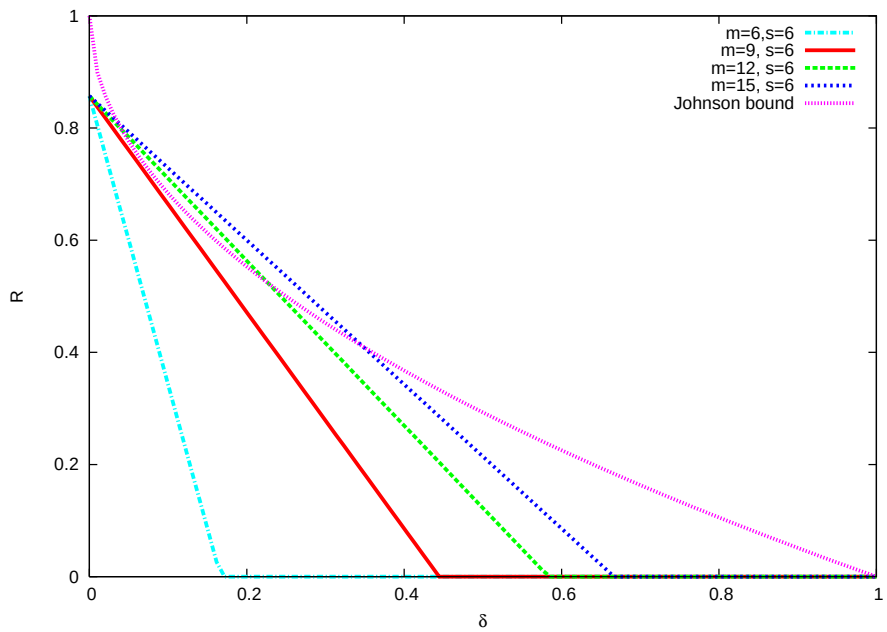


Figure 16.9: The tradeoff between rate R and the fraction of errors that can be corrected by Algorithm 27 for $s = 6$ and folding parameter $m = 6, 9, 12$ and 15. The Johnson bound is also plotted for comparison.

16.3.2 Bounding the Output List Size

We finally address the question of bounding the output list size in the root finding step of the algorithm. We will present a proof that will immediately lead to an algorithm to implement the root finding step. We will show that there are at most q^{s-1} possible solutions for the root finding step.

The main idea is the following: think of the coefficients of the output polynomial $f(X)$ as variables. Then the constraint (16.6) implies $D+k$ linear equations on these k variables. It turns out that if one picks only k out of these $D+k$ constraints, then the corresponding constraint matrix has rank at least $k-s+1$, which leads to the claimed bound. Finally, the claim on the rank

of the constraint matrix follows by observing (and this is the crucial insight) that the constraint matrix is upper triangular. Further, the diagonal elements are evaluation of a non-zero polynomial of degree at most $s-1$ in k distinct elements. By the degree mantra (Proposition 5.2.3), this polynomial can have at most $s-1$ roots, which implies that at least $k-s+1$ elements of the diagonal are non-zero, which then implies the claim. See Figure 16.10 for an illustration of the upper triangular system of linear equations.

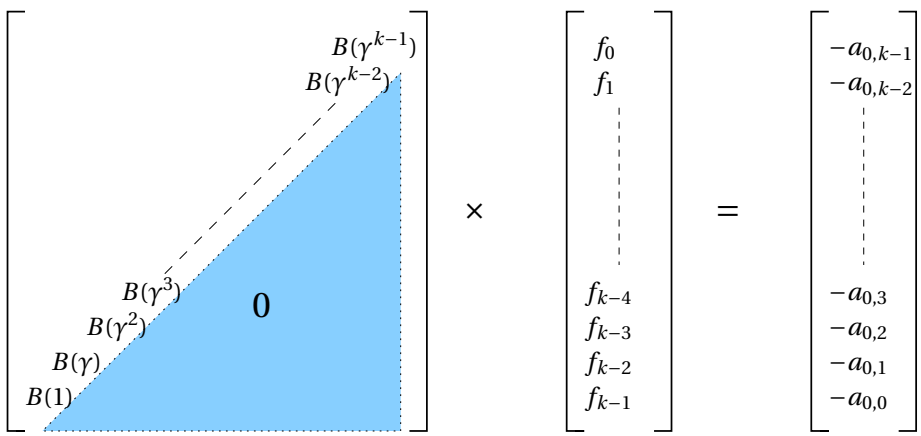


Figure 16.10: The system of linear equations with the variables $f_0, \ldots, f_{k-1}$ forming the coefficients of the polynomial $f(X) = \sum_{i=0}^{k-1} f_i X^i$ that we want to output. The constants $a_{j,0}$ are obtained from the interpolating polynomial from Step 1. $B(X)$ is a non-zero polynomial of degree at most $s-1$.

Next, we present the argument above in full detail. (Note that the constraint on (16.8) is the same as the one in (16.6) because of the constraint on the structure of Q imposed by Step 1.)

Lemma 16.3.4. *There are at most q^{s-1} solutions to $f_0, f_1, .., f_{k-1}$ (where $f(X) = f_0 + f_1 X + \ldots + f_{k-1} X^{k-1}$) to the equations*

$$A_0(X) + A_1(X) f(X) + A_2(X) f\left(\gamma X\right) + \ldots + A_s(X) f\left(\gamma^{s-1} X\right) \equiv 0 \tag{16.8}$$

Proof. First we assume that X does not divide all of the polynomials $A_0, A_1, \ldots, A_s$. Then it implies that there exists $i^* > 0$ such that the constant term of the polynomial $A_{i^*}(X)$ is not zero. (Because otherwise, since $X | A_1(X), \ldots, A_s(X)$, by (16.8), we have X divides $A_0(X)$ and hence X divide all the $A_i(X)$ polynomials, which contradicts the assumption.)

To facilitate the proof, we define few auxiliary variables a_{ij} such that

$$A_i(X) = \sum_{j=0}^{D+k-1} a_{ij} X^j \text{ for every } 0 \le i \le s,$$

and define the following univariate polynomial:

$$B(X) = a_{1,0} + a_{2,0} X + a_{3,0} X^2 + \ldots + a_{s,0} X^{s-1}. \tag{16.9}$$

Notice that $a_{i^*,0} \neq 0$, so $B(X)$ is non-zero polynomial. And because degree of $B(X)$ is at most $s-1$, by the degree mantra (Proposition 5.2.3), $B(X)$ has at most $s-1$ roots. Next, we claim the following:

Claim 16.3.5. *For every $0 \leq j \leq k-1$:*

- *If $B(\gamma^j) \neq 0$, then f_j is uniquely determined by $f_{j-1}, f_{j-2}, \ldots, f_0$.*
- *If $B(\gamma^j) = 0$, then f_j is unconstrained, i.e. f_j can take any of the q values in $\mathbb{F}_q$.*

We defer the proof of the claim above for now. Suppose that the above claim is correct. Then as γ is a generator of $\mathbb{F}_q$, $1, \gamma, \gamma^2, \ldots, \gamma^{k-1}$ are distinct (since $k-1 \leq q-2$). Further, by the degree mantra (Proposition 5.2.3) at most $s-1$ of these elements are roots of the polynomial $B(X)$. Therefore by Claim 16.3.5, the number of solutions to $f_0, f_1, \ldots, f_{k-1}$ is at most q^{s-1}. [2]

We are almost done except we need to remove our earlier assumption that X does not divide every A_i. Towards this end, we essentially just factor out the largest common power of X from all of the A_i's, and proceed with the reduced polynomial. Let $l \geq 0$ be the largest l such that $A_i(X) = X^l A'_i(X)$ for $0 \leq i \leq s$; then X does not divide all of $A'_i(X)$ and we have:

$$X^l \left(A'_0(X) + A'_1(X) f(X) + \cdots + A'_s(X) f(\gamma^{s-1} X)\right) \equiv 0.$$

Thus we can do the entire argument above by replacing $A_i(X)$ with $A'_i(X)$ since the above constraint implies that $A'_i(X)$'s also satisfy (16.8). □

Next we prove Claim 16.3.5.

Proof of Claim 16.3.5. Recall that we can assume that X does not divide all of $\{A_0(X), \ldots, A_s(X)\}$.

Let $C(X) = A_0(X) + A_1(X) f(X) + \cdots + A_s f\left(\gamma^{s-1} X\right)$. Recall that we have $C(X) \equiv 0$. If we expand out each polynomial multiplication, we have:

$$\begin{aligned}
C(X) = & a_{0,0} + a_{0,1} X + \cdots + a_{0,D+k-1} X^{D+k-1} \\
& + \left(a_{1,0} + a_{1,1} X + \cdots + a_{1,D+k-1} X^{D+k-1}\right)\left(f_0 + f_1 X + f_2 X^2 + \cdots + f_{k-1} X^{k-1}\right) \\
& + \left(a_{2,0} + a_{2,1} X + \cdots + a_{2,D+k-1} X^{D+k-1}\right)\left(f_0 + f_1 \gamma X + f_2 \gamma^2 X^2 + \cdots + f_{k-1} \gamma^{k-1} X^{k-1}\right) \\
& \vdots \\
& + \left(a_{s,0} + a_{s,1} X + \cdots + a_{s,D+k-1} X^{D+k-1}\right)\left(f_0 + f_1 \gamma^{s-1} X + f_2 \gamma^{2(s-1)} X^2 + \cdots + f_{k-1} \gamma^{(k-1)(s-1)} X^{k-1}\right)
\end{aligned} \tag{16.10}$$

Now if we collect terms of the same degree, we will have a polynomial of the form:

$$C(X) = c_0 + c_1 X + c_2 X^2 + \cdots + c_{D+k-1} X^{D+k-1}.$$

[2] Build a "decision tree" with f_0 as the root and f_j in the jth level: each edge is labeled by the assigned value to the parent node variable. For any internal node in the jth level, if $B(\gamma^j) \neq 0$, then the node has a single child with the edge taking the unique value promised by Claim 16.3.5. Otherwise the node has q children with q different labels from $\mathbb{F}_q$. By Claim 16.3.5, the number of solutions to $f(X)$ is upper bounded by the number of nodes in the kth level in the decision tree, which by the fact that B has at most $s-1$ roots is upper bounded by q^{s-1}.

So we have $D+k$ linear equations in variables $f_0, \ldots, f_{k-1}$, and we are seeking those solutions such that $c_j = 0$ for every $0 \le j \le D+k-1$. We will only consider the $0 \le j \le k-1$ equations. We first look at the equation for $j = 0$: $c_0 = 0$. This implies the following equalities:

$$0 = a_{0,0} + f_0 a_{1,0} + f_0 a_{2,0} + \cdots + f_0 a_{s,0} \tag{16.11}$$
$$0 = a_{0,0} + f_0 \left(a_{1,0} + a_{2,0} + \cdots + a_{s,0}\right) \tag{16.12}$$
$$0 = a_{0,0} + f_0 B(1). \tag{16.13}$$

In the above (16.11) follows from (16.10), (16.12) follows by simple manipulation while (16.13) follows from the definition of $B(X)$ in (16.9).

Now, we have two possible cases:

- **Case 1:** $B(1) \neq 0$. In this case, (16.13) implies that $f_0 = \frac{-a_{0,0}}{B(1)}$. In particular, f_0 is fixed.
- **Case 2:** $B(1) = 0$. In this case f_0 has no constraint (and hence can take on any of the q values in $\mathbb{F}_q$).

Now consider the equation for $j = 1$: $c_1 = 0$. Using the same argument as we did for $j = 0$, we obtain the following sequence of equalities:

$$\begin{aligned}
0 &= a_{0,1} + f_1 a_{1,0} + f_0 a_{1,1} + f_1 a_{2,0}\gamma + f_0 a_{2,1} + \cdots + f_1 a_{s,0}\gamma^{s-1} + f_0 a_{s,1} \\
0 &= a_{0,1} + f_1 \left(a_{1,0} + a_{2,0}\gamma + \cdots + a_{s,0}\gamma^{s-1}\right) + f_0 \left(\sum_{l=1}^{s} a_{l,1}\right) \\
0 &= a_{0,1} + f_1 B(\gamma) + f_0 b_0^{(1)} \qquad (16.14)
\end{aligned}$$

where $b_0^{(1)} = \sum_{l=1}^{s} a_{l,1}$ is a constant. We have two possible cases:

- **Case 1:** $B(\gamma) \neq 0$. In this case, by (16.14), we have $f_1 = \frac{-a_{0,1} - f_0 b_0^{(1)}}{B(\gamma)}$ and there is a unique choice for f_1 given fixed f_0.
- **Case 2:** $B(\gamma) = 0$. In this case, f_1 is unconstrained.

Now consider the case of arbitrary j: $c_j = 0$. Again using similar arguments as above, we get:

$$\begin{aligned}
0 = a_{0,j} &+ f_j(a_{1,0} + a_{2,0}\gamma^j + a_{3,0}\gamma^{2j} + \cdots + a_{s,0}\gamma^{j(s-1)}) \\
&+ f_{j-1}(a_{1,1} + a_{2,1}\gamma^{j-1} + a_{3,1}\gamma^{2(j-1)} + \cdots + a_{s,1}\gamma^{(j-1)(s-1)}) \\
&\qquad \vdots \\
&+ f_1(a_{1,j-1} + a_{2,j-1}\gamma + a_{3,j-1}\gamma^2 + \cdots + a_{s,j-1}\gamma^{s-1}) \\
&+ f_0(a_{1,j} + a_{2,j} + a_{3,j} + \cdots + a_{s,j}) \\
0 = a_{0,j} &+ f_j B(\gamma^j) + \sum_{l=0}^{j-1} f_l b_l^{(j)} \qquad (16.15)
\end{aligned}$$

where $b_l^{(j)} = \sum_{\iota=1}^{s} a_{\iota,j-l} \cdot \gamma^{l(\iota-1)}$ are constants for $0 \le j \le k-1$.

We have two possible cases:

- **Case 1:** $B(\gamma^j) \neq 0$. In this case, by (16.15), we have

$$f_j = \frac{-a_{0,j} - \sum_{l=0}^{j-1} f_l b_l^{(j)}}{B(\gamma^j)} \tag{16.16}$$

  and there is a unique choice for f_j given fixed $f_0, \ldots, f_{j-1}$.

- **Case 2:** $B(\gamma^j) = 0$. In this case f_j is unconstrained.

This completes the proof. □

We now revisit the proof above and make some algorithmic observations. First, we note that to compute all the tuples $(f_0, \ldots, f_{k-1})$ that satisfy (16.8) one needs to solve the linear equations (16.15) for $j = 0, \ldots, k-1$. One can state this system of linear equation as (see also Figure 16.10)

$$C \cdot \begin{pmatrix} f_0 \\ \vdots \\ f_{k-1} \end{pmatrix} = \begin{pmatrix} -a_{0,k-1} \\ \vdots \\ -a_{0,0} \end{pmatrix},$$

where C is a $k \times k$ upper triangular matrix. Further each entry in C is either a 0 or $B(\gamma^j)$ or $b_l^{(j)}$– each of which can be computed in $O(s \log s)$ operations over $\mathbb{F}_q$. Thus, we can setup this system of equations in $O\left(s \log s k^2\right)$ operations over $\mathbb{F}_q$.

Next, we make the observation that all the solutions to (16.8) form an affine subspace. Let $0 \leq d \leq s-1$ denote the number of roots of $B(X)$ in $\{1, \gamma, \ldots, \gamma^{k-1}\}$. Then since there will be d unconstrained variables among $f_0, \ldots, f_{k-1}$ (one of every j such that $B(\gamma^j) = 0$), it is not too hard to see that all the solutions will be in the set $\left\{M \cdot \mathbf{x} + \mathbf{z} | \mathbf{x} \in \mathbb{F}_q^d\right\}$, for some $k \times d$ matrix M and some $\mathbf{z} \in \mathbb{F}_q^k$. Indeed every $\mathbf{x} \in \mathbb{F}_q^d$ corresponds to an assignment to the d unconstrained variables among $f_0, \ldots, f_j$. The matrix M and the vector $\mathbf{z}$ are determined by the equations in (16.16). Further, since C is upper triangular, both M and $\mathbf{z}$ can be computed with $O\left(k^2\right)$ operations over $\mathbb{F}_q$.

The discussion above implies the following:

Corollary 16.3.6. *The set of solutions to (16.8) are contained in an affine subspace* $\left\{M \cdot \mathbf{x} + \mathbf{z} | \mathbf{x} \in \mathbb{F}_q^d\right\}$ *for some* $0 \leq d \leq s-1$ *and* $M \in \mathbb{F}_q^{k \times d}$ *and* $\mathbf{z} \in \mathbb{F}_q^k$. *Further,* M *and* $\mathbf{z}$ *can be computed from the polynomials* $A_0(X), \ldots, A_s(X)$ *with* $O(s \log s k^2)$ *operations over* $\mathbb{F}_q$.

16.3.3 Algorithm Implementation and Runtime Analysis

In this sub-section, we discuss how both the interpolation and root finding steps of the algorithm can be implemented and analyze the run time of each step.

Step 1 involves solving Nm linear equation in $O(Nm)$ variables and can e.g. be solved by Gaussian elimination in $O((Nm)^3)$ operations over $\mathbb{F}_q$. This is similar to what we have seen for Algorithms 22, 23 and 24. However, the fact that the interpolation polynomial has total degree

of one in the variables $Y_1, \ldots, Y_s$ implies a much faster algorithm. In particular, one can perform the interpolation in $O(Nm \log^2(Nm) \log\log(Nm))$ operations over $\mathbb{F}_q$.

The root finding step involves computing all the "roots" of Q. The proof of Lemma 16.3.4 actually suggests Algorithm 28.

Algorithm 28 The Root Finding Algorithm for Algorithm 27

INPUT: $A_0(X), \ldots, A_s(X)$
OUTPUT: All polynomials $f(X)$ of degree at most $k-1$ that satisfy (16.8)

1: Compute ℓ such that X^ℓ is the largest common power of X among $A_0(X), \ldots, A_s(X)$.
2: FOR every $0 \le i \le s$ DO
3: $\quad A_i(X) \leftarrow \frac{A_i(X)}{X^\ell}$.
4: Compute $B(X)$ according to (16.9)
5: Compute d, $\mathbf{z}$ and M such that the solutions to the k linear system of equations in (16.15) lie in the set $\left\{M \cdot \mathbf{x} + \mathbf{z} | \mathbf{x} \in \mathbb{F}_q^d\right\}$.
6: $\mathbb{L} \leftarrow \emptyset$
7: FOR every $\mathbf{x} \in \mathbb{F}_q^d$ DO
8: $\quad (f_0, \ldots, f_{k-1}) \leftarrow M \cdot \mathbf{x} + \mathbf{z}$.
9: $\quad f(X) \leftarrow \sum_{i=0}^{k-1} f_i \cdot X^i$.
10: $\quad$ IF $f(X)$ satisfies (16.8) THEN
11: $\quad\quad$ Add $f(X)$ to $\mathbb{L}$.
12: RETURN $\mathbb{L}$

Next, we analyze the run time of the algorithm. Throughout, we will assume that all polynomials are represented in their standard coefficient form.

Step 1 just involves figuring out the smallest power of X in each $A_i(X)$ that has a non-zero coefficient from which one can compute the value of ℓ. This can be done with $O(D + k + s(D + 1)) = O(Nm)$ operations over $\mathbb{F}_q$. Further, given the value of ℓ one just needs to "shift" all the coefficients in each of the $A_i(X)$'s to the right by ℓ, which again can be done with $O(Nm)$ operations over $\mathbb{F}_q$.

Now we move to the root finding step. The run time actually depends on what it means to "solve" the linear system. If one is happy with a succinct description of a set of possible solution that contains the actual output then one can halt Algorithm 28 after Step 5 and Corollary 16.3.6 implies that this step can be implemented in $O\left(s \log s k^2\right) = O\left(s \log s (Nm)^2\right)$ operations over $\mathbb{F}_q$. However, if one wants the actual set of polynomials that need to be output, then the only known option so far is to check all the q^{s-1} potential solutions as in Steps 7-11. (However, we'll see a twist in Section 16.4.) The latter would imply a total of $O(s \log s (Nm)^2) + O(q^{s-1} \cdot (Nm)^2)$ operations over $\mathbb{F}_q$.

Thus, we have the following result:

Lemma 16.3.7. *With $O(s \log s (Nm)^2)$ operations over $\mathbb{F}_q$, the algorithm above can return an affine subspace of dimension $s-1$ that contains all the polynomials of degree at most $k-1$*

that need to be output. Further, the exact set of solution can be computed in with additional $O(q^{s-1} \cdot (Nm)^2)$ operations over $\mathbb{F}_q$.

16.3.4 Wrapping Up

By Theorem 16.3.3, we know that we can list decode a folded Reed-Solomon code with folding parameter $m \geq 1$ up to

$$\frac{s}{s+1} \cdot \left(1 - \frac{m}{m-s+1} \cdot R\right) \tag{16.17}$$

fraction of errors for any $1 \leq s \leq m$.

To obtain our desired bound $1 - R - \varepsilon$ fraction of errors, we instantiate the parameter s and m such that

$$\frac{s}{s+1} \geq 1 - \varepsilon \text{ and } \frac{m}{m-s+1} \leq 1 + \varepsilon. \tag{16.18}$$

It is easy to check that one can choose

$$s = \Theta(1/\varepsilon) \text{ and } m = \Theta(1/\varepsilon^2)$$

such that the bounds in (16.18) are satisfied. Using the bounds from (16.18) in (16.17) implies that the algorithm can handle at least

$$(1-\varepsilon)(1-(1+\varepsilon)R) = 1 - \varepsilon - R + \varepsilon^2 R > 1 - R - \varepsilon$$

fraction of errors, as desired.

We are almost done since Lemma 16.3.7 shows that the run time of the algorithm is $q^{O(s)}$. The only thing we need to choose is q: for the final result we pick q to be the smallest power of 2 that is larger than $Nm + 1$. Then the discussion above along with Lemma 16.3.7 implies the following result (the claim on strong explicitness follows from the fact that Reed-Solomon codes are strongly explicit):

Theorem 16.3.8. *There exist* strongly explicit *folded Reed-Solomon codes of rate R that for large enough block length N can be list decoded from $1 - R - \varepsilon$ fraction of errors (for any small enough $\varepsilon > 0$) in time $\left(\frac{N}{\varepsilon}\right)^{O(1/\varepsilon)}$. The worst-case list size is $\left(\frac{N}{\varepsilon}\right)^{O(1/\varepsilon)}$ and the alphabet size is $\left(\frac{N}{\varepsilon}\right)^{O(1/\varepsilon^2)}$.*

16.4 Bibliographic Notes and Discussion

There was no improvement to the Guruswami-Sudan result (Theorem 15.2.6) for about seven years till Parvaresh and Vardy showed that "Correlated" Reed-Solomon codes can be list-decoded up to $1-(mR)^{\frac{1}{m+1}}$ fraction of errors for $m \geq 1$ [77]. Note that for $m = 1$, correlated Reed-Solomon codes are equivalent to Reed-Solomon codes and the result of Parvaresh and Vardy recovers Theorem 15.2.6. Immediately, after that Guruswami and Rudra [44] showed that Folded Reed-Solomon codes can achieve the list-decoding capacity of $1 - R - \varepsilon$ and hence, answer Question 16.0.1 in the affirmative. Guruswami [39] reproved this result but with a much simpler

proof. In this chapter, we studied the proof due to Guruswami. Guruswami in [39] credits Salil Vadhan for the the interpolation step. An algorithm presented in Brander's thesis [10] shows that for the special interpolation in Algorithm 27, one can perform the interpolation in $O(Nm\log^2(Nm)\log\log(Nm))$ operations over $\mathbb{F}_q$. The idea of using the "sliding window" for list decoding Folded Reed-Solomon codes is originally due to Guruswami and Rudra [43].

The bound of q^{s-1} on the list size for folded Reed-Solomon codes was first proven in [43] by roughly the following argument. One reduced the problem of finding roots to finding roots of a *univariate* polynomial related to Q over $\mathbb{F}_{q^k}$. (Note that each polynomial in $\mathbb{F}_q[X]$ of degree at most $k-1$ has a one to one correspondence with elements of $\mathbb{F}_{q^k}$– see e.g. Theorem 12.2.1.) The list size bound follows from the fact that this new univariate polynomial had degree q^{s-1}. Thus, implementing the algorithm entails running a root finding algorithm over a big extension field, which in practice has terrible performance.

Discussion. For constant ε, Theorem 16.3.8 answers Question 16.0.1 in the affirmative. However, from a practical point of view, there are three issues with the result: alphabet, list size and run time. Below we tackle each of these issues.

Large Alphabet. Recall that one only needs an alphabet of size $2^{O(1/\varepsilon)}$ to be able to list decode from $1-R-\varepsilon$ fraction of errors, which is independent of N. It turns out that combining Theorem 16.3.8 along with code concatenation and expanders allows us to construct codes over alphabets of size roughly $2^{O(1/\varepsilon^4)}$ [43]. (The idea of using expanders and code concatenation was not new to [43]: the connection was exploited in earlier work by Guruswami and Indyk [42].)

The above however, does not answer the question of achieving list decoding capacity for *fixed* q, say e.g. $q=2$. We know that there exists binary code of rate R that are $(H^{-1}(1-R-\varepsilon), O(1/\varepsilon))$-list decodable codes (see Theorem 7.4.1). The best known explicit codes with efficient list decoding algorithms are those achieved by concatenating folded Reed-Solomon codes with suitable inner codes achieve the so called *Blokh-Zyablov* bound [45]. However, the tradeoff is far from the list decoding capacity. As one sample point, consider the case when we want to list decode from $\frac{1}{2}-\varepsilon$ fraction of errors. Then the result of [45] gives codes of rate $\Theta(\varepsilon^3)$ while the codes on list decoding capacity has rate $\Omega(\varepsilon^2)$. The following fundamental question is still very much wide open:

Open Question 16.4.1. *Do there exist explicit binary codes with rate R that can be list decoded from $H^{-1}(1-R-\varepsilon)$ fraction of errors with polynomial list decoding algorithms?*

The above question is open even if we drop the requirement on efficient list decoding algorithm or we only ask for a code that can list decode from $1/2-\varepsilon$ fraction of errors with rate $\Omega(\varepsilon^a)$ for some $a<3$. It is known (combinatorially) that concatenated codes can achieve the list decoding capacity but the result is via a souped up random coding argument and does not give much information about an efficient decoding algorithm [46].

List Size. It is natural to wonder if the bound on the list size in Lemma 16.3.4 above can be improved as that would show that folded Reed-Solomon codes can be list decoded up to the list decoding capacity but with a smaller output list size than Theorem 16.3.8. Guruswami showed that in its full generality the bound cannot be improved [39]. In particular, he exhibits explicit polynomials $A_0(X), \ldots, A_s(X)$ such that there are at least q^{s-2} solutions for $f(X)$ that satisfy (16.8). However, these $A_i(X)$'s are not known to be the output for an actual interpolation instance. In other words, the following question is still open:

Open Question 16.4.2. *Can folded Reed-Solomon codes of rate R be list decoded from $1 - R - \varepsilon$ fraction of errors with list size $f(1/\varepsilon) \cdot N^c$ for some increasing function $f(\cdot)$ and absolute constant c?*

Even the question above with $N^{(1/\varepsilon)^{o(1)}}$ is still open.

However, if one is willing to consider codes other than folded Reed-Solomon codes in order to answer to achieve list decoding capacity with smaller list size (perhaps with one only dependent on ε), then there is good news. Guruswami in the same paper that presented the algorithm in this chapter also present a *randomized* construction of codes of rate R that are $(1 - R - \varepsilon, O(1/\varepsilon^2))$-list decodable codes [39]. This is of course worse than what we know from the probabilistic method. However, the good thing about the construction of Guruswami comes with an $O(N/\varepsilon)^{O(1/\varepsilon)}$-list decoding algorithm.

Next we briefly mention the key ingredient in the result above. To see the potential for improvement consider Corollary 16.3.6. The main observation is that all the potential solutions lie in an affine subspace of dimension $s - 1$. The key idea in [39] was use the folded Reed-Solomon encoding for a special subset of the message space $\mathbb{F}_q^k$. Call a subspace $S \subseteq \mathbb{F}_q^k$ to be a $(q, k, \varepsilon, \ell, L)$-*subspace evasive subset* if

1. $|S| \geq q^{k(1-\varepsilon)}$; and
2. For any (affine) subspace $T \subseteq \mathbb{F}_q^k$ of dimension ℓ, we have $|S \cap T| \leq L$.

The code in [39], applies the folded Reed-Solomon encoding on a $\left(q, k, s, O\left(s^2\right)\right)$-subspace evasive subset (such a subset can be shown to exist via the probabilistic method). The reason why this sub-code of folded Reed-Solomon code works is as follows: Condition (1) ensures that the new code has rate at least $R(1-\varepsilon)$, where R is the rate of the original folded Reed-Solomon code and condition (2) ensures that the number of output polynomial in the root finding step of the algorithm we considered in the last section is at most L. (This is because by Corollary 16.3.6 the output message space is an affine subspace of dimension $s - 1$ in $\mathbb{F}_Q^k$. However, in the new code by condition 2, there can be at most $O\left(s^2\right)$ output solutions.)

The result above however, has two shortcomings: (i) the code is no longer explicit and (ii) even though the worst case list size is $O\left(\frac{1}{\varepsilon^2}\right)$, it was not know how to obtain this output without listing all the q^{s-1} possibilities and pruning them against S. The latter meant that the decoding runtime did not improve over the one achieved in Theorem 16.3.8.

Large Runtime. We finally address the question of the high run time of all the list decoding algorithms so far. Dvir and Lovett [25], presented a construction of an *explicit* $(q, k, \varepsilon, s, s^{O(s)})$-subspace evasive subset S^*. More interestingly, given any affine subspace T of dimension at most s, it can compute $S \cap T$ in time proportional to the output size. Thus, this result along with the discussion above implies the following result:

Theorem 16.4.1. *There exist* strongly explicit *codes of rate R that for large enough block length N can be list decoded from $1-R-\varepsilon$ fraction of errors (for any small enough $\varepsilon > 0$) in time $O\left(\left(\frac{N}{\varepsilon^2}\right)^2\right) + \left(\frac{1}{\varepsilon}\right)^{O(1/\varepsilon)}$. The worst-case list size is $\left(\frac{1}{\varepsilon}\right)^{O(1/\varepsilon)}$ and the alphabet size is $\left(\frac{N}{\varepsilon}\right)^{O(1/\varepsilon^2)}$.*

The above answers Question 16.0.1 pretty satisfactorily. However, to obtain a completely satisfactory answer one would have to solve the following open question:

Open Question 16.4.3. *Are there explicit codes of rate $R > 0$ that are $\left(1-R-\varepsilon, (1/\varepsilon)^{O(1)}\right)$-list decodable that can be list-decoded in time* $\mathrm{poly}(N, 1/\varepsilon)$ *over alphabet of size $q \leq poly(n)$?*

The above question, without the requirement of explicitness, has been answered by Guruswami and Xing [51].

Part V

The Applications

Chapter 17

Cutting Data Down to Size: Hashing

In this chapter, we will study hashing, which is a method to compute a small digest of data that can be used as a surrogate to later perform quick checks on the data. We begin with brief descriptions of three practical applications where hashing is useful. We then formally state the definition of hash functions that are needed in these applications (the so called “universal" hash functions). Next, we will show how in some sense good hashing functions and good codes are equivalent. Finally, we will see how hashing can solve a problem motivated by outsourced storage in the “cloud."

17.1 Why Should You Care About Hashing?

Hashing is one of the most widely used objects in computer science. In this section, we outline three practical applications that heavily use hashing. While describing the applications, we will also highlight the properties of hash functions that these applications need.

Before we delve into the applications, let us first formally define a hash function.

Definition 17.1.1 (Hash Function)**.** Given a domain $\mathbb{D}$ and a range Σ, (typically, with $|\Sigma| < |\mathbb{D}|$), a hash function is a map

$$h : \mathbb{D} \to \Sigma.$$

Of course, the definition above is too general and we will later specify properties that will make the definition more interesting.

Integrity Checks on Routers. Routers on the Internet process a lot of packets in a very small amount of time. Among other tasks, router has to perform an “integrity check" on the packet to make sure that the packet it is processing is not corrupted. Since the packet has well defined fields, the router could check if all the field have valid entries. However, it is possible that one of the valid entry could be turned into another valid entry. However, the packet as a whole could still be invalid.

If you have progressed so far in the book, you will recognize that the above is the error detection problem and we know how to do error detection (see e.g., Proposition 2.3.3). However, the

algorithms that we have seen in this book are too slow to implement in routers. Hence, Internet protocols use a hash function on a domain $\mathbb{D}$ that encodes all the information that needs to go into a packet. Thus, given an $\mathbf{x} \in \mathbb{D}$, the packet is the pair $(\mathbf{x}, h(\mathbf{x}))$. The sender sends the packet $(\mathbf{x}, h(\mathbf{x}))$ and the receiver gets $(\mathbf{x}', y)$. In order to check if any errors occurred during transmission, the receiver checks if $h(\mathbf{x}') = y$. If the check fails, the receiver asks for a re-transmission otherwise it assumes there were no errors during transmission. There are two requirements from the hash function: (i) It should be super efficient to compute $h(\mathbf{x})$ given $\mathbf{x}$ and (ii) h should avoid "collisions," i.e. if $\mathbf{x} \neq \mathbf{x}'$, then $h(\mathbf{x}) \neq h(\mathbf{x}')$.[1]

Integrity Checks in Cloud Storage. Say, you (as a client) have data $\mathbf{x} \in \mathbb{D}$ that you want to outsource $\mathbf{x}$ to a cloud storage provider. Of course once you "ship" off $\mathbf{x}$ to the cloud, you do not want to store it locally. However, you do not quite trust the cloud either. If you do not audit the cloud storage server in any way, then nothing stops the storage provider from throwing away $\mathbf{x}$ and send you some other data $\mathbf{x}'$ when you ask for $\mathbf{x}$. The problem of designing an auditing protocol that can verify whether the server has the data $\mathbf{x}$ is called the *data possession* problem.

We consider two scenarios. In the first scenario, you access the data pretty frequently during "normal" operation. In such cases, here is a simple check you can perform. When you ship off $\mathbf{x}$ to the cloud, compute $z = h(\mathbf{x})$ and store it. Later when you access $\mathbf{x}$ and the storage provider send you $\mathbf{x}'$, you compute $h(\mathbf{x}')$ and check if it is the same as the stored $h(\mathbf{x})$. This is exactly the same solution as the one for packet verification mentioned above.

Now consider the scenario, where the cloud is used as an archival storage. In such a case, one needs an "auditing" process to ensure that the server is indeed storing $\mathbf{x}$ (or is storing some massaged version from which it can compute $\mathbf{x}$– e.g. the storage provider can compress $\mathbf{x}$). One can always ask the storage provider to send back $\mathbf{x}$ and then use the scheme above. However, if $\mathbf{x}$ is meant to be archived in the cloud, it would be better to resolve the following question:

Question 17.1.1. *Is there an auditing protocol with small client-server communication*[a]*, which if the server passes then the client should be able to certain (with some high confidence) that the server is indeed storing* $\mathbf{x}$*?*

[a] In particular, we rule out solutions where the server sends $\mathbf{x}$ to the client.

We will see later how this problem can be solved using hashing.

Fast Table Lookup. One of the most common operations on databases is the following. Assume there is a table with entries from $\mathbb{D}$. One would like to decide on a data structure to store

[1] Note that in the above example, one could have $\mathbf{x} \neq \mathbf{x}'$ and $h(\mathbf{x}) \neq h(\mathbf{x}')$ but it is still possible that $y = h(\mathbf{x}')$ and hence the corrupted packet $(\mathbf{x}', y)$ would pass the check above. Our understanding is that such occurrences are rare.

the table so that later on given an element $\mathbf{x} \in \mathbb{D}$, one would quickly like to decide whether $\mathbf{x}$ is in the table or now.

Let us formalize the problem a bit: assume that the table needs to store N values $a_1, \ldots, a_N \in \mathbb{D}$. Then later given $\mathbf{x} \in \mathbb{D}$ one needs to decide if $\mathbf{x} = a_i$ for some i. Here is one simple solution: sort the n elements in an array T and given $\mathbf{x} \in \mathbb{D}$ use binary search to check if $\mathbf{x}$ is in T or not. This solution uses $\Theta(N)$ amounts of storage and searching for $\mathbf{x}$ takes $\Theta(\log N)$ time. Further, the pre-processing time (i.e. time taken to build the array T) is $\Theta(N \log N)$. The space usage of this scheme is of course optimal but one would like the lookup to be faster: ideally we should be able to perform the search in $O(1)$ time. Also it would be nice to get the pre-processing time closer to the optimal $O(N)$. Further, this scheme is very bad for dynamic data: inserting an item to and deleting an item from T takes $\Theta(N)$ time in the worst-case.

Now consider the following solution: build a boolean array B with one entry for each $z \in \mathbb{D}$ and set $B[a_i] = 1$ for every $i \in [N]$ (and every other entry is 0).[2] Then searching for $\mathbf{x}$ is easy: just lookup $B[\mathbf{x}]$ and check if $B[\mathbf{x}] \neq 0$. Further, this data structure can easily handle addition and deletion of elements (by incrementing and decrementing the corresponding entry of B respectively). However, the amount of storage and pre-processing time are both $\Theta(|\mathbb{D}|)$, which can be much much bigger than the optimal $O(N)$. This is definitely true for tables stored in real life databases. This leads to the following question:

Question 17.1.2. *Is there a data structure that supports searching, insertion and deletion in $O(1)$ time but only needs $O(N)$ space and $O(N)$ pre-processing time?*

We will see later how to solve this problem with hashing.

17.2 Avoiding Hash Collisions

One of the properties that we needed in the applications outlined in the previous section was that the hash function $h : \mathbb{D} \to \Sigma$ should avoid collisions. That is, given $\mathbf{x} \neq y \in \mathbb{D}$, we want $h(\mathbf{x}) \neq h(y)$. However, since we have assumed that $|\Sigma| < |\mathbb{D}|$, this is clearly impossible. A simple counting argument shows that there will exist an $\mathbf{x} \neq y \in \mathbb{D}$ such that $h(\mathbf{x}) = h(y)$. There are two ways to overcome this hurdle.

The first is to define a cryptographic collision resistant hash function h, i.e. even though there exists collisions for the hash function h, it is computationally hard for an adversary to compute $\mathbf{x} \neq y$ such that $h(\mathbf{x}) = h(y)$.[3] This approach is out of the scope of this book and hence, we will not pursue this solution.

[2] If one wants to handle duplicates, one could store the number of occurrences of y in $B[y]$.

[3] This is a very informal definition. Typically, an adversary is modeled as a randomized polynomial time algorithm and there are different variants on whether the adversary is given $h(\mathbf{x})$ or $\mathbf{x}$ (or both). Also there are variants where one assumes a distribution on $\mathbf{x}$. Finally, there are no unconditionally collision resistant hash function but there exists provably collision resistant hash function under standard cryptographic assumptions: e.g. factoring is hard.

The second workaround is to define a family of hash functions and then argue that the probability of collision is small for a hash function chosen randomly from the family. More formally, we define a hash family:

Definition 17.2.1 (Hash Family)**.** Given $\mathbb{D}, \Sigma$ and an integer $m \geq 1$, a hash family $\mathscr{H}$ is a set $\{h_1, \ldots, h_m\}$ such that for each $i \in [m]$,

$$h_i : \mathbb{D} \to \Sigma.$$

Next we define the notion of (almost) universal hash function (family).

Definition 17.2.2 (Almost Universal Hash Family)**.** A hash family $\mathscr{H} = \{h_1, \ldots, h_m\}$ defined over the domain $\mathbb{D}$ and range Σ is said to be ε-almost universal hash function (family) for some $0 < \varepsilon \leq 1$ if for every $\mathbf{x} \neq y \in \mathbb{D}$,

$$\Pr_i \left[h_i(\mathbf{x}) = h_i(y) \right] \leq \varepsilon,$$

where in the above i is chosen uniformly at random from $[m]$.

We will show in the next section that ε-almost universal hash functions are equivalent to codes with (large enough) distance. In the rest of the section, we outline how these hash families provides satisfactory solutions to the problems considered in the previous section.

Integrity Checks. For the integrity check problem, one pick random $i \in [m]$ and chooses $h_i \in \mathscr{H}$, where $\mathscr{H}$ is an ε-almost universal hash function. Thus, for any $\mathbf{x} \neq y$, we're guaranteed with probability at least $1 - \varepsilon$ (over the random choice of i) that $h_i(\mathbf{x}) \neq h_i(y)$. Thus, this gives a randomized solution to the integrity checking problem in routers and cloud storage (where we consider the first scenario in which the cloud is asked to return the original data in its entirety).

It is not clear whether such hash functions can present a protocol that answers Question 17.1.1. There is a very natural protocol to consider though. When the client ships off data $\mathbf{x}$ to the cloud, it picks a random hash function $h_i \in \mathscr{H}$, where again $\mathscr{H}$ is an ε-universal hash function, and computes $h_i(\mathbf{x})$. Then it stores $h_i(\mathbf{x})$ and ships off $\mathbf{x}$ to the cloud. Later on, when the client wants to audit, it asks the cloud to send $h_i(\mathbf{x})$ back to it. Then if the cloud returns with z, the client checks if $z = h_i(\mathbf{x})$. If so, it assumes that the storage provider is indeed storing $\mathbf{x}$ and otherwise it concludes that the cloud is not storing $\mathbf{x}$.

Note that it is crucial that the hash function be chosen randomly: if the client picks a deterministic hash function h, then the cloud can store $h(\mathbf{x})$ and throw away $\mathbf{x}$ because it knows that the client is only going to ask for $h(\mathbf{x})$. Intuitively, the above protocol might work since the random index $i \in [m]$ is not known to the cloud till the client asks for $h_i(\mathbf{x})$, it seems "unlikely" that the cloud can compute $h_i(\mathbf{x})$ without storing $\mathbf{x}$. We will see later how the coding view of almost universal hash functions can make this intuition rigorous.

Fast Table Lookup. We now return to Question 17.1.2. The basic idea is simple: we will modify the earlier solution that maintained an entry for each element in the domain $\mathbb{D}$. The new solution will be to keep an entry for all possible hash values (instead of all entries in $\mathbb{D}$).

More formally, let $\mathscr{H} = \{h_1, \ldots, h_m\}$ be an ε-almost hash family with domain $\mathbb{D}$ and range Σ. Next we build an array of link list with one entry in the array for each value $v \in \Sigma$. We pick a random hash function $h_i \in \mathscr{H}$. Then for each a_j ($j \in [N]$) we add it to the link list corresponding to $h_i(a_j)$. Now to determine whether $\mathbf{x} = a_j$ for some j, we scan the link list corresponding to $h_i(\mathbf{x})$ and check if $\mathbf{x}$ is in the list or not. Before we analyze the space and time complexity of this data structure, we point out that insertion and deletion are fairly easy. For inserting an element $\mathbf{x}$, we compute $h_i(\mathbf{x})$ and add $\mathbf{x}$ to the link list corresponding to $h_i(\mathbf{x})$. For deletion, we first perform the search algorithm and then remove $\mathbf{x}$ from the list corresponding to $h_i(\mathbf{x})$, if it is present. It is easy to check that the algorithms are correct.

Next we analyze the space complexity. Note that for a table with N elements, we will use up $O(N)$ space in the linked lists and the array is of size $O(|\Sigma|)$. That is, the total space usage is $O(N + |\Sigma|)$. Thus, if we can pick $|\Sigma| = O(N)$, then we would match the optimal $O(N)$ bound for space.

Now we analyze the time complexity of the various operations. We first note that insertion is $O(1)$ time (assuming computing the hash value takes $O(1)$ time). Note that this also implies that the pre-processing time is $O(N + |\Sigma|)$, which matches the optimal $O(N)$ bound for $|\Sigma| \leq O(N)$. Second, for deletion, the time taken after performing the search algorithm is $O(1)$, assuming the lists as stored as doubly linked lists. (Recall that deleting an item from a doubly linked list if one has a pointer to the entry can be done in $O(1)$ time.)

Finally, we consider the search algorithm. Again assuming that computing a hash value takes $O(1)$ time, the time complexity of the algorithm to search for $\mathbf{x}$ is dominated by size of the list corresponding to $h_i(\mathbf{x})$. In other words, the time complexity is determined by the number of a_j that collide with $\mathbf{x}$, i.e., $h_i(\mathbf{x}) = h_i(a_j)$. We bound this size by the following simple observation.

Claim 17.2.1. *Let $\mathscr{H} = \{h_1, \ldots, h_m\}$ with domain $\mathbb{D}$ and range Σ be an ε-almost universal hash family. Then the following is true for any (distinct) $\mathbf{x}, a_1, a_2, \ldots, a_N \in \mathbb{D}$:*

$$\mathbb{E}_i \left[|\{a_j | h_i(\mathbf{x}) = h_i(a_j)\}| \right] \leq \varepsilon \cdot N,$$

where the expectation is taken over a uniformly random choice of $i \in [m]$.

Proof. Fix a $j \in [N]$. Then by definition of an ε-almost universal hash family, we have that

$$\Pr_i [h_i(\mathbf{x}) = h_i(a_j)] \leq \varepsilon.$$

Note that we want to bound $\mathbb{E}\left[\sum_{j=1}^{N} \mathbb{1}_{h_i(a_j)=h_i(\mathbf{x})}\right]$. The probability bound above along with the linearity of expectation (Proposition 3.1.2) and Lemma 3.1.1 completes the proof. □

The above discussion then implies the following:

Proposition 17.2.2. *Given an $O\left(\frac{1}{N}\right)$-almost universal hash family with domain $\mathbb{D}$ and range Σ such that $|\Sigma| = O(N)$, there exists a randomized data structure that given N elements $a_1, \ldots, a_N \in \mathbb{D}$, supports searching, insertion and deletion in expected $O(1)$ time while using $O(N)$ space in the worst-case.*

Thus, Proposition 17.2.2 answers Question 17.1.2 in the affirmative if we can answer the following question in the affirmative:

Question 17.2.1. *Given a domain $\mathbb{D}$ and an integer $N \geq 1$, does there exist an $O\left(\frac{1}{N}\right)$-almost universal hash function with domain $\mathbb{D}$ and a range Σ such that $|\Sigma| = O(N)$?*

We will answer the question above (spoiler alert!) in the affirmative in the next section.

17.3 Almost Universal Hash Function Families and Codes

In this section, we will present a very strong connection between almost universal hash families and codes with good distance: in fact, we will show that they are in fact *equivalent*.

We first begin by noting that any hash family has a very natural code associated with it and that every code has a very natural hash family associated with it.

Definition 17.3.1. Given a hash family $\mathscr{H} = \{h_1, \ldots, h_n\}$ where for each $i \in [n]$, $h_i : \mathbb{D} \to \Sigma$, consider the following associated code

$$C_{\mathscr{H}} : \mathbb{D} \to \Sigma^n,$$

where for any $\mathbf{x} \in \mathbb{D}$, we have

$$C_{\mathscr{H}}(\mathbf{x}) = (h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_n(\mathbf{x})).$$

The connection also goes the other way. That is, given an $(n,k)_\Sigma$ code C, we call the associated hash family $\mathscr{H}_C = \{h_1, \ldots, h_n)$, where for every $i \in [n]$,

$$h_i : \Sigma^k \to \Sigma$$

such that for every $\mathbf{x} \in \Sigma^k$ and $i \in [n]$,

$$h_i(\mathbf{x}) = C(\mathbf{x})_i.$$

Next we show that an ε-almost universal hash family is equivalent to a code with good distance.

Proposition 17.3.1. *Let $\mathscr{H} = \{h_1, \ldots, h_n\}$ be an ε-almost universal hash function, then the code $C_{\mathscr{H}}$ has distance at least $(1-\varepsilon)n$. On the other hand if C is an $(n,k,\delta n)$-code, then $\mathscr{H}_C$ is a $(1-\delta)$-almost universal hash function.*

Proof. We will only prove the first claim. The proof of the second claim is essentially identical and is left as an exercise.

Let $\mathscr{H} = \{h_1, \ldots, h_n\}$ be an ε-almost universal hash function. Now fix arbitrary $\mathbf{x} \neq \mathbf{y} \in \mathbb{D}$. Then by definition of $C_{\mathscr{H}}$, we have

$$\{i | h_i(\mathbf{x}) = h_i(\mathbf{y})\} = \{i | C_{\mathscr{H}}(\mathbf{x})_i = C_{\mathscr{H}}(\mathbf{y})_i\}.$$

This implies that

$$\Pr_i\left[h_i(\mathbf{x}) = h_i(\mathbf{y})\right] = \frac{|\{i|h_i(\mathbf{x}) = h_i(\mathbf{y})\}|}{n} = \frac{n - \Delta(C_{\mathscr{H}}(\mathbf{x}), C_{\mathscr{H}}(\mathbf{y}))}{n} = 1 - \frac{\Delta(C_{\mathscr{H}}(\mathbf{x}), C_{\mathscr{H}}(\mathbf{y}))}{n},$$

where the second equality follows from the definition of the Hamming distance. By the definition of ε-almost universal hash family the above probability is upper bounded by ε, which implies that

$$\Delta(C_{\mathscr{H}}(\mathbf{x}), C_{\mathscr{H}}(\mathbf{y})) \geq n(1-\varepsilon).$$

Since the choice of $\mathbf{x}$ and $\mathbf{y}$ was arbitrary, this implies that $C_{\mathscr{H}}$ has distance at least $n(1-\varepsilon)$ as desired. □

17.3.1 The Polynomial Hash Function

We now consider the hash family corresponding to a Reed-Solomon code. In particular, let C be a $[q, k, q-k+1]_q$ Reed-Solomon code. By Proposition 17.3.1, the hash family $\mathscr{H}_C$ is an $\frac{k-1}{q}$-almost universal hash family– this hash family in the literature is called the polynomial hash. Thus, if we pick q to be the smallest power of 2 larger than N and pick $k = O(1)$, then this leads to an $O(1/N)$-universal hash family that satisfies all the required properties in Question 17.2.1.

Note that the above implies that $|\mathbb{D}| = N^{O(1)}$. One might wonder if we can get an $O(1/N)$-almost universal hash family with the domain size being $N^{\omega(1)}$. We leave the resolution of this question as an exercise.

17.4 Data Possession Problem

In this section, we return to Question 17.1.1. Next we formalize the protocol for the data possession problem that we outlined in Section 17.2. Algorithm 29 presents the pre-processing step.

Algorithm 29 Pre-Processing for Data Possession Verification

INPUT: Data $\mathbf{x} \in \mathbb{D}$, hash family $\mathscr{H} = \{h_1, \ldots, h_m\}$ over domain $\mathbb{D}$

1: Client computes an index i for $\mathbf{x}$.
2: Client picks a random $j \in [m]$.
3: Client computes $z \leftarrow h_j(\mathbf{x})$ and stores (i, j, z).
4: Client sends $\mathbf{x}$ to the server.

Algorithm 30 formally states the verification protocol. Note that if the server has stored $\mathbf{x}$ (or is able to re-compute $\mathbf{x}$ from what it had stored), then it can pass the protocol by returning $a \leftarrow h_j(\mathbf{x})$. Thus, for the remainder of the section, we will consider the case when the server tries to cheat. We will show that if the server is able to pass the protocol in Algorithm 30 with high enough probability, then the server indeed has stored $\mathbf{x}$.

Algorithm 30 Verification for Data Possession Verification

INPUT: Index i of data $\mathbf{x} \in \mathbb{D}$
OUTPUT: 1 if Server has $\mathbf{x}$ and 0 otherwise

```
Client sends a challenge (i, j) to the server.
Client receives an answer a.
IF a = z THEN
    RETURN 1
ELSE
    RETURN 0
```

Before we formally prove the result, we state our assumptions on what the server can and cannot do. We assume that the server follows the following general protocol. First, when the server receives $\mathbf{x}$, it does performs some computation (that terminates) on $\mathbf{x}$ to produce $\mathbf{y}$ and then it stores $\mathbf{y}$. (E.g., the server could store $\mathbf{y} = \mathbf{x}$ or $\mathbf{y}$ could be a compressed version of $\mathbf{x}$.) Then when it receives the challenge (i, j) for $\mathbf{x}$, it uses another algorithm $\mathscr{A}$ and returns the answers $a \leftarrow \mathscr{A}(\mathbf{y}, j)$. We assume that $\mathscr{A}$ always terminates on any input.[4] Note that the server is allowed to use arbitrary (but finite) amount of time to compute its answer. Next, we will prove that under these assumptions, the server cannot cheat with too high a probability.

Theorem 17.4.1. *Assume that the hash family $\mathscr{H}$ is an ε-almost universal hash family. Then if the server passes the protocol in Algorithm 30 with probability $> \frac{1}{2} + \frac{\varepsilon}{2}$, then the server has enough information to recreate $\mathbf{x}$.*

Proof. To prove the claim, we present an algorithm that can compute $\mathbf{x}$ from $\mathbf{y}$. (Note that we do not need this algorithm to be efficient: it just needs to terminate with $\mathbf{x}$.) In particular, consider Algorithm 31.

Algorithm 31 Decompression Algorithm

INPUT: $\mathscr{A}, \mathbf{y}$
OUTPUT: $\mathbf{x}'$

```
z ← (A(y, j))_{j∈[m]}.
Run the MLD algorithm (Algorithm 2) for C_H on z and let C_H(x') be its output.
RETURN x'
```

To complete the proof, we will show that $\mathbf{x}' = \mathbf{x}$. Towards this end we claim that $\Delta(\mathbf{z}, C_{\mathscr{H}}(\mathbf{x})) < \frac{m}{2} \cdot (1 - \varepsilon)$. Assuming this is true, we complete the proof. Note that Proposition 17.3.1 implies that $C_{\mathscr{H}}$ has distance at least $m(1 - \varepsilon)$. Thus, Proposition 1.4.1 (in particular, its proof) implies that Algorithm 2 will return $C_{\mathscr{H}}(\mathbf{x})$ and thus, $\mathbf{x}' = \mathbf{x}$, as desired.

[4] We have stated the algorithm to be independent of $\mathbf{y}$ and j but that need not be the case. However later in the section, we will need the assumption that $\mathscr{A}$ is independent of $\mathbf{y}$ and j, so we will keep it that way.

Finally, we argue that $\Delta(\mathbf{z}, C_{\mathscr{H}}(\mathbf{x})) < m(1-\varepsilon)/2$. To see this note that if the server passes the protocol in Algorithm 30 (i.e. the client outputs 1), then it has to be the case that $z_j \stackrel{\text{def}}{=} \mathscr{A}(\mathbf{y}, j) = h_j(\mathbf{x})$. Recall that by definition of $C_{\mathscr{H}}$, $h_j(\mathbf{x}) = C_{\mathscr{H}}(\mathbf{x})_j$ and that the server passes the protocol with probability $> 1/2+\varepsilon/2$. Since j is chosen uniformly from $[m]$, this implies that for $> m(1/2+\varepsilon/2)$ positions j, $z_j = C_{\mathscr{H}}(\mathbf{x})_j$, which then implies the claim. □

17.4.1 Driving Down the Cheating Probability

One of the unsatisfying aspects of Theorem 17.4.1 is that the probability of catching a "cheating" server is strictly less than 50%.[5] It is of course desirable to drive this up as close to 100% as possible. One way to obtain this would be to "repeat" the protocol: i.e. the client can choose multiple random hashes and store the corresponding values (in Algorithm 29) and (in Algorithm 30) asks the server to send back all the hash values and accepts if and only if all the returned answers match with the stored hash values. This however, comes at a cost: the client has to store more hash values (and also the communication cost between the client and the server goes up accordingly.)

Next we argue using list decoding that the protocol in Algorithm 29 and 30 (without any modification) gives a more powerful guarantee than the one in Theorem 17.4.1. To see why list decoding might buy us something, let us look back at Algorithm 31. In particular, consider Step 2: since we run MLD, we can only guarantee unique decoding up to half the (relative) distance of $C_{\mathscr{H}}$. This in turn leads to the bound of $1/2+\varepsilon/2$ in Theorem 17.4.1. We have seen that list decoding allows us to go beyond half the distance number of errors. So maybe, running a list decoding algorithm instead of MLD in Step 2 of Algorithms 31 would help us get better results. There are two potential issues that we'll need to tackle:

- We will need a general bound that shows that list decoding (arbitrarily) close to 100% is possible for any $C_{\mathscr{H}}$ for an ε-almost universal hash family; and
- Even if the above is possible, what will we do when a list decoding algorithm returns a *list* of possible data?

We will get around the first concern by using the Johnson bound 7.3.1. To get around the second issue we will indirectly use "side information" (like we mentioned in Section 7.2). For the latter, we will need the notion of Kolmogorov complexity, which captures the amount of information content in any given string. For our purposes, the following informal description is enough:

Definition 17.4.1. Given a string $\mathbf{x}$, its Kolmogorov complexity, denoted by $\mathscr{K}(\mathbf{x})$ is the minimum of $|\mathbf{y}|+|\mathscr{D}|$, where $\mathscr{D}$ is a decompression algorithm that on input $\mathbf{y}$ outputs $\mathbf{x}$ (where $|\mathbf{x}|$ and $|\mathscr{D}|$ are the length of $\mathbf{x}$ and (a description of) $\mathscr{D}$ in bits).

Informally, $\mathscr{K}(\mathbf{x})$ is the amount of information that can be obtained algorithmically from $\mathbf{x}$. Kolmogorov complexity is a fascinating topic that it outside the scope of this book. Here we will

[5] Note that Theorem 17.4.1 states that a server that cannot recreate $\mathbf{x}$ can pass the test with probablity at most $1/2+\varepsilon/2$. In other words, the probability that such a server is caught is at most $1/2-\varepsilon/2 < 1/2$.

only need to use the definition of $\mathcal{K}(\mathbf{x})$. We are now ready to prove the following list decoding counterpart of Theorem 17.4.1:

Theorem 17.4.2. *Assume that the hash family $\mathcal{H}$ is an ε-almost universal hash family. Then if the server passes the protocol in Algorithm 30 with probability $> \sqrt{\varepsilon}$, then the amount of information server has stored for $\mathbf{x}$ is at least $\mathcal{K}(\mathbf{x}) - O(\log|\mathbf{x}|)$.*

We note that the above is not a strict generalization of Theorem 17.4.1, as even though probability of catching a cheating server has gone up our guarantee is weaker. Unlike Theorem 17.4.1, where we can guarantee that the server can re-create $\mathbf{x}$, here we can only guarantee "storage enforceability"– i.e. we can only force a server to store close to $\mathcal{K}(\mathbf{x})$ amounts of memory.

Proof of Theorem 17.4.2. Here is the main idea of the proof. We first assume for the sake of contradiction that $|\mathbf{y}| < \mathcal{K}(\mathbf{x}) - O(\log(|\mathbf{x}|))$. Then using we construct a decompression algorithm $\mathcal{D}$ that on given input $\mathbf{y}$ and $O(\log(|\mathbf{x}|))$ extra information (or "advice"), outputs $\mathbf{x}$. Then we will show this overall contradicts the definition of $\mathcal{K}(\mathbf{x})$ (as this gives an overall smaller description of $\mathbf{x}$).

Before we state the decompression algorithm, we recall some facts. First note that $C_{\mathcal{H}}$ by Proposition 17.3.1 is a q-ary code (with $|\Sigma| = q$) with distance $m(1-\varepsilon)$. Further, by the Johnson bound (Theorem 7.3.1), $C_{\mathcal{H}}$ is a $(1-\sqrt{\varepsilon}, L)$-list decodable, where

$$L \le qm^2. \tag{17.1}$$

Next, in Algorithm 32, we present a decompression algorithm that can compute $\mathbf{x}$ from $\mathbf{y}$ and an advice string $a \in [L]$. (As before, we do not need this algorithm to be efficient: it just needs to terminate with $\mathbf{x}$.)

Algorithm 32 Decompression Algorithm Using List Decoding

INPUT: $\mathcal{A}, \mathbf{y}, a$
OUTPUT: $\mathbf{x}$

```
z ← (A(y, j))_{j∈[m]}.
L ← ∅.
FOR x' ∈ D DO
    IF Δ(C_H(x'), z) ≤ (1 − √ε)m THEN
        Add x' to L
RETURN The ath element in L
```

To complete the proof, we claim that there exists a choice of $a \in [L]$ such that Algorithm 32 outputs $\mathbf{x}$. Note that this implies that $(\mathbf{y}, a)$ along with Algorithm 32 gives a complete description of $\mathbf{x}$. Now note that Algorithm 32 can be described in $O(1)$ bits. This implies that the size of this description is $|\mathbf{y}| + \log L + O(1)$, which by Definition 17.4.1 has to be at least $\mathcal{K}(\mathbf{x})$. This implies that

$$|\mathbf{y}| \ge \mathcal{K}(\mathbf{x}) - |a| - O(1) = \mathcal{K}(\mathbf{x}) - \log L - O(1) \ge \mathcal{K}(\mathbf{x}) - O(\log|\mathbf{x}|),$$

where the last inequality follows from (17.1).

Next, we argue the existence of an appropriate $a \in [L]$. Towards this end we claim that $\Delta(\mathbf{z}, C_{\mathscr{H}}(\mathbf{x})) < m(1-\sqrt{\varepsilon})$. Note that this implies that $\mathbf{x} \in \mathscr{L}$. Since $|\mathscr{L}| \leq L$, then we can just assign a to be the index of $\mathbf{x}$ in $\mathscr{L}$. Finally, we argue that $\Delta(\mathbf{z}, C_{\mathscr{H}}(\mathbf{x})) < m(1-\sqrt{\varepsilon})$. To see this note that if the server passes the protocol in Algorithm 30 (i.e. the client outputs 1), then it has to be the case that $z_j \stackrel{\text{def}}{=} \mathscr{A}(\mathbf{y}, j) = h_j(\mathbf{x})$. Recall that by definition of $C_{\mathscr{H}}$, $h_j(\mathbf{x}) = C_{\mathscr{H}}(\mathbf{x})_j$ and that the server passes the protocol with probability $> \sqrt{\varepsilon}$. Since j is chosen uniformly from $[m]$, this implies that for $> m\sqrt{\varepsilon}$ positions j, $z_j = C_{\mathscr{H}}(\mathbf{x})_j$, which then implies the claim. □

17.5 Bibliographic Notes

Universal hash functions were defined in the seminal paper of Carter and Wegman [12]. Almost universal hash function family was defined by Stinson [95].

Kolmogorov complexity was defined by Kolmogorov [63]. For a thorough treatment see the textbook by Li and Vitányi [68].

Chapter 18

Securing Your Fingerprints: Fuzzy Vaults

String-based passwords are the dominant mode of authentication in today's information-reliant world. Indeed, all of us use passwords on a regular basis to authenticate ourselves in almost any online activity. Strings, a primitive data type, have become widespread due to several nice mathematical properties. First, matching two strings (that is, checking if two strings are exactly the same) is computationally very fast (and easy). Second, and more importantly, there exist secure hash functions that map a string x to another string $h(x)$ such that, given $h(x)$, determining x is hard. Furthermore, since $h(x)$ is itself a string, we can check if a claimed password y is the same as the original string x by comparing $h(y)$ and $h(x)$, which (as we just observed) is easy to do. This implies that the server performing the authentication only needs to store the hashes $h(x)$ of the original passwords. Hence, even if the list of hashed passwords were compromised, the passwords themselves would remain secure.

The above scheme is perfect as long as the passwords x are "random enough," and this can be achieved if the passwords were generated randomly by some automated process. However, in real life passwords are generated by humans and are not really random. (One of the most quoted facts is that the most commonly used password is the string "password" itself.) Further, we tend to forget passwords, which has lead to the near ubiquity of "Forgot passwords" links in places where we need to login.

One alternative that has been gaining traction in the last decade or so is to use a user's fingerprint (and more generally their biometrics: e.g. their face) as their password. The big advantage is that it is hard to "forget" one's fingerprint. In this chapter, we will look at the issues in using fingerprints as passwords and see how Reed-Solomon codes can help.

18.1 Some quick background on fingerprints

First we give a very short (and necessarily incomplete) description of how a fingerprint (image/sensor reading) is converted in to a string f. The standard way to store a fingerprint (see Figure 18.1 for images of two fingerprints) is as a collection of triples, called minutia. Each minutia point is located where one ridge[1] splits into two, or where one ridge ends. The ith

[1] In Figure 18.1, the ridges are the lines.

minutia is the triple (x_i, y_i, θ_i), where x_i and y_i are the x and y coordinates of a point on the finger, and θ_i indicates the direction of the ridge that created the minutia point relative to the plane.

Given that we now have a string representation of a fingerprint, we have the following naive solution for designing a hash function for the fingerprint f:

Naive Solution. Use any off-the-shelf hash function h for strings and then store $h(f)$ instead of f.

To see the issues with the naive solution, we first need to know a little bit about how fingerprints are stored.

The main issue with our naive solution is that two fingerprint readings will never be exactly the same, even if the same finger is used. For any two fingerprint readings, the following issues may produce errors:

1. Translation and rotation, even when using the same finger.

2. Varying pressure.

3. Each reading may not completely overlap with the ideal fingerprint region (i.e., the finger may be slightly tilted).

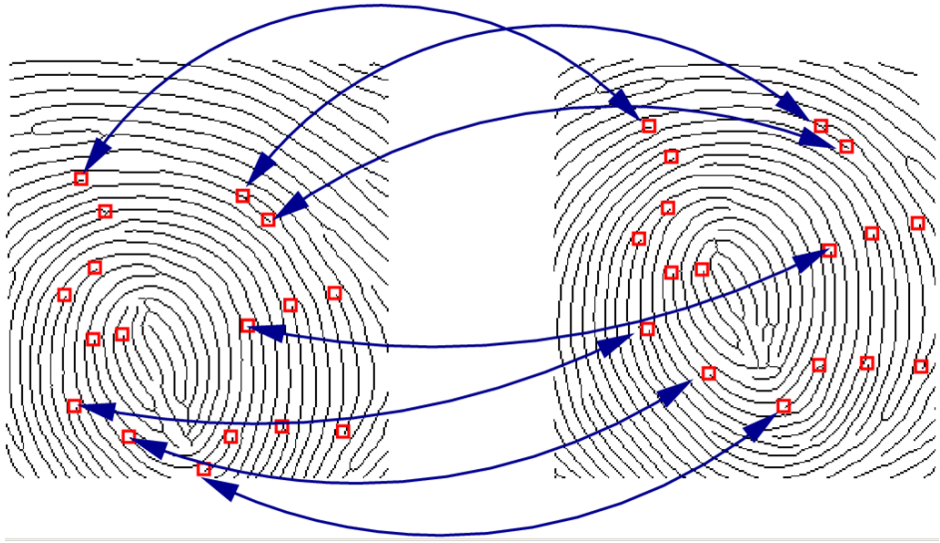
4. The minutiae are not ordered, so they form a set instead of a vector. Of course one can sort the set by the lexicographic order to produce a string. But the earlier issue (especially points 1 and 2) imply that the specific values of (x_i, y_i, θ_i) are not that important individually. Furthermore, the fact that any two readings might not have complete overlap means that we are interested in matching readings that have significant overlap, so it turns out that the set notation is ideal to theoretically deal with these issues.

We can now see that the naive solution is inadequate. Even if we could somehow correct the first three issues, existing hash functions for strings require a vector, not a set, so our naive solution will fail.

Remark 18.1.1. The four problems that came up in our naive solution will come up in any solution we propose. Technology has not yet developed to the point where we can securely eliminate these issues, which is why there are no prevalent secure commercial systems that safeguard secrets using fingerprints. (The reason government agencies, such as the police or FBI, use fingerprinting is because there is an inherent trust that the government will keep your data secure, even when it does not apply a good hash function to it.)

Thus, we need secure hash functions designed to handle the additional challenges posed by fingerprints. We would like to mention that for fingerprints to replace strings as passwords, the hash function needs to satisfy both of these properties *simultaneously*: (i) we should be able to match hashes from the “same" fingerprint and (ii) an adversary should not be able to “break" the hash function.

Figure 18.1: The minutiae are unordered and form a set, not a vector.



18.2 The Fuzzy Vault Problem

We begin with the fuzzy vault problem, which is slightly different from the one we have been studying so far. Say you have a secret string s, and you want to store it in a secure way. Instead of using a password, you want to use your fingerprint, f, to "lock" the secret s. You want the locked version of your secret to have two properties:

1. You should be able to "unlock" the locked version of s
2. No one else should be able to "unlock" s

We claim that if we can solve the above problem, then we can solve the problem of designing a secure hash function for fingerprints. We leave the details as an exercise. (Hint: pick s at random and then in addition to the locked version of s also store $h(s)$, where h is an off-the-shelf secure hash function for strings.)

We will now formalize the fuzzy vault problem.

18.2.1 Formal Problem Statement

The first thing we need to do is quantize the measurements of the minutiae. We cannot be infinitely precise in our measurements anyways, so let us assume that all quantized minutiae, (x_i, y_i, θ_i), can be embedded into $\mathbb{F}_q$ for some large enough prime power q. Theoretically, this can also help to correct the first two issues from our naive solution. We could go through all possible values in $\mathbb{F}_q$ to get rid of translation and rotation errors (e.g., for every $(\Delta x, \Delta y, \Delta z) \in \mathbb{F}_q$, we rotate and translate each minutia (x_i, y_i, z_i) to $(x_i + \Delta x, y_i + \Delta y, z_i + \Delta z)$). [2] We could also

[2] To be more precise we first perform the translation and rotation over the reals and then quantize and map to the appropriate $\mathbb{F}_q$ value.

do some local error correction to a quantized value to mitigate the effect of varying pressure. Going over all possible shifts is not a practical solution, but theoretically this can still lead to a polynomial-time solution.

We now formally define our problem, which primarily captures issues 3 and 4. (Below for any integers $t \geq 1$, $\binom{\mathbb{F}_q}{t}$ denotes the set of all subsets of $\mathbb{F}_q$ of size exactly t.) The following are the components of the problem:

- Integers $k \geq 1$, $n \geq t \geq 1$
- Secret $s \in \mathbb{F}_q^k$
- Fingerprint $f \in \binom{\mathbb{F}_q}{t}$
- $\textsc{lock} : \mathbb{F}_q^k \times \binom{\mathbb{F}_q}{t} \to \binom{\mathbb{F}_q}{n}$
- $\textsc{unlock} : \binom{\mathbb{F}_q}{t} \times \binom{\mathbb{F}_q}{n} \to \mathbb{F}_q^k$

The goal of the problem is to define the functions LOCK and UNLOCK such that they satisfy these two properties (for some $c < t$):

1. (c-completeness.) For any $f, f' \in \binom{\mathbb{F}_q}{t}$ such that $|f - f'| \leq c$, the following should hold:

$$\textsc{unlock}\big(\textsc{lock}(s, f), f'\big) = s.$$

2. (Soundness.) It should be "hard" for an adversary to get s from $\textsc{lock}(s, f)$. (The notion of "hard" will be more formally defined later.)

Note that the completeness property corresponds to the matching property we need from our hash function, while the soundness property corresponds to the security property of the hash function.

18.2.2 Two Futile Attempts

We begin with two attempts at designing the LOCK and UNLOCK functions, which will not work. However, later we will see how we can combine both to get our final solution.

For this section, unless mentioned otherwise, we will assume that the original fingerprint f is given by the set $\{\alpha_1, \ldots, \alpha_t\}$.

Attempt 1. We begin with a scheme that focuses on the soundness property. A very simple idea, which is what we will start off with, would be to just add $n-t$ random values to f to get our vault. The intuition, which can be made precise, is that an adversary just looking at the vault will just see random points and will not be able to recover f from the random set of points. The catch of course that this scheme has terrible completeness. In particular, if we get a match between a value in the second fingerprint f' and the vault, we have no way to know whether the match is to one of the original values in f or if the match is with one of the random "chaff" points there were added earlier.

Attempt 2. Next, we specify a scheme that has good completeness (but has pretty bad soundness).

We begin with the LOCK function:

$$\text{LOCK}_2(s, f) = \{(\alpha_1, P_s(\alpha_1)), \ldots, (\alpha_t, P_s(\alpha_t))\},$$

where $P_s(X) = \sum_{i=0}^{k-1} s.X^i$ and recall $f = \{\alpha_1, \ldots, \alpha_t\}$. (Note that we have $n = t$.)

The main intuition behind LOCK_2 is the following. Given another fingerprint $f' = \{\beta_1, \ldots, \beta_t\}$ such that it is close enough to f, i.e. $|f \setminus f'| \le c$, for every value in $f \cap f'$, we will know the corresponding P_s value and thus, we can use the fact that we can decode Reed-Solomon codes from erasures to recover the secret s. We formally present UNLOCK_2 as Algorithm 33.

Algorithm 33 UNLOCK_2

INPUT: Vault $\{(\alpha_1, y_1), \ldots, (\alpha_t, y_t)\} = \text{LOCK}(s, f)$ and another fingerprint $f' = \{\beta_1, \ldots, \beta_t\}$
OUTPUT: s if $|f \setminus f'| \le c$

1: FOR $i = 1, \ldots, t$ DO
2: IF there exists a $j \in [t]$ such that $\alpha_i = \beta_j$ THEN
3: $z_j \leftarrow y_i$
4: ELSE
5: $z_j \leftarrow ?$
6: $\mathbf{z} \leftarrow (z_1, \ldots, z_t)$
7: Run Algorithm from Theorem 12.2.1 to correct $\mathbf{z}$ from erasures for RS codes with evaluation points $\{\beta_1, \ldots, \beta_t\}$ and output resulting message as s.

The following result is fairly simple to argue.

Lemma 18.2.1. *The pair* $(\text{LOCK}_2, \text{UNLOCK}_2)$ *of functions is* $(t-k)$*-complete. Further, both functions can be implemented in polynomial time.*

Proof. Let us assume $|f \setminus f'| \le t-k$. Now as both f and f' have exactly t values, this means that $\mathbf{z}$ has at most $t-k$ erasures. Thus, by Theorem 12.2.1, Step 6 will output s and UNLOCK_2 can be implemented in polynomial time. Further, the claim on the polynomial run time of LOCK_2 follows from the fact that one can do encoding of Reed-Solomon code in polynomial time. □

Unfortunately, $(\text{LOCK}_2, \text{UNLOCK}_2)$ pair has terrible soundness. This is because the vault $\{(\alpha_1, y_1), \ldots, (\alpha_t, y_t)\}$ has f in the first component in each pair. This an adversary can just read off those values and present $f' = \{\alpha_1, \ldots, \alpha_t\}$, which would imply that $\text{UNLOCK}_2(\text{LOCK}_2(s, f), f') = s$, which means that the vault would be "broken."

18.3 The Final Fuzzy Vault

So far we have seen two attempts: one that (intuitively) has very good soundness but no completeness and another which has good completeness but terrible soundness. It is natural to consider if we can combine both of these attempts and get the best of both worlds. Indeed, it turns we can easily combine both of our previous attempts to get the final fuzzy vault.

Algorithm 34 presents the new LOCK_3 function.

Algorithm 34 LOCK_3

INPUT: Fingerprint $f = \{\alpha_1, \ldots, \alpha_t\}$ and secret $s = (s_0, \ldots, s_{k-1}) \in \mathbb{F}_q^k$
OUTPUT: Vault with f locking s

1: $R, T \leftarrow \emptyset$
2: $P_s(X) \leftarrow \sum_{i=0}^{k-1} s_i \cdot X^i$
3: FOR $i = 1, \ldots, t$ DO
4: $\quad T \leftarrow T \cup \{\alpha_i\}$
5: FOR $i = t+1, \ldots, n$ DO
6: $\quad \alpha_i$ be a random element from $\mathbb{F}_q \setminus T$
7: $\quad T \leftarrow T \cup \{\alpha_i\}$
8: FOR every $\alpha \in T$ DO
9: $\quad \gamma$ be a random element from $\mathbb{F}_q \setminus P_s(\alpha)$
10: $\quad R \leftarrow R \cup \{(\alpha, \gamma)\}$
11: Randomly permute R
12: RETURN R

Algorithm 35 presents the new UNLOCK_3 function.

The following result is a simple generalization of Lemma 18.2.1.

Lemma 18.3.1. *The pair* $(\text{LOCK}_3, \text{UNLOCK}_3)$ *of functions is* $(t-k)/2$*-complete. Further, both functions can be implemented in polynomial time.*

Proof. Let us assume $|f \setminus f'| \le (t-k)/2$. Now as both f and f' have exactly t values, it implies that $|f \cap f'| \ge (t+k)/2$. Further for each $j \in [t]$ such that $\beta_j \in f \cap f'$, we have that $z_j = P_s(\beta_j)$. In other words, this means that $\mathbf{z}$ has at most $(t-k)/2$ errors.[3] Thus, by Theorem 12.2.2, Step 6 will output s and UNLOCK_3 can be implemented in polynomial time. Further, the claim on the polynomial run time of LOCK_3 follows from the fact that one can do encoding of Reed-Solomon code in polynomial time. □

[3] To be more precise if $\mathbf{z}$ has e errors and s erasures w.r.t. the codeword corresponding to s, then $2e + s \le t - k$.

Algorithm 35 UNLOCK$_2$

INPUT: Vault $\{(\alpha_1, y_1), \ldots, (\alpha_n, y_n)\} = \text{LOCK}(s, f)$ and another fingerprint $f' = \{\beta_1, \ldots, \beta_t\}$
OUTPUT: s if $|f \setminus f'| \leq c$

```
FOR i = 1,...,t DO
    IF there exists a j ∈ [n] such that α_i = β_j THEN
        z_j ← y_i
    ELSE
        z_j ←?
z ← (z_1,...,z_t)
Run Algorithm from Theorem 12.2.2 to correct z from errors and erasures for RS codes with
evaluation points {β_1,...,β_t} and output resulting message as s.
```

18.3.1 Soundness

To avoid getting into too much technical details, we will present a high level argument for why the proposed fuzzy vault scheme has good soundness. Given a vault $\{(\alpha_1, y_1), \ldots, (\alpha_n, y_n)\} = \text{LOCK}_3(s, f)$, we know that there are exactly t values (i.e. those $\alpha_j \in f$) such that the polynomial $P_s(X)$ agrees with the vault on exactly those t points. Thus, an intuitive way to argue the soundness of the vault would be to argue that there exists a lot other secrets $s' \in \mathbb{F}_q^k$ such that $P_{s'}(X)$ also agrees with the vault in exactly t positions. (One can formalize this intuition and argue that the vault satisfies a more formal definition of soundness but we will skip those details.)

We will formalize the above argument by proving a slightly different result (and we will leave the final proof as an exercise).

Lemma 18.3.2. *Let $V = \{(x_1, y_1), \ldots, (x_n, y_n)$ be n independent random points from $\mathbb{F}_q \times \mathbb{F}_q$. Then, in expectation, there are at least $\frac{1}{3} \cdot q^k \cdot \left(\frac{n}{qt}\right)^t$ polynomials $P(X)$ of degree at most $k-1$ such that for exactly t values of $j \in [n]$, we have $P(x_j) = y_j$.*

Proof. Consider a fixed polynomial $P(X)$ and a $j \in [n]$. Then for any $x_j \in \mathbb{F}_q$, the probability that for a random $y_j \in F_q$, $P(x_j) = y_j$ is exactly $1/q$. Further, these probabilities are all independent. This implies that the probability that $P(X)$ agrees with V in exactly t positions is given by

$$\binom{n}{t} \cdot \left(\frac{1}{q}\right)^t \cdot \left(1 - \frac{1}{q}\right)^{n-t} \geq \frac{1}{3}\left(\frac{n}{qt}\right)^t.$$

Since there are q^k such polynomials, the claimed result follows. □

We note that there are two aspects of the above lemma that are not satisfactory. (i) The result above is for a vault V with completely random points whereas we would like to prove a similar result but with $V = \text{LOCK}_3(s, f)$ and (ii) Instead of a bound in expectation, we would like to prove a similar exponential lower bound but with high probability. We leave the proof that these can be done as an exercise. (Hint: Use the "Inverse Markov Inequality.")

18.4 Bibliographic Notes

The fuzzy vault presented in this chapter is due to Juels and Sudan [57]. The “inverse Markov inequality" first appeared in Dumer et al. [24].

Chapter 19

Finding Defectives: Group Testing

Consider the following situation that arises very frequently in *biological screens*. Say there are N individuals and the objective of the study is to identify the individuals with a certain "biomarker" that could e.g. indicate that the individual has some specific disease or is at risk for a certain health condition. The naive way to do this would be to test each person individually, that is:

1. Draw sample (e.g. a blood or DNA sample) from a given individual,
2. Perform required tests, then
3. Determine presence or absence of the biomarker.

This method of one test per person will gives us a total of N tests for a total of N individuals. Say we had more than $70-75\%$ of the population infected. At such large numbers, the use of the method of individual testing is reasonable. However, our goal is to achieve effective testing in the more likely scenario where it doesn't make sense to test $100,000$ people to get just (say) 10 positives.

The feasibility of a more effective testing scheme hinges on the following property. We can combine blood samples and test a combined sample together to check if at least one individual has the biomarker.

The main question in group testing is the following: If we have a very large number of items to test, and we know that only certain few will turn out positive, what is a nice and efficient way of performing the tests?

19.1 Formalization of the problem

We now formalize the group testing problem. The input to the problem consists of the following:

- The total number of individuals: N.
- An upper bound on the number of infected individuals d.

- The input can be described as a vector $\mathbf{x} = (x_1, x_2, ..., x_n)$ where $x_i = 1$ if individual i has the biomarker, else $x_i = 0$.

Note that $wt(\mathbf{x}) \leq d$. More importantly, notice that the vector $\mathbf{x}$ is an implicit input since we do not know the positions of 1s in the input. The only way to find out is to run the *tests*. Now, we will formalize the notion of a test.

A query/test S is a subset of $[N]$. The answer to the query $S \subseteq [N]$ is defined as follows:

$$A(S) = \begin{cases} 1 & \text{if } \sum_{i \in S} x_i \geq 1; \\ 0 & \text{otherwise.} \end{cases}$$

Note that the answer to the S is the logical-OR of all bits in S, i.e. $A(S) = \bigvee_{i \in S} x_i$.

The goal of the problem is to compute $\mathbf{x}$ and minimize the number of tests required to determine $\mathbf{x}$.

Testing methods. There is another aspect of the problem that we need specify. In particular, we might need to restrict how the tests interact with each other. Below are two commons ways to carry out tests:

1. *Adaptive group testing* is where we test a given subset of items, get the answer and base our further tests on the outcome of the previous set of tests and their answers.

2. *Non-Adaptive group testing* on the other hand is when all our tests are set even before we perform our first test. That is, all our tests are decided a priori.

Non-adaptive group testing is crucial for many applications. This is because the individuals could be geographically spread pretty far out. Due to this, adaptive group testing will require a very high degree of co-ordination between the different groups. This might actually increase the cost of the process.

Notation. We will also need notation to denote the minimum number of tests needed in group testing. Towards this end, we present the following two definitions.

Definition 19.1.1 ($t(d, N)$)**.** Given a subset of N items with d defects represented as $\mathbf{x} \in \{0,1\}^N$, the minimum number of *non-adaptive* tests that one would have to make is defined as $t(d, N)$.

Definition 19.1.2. $t^a(d, N)$: Given a set of N items with d defects, $t^a(d, N)$ is defined as the number of *adaptive* tests that one would have to make to detect all the defective items.

The obvious questions are to prove bounds on $t(d, N)$ and $t^a(d, N)$:

Question 19.1.1. *Prove asymptotically tight bounds on $t(d, N)$.*

Question 19.1.2. *Prove asymptotically tight bounds on $t^a(d,N)$.*

We begin with some simple bounds:

Proposition 19.1.1. *For every $1 \le d \le N$, we have*

$$1 \le t^a(d,N) \le t(d,N) \le N.$$

Proof. The last inequality follows from the naive scheme of testing all individuals with singleton tests while the first inequality is trivial. The reason for $t^a(d,N) \le t(d,N)$ is due to the fact that any non-adaptive test can be performed by an adaptive test by running all of the tests in the first step of the adaptive test. Adaptive tests can be faster than non-adaptive tests since the test can be changed after certain things are discovered. □

Representing the set of tests as a matrix. It turns out that is is useful to represent a non-adaptive group testing scheme as a matrix. Next, we outline the natural such representation. For, $S \subseteq [N]$, define $\chi_S \in \{0,1\}^N$ such that $i \in S$ if and only if $\chi_S(i) = 1$. Consider a non-adaptive group testing scheme with t test $S_1, \ldots, S_t$. The corresponding matrix M is a $t \times N$ matrix where the ith row is χ_{S_i}. (Note that the trivial solution of testing each individual as a singleton set would just be the $N \times N$ identity matrix.) In other words, $M = \{M_{ij}\}_{i \in [t], j \in [N]}$ such that $M_{ij} = 1$ if $j \in S_i$ and $M_{ij} = 0$ otherwise.

If we assume that multiplication is logical AND ($\wedge$) and addition is logical OR ($\vee$), then we have $M \times \mathbf{x} = \mathbf{r}$ where $\mathbf{r} \in \{0,1\}^t$ is the vector of the answers to the t tests. We will denote this operation as $M \odot \mathbf{x}$. To think of this in terms of testing, it is helpful to visualize the matrix-vector multiplication. Here, $\mathbf{r}$ will have a 1 in position i if and only if there was a 1 in that position in both M and $\mathbf{x}$ i.e. if that person was tested with that particular group and if he tested out to be positive.

Thus, our goal is to get to compute $\mathbf{x}$ from $M \odot \mathbf{x}$ with as small a t as possible.

19.2 Bounds on $t^a(d,N)$

In this section, we will explore lower and upper bounds on $t^a(d,N)$ with the ultimate objective of answering Question 19.1.2.

We begin with a lower bound that follows from a simple counting argument.

Proposition 19.2.1. *For every $1 \le d \le N$,*

$$t^a(d,N) \ge d \log \frac{N}{d}.$$

Proof. Fix any valid adaptive group testing scheme with t tests. Observe that if $\mathbf{x} \ne \mathbf{y} \in \{0,1\}^N$, with $wt(\mathbf{x}), wt(\mathbf{y}) \le d$ then $\mathbf{r}(\mathbf{x}) \ne \mathbf{r}(\mathbf{y})$, where $\mathbf{r}(\mathbf{x})$ denotes the result vector for running the tests

on $\mathbf{x}$ and similarly for $\mathbf{r}(\mathbf{y})$. The reason for this is because two valid inputs cannot give the same result. If this were the case and the results of the tests gave $\mathbf{r}(\mathbf{x}) = \mathbf{r}(\mathbf{y})$ then it would not be possible to distinguish between $\mathbf{x}$ and $\mathbf{y}$.

The above observation implies that total number of distinct test results is the number distinct binary vectors with Hamming weight at most d, i.e. $Vol_2(d,N)$. On the other hand, the number of possible distinct t-bit vectors is at most 2^t, which with the previous argument implies that

$$2^t \geq Vol_2(d,N)$$

and hence, it implies that

$$t \geq \log Vol_2(d,N).$$

Recall that

$$Vol_2(d,N) \geq \binom{N}{d} \geq \left(\frac{N}{d}\right)^d,$$

where the first inequality follows from (3.23) and the second inequality follows from Lemma B.1.1. So $t \geq d\log(N/d)$, as desired. □

It turns out that $t^a(d,N)$ is also $O\left(d\log\left(\frac{N}{d}\right)\right)$. (See Exercise 19.1.) This along with Proposition 19.2.1 implies that $t^a(d,N) = \Theta\left(d\log\left(\frac{N}{d}\right)\right)$, which answers Question 19.1.2. The upper bound on $t^a(d,N)$ follows from an adaptive group testing scheme and hence does not say anything meaningful for Question 19.1.1. (Indeed, we will see later that $t(d,N)$ cannot be $O(d\log(N/d))$.) Next, we switch gears to talk about non-adaptive group testing.

19.3 Bounds on $t(d,N)$

We begin with the simplest case of $d = 1$. In this case it is instructive to recall our goal. We want to define a matrix M such that given any $\mathbf{x}$ with $wt(\mathbf{x}) \leq 1$, we should be able to compute $\mathbf{x}$ from $M \odot \mathbf{x}$. In particular, let us consider the case when $\mathbf{x} = \mathbf{e}_i$ for some $i \in [N]$. Note that in this case $M \odot \mathbf{x} = M^i$, where M^i is the ith column of M. Hence we should design M such that M^i uniquely defined i. We have already encountered a similar situation before in Section 2.6 when trying to decode the Hamming code. It turns out that is suffices to pick M as the parity check matrix of a Hamming code. In particular, we can prove the following result:

Proposition 19.3.1. $t(1,N) \leq \lceil \log(N+1) \rceil + 1$

Proof. We prove the upper bound by exhibiting a matrix that can handle non adaptive group testing for $d = 1$. The group test matrix M is the parity check matrix for $[2^m - 1, 2^m - m - 1, 3]_2$, i.e. H_m where the i-th column is the binary representation of i (recall Section 2.4). This works because when performing $H_m \odot \mathbf{x} = \mathbf{r}$, if $wt(\mathbf{v}) \leq 1$ then $\mathbf{r}$ will correspond to the binary representation of i. Further, note that if $wt(\mathbf{x}) = 0$, then $\mathbf{r} = \mathbf{0}$, which is exactly $\mathbf{x}$. Hence, $M \odot \mathbf{x}$ uniquely identifies $\mathbf{x}$ when $wt(\mathbf{x}) \leq 1$, as desired.

If $N \neq 2^m - 1$ for any m, the matrix H_m corresponding to the m such that $2^{m-1} - 1 < N < 2^m - 1$ can be used by adding 0s to the end of $\mathbf{x}$. By doing this, decoding is "trivial" for both

cases since the binary representation is given for the location. So the number of tests is at most $\lceil \log(N+1) \rceil + 1$, which completes the proof. $\square$

Note that Propositions 19.1.1 and 19.2.1 imply that $t(d,N) \geq \log N$, which with the above result implies that $t(1,N) = \Theta(\log N)$. This answers Question 19.1.1 for the case of $d = 1$. We will see later that such a tight answer is not known for larger d. However, at the very least we should try and extend Proposition 19.3.1 for larger values of d. In particular,

Question 19.3.1. *Prove asymptotic upper bounds on $t(d,N)$ that hold for every $1 < d \leq N$.*

We would like to point out something that was implicitly used in the proof of Proposition 19.3.1. In particular, we used the implicitly understanding that a non-adaptive group testing matrix M should have the property that given any $\mathbf{x} \in \{0,1\}^N$ such that $wt(\mathbf{x}) \leq d$, the result vector $M \odot \mathbf{x}$ should uniquely determine $\mathbf{x}$. This notion is formally captured in the following definition of non-adaptive group testing matrix:

Definition 19.3.1. A $t \times N$ matrix M is *d-separable* if and only if for every $S_1 \neq S_2 \subseteq [N]$ such that $|S_1|, |S_2| \leq d$, we have

$$\bigcup_{j \in S_1} M^j \neq \bigcup_{i \in S_2} M^i.$$

In the above we think of a columns $M^i \in \{0,1\}^t$ as a subset of $[t]$ and for the rest of this chapter we will use both views of the columns of a group testing matrix. Finally, note that the above definition is indeed equivalent to our earlier informal definition since for any $\mathbf{x} \in \{0,1\}^N$ with $wt(\mathbf{x}) \leq d$, the vector $M \odot \mathbf{x}$ when thought of as its equivalent subset of $[t]$ is exactly the set $\cup_{i \in S} M^i$, where S is the support of $\mathbf{x}$, i.e. $S = \{i | x_i = 1\}$.

Like in the coding setting, where we cared about the run time of the decoding algorithm, we also care about the time complexity of the decoding problem (given $M \odot \mathbf{x}$ compute $\mathbf{x}$) for group testing. We will now look at the obvious decoding algorithm for d-separable matrices: just check all possible sets that could form the support of $\mathbf{x}$. Algorithm 36 has the details.

The correctness of Algorithm 36 follows from Definition 19.3.1. Further, it is easy to check that this algorithm will run in $N^{\Theta(d)}$ time, which is not efficient for even moderately large d. This naturally leads to the following question:

Question 19.3.2. *Do there exists d-separable matrices that can be efficient decoded?*

We would like to remark here that the matrices that we seek in the answer to Question 19.3.2 should have small number of tests (as otherwise the identity matrix answers the question in the affirmative).

Algorithm 36 Decoder for Separable Matrices

INPUT: Result vector $\mathbf{r}$ and d-separable matrix M
OUTPUT: $\mathbf{x}$ if $\mathbf{r} = M \odot \mathbf{x}$ else `Fail`

```
R ← {i | r_i = 1}.
FOR Every T ⊆ [N] such that |T| ≤ d DO
    S_T ← ∪_{i∈T} M^i
    IF R = S_T THEN
        x ← (x_1, ..., x_N) ∈ {0,1}^N such that x_i = 1 if and only i ∈ T.
        RETURN x
RETURN Fail
```

19.3.1 Disjunct Matrices

We now define a stronger notion of group testing matrices that have a more efficient decoding algorithm than d-separable matrices.

Definition 19.3.2. A $t \times N$ matrix M is d-disjunct if and only if for every $S \subset [N]$ with $|S| \leq d$ and for every $j \notin S$, there exist an $i \in [t]$ such that $M_{ij} = 1$ but for all $k \in S$, $M_{ik} = 0$. Or equivalently

$$M^j \not\subseteq \bigcup_{k \in S} M^k.$$

For an illustration of the definition, see Figure 19.3.2.

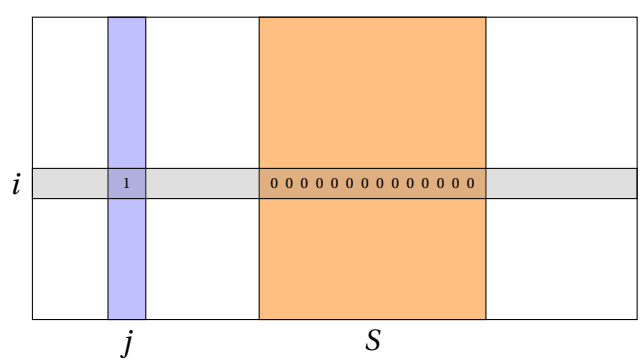


Figure 19.1: Pick a subset S (not necessarily contiguous). Then pick a column j that is not present in S. There will always be i such that row i has a 1 in column j and all zeros in S.

Next we argue that disjunctness is a sufficient condition for separability.

Lemma 19.3.2. *Any d-disjunct matrix is also d-separable.*

Proof. For contradiction, assume M is d-disjunct but not d-separable. Since M is not d-separable, then union of two subset $S \neq T \subset [N]$ of size at most d each are the same; i.e.

$$\bigcup_{k \in S} M^k = \bigcup_{k \in T} M^k.$$

Since $S \neq T$, there exists $j \in T \setminus S$. But we have

$$M^j \subseteq \bigcup_{k \in T} M^k = \bigcup_{k \in S} M^k,$$

where the last equality follows from the previous equality. However, since by definition $j \notin S$, the above contradicts the fact that M is d-disjunct. □

In fact, it turns out that disjunctness is also almost a necessary condition: see Exercise 19.2.

Next, we show the real gain of moving to the notion of disjunctness from the notion of separability.

Lemma 19.3.3. *There exists a $O(tN)$ time decoding algorithm for any $t \times N$ matrix that is d-disjunct.*

Proof. The proof follows from the following two observations.

First, say we have a matrix M and a vector $\mathbf{x}$ and $\mathbf{r} = M \odot \mathbf{x}$ such that $r_i = 1$. Then there exists a column j in matrix that made it possible i.e. if $r_i = 1$, then there exists a j such that $M_{ij} = 1$ and $x_j = 1$.

Second, let T be a subset and j be a column not in T where $T = \{\ell \mid x_\ell = 1\}$ and $|T| \leq d$. Consider the ith row such that T has all zeros in the ith row, then $r_i = 0$. Conversely, if $r_i = 0$, then for every $j \in [N]$ such that $M_{ij} = 1$, it has to be the case that $x_j = 0$. This naturally leads to the decoding algorithm in Algorithm 37.

The correctness of Algorithm 37 follows from the above observation and it can be checked that the algorithm runs in time $O(tN)$– see Exercise 19.3. □

Modulo the task of exhibiting the existence of d-disjunct matrices, Lemmas 19.3.3 and 19.3.2 answer Question 19.3.2 in the affirmative. Next, we will tackle the following question:

Question 19.3.3. *Design d-disjunct matrices with few rows.*

As we will see shortly answering the above question will make connection to coding theory becomes even more explicit.

Algorithm 37 Naive Decoder for Disjunct Matrices

INPUT: Result vector $\mathbf{r}$ and d-disjunct matrix M
OUTPUT: $\mathbf{x}$ if $M \odot \mathbf{x} = \mathbf{r}$ else `Fail`

```
Initialize x ∈ {0,1}^N to be the all ones vector
FOR every i ∈ [t] DO
    IF r_i = 0 THEN
        FOR Every j ∈ [N] DO
            IF M_ij = 1 THEN
                x_j ← 0
IF M ⊙ x = r THEN
    RETURN x
ELSE
    RETURN Fail
```

19.4 Coding Theory and Disjunct Matrices

In this section, we present the connection between coding theory and disjunct matrices with the final goal of answering Question 19.3.3. First, we present a sufficient condition for a matrix to be d-disjunct.

Lemma 19.4.1. *Let $1 \le d \le N$ be an integer and M be a $t \times N$ matrix, such that*

(i) For every $j \in [N]$, the jth column has at least $w_{\min}$ ones in it, i.e. $\left|M^j\right| \ge w_{\min}$ and

(ii) For every $i \ne j \in [N]$, the i and j'th columns have at most $a_{\max}$ ones in common, i.e. $\left|M^i \cap M^j\right| \le a_{\max}$

for some integers $a_{\max} \le w_{\min} \le t$. Then M is a $\left\lfloor \frac{w_{\min}-1}{a_{\max}} \right\rfloor$-disjunct.

Proof. For notational convenience, define $d = \left\lfloor \frac{w_{\min}-1}{a_{\max}} \right\rfloor$. Fix an arbitrary $S \subset [N]$ such that $|S| \le d$ and a $j \notin S$. Note we have to show that

$$M^j \not\subseteq \cup_{i \in S} M^i,$$

or equivalently

$$M^j \not\subseteq \cup_{i \in S} \left(M^i \cap M^j\right).$$

We will prove the above by showing that

$$\left|M^j \setminus \cup_{i \in S} \left(M^i \cap M^j\right)\right| > 0.$$

Indeed, consider the following sequence of relationships:

$$\begin{aligned}
\left|M^j \setminus \cup_{i\in S}\left(M^i \cap M^j\right)\right| &= \left|M^j\right| - \left|\cup_{i\in S}\left(M^i \cap M^j\right)\right| \\
&\geq \left|M^j\right| - \sum_{i\in S}\left|\left(M^i \cap M^j\right)\right| && (19.1)\\
&\geq w_{\min} - |S|\cdot a_{\max} && (19.2)\\
&\geq w_{\min} - d\cdot a_{\max} && (19.3)\\
&\geq w_{\min} - \frac{w_{\min}-1}{a_{\max}}\cdot a_{\max} && (19.4)\\
&= 1.
\end{aligned}$$

In the above, (19.1) follows from the fact that size of the union of sets is at most the sum of their sizes. (19.2) follows from the definitions of $w_{\min}$ and $a_{\max}$. (19.3) follows from the fact that $|S| \leq d$ while (19.4) follows from the definition of d. The proof is complete. □

Next, we present a simple way to convert a code into a matrix. Let $C \subseteq [q]^t$ be a code such that $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$. Then consider the matrix M_C whose i'th column is $\mathbf{c}_i$, i.e.

$$M_C = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}.$$

Thus, by Lemma 19.4.1, to construct an $\left\lfloor \frac{w_{\min}-1}{a_{\max}} \right\rfloor$-disjunct matrix, it is enough to design a binary code $C^* \subseteq \{0,1\}^t$ such that (i) for every $\mathbf{c} \in C^*$, $wt(\mathbf{c}) \geq w_{\min}$ and (ii) for every $\mathbf{c}^1 \neq \mathbf{c}^2 \in C^*$, we have $|\{i | c_i^1 = c_i^2 = 1\}| \leq a_{\max}$. Next, we look at the construction of such a code.

19.4.1 Kautz-Singleton Construction

In this section, we will prove the following result:

Theorem 19.4.2. *For every integer $d \geq 1$ and large enough $N \geq d$, there exists a $t \times N$ matrix is d-disjunct where $t = O\left(d^2 \left(\log_d N\right)^2\right)$.*

Note that the above result answers Question 19.3.3. It turns out that one can do a bit better: see Exercise 19.4.

Towards this end, we will now study a construction of C^* as in the previous section due to Kautz and Singleton. As we have already seen in Chapter 10, concatenated codes are a way to design binary codes. For our construction of C^*, we will also use code concatenation. In particular, we will pick $C^* = C_{\text{out}} \circ C_{\text{in}}$, where C_{out} is a $[q, k, q-k+1]_q$ Reed-Solomon code (see Chapter 5) while the inner code $C_{\text{in}} : \mathbb{F}_q \to \{0,1\}^q$ is defined as follows. For any $i \in \mathbb{F}_q$, define $C_{\text{in}}(i) = \mathbf{e}_i$. Note that $M_{C_{\text{in}}}$ is the identity matrix and that $N = q^k$ and $t = q^2$.

Example 19.4.3. *Let $k = 1$ and $q = 3$. Note that by our choice of $[3,1]_3$ Reed-Solomon codes, we have $C_{\text{out}} = \{(0,0,0),(1,1,1),(2,2,2)\}$. In other words,*

$$M_{C_{\text{out}}} = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 1 & 2 \end{bmatrix}.$$

Then the construction of M_{C^*} can be visualized as in Figure 19.4.3.

$$\underbrace{\begin{bmatrix} 0 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 1 & 2 \end{bmatrix}}_{M_{C_{\text{out}}}} \circ \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{M_{C_{\text{in}}}} \rightarrow \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{M_{C^*}}$$

Figure 19.2: Construction of the final matrix M_{C^*} from $M_{C_{\text{out}}}$ and $M_{C_{\text{in}}}$ from Example 19.4.3. The rows in M_{C^*} that correspond to the same row in $M_{C_{\text{out}}}$ have the same color.

Next, we instantiate parameters in the Kautz-Singleton construction to prove Theorem 19.4.2.

Proof of Theorem 19.4.2. We first analyze the construction to determine the parameters $w_{\min}$ and $a_{\max}$. Then we pick the parameters q and k in terms of d to complete the proof.

Recall that $N = q^k$ and $t = q^2$. It is easy to check that every column of M_{C^*} has exactly q ones in it. In other words, $w_{\min} = q$. Next, we estimate $a_{\max}$.

Divide the rows into q sized chunks, and index the $t = q^2$ rows by pairs in $[q] \times [q]$. Recall that each column in M_{C^*} corresponds to a codeword in C^*. For notational convenience, we will use M for M_{C^*}. Note that for any row $(i,j) \in [q] \times [q]$ and a column index $\ell \in [N]$, we have $M_{(i,j),\ell} = 1$ if and only if $\mathbf{c}_\ell(j) = j$ (where we use some fixed bijection between $\mathbb{F}_q$ and $[q]$ and $\mathbf{c}_\ell$ is the ℓ'th codeword in C_{out}). In other words, the number of rows where the ℓ_1th and ℓ_2th columns both have a one is exactly the number of positions where $\mathbf{c}_{\ell_1}$ and $\mathbf{c}_{\ell_2}$ agree, which is exactly $q - \Delta(\mathbf{c}_{\ell_1}, \mathbf{c}_{\ell_2})$. Since C_{out} is a $[q,k,q-k+1]_q$ code, the number of rows where any two columns agree is at most $k-1$. In other words, $a_{\max} = k-1$.[1]

Lemma 19.4.1 implies that M_{C^*} is d-disjunct if we pick

$$d = \left\lfloor \frac{q-1}{k-1} \right\rfloor.$$

[1]The equality is obtained due to columns that corresponds to codewords that agree in exactly $k-1$ positions.

Thus, we pick q and k such that the above is satisfied. Note that we have $q = O(kd)$. Further, since we have $N = q^k$, we have

$$k = \log_q N.$$

This implies that $q = O(d \cdot \log_q N)$, or $q \log q = O(d \log N)$. In other words we have

$$q = O(d \log_d N).$$

Recalling that $t = q^2$ completes the proof. □

An inspection of the proof of Theorem 19.4.2 shows that we only used the distance of the Reed-Solomon code and in particular, any C_{out} with large enough distance suffices. In particular, if we pick C_{out} to be a random code over an appropriate sized alphabet then one can obtain $t = O(d^2 \log N)$. (See Exercise 19.5 for more on this.) Note that this bound is incomparable to the bound in Theorem 19.4.2. It turns out that these two are the best known upper bounds on $t(d, N)$. In particular,

Open Question 19.4.1. *Can we beat the upper bound of* $O\left(d^2 \cdot \min\left(\log(N/d), \log_d^2 N\right)\right)$ *on* $t(d, N)$*?*

It turns out that the quadratic dependence on d in the upper bounds is tight. In fact it is known that $t(d, N) \geq \Omega(d^2 \log_d N)$. (See Exercises 19.7 and 19.8.)

Next, we present an application of group testing in the field of data stream algorithms, which in turn will bring out another facet of the connection between coding theory and group testing.

19.5 An Application in Data Stream Algorithms

Let us consider the problem of tracking updates on stock trades. Given a set of trades $(i_1, u_1), \cdots, (i_m, u_m)$, where i_j is the stock id for the jth trade, u_j is the amount of the stocks in the jth trade. The problem is to keep track of the top d stocks. Such a problem is also called hot items/ heavy hitters problem.

Let N be the total number of stocks in the market. This problem could be solved in $O(m) + O(N \log N) \approx O(m)$ time and $O(N)$ space by setting a $O(N)$ size buffer to record the total number of trading for each stock and then sort the buffer later. However, m could be of the order of millions for one minute's trading, e.g. in the first minute of April 19, 2010, there are 8077600 stocks were traded. Taking the huge amount of trades into consideration, such an algorithm is not practical.

A more practical model of efficient algorithms in this context is one of *data stream algorithm*, which is defined as follows.

Definition 19.5.1. A data stream algorithm has four requirements listed below:

1. The algorithm should make one sequential pass over the input.

2. The algorithm should use poly-log space. (In particular, it cannot store the entire input.)

3. The algorithm should have poly-log update time, i.e. whenever a new item appears, the algorithm should be able to update its internal state in poly-log time.

4. The algorithm should have poly-log reporting time, i.e. at any point of time the algorithm should be able to return the required answers in poly-log time.

Thus, ideally we would like to design a data stream algorithm to solve the hot items problem that we discussed earlier. Next, we formally define the hot items problem. We begin with the definition of frequencies of each stock:

Definition 19.5.2. Let f_ℓ denote the total count for the stock id ℓ. Initially $f_\ell = 0$, given (ℓ, u_ℓ), $f_\ell \leftarrow f_\ell + u_\ell$.

Next, we define the hot items problem.

Definition 19.5.3 (Hot Items Problem)**.** Given N different items, for m input pairs of data (i_ℓ, u_ℓ) for $1 \le \ell \le m$, where $i_\ell \in [N]$ indicates the item index and u_ℓ indicates corresponding count. The problem requires updating the count $f_\ell (1 \le \ell \le m)$ for each item, and to output all item indices j such that $f_j > \frac{\sum_{\ell=1}^{N} u_\ell}{d}$. (Any such item is called a *hot item*.)

Note that there can be at most d hot items. In this chapter, we will mostly think of d as $O(\log N)$. Hot items problem is also called heavy hitters problems. We state the result below without proof:

Theorem 19.5.1. *Computing hot items exactly by a deterministic one pass algorithm needs* $\Omega(n)$ *space (even with exponential time).*

This theorem means that we cannot solve the hot items problem in poly-log space as we want. However, we could try to find solutions for problems around this. The first one is to output an approximate solution, which will output a set that contains all hot items and some non-hot items. For this solution, we want to make sure that the size of the output set is not too large (e.g. outputting $[N]$ is not a sensible solution).

Another solution is to make some assumptions on the input. For example, we can assume Zipf-like distribution of the input data, which means only a few items appear frequently. More specifically, we can assume *heavy-tail distribution* on the input data, i.e.:

$$\sum_{\ell:\text{not hot}} f_\ell \le \frac{m}{d}. \tag{19.5}$$

This is reasonable for many applications, such as hot stock finding, where only a few of them have large frequency. Next, we will explore the connection between group testing and hot items problem based on this assumption.

19.5.1 Connection to Group Testing

Let us recall the naive solution that does not lead to a data stream algorithm: for each item $j \in [N]$, we maintain the actual count of number of trades for stock j. In other words, at any point of time, if C_j is the count for stock j, we have $C_j = f_j$. Another way to represent this is if M is the $N \times N$ identity matrix, then we maintain the vector of counts via $M \cdot \mathbf{f}$, where $\mathbf{f}$ is the vector of the frequencies of the items. Paralleling the story in group testing where we replace the identity matrix with a matrix with fewer rows, a natural idea here would be to replace M by matrix with fewer rows that utilizes the fact that there can be at most d hot items. Next, we show that this idea works if the heavy tail distribution holds. In particular, we will reduce the hot items problem to the group testing problem.

We now show how we solve the hot items problem from Definition 19.5.3. Let M be an $t \times N$ matrix that is d-disjunct. We maintain counters $C_1, \ldots, C_t$, where each C_i is the total count of any item that is present in the ith row. We also maintain the total number of items m seen so far. Algorithm 38 and Algorithm 39 present the initialization and update algorithms. The reporting algorithm then needs to solve the following problem: at any point of time, given the counts $C_1, \ldots, C_t$ and m output the at most d hot items.

Algorithm 38 Initialization

OUTPUT: Initialize the counters

```
m ← 0
FOR every j ∈ [t] DO
    C_j ← 0
```

Algorithm 39 Update

INPUT: Input pair (i, u), $i \in [N]$ and $u \in \mathbb{Z}$
OUTPUT: Update the Counters

```
m ← m + 1,
FOR every j ∈ [t] DO
    IF M_ij = 1 THEN
        C_j ← C_j + u
```

Next, we reduce the problem of reporting hot items to the decoding problem of group testing. The reduction essentially follows from the following observations.

Observation 19.5.2. *If j is a hot item and $M_{ij} = 1$, then $C_i > \frac{m}{d}$.*

Proof. Let $i \in [t]$ be such that $M_{ij} = 1$. Then note that at any point of time,

$$C_i = \sum_{k:M_{ik}=1} f_k \geq f_j.^{2}$$

Since j is a hot item, we have $f_j > \frac{m}{d}$, which completes the proof. □

Observation 19.5.3. *For any $1 \leq i \leq t$, if all j with $M_{ij} = 1$ are not hot items, then we have $C_i \leq \frac{m}{d}$.*

Proof. Fix an $\in [t]$ such that every $j \in [N]$ such that $M_{ij} = 1$ is not a hot item. Then by the same argument as in proof of Observation 19.5.2, we have

$$C_i = \sum_{k:M_{ik}=1} f_k.$$

The proof then follows by the choice of i and (19.5). □

Armed with these observations, we now present the reduction. Define $\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \{0,1\}^N$ with $x_j = 1$ if and only if j is a hot item, and $\mathbf{r} = (r_1, r_2, \ldots, r_t) \in \{0,1\}^t$ with $r_i = 1$ if and only if $C_i > \frac{m}{d}$, we will have $r_i = \vee_{j:M_{ij}=1} x_j$. The latter claim follows from Observations 19.5.2 and 19.5.3 above. This means we have:

$$M \odot \mathbf{x} = \mathbf{r}. \tag{19.6}$$

Note that by definition, $wt(\mathbf{x}) < d$. Thus reporting the hot items is the same as decoding to compute $\mathbf{x}$ given M and $\mathbf{r}$, which successfully changes the hot items problem into group testing problem. Algorithm 40 has the formal specification of this algorithm.

Algorithm 40 Report Heavy Items

INPUT: Counters m and $C_1, \ldots, C_t$
OUTPUT: Output the heavy items

```
FOR every j ∈ [t] DO
    IF C_t > m/d THEN
        r_j ← 1
    ELSE
        r_j ← 0
Let x be the result of decoding (for group testing) r
RETURN {i | x_i = 1}
```

Next, we will design and analyze the algorithm above and check if the conditions in Definition 19.5.1 are met.

[2] The equality follows e.g. by applying induction on Algorithm 39.

Analysis of the Algorithm

In this part, we will review the requirements on data stream algorithm one by one and check if the algorithm for the hot items problem based on group testing satisfies them. In particular, we will need to pick M and the decoding algorithm. We will pick M to be the d-disjunct matrix from Theorem 19.4.2.

1. *One-pass requirement*
   If we use non-adaptive group testing, the algorithm for the hot items problem above can be implemented in one pass, which means each input is visited only once. (If adaptive group testing is used, the algorithm is no longer one pass, therefore we choose non-adaptive group testing.) We note that by definition, our choice of M satisfies this condition.

2. *Poly-log space requirement*
   In the algorithm, we have to maintain the counters C_i and m. The maximum value for them is mN, thus we can represent each counter in $O(\log N + \log m)$ bits. This means we need $O((\log N + \log m)\, t)$bits to maintain the counters. Theorem 19.4.2 implies that $t = O(d^2 \log_d^2 N)$. Thus, the total space we need to maintain the counters is $O(d^2 \log_d^2 N(\log N + \log m))$.

   On the other hand, if we need to store the matrix M, we will need $\Omega(tN)$ space. Therefore, poly-log space requirement can be achieved only if matrix M is not stored directly. (We will tackle this issues in the next point.)

3. *Poly-log update time*
   As mentioned in the previous part, we cannot store the matrix M directly in order to have poly-log space. Since RS code is strongly explicit (see Exercise 6.9), we do not need to explicitly store M (we just need to store the parameters of the code C_{out} and C_{in}, which can be done in poly-log space). In the following, we will argue that the runtime of Algorithm 39 is $O(t \times \text{polylog}\, t)$. It is easy to check the claimed time bound is correct as long as we can perform the check in Step 3 in $\text{polylog}(t)$ time. In particular, we would be done if given $j \in [N]$, we can compute the column M^j in $O(t \times \text{polylog}\, t)$ time. Next, we argue that the latter claim is correct.

   Recall that $M = M_{C^*}$, with $C^* = C_{\text{out}} \circ C_{\text{in}}$, where C_{out} is a $[q, k, q-k+1]_q$ RS code and C_{in} chosen such that $M_{C_{\text{in}}}$ is the $q \times q$ identity matrix. Recall that codewords of C^* are columns of the matrix M, and we have $n = q^k$, $t = q^2$.

   Since every column of M corresponds to a codeword of C^*, we can think of j equivalently as a message $\mathbf{m} \in \mathbb{F}_q{}^k$. In particular, M^j then corresponds to the codeword $C_{\text{out}}(\mathbf{m})$. On the other hand, the column M^j can be partitioned into q chunks, each chunk is of length q. Notice that $(C_{\text{out}}(\mathbf{m}))_{i_1} = i_2$ if and only if the i_1th chunk has 1 on its i_2th position and 0 on other positions (recall the definition of C_{in}). Therefore, we can compute M^j by computing $C_{\text{out}}(\mathbf{m})$. Because C_{out} is a linear code, $C_{\text{out}}(\mathbf{m})$ can be computed in

$O(q^2 \times \text{polylog}\, q)$ time,[3] implies that M^j can be computed in $O(q^2 \times \text{polylog}\, q)$ time. Since we have $t = q^2$, the update process can be finished with $O(t \times \text{polylog}\, t)$ time. (We do not need C_{out} to be strongly explicit: as long as C_{out} is linear the arguments so far work just as well.)

4. *Reporting time*
   It is easy to check that the run time of Algorithm 40 is dominated by Step 6. So far, the only decoding algorithm for M that we have seen is Algorithm 37, which runs in time $\Omega(tN)$, which does not satisfy the required reporting time requirement. In Exercise 19.11, we show that using the fact that C_{out} is the Reed-Solomon code, one can solve the decoding problem in $\text{poly}(t)$.

Thus, we have argued that

Theorem 19.5.4. *There exists a data streaming algorithm that computes d hot items with one pass, $O(t \log N)$ space for $t = O(d^2 \log_d^2 N)$, $O(t\,\text{polylog}\, t)$ update time and $\text{poly}(t)$ reporting time.*

19.6 Summary of best known bounds

We conclude the chapter by collecting the best known bounds on both adaptive and non-adaptive group testing. First, we know the correct bound on the best possible number of adaptive tests:

Theorem 19.6.1.

$$t^a(d,N) = \Theta\left(d \log(N/d)\right).$$

The upper bound follows from Exercise 19.1 while the lower bound follows from Proposition 19.2.1.

There is a gap between the best known upper and lower bound on the number of non-adaptive tests:

Theorem 19.6.2.

$$\Omega\left(d^2 \log_d N\right) \le t(d,N) \le O\left(d^2 \min\left(\log(N/d), \log_d^2 N\right)\right).$$

The upper bounds follow from Theorem 19.4.2 and Exercise 19.5 while the lower bound follows from Exercise 19.8.

Finally, note that Theorem 19.6.1 and 19.6.2 imply that there is a gap between the minimum number of tests needed for adaptive and non-adaptive group testing:

Corollary 19.6.3.

$$\frac{t(d,N)}{t^a(d,N)} \ge \Omega\left(\frac{d}{\log d}\right).$$

[3] This follows from Proposition 2.3.2 and the fact that C_{out} is strongly explicit

19.7 Exercises

Exercise 19.1 (Upper bound on $t^a(d,N)$). In this problem we will show that $t^a(d,N) = O(d\log(N/d))$. We begin by trying to prove a weaker bound of $O(d\log N)$:

- Show that one can identify at least one i such that $x_i = 1$ (or report none exist) with $O(\log N)$ adaptive tests.

  (*Hint:* Use binary search.)

- Using the scheme above argue that one can compute $\mathbf{x}$ with $O(wt(\mathbf{x})\cdot\log N)$ adaptive tests. Conclude that $t^a(d,N) \le O(d\log N)$.

Next we argue that we can tighten the bound to the optimal bound of $O(d\log(N/d))$:

- Argue that any scheme that computes $\mathbf{x} \in \{0,1\}^N$ with $O(wt(\mathbf{x})\cdot\log N)$ adaptive tests can be used to compute $\mathbf{x}$ with $O(d\log(N/d))$ adaptive tests where $wt(\mathbf{x}) \le d$.

- Conclude that $t^a(d,N) \le O(d\log(N/d))$.

Exercise 19.2. Show that every d-separable matrix is also $(d-1)$-disjunct.

Exercise 19.3. Prove that Algorithm 37 is correct and runs in time $O(tN)$.

Exercise 19.4. For every integer $d \ge 1$ and large enough integer $N \ge d$ show that there exists a d-disjunct matrix with $O(d^2\log(N/d))$ rows.

(*Hint:* Use the probabilistic method. It might help to pick each of tN bits in the matrix independently at random with the same probability.)

Exercise 19.5. We first begin by generalizing the argument of Theorem 19.4.2:

- Let C_{out} be an $(n,k,D)_q$ code. Let C_{in} be defined such that $M_{C_{\text{in}}}$ is the $q\times q$ identity matrix. Let $M_{C_{\text{out}}\circ C_{\text{in}}}$ be a $t\times N$ matrix that is d-disjunct. Derive the parameters d, t and N.

Next argue that it is enough to pick an outer random code to obtain a d-disjunct matrix with the same parameters obtained in Exercise 19.4:

- Pick $q = \Theta(d)$. Then using the previous part or otherwise show that if C_{out} is a random $[n,k,D]_q$ code, then the resulting $t\times N$ matrix $M_{C_{\text{out}}\circ C_{\text{in}}}$ is d-disjunct with $t = O(d^2\log N)$ for large enough N.

  (*Hint:* Use Theorem 4.2.1 and Proposition 3.3.5.)

Exercise 19.6. For every integer $d \ge 1$ and large enough $N \ge d$, construct a d-disjunct matrix with $O(d^2\log N)$ rows in (deterministic) time poly(N).

Hint: Recall Exercise 4.7.

Exercise 19.7 (Lower Bound on $t(d,N)$ due to Bassalygo). In this problem we will show that $t(d,N) \ge \min\left\{\binom{d+2}{2}, N\right\}$. In what follows let M be a $t\times N$ matrix that is d-disjunct.

(a) Argue that if $wt(M^j) < d$ then M^j has a *private row* i.e. there exists a row $i \in [t]$ such that $M_{ij} = 1$ but $M_{ij'} = 0$ for every $j' \neq j$.

(b) Using part (a) or otherwise, argue that if all columns of M have Hamming weight at most $d-1$, then $t \geq N$.

(c) Let M^{-j} for $j \in [N]$ be the matrix M with M^j as well as all rows $i \in [t]$ such that $M_{ij} = 1$ removed. Then argue that M^{-j} is $(d-1)$-disjunct.

(d) Argue that $t(1, N) \geq \min\{3, N\}$.

(e) Using induction with parts (b)-(d) or otherwise, argue that $t \geq \min\left\{\binom{d+2}{2}, N\right\}$.

Exercise 19.8 (Lower Bound on $t(d, N)$ due to Ruszinkó and Alon-Asodi). In this problem, we will show that

$$t(d, N) \geq \Omega\left(\min\left\{d^2 \log_d N, N\right\}\right). \tag{19.7}$$

In what follows let M be a $t \times N$ matrix that is d-disjunct.

(a) Argue that any $j \in [N]$ such that $wt(M^j) < \frac{2t}{d}$ has a *private subset* of size $\lceil 4t/d^2 \rceil$, i.e. there exists a subset $S \subseteq [N]$ with $|S| = \lceil 4t/d^2 \rceil$ such that M^j has ones in all $i \in S$ but for every $j \neq j'$, $M^{j'}$ has at least one row $i' \in S$ such that $M_{i'j'} = 0$.

(b) Using part (a) or otherwise, argue:

$$N - \frac{d}{2} \leq \binom{t}{\lceil 4t/d^2 \rceil}.$$

(c) Using Exercise 19.7 and part (b) or otherwise argue (19.7).

Exercise 19.9. In this exercise and the ones that follow it, we will consider the following equivalent version of the decoding problem: given $\mathbf{r} = M \odot \mathbf{x}$ with $wt(\mathbf{x}) \leq d$, output $\{i | x_i = 1\}$. Now consider the following easier version of the problem. In addition to $\mathbf{r}$ and M assume that one is also given a set S such that $\{i | x_i = 1\} \subseteq S$. Modify Algorithm 37 to design a decoding algorithm that computes $\{i | x_i = 1\}$ in time $O(t \cdot |S|)$.

Exercise 19.10. A $t \times N$ matrix M is called (d, L)-list disjunct if and only if the following holds for every disjoint subsets $S, T \subset [N]$ such that $|S| = d$ and $|T| = L - d$, there is a row in M where all columns in S have a 0 but at least one column in T has a 1.

- What is a $(d, d+1)$-list disjunct matrix?
- Let C_{out} be an $(n, k)_q$ code that is $(0, d, L)$-list recoverable code (recall Definition 15.3.2). Let C_{in} be the inner code such that $M_{C_{\text{in}}}$ is the $q \times q$ identity matrix. Argue that $M_{C_{\text{out}} \circ C_{\text{in}}}$ is (d, L) list disjunct.

Exercise 19.11. Using Exercises 19.9 and 19.10 or otherwise prove the following. Let M_{C^*} be the Kautz-Singleton matrix from Section 19.4.1. Then given $M_{C^*} \odot \mathbf{x}$ with $wt(\mathbf{x}) \leq d$, one can compute $\{i | x_i = 1\}$ in poly(t) time.
(*Hint:* Theorem 15.3.2 could be useful.)

19.8 Bibliographic Notes

Robert Dorfman's paper in 1943 [22] introduced the field of (Combinatorial) Group Testing. It must be noted that though this book covers group testing as an application of coding theory, it took off long before coding theory itself.

The original motivation arose during the Second World War when the United States Public Health service and the Selective Service embarked upon a large scale project. The objective was to weed out all syphilitic men called up for induction. [22].

The connection between group testing and hot items problem considered in Section 19.5 was established by Cormode and Muthukrishnan [20]. More details on data stream algorithms can be bound in the survey by Muthukrishnan [76].

Bibliography

[1] Manindra Agrawal, Neeraj Kayal, and Nitin Saxena. PRIMES Is in P. *Annals of Mathematics*, 160(2):781–793, 2004.

[2] M. Alekhnovich. Linear diophantine equations over polynomials and soft decoding of reed-solomon codes. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 439–448, Nov 2002.

[3] Noga Alon and Joel Spencer. *The Probabilistic Method*. John Wiley, 1992.

[4] Erdal Arıkan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory*, pages 3051–3073, July 2009.

[5] Jarosław Bł asiok, Venkatesan Guruswami, Preetum Nakkiran, Atri Rudra, and Madhu Sudan. General strong polarization. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing*, pages 485–492, 2018.

[6] P.G.H. Bachmann. *Die analytische Zahlentheorie*. Number v. 2 in Zahlentheorie. Versuch einer Gesammtdarstellung dieser Wissenschaft in ihren Hauptheilen. 2. th. Teubner, 1894.

[7] Donald Beaver and Joan Feigenbaum. Hiding instances in multioracle queries. In Choffrut and Lengauer [17], pages 37–48.

[8] Eli Ben-Sasson, Swastik Kopparty, and Jaikumar Radhakrishnan. Subspace polynomials and limits to list decoding of reed-solomon codes. *IEEE Trans. Information Theory*, 56(1):113–120, 2010.

[9] E. R. Berlekamp. Factoring polynomials over large finite fields. *Mathematics of Computation*, 24:713–735, 1970.

[10] Kristian Brander. *Interpolation and list decoding of algebraic codes*. PhD thesis, Technical University of Denmark, 2010.

[11] P.S. Bullen. *Handbook of Means and Their Inequalities*. Mathematics and Its Applications. Springer Netherlands, 2010.

[12] Larry Carter and Mark N. Wegman. Universal classes of hash functions. *J. Comput. Syst. Sci.*, 18(2):143–154, 1979.

[13] Donald G. Chandler, Eric P. Batterman, and Govind Shah. Hexagonal, information encoding article, process and system. *US Patent Number 4,874,936*, October 1989.

[14] C. L. Chen and M. Y. Hsiao. Error-correcting codes for semiconductor memory applications: A state-of-the-art review. *IBM Journal of Research and Development*, 28(2):124–134, 1984.

[15] Peter M. Chen, Edward K. Lee, Garth A. Gibson, Randy H. Katz, and David A. Patterson. RAID: High-performance, reliable secondary storage. *ACM Computing Surveys*, 26(2):145–185, 1994.

[16] Q. Cheng and D. Wan. On the list and bounded distance decodability of reedâĂŞsolomon codes. *SIAM Journal on Computing*, 37(1):195–209, 2007.

[17] Christian Choffrut and Thomas Lengauer, editors. *STACS 90, 7th Annual Symposium on Theoretical Aspects of Computer Science, Rouen, France, February 22-24, 1990, Proceedings*, volume 415 of *Lecture Notes in Computer Science*. Springer, 1990.

[18] Alan Cobham. The Intrinsic Computational Difficulty of Functions. In Y. Bar-Hillel, editor, *Logic, Methodology and Philosophy of Science, proceedings of the second International Congress, held in Jerusalem, 1964*, Amsterdam, 1965. North-Holland.

[19] Stephen A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 151–158, 1971.

[20] Graham Cormode and S. Muthukrishnan. What's hot and what's not: tracking most frequent items dynamically. *ACM Trans. Database Syst.*, 30(1):249–278, 2005.

[21] Eren Şaşoğlu. Polarization and polar codes. *Foundations and Trends in Communications and Information Theory*, 8(4):259–381, 2012.

[22] Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, December 1943.

[23] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[24] Ilya Dumer, Daniele Micciancio, and Madhu Sudan. Hardness of approximating the minimum distance of a linear code. *IEEE Transactions on Information Theory*, 49(1):22–37, 2003.

[25] Zeev Dvir and Shachar Lovett. Subspace evasive sets. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:139, 2011.

[26] Jack Edmonds. Paths, trees, and flowers. In Ira Gessel and Gian-Carlo Rota, editors, *Classic Papers in Combinatorics*, Modern BirkhÃd'user Classics, pages 361–379. BirkhÃd'user Boston, 1987.

[27] Peter Elias. Error-free coding. *IEEE Transactions on Information Theory*, 4(4):29–37, 1954.

[28] Peter Elias. List decoding for noisy channels. *Technical Report 335, Research Laboratory of Electronics, MIT*, 1957.

[29] P. Erdös. On extremal problems of graphs and generalized graphs. *Israel Journal of Mathematics*, 2(3):183–190, 1964.

[30] G. David Forney. *Concatenated Codes*. MIT Press, Cambridge, MA, 1966.

[31] G. David Forney. Generalized Minimum Distance decoding. *IEEE Transactions on Information Theory*, 12:125–131, 1966.

[32] Peter Gemmell, Richard J. Lipton, Ronitt Rubinfeld, Madhu Sudan, and Avi Wigderson. Self-testing/correcting for polynomials and for approximate functions. In Cris Koutsougeras and Jeffrey Scott Vitter, editors, *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing, May 5-8, 1991, New Orleans, Louisiana, USA*, pages 32–42. ACM, 1991.

[33] Peter Gemmell and Madhu Sudan. Highly resilient correctors for multivariate polynomials. *Information Processing Letters*, 43(4):169–174, 1992.

[34] Peter Gemmell and Madhu Sudan. Highly resilient correctors for polynomials. *Inf. Process. Lett.*, 43(4):169–174, 1992.

[35] E. N. Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31:504–522, 1952.

[36] M. J. E. Golay. Notes on digital coding. *Proceedings of the IRE*, 37:657, 1949.

[37] Venkatesan Guruswami. Limits to list decodability of linear codes. In *Proceedings of the 34th ACM Symposium on Theory of Computing (STOC)*, pages 802–811, 2002.

[38] Venkatesan Guruswami. *List decoding of error-correcting codes*. Number 3282 in Lecture Notes in Computer Science. Springer, 2004. (Winning Thesis of the 2002 ACM Doctoral Dissertation Competition).

[39] Venkatesan Guruswami. Linear-algebraic list decoding of folded reed-solomon codes. In *Proceedings of the 26th Annual IEEE Conference on Computational Complexity (CCC)*, pages 77–85, 2011.

[40] Venkatesan Guruswami, Johan Håstad, and Swastik Kopparty. On the list-decodability of random linear codes. *IEEE Transactions on Information Theory*, 57(2):718–725, 2011.

[41] Venkatesan Guruswami, Johan Håstad, Madhu Sudan, and David Zuckerman. Combinatorial bounds for list decoding. *IEEE Transactions on Information Theory*, 48(5):1021–1035, 2002.

[42] Venkatesan Guruswami and Piotr Indyk. Linear-time encodable/decodable codes with near-optimal rate. *IEEE Transactions on Information Theory*, 51(10):3393–3400, 2005.

[43] Venkatesan Guruswami and Atri Rudra. Limits to list decoding reed-solomon codes. *IEEE Transactions on Information Theory*, 52(8):3642–3649, August 2006.

[44] Venkatesan Guruswami and Atri Rudra. Explicit codes achieving list decoding capacity: Error-correction with optimal redundancy. *IEEE Transactions on Information Theory*, 54(1):135–150, 2008.

[45] Venkatesan Guruswami and Atri Rudra. Better binary list decodable codes via multilevel concatenation. *IEEE Transactions on Information Theory*, 55(1):19–26, 2009.

[46] Venkatesan Guruswami and Atri Rudra. The existence of concatenated codes list-decodable up to the hamming bound. *IEEE Transactions on Information Theory*, 56(10):5195–5206, 2010.

[47] Venkatesan Guruswami and Igor Shparlinski. Unconditional proof of tightness of johnson bound. In *Proceedgins of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 754–755, 2003.

[48] Venkatesan Guruswami and Madhu Sudan. Improved decoding of Reed-Solomon and algebraic-geometry codes. *IEEE Transactions on Information Theory*, 45(6):1757–1767, 1999.

[49] Venkatesan Guruswami and Ameya Velingker. An entropy sumset inequality and polynomially fast convergence to Shannon capacity over all alphabets. In *Proceedings of 30th Conference on Computational Complexity*, pages 42–57, 2015.

[50] Venkatesan Guruswami and Patrick Xia. Polar codes: Speed of polarization and polynomial gap to capacity. *IEEE Trans. Information Theory*, 61(1):3–16, 2015. Preliminary version in Proc. of FOCS 2013.

[51] Venkatesan Guruswami and Chaoping Xing. Folded codes from function field towers and improved optimal rate list decoding. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:36, 2012.

[52] Richard W. Hamming. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 29:147–160, April 1950.

[53] G.H. Hardy and J.E. Littlewood. Some problems of diophantine approximation. *Acta Mathematica*, 37(1):193–239, 1914.

[54] Seyed Hamed Hassani, Kasra Alishahi, and Rüdiger L. Urbanke. Finite-length scaling for polar codes. *IEEE Trans. Information Theory*, 60(10):5875–5898, 2014.

[55] Johan Håstad, Steven Phillips, and Shmuel Safra. A well-characterized approximation problem. *Inf. Process. Lett.*, 47(6):301–305, 1993.

[56] Tom Høholdt, J. H. van Lint, and Ruud Pellikaan. Algebraic geometry codes. In W. C. Huffamn V. S. Pless and R. A.Brualdi, editors, *Handbook of Coding Theory*. North Holland, 1998.

[57] Ari Juels and Madhu Sudan. A fuzzy vault scheme. *Des. Codes Cryptography*, 38(2):237–257, 2006.

[58] J. Justesen. Class of constructive asymptotically good algebraic codes. *IEEE Trans. Inform. Theory*, pages 652–656, Sep 1972.

[59] Erich Kaltofen. Polynomial-time reductions from multivariate to bi- and univariate integral polynomial factorization. *SIAM J. Comput.*, 14(2):469–489, 1985.

[60] Richard M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103, 1972.

[61] John Y. Kim and Swastik Kopparty. Decoding reed-muller codes over product sets. *Theory of Computing*, 13(1):1–38, 2017.

[62] Donald E. Knuth. Big omicron and big omega and big theta. *SIGACT News*, 8(2):18–24, April 1976.

[63] Andrei N. Kolmogorov. Three Approaches to the Quantitative Definition of Information. *Problems of Information Transmission*, 1(1):1–7, 1965.

[64] Satish Babu Korada, Eren Sasoglu, and Rüdiger L. Urbanke. Polar codes: Characterization of exponent, bounds, and constructions. *IEEE Transactions on Information Theory*, 56(12):6253–6264, 2010.

[65] E. Landau. *Handbuch der lehre von der verteilung der primzahlen*. Number v. 1 in Handbuch der lehre von der verteilung der primzahlen. B. G. Teubner, 1909.

[66] Amos Lapidoth and P. Narayan. Reliable communication under channel uncertainty. *IEEE Transactions on Information Theory*, 44(6):2148–2177, 1998.

[67] Leonid A Levin. Universal sequential search problems. *Problemy Peredachi Informatsii*, 9(3):115–116, 1973.

[68] Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Graduate Texts in Computer Science. Springer, New York, NY, USA, third edition, 2008.

[69] Rudolf Lidl and Harald Niederreiter. *Introduction to Finite Fields and their applications*. Cambridge University Press, Cambridge, MA, 1986.

[70] Richard J. Lipton. Efficient checking of computations. In Choffrut and Lengauer [17], pages 207–215.

[71] Michael Luby, Michael Mitzenmacher, Mohammad Amin Shokrollahi, and Daniel A. Spielman. Efficient erasure correcting codes. *IEEE Transactions on Information Theory*, 47(2):569–584, 2001.

[72] Robert J. McEliece. On the average list size for the Guruswami-Sudan decoder. In *7th International Symposium on Communications Theory and Applications (ISCTA)*, July 2003.

[73] Robert J. McEliece, Eugene R. Rodemich, Howard Rumsey Jr., and Lloyd R. Welch. New upper bounds on the rate of a code via the Delsarte-Macwilliams inequalities. *IEEE Transactions on Information Theory*, 23:157–166, 1977.

[74] Ryuhei Mori and Toshiyuki Tanaka. Source and channel polarization over finite fields and Reed-Solomon matrices. *IEEE Trans. Information Theory*, 60(5):2720–2736, 2014.

[75] David E. Muller. Application of boolean algebra to switching circuit design and to error detection. *Trans. I.R.E. Prof. Group on Electronic Computers*, 3(3):6–12, 1954.

[76] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.

[77] Farzad Parvaresh and Alexander Vardy. Correcting errors beyond the guruswami-sudan radius in polynomial time. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 285–294, 2005.

[78] Ruud Pellikaan and Xin-Wen Wu. List decoding of q-ary reed-muller codes. *IEEE Trans. Information Theory*, 50(4):679–682, 2004.

[79] Larry L. Peterson and Bruce S. Davis. *Computer Networks: A Systems Approach*. Morgan Kaufmann Publishers, San Francisco, 1996.

[80] W. Wesley Peterson. Encoding and error-correction procedures for Bose-Chaudhuri codes. *IEEE Transactions on Information Theory*, 6:459–470, 1960.

[81] Michael O. Rabin. Probailistic algorithms. In J. F. Traub, editor, *Algorithms and Complexity, Recent Results and New Directions*, pages 21–39, 1976.

[82] I. Reed, R. Scholtz, , and L. Welch. The fast decoding of reed-solomon codes using fermat theoretic transforms and continued fractions. *IEEE Transactions on Information Theory*, 24(1):100–106, January 1978.

[83] Irving S. Reed. A class of multiple-error-correcting codes and the decoding scheme. *Trans. of the IRE Professional Group on Information Theory (TIT)*, 4:38–49, 1954.

[84] Irving S. Reed and Gustav Solomon. Polynomial codes over certain finite fields. *SIAM Journal on Applied Mathematics*, 8(2):300–304, 1960.

[85] Herbert Robbins. A remark on Stirling's formula. *Amer. Math. Monthly*, 62:26–29, 1955.

[86] Atri Rudra and Steve Uurtamo. Two theorems on list decoding. In *Proceedings of the 14th Intl. Workshop on Randomization and Computation (RANDOM)*, pages 696–709, 2010.

[87] Atri Rudra and Mary Wootters. Every list-decodable code for high noise has abundant near-optimal rate puncturings. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 764–773, 2014.

[88] E. Sasoglu, E. Telatar, and E. Arikan. Polarization for arbitrary discrete memoryless channels. In *2009 IEEE Information Theory Workshop*, pages 144–148, Oct 2009.

[89] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[90] Victor Shoup. New algorithms for finding irreducible polynomials over finite fields. *Math. Comp.*, 54:435–447, 1990.

[91] Victor Shoup. *A computational introduction to number theory and algebra*. Cambridge University Press, 2006.

[92] R. Singleton. Maximum distance q -nary codes. *Information Theory, IEEE Transactions on*, 10(2):116 – 118, apr 1964.

[93] Michael Sipser. The history and status of the p versus np question. In *Proceedings of the Twenty-fourth Annual ACM Symposium on Theory of Computing*, STOC '92, pages 603–618, New York, NY, USA, 1992. ACM.

[94] Daniel A. Spielman. Linear-time encodable and decodable error-correcting codes. *IEEE Transactions on Information Theory*, 42(6):1723–1731, 1996.

[95] Douglas R. Stinson. Universal hashing and authentication codes. *Des. Codes Cryptography*, 4(4):369–380, 1994.

[96] Madhu Sudan. Decoding of Reed Solomon codes beyond the error-correction bound. *J. Complexity*, 13(1):180–193, 1997.

[97] Madhu Sudan. List decoding: Algorithms and applications. *SIGACT News*, 31:16–27, 2000.

[98] Ido Tal and Alexander Vardy. How to construct polar codes. *IEEE Trans. Information Theory*, 59(10):6562–6582, 2013.

[99] Robert Endre Tarjan. Algorithmic design. *Commun. ACM*, 30(3):204–212, 1987.

[100] Aimo Tietavainen. On the nonexistence theorems for perfect error-correcting codes. *SIAM Journal of Applied Mathematics*, 24(1):88–96, 1973.

[101] Jacobus H. van Lint. Nonexistence theorems for perfect error-correcting codes. In *Proceedings of the Symposium on Computers in Algebra and Number Theory*, pages 89–95, 1970.

[102] R. R. Varshamov. Estimate of the number of signals in error correcting codes. *Doklady Akadamii Nauk*, 117:739–741, 1957.

[103] Joachim von zur Gathen and Jürgen Gerhard. *Modern Computer Algebra*. Cambridge University Press, 3 edition, 2013.

[104] Lloyd R. Welch and Elwyn R. Berlekamp. Error correction of algebraic block codes. *US Patent Number 4,633,470*, December 1986.

[105] John M. Wozencraft. List Decoding. *Quarterly Progress Report, Research Laboratory of Electronics, MIT*, 48:90–95, 1958.

Appendix A

Notation Table

$\mathbb{R}$	The set of real numbers	
$\mathbb{Z}$	The set of integers	
$\neg E$	Negation of the event E	
$\log x$	Logarithm to the base 2	
Σ^m	Vectors of length m with symbols from Σ	
$\mathbf{v}$	A vector	
$\mathbf{0}$	The all zero vector	
$\mathbf{e}_i$	The ith standard vector, i.e. 1 in position i and 0 everywhere else	
$\mathbf{v}_S$	Vector $\mathbf{v}$ projected down to indices in S	
$\langle \mathbf{u}, \mathbf{v} \rangle$	Inner-product of vectors $\mathbf{u}$ and $\mathbf{v}$	
$[a, b]$	$\{x \in \mathbb{R} \mid a \leq x \leq b\}$	
$[x]$	The set $\{1, \ldots, x\}$	Secti
n	Block length of a code	Defin
Σ	Alphabet of a code	Defin
q	$q = \lvert\Sigma\rvert$	Defin
k	Dimension of a code	Defin
R	Rate of a code	Defin
$\Delta(\mathbf{u}, \mathbf{v})$	Hamming distance between $\mathbf{u}$ and $\mathbf{v}$	Defin
d	Minimum distance of a code	Defin
$wt(\mathbf{v})$	Hamming weight of $\mathbf{v}$	Defin
$B(\mathbf{x}, r)$	Hamming ball of radius r centered on $\mathbf{x}$	Defin
$Vol_q(r, n)$	Volume Hamming ball of radius r	Defin
$(n, k, d)_\Sigma$	A code with block length n, dimension k, distance d and alphabet Σ	Defin
$(n, k, d)_q$	A code with block length n, dimension k, distance d and alphabet size q	Defin
$[n, k, d]_q$	A linear $(n, k, d)_q$ code	Defin
$\mathbb{F}_q$	The finite field with q elements (q is a prime power)	Secti
$\mathbb{F}^*$	The set of non-zero elements in the field $\mathbb{F}$	
$\mathbb{F}_q^{m \times N}$	The set of all $m \times N$ matrices where each entry is from $\mathbb{F}_q$	
$\mathbb{F}_q[X_1, \ldots, X_m]$	The set of all m-variate polynomials with coefficients from $\mathbb{F}_q$	

$R(C)$	Rate of a code family C	Defin
$\delta(C)$	Relative distance of a code family C	Defin
$\mathscr{U}$	The uniform distribution	Defin
$\mathbb{E}[V]$	Expectation of a random variable V	Defin
$\mathbb{1}_E$	Indicator variable for event E	Secti
$H_q(x)$	$x\log_q(q-1) - x\log_q x - (1-x)\log_q(1-x)$	Defin
$H_q^{-1}(y)$	Unique $x \in [0, 1-1/q]$ such that $H_q(x) = y$	Secti
$\deg(P)$	Degree of polynomial $P(X)$	Defin
$\mathbb{F}_q[X]$	The set of all univariate polynomials in X over $\mathbb{F}_q$	Secti
$J_q(x)$	$(1-1/q)(1-\sqrt{1-qx/(q-1)})$	Theo
$\binom{S}{t}$	$\{T \subseteq S \mid \lvert T\rvert = t\}$	
$M \odot \mathbf{x}$	Binary matrix-vector multiplication where multiplication is AND and addition is OR	

Appendix B

Some Useful Facts

B.1 Some Useful Inequalities

Recall that the binomial coefficient for integers $a \le b$, defined as

$$\binom{b}{a} = \frac{b!}{a!(b-a)!}.$$

We begin with a simple lower bound on the binomial coefficient:

Lemma B.1.1. *For every integers* $1 \le a \le b$*, we have*

$$\binom{b}{a} \ge \left(\frac{b}{a}\right)^a.$$

Proof. The following sequence of relations completes the proof:

$$\binom{a}{b} = \prod_{i=0}^{a-1} \frac{b-i}{a-i} \ge \prod_{i=0}^{a-1} \frac{b}{a} = \left(\frac{b}{a}\right)^a.$$

In the above, the first equality follows from definition and the inequality is true since $b \ge a$ and $i \ge 0$. □

We state the next set of inequalities without proof (see [85] for a proof):

Lemma B.1.2 (Stirling's Approximation)**.** *For every integer* $n \ge 1$*, we have*

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\lambda_1(n)} < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\lambda_2(n)},$$

where

$$\lambda_1(n) = \frac{1}{12n+1} \text{ and } \lambda_2(n) = \frac{1}{12n}.$$

We prove another inequality involving Binomial coefficient.

Lemma B.1.3. *For every integers* $1 \le a \le b$*, we have*

$$\binom{b}{a} \le \left(\frac{eb}{a}\right)^a.$$

Proof. First note that

$$\binom{a}{b} = \frac{a(a-1)\cdots(a-b+1)}{b!} \le \frac{a^b}{b!}.$$

The final bound follows from the fact that

$$b! > \left(\frac{b}{e}\right)^b,$$

which in turns follows from the lower bound in Lemma B.1.2. □

We next state Bernoulli's inequality:

Lemma B.1.4 (Bernoulli's Inequality)**.** *For every real numbers* $k \ge 1$ *and* $x \ge -1$*, we have*

$$(1+x)^k \ge 1 + kx.$$

Proof Sketch. We only present the proof for integer k. For the full proof see e.g. [11].

For the base case of $k = 1$, the inequality holds trivially. Assume that the inequality holds for some integer $k \ge 1$ and to complete the proof, we will prove it for $k+1$. Now consider the following inequalities:

$$\begin{aligned}
(1+x)^{k+1} &= (1+x)\cdot(1+x)^k \\
&\ge (1+x)\cdot(1+kx) \\
&= 1 + (k+1)x + kx^2 \\
&\ge 1 + (k+1)x,
\end{aligned}$$

as desired. In the above, the first inequality follows from the inductive hypothesis and the second inequality follows from the fact that $k \ge 1$. □

Lemma B.1.5. *For* $|X|le1$*,*

$$\sqrt{1+x} \le 1 + \frac{x}{2} - \frac{x^2}{16}.$$

Proof. Squaring the RHS we get

$$\left(1 + \frac{x}{2} - \frac{x^2}{16}\right)^2 = 1 + \frac{x^2}{4} + \frac{x^4}{256} + x - \frac{x^2}{16} - \frac{x^3}{32} = 1 + x + \frac{3x^2}{16} - \frac{x^3}{32} + \frac{x^4}{256} \ge 1 + x,$$

as desried. □

B.2 Some Useful Identities and Bounds

We start off with an equivalence between two inequalities.

Lemma B.2.1. *Let $a, b, c, d > 0$. Then $\frac{a}{b} \le \frac{c}{d}$ if and only if $\frac{a}{a+b} \le \frac{c}{c+d}$.*

Proof. Note that $\frac{a}{b} \le \frac{c}{d}$ if and only if

$$\frac{b}{a} \ge \frac{d}{c}.$$

The above is true if and only if

$$\frac{b}{a} + 1 \ge \frac{d}{c} + 1,$$

which is same as $\frac{a}{a+b} \le \frac{c}{c+d}$. □

Next, we state some infinite sums that are identical to certain logarithms (the proofs are standard and are omitted).

Lemma B.2.2. *For $|x| < 1$,*

$$\ln(1+x) = x - \frac{x^2}{2!} + \frac{x^3}{3!} - \cdots.$$

We can use the above to prove some bounds on $\ln(1+x)$ (we omit the proof):

Lemma B.2.3. *For $0 \le x < 1$, we have*

$$x - x^2/2 \le \ln(1+x) \le x,$$

and for $0 \le x \le 1/2$, we have

$$-x - x^2 \le \ln(1-x) \le -x.$$

We can use the above bounds to further prove boounds on the (binary) entropy function:

Lemma B.2.4. *For $x \le 1/4$, we have*

$$1 - 5x^2 \le H(1/2 - x) \le 1 - x^2.$$

Proof. By definition $H(1/2 - x) = 1 - 1/2\log(1-4x^2) + x\log(1-2x)/(1+2x)$, and using the approximations for $\ln(1+x)$ from Lemma B.2.3, we have, for $x < 1/4$,

$$\begin{aligned} H(1/2 - x) &\le 1 + \frac{1}{2\ln 2} \cdot (4x^2 + 16x^4) + \frac{1}{\ln 2} \cdot (-2x^2) - \frac{1}{\ln 2} \cdot (2x^2 - 2x^3) \\ &= 1 - \frac{2}{\ln 2} \cdot x^2 + \frac{2}{\ln 2} \cdot x^3 + \frac{8}{\ln 2} \cdot x^4 \\ &\le 1 - \frac{x^2}{\ln 2} \qquad \text{(B.1)} \\ &\le 1 - x^2. \end{aligned}$$

In the above, (B.1) follows by using our assumption that $x \le 1/4$.

Using the other sides of the approximations we also have:

$$\begin{aligned} H(1/2 - x) &\geq 1 + \frac{1}{2\ln 2} \cdot (4x^2) + \frac{1}{\ln 2} \cdot (-2x^2 - 4x^3) - \frac{1}{\ln 2} \cdot (2x^2) \\ &\geq 1 - \frac{3x^2}{\ln 2} \\ &\geq 1 - 5x^2, \end{aligned}$$

where the second inequality uses our assumption that $x \leq 1/4$. □

The following fact follows from the well-known fact that $\lim_{x\to\infty}(1 + 1/x)^x = e$:

Lemma B.2.5. *For every real $x > 0$,*

$$\left(1 + \frac{1}{x}\right)^x \leq e.$$

Appendix C

Background on Asymptotic notation, Algorithms and Complexity

In this chapter, we collect relevant background on algorithms and their analysis (as well as their limitations). We begin with notation that we will use to bound various quantities when we do not pay close attention to the constants.

C.1 Asymptotic Notation

Throughout the book, we will encounter situations where we would be interested in how a function $f(N)$ grows as the input parameter N grows. (We will assume that the real valued function f is monotone.) The most common such situation is when we would like to bound the runtime of an algorithm we are analyzing– we will consider this situation in some detail shortly. In particular, we will interested in bounds on $f(N)$ that are "oblivious" to constants. E.g. given that an algorithm takes $24N^2 + 100N$ steps to terminate, we would be interested in the fact that the dominating term is the N^2 (for large enough N). Technically, speaking we are interested in the *asymptotic* growth of $f(N)$. Throughout this chapter, we will assume that all functions are monotone.

The first definition is when we are interested in an upper bound on the function $f(N)$. When talking about numbers, we say b is an upper bound on a if $a \leq b$. We will consider a similar definition for functions that in some sense ignores constants.

Definition C.1.1. We say $f(N)$ is $O(g(N))$ (to be read as $f(N)$ is "Big-Oh" of $g(N)$) if there exists constants $c, N_0 \geq 0$ that are independent of N such that for every large enough $N \geq N_0$:

$$f(N) \leq c \cdot g(N).$$

Alternatively $f(N)$ is $O(g(N))$ if and only if

$$\lim_{N \to \infty} \frac{f(N)}{g(N)} \leq C,$$

for some absolute constant C. (See Exercise C.1.) So for example both $24N^2+100N$ and $N^2/2-N$ are $O(N^2)$ as well as $O(N^3)$. However, neither of them are $O(N)$ or $O(N^{3/2})$.

The second definition is when we are interested in a lower bound on the function $f(N)$. When talking about numbers, we say b is a lower bound on a if $a \geq b$. We will consider a similar definition for functions that in some sense ignores constants.

Definition C.1.2. We say $f(N)$ is $\Omega(g(N))$ (to be read as $f(N)$ is "Big-Omega" of $g(N)$) if there exists constants $\varepsilon, N_0 \geq 0$ that are independent of n such that for every large enough $N \geq N_0$:

$$f(N) \geq \varepsilon \cdot g(N).$$

Alternatively $f(N)$ is $\Omega(g(N))$ if and only if

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} \geq C,$$

for some absolute constant C. (See Exercise C.2.) So for example both $24N^2+100N$ and $N^2/2-N$ are $\Omega(N^2)$ as well as $\Omega(N^{3/2})$. However, neither of them are $\Omega(N^3)$ or $\Omega(N^{5/2})$.

The third definition is when we are interested in a tight bound on the function $f(N)$. When talking about numbers, we say b is same as a if $a = b$. We will consider a similar definition for functions that in some sense ignores constants.

Definition C.1.3. We say $f(N)$ is $\Theta(g(N))$ (to be read as $f(N)$ is "Theta" of $g(N)$) if and only if $f(N)$ is $O(g(N))$ and is also $\Omega(g(N))$.

Alternatively $f(N)$ is $\Theta(g(N))$ if and only if

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} = C,$$

for some absolute constant C. (See Exercise C.3.) So for example both $24N^2+100N$ and $N^2/2-N$ are $\Theta(N^2)$. However, neither of them are $\Theta(N^3)$ or $\Theta(N)$.

The fourth definition is when we are interested in a strict upper bound on the function $f(N)$. When talking about numbers, we say b is a strict upper bound on a if $a < b$. We will consider a similar definition for functions that in some sense ignores constants.

Definition C.1.4. We say $f(N)$ is $o(g(N))$ (to be read as $f(N)$ is "little-oh" of $g(N)$) if $f(N)$ is $O(g(N))$ but $f(N)$ is not $\Omega(g(N))$.

Alternatively $f(N)$ is $o(g(N))$ if and only if

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} = 0.$$

(See Exercise C.4.) So for example both $24N^2+100N$ and $N^2/2-N$ are $o(N^3)$ as well as $o(N^{5/2})$. However, neither of them are $o(N^2)$ or $o(N^{3/2})$.

The final definition is when we are interested in a strict lower bound on the function $f(N)$. When talking about numbers, we say b is a strict lower bound on a if $a > b$. We will consider a similar definition for functions that in some sense ignores constants.

Definition C.1.5. We say $f(N)$ is $\omega(g(N))$ (to be read as $f(N)$ is "little-omega" of $g(N)$) if $f(N)$ is $\Omega(g(N))$ but $f(N)$ is not $O(g(N))$.

Alternatively $f(N)$ is $\omega(g(N))$ if and only if

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} = \infty.$$

(See Exercise C.5.) So for example both $24N^2 + 100N$ and $N^2/2 - N$ are $\omega(N)$ as well as $\omega(N^{3/2})$. However, neither of them are $\omega(N^2)$ or $\omega(N^{5/2})$.

C.1.1 Some Properties

We now collect some properties of asymptotic notation that we will be useful in this book.

First all the notations are transitive:

Lemma C.1.1. *Let $\alpha \in \{O, \Omega, \Theta, o, \omega\}$. Then if $f(N)$ is $\alpha(g(N))$ and $g(N)$ is $\alpha(h(N))$, then $f(N)$ is $\alpha(h(N))$.*

Second, all the notations are additive:

Lemma C.1.2. *Let $\alpha \in \{O, \Omega, \Theta, o, \omega\}$. Then if $f(N)$ is $\alpha(h(N))$ and $g(N)$ is $\alpha(h(N))$, then $f(N) + g(N)$ is $\alpha(h(N))$.*

Finally, all the notations are multiplicative:

Lemma C.1.3. *Let $\alpha \in \{O, \Omega, \Theta, o, \omega\}$. Then if $f(N)$ is $\alpha(h_1(N))$ and $g(N)$ is $\alpha(h_2(N))$, then $f(N) \cdot g(N)$ is $\alpha(h_1(N) \cdot h_2(N))$.*

The proofs of the above properties are left as an exercise (see Exercise C.6).

C.2 Bounding Algorithm run time

Let $\mathscr{A}$ be the algorithm we are trying to analyze. Then we will define $T(N)$ to be the worst-case run-time of $\mathscr{A}$ over all inputs of size N. Slightly more formally, let $t_{\mathscr{A}}(\mathbf{x})$ be the number of steps taken by the algorithm $\mathscr{A}$ on input $\mathbf{x}$. Then

$$T(N) = \max_{\mathbf{x}:\mathbf{x} \text{ is of size } N} t_{\mathscr{A}}(\mathbf{x}). \tag{C.1}$$

In this section, we present two useful strategies to prove statements like $T(N)$ is $O(g(N))$ or $T(N)$ is $\Omega(h(N))$. Then we will analyze the run time of a very simple algorithm. However, before that we digress to clarify the following: (i) For most of the book, we will be interested in deterministic algorithms (i.e. algorithm whose execution is fixed given the input). However, we will consider randomized algorithms (see Section C.3 for more on this). (ii) One needs to clarify what constitutes a "step" in the definition of $T(N)$ above. We do so next.

C.2.1 RAM model

In this book, unless specified otherwise we will assume that the algorithms run on the RAM model. Informally, this computation model is defined as follows. For an input with n items, the memory consists of registers with $O(\log n)$ bits. For simplicity, we can assume that the input and output have separate dedicated registers. Note that the input will have n dedicated registers.

Any (atomic) step an algorithm can take are essentially any basic operations on constant such registers which can be implemented in $O(\log n)$ bit operations. In particular, the following operations are considered to take one step: loading $O(\log n)$ from a register or storing $O(\log n)$ bits in a register, initializing the contents of a register, bit-wise operations among registers, e.g. taking bit-wise XOR of the bits of two registers, adding numbers stored in two registers, incrementing the value stored in a register, comparing the values stored in two registers. Some examples of operations that are *not* single step operations: multiplying numbers or exponentiation (where the operands fit into one register each).

C.2.2 Proving $T(N)$ is $O(f(N))$

We start off with an analogy. Say you wanted prove that given m numbers $a_1, \ldots, a_m$, $\max_i a_i \leq U$. Then how would you go about doing so? One way is to argue that the maximum value is attained at i^* and then show that $a_{i^*} \leq U$. Now this is a perfectly valid way to prove the inequality we are after but note that you will *also* have to prove that the maximum value is attained at i^*. Generally, this is a non-trivial task. However, consider the following strategy:

> Show that for *every* $1 \leq i \leq m$, $a_i \leq U$. Then conclude that $\max_i a_i \leq U$.

Let us consider an example to illustrate the two strategies above. Let us say for whatever reason we are interested in showing that the age of the oldest person in your coding theory lectures is at most 100. Assume there are 98 students registered and the instructor is always present in class. This implies that there are at most $m = 99$ folks in the class. Let us order them somehow and let a_i denote the age of the i'th person. Then we want to show that $\max\{a_1, \ldots, a_{99}\} \leq 100$ (i.e. $U = 100$). The first strategy above would be to first figure out who is the oldest person in room: say that is the i^*'th person (where $1 \leq i^* \leq 99$) and then check if $a_{i^*} \leq 100$. However, this strategy is somewhat invasive: e.g. the oldest person might not want to reveal that he or she is the oldest person in the room. This is where the second strategy works better: we ask every person in the room if their age is ≤ 100: i.e. we check if for every $1 \leq i \leq 99$, $a_i \leq 100$. If everyone says yes, then we have proved that $\max_i a_i \leq 100$ (without necessarily revealing the identity of the oldest person).

Mathematically the above two strategies are the same. However, in "practice," using the second strategy turns out to be much easier. (E.g. this was true in the age example above.) Thus, here is the strategy to prove that $T(N)$ is $O(f(N))$:

> For every large enough N, show that for **every** input $\mathbf{x}$ of size N, $t_{\mathscr{A}}(\mathbf{x})$ is $O(f(N))$. Then conclude that $T(N)$ is $O(f(N))$.

C.2.3 Proving $T(N)$ is $\Omega(f(N))$

We start off with the same analogy as in the previous section. Say you wanted prove that given m numbers $a_1, \ldots, a_m$, $\max_i a_i \geq L$. Then how would you go about doing so? Again, one way is to argue that the maximum value is attained at i^* and then show that $a_{i^*} \geq L$. Now this is a perfectly valid way to prove the inequality we are after but note that you will *also* have to prove that the maximum value is attained at i^*. Generally, this is a non-trivial task. However, consider the following strategy:

> Show that there *exists* an $1 \leq i \leq m$, such that $a_i \geq L$. Then conclude that $\max_i a_i \geq L$.

Let us go back to the class room example. Now let us say we are interesting in proving that the oldest person in the room is at least 25 years old. (So $a_1, \ldots, a_m$ is as in Section C.2.2 but now $L = 25$.) Again, the first strategy would be to first figure out the oldest person, say i^* and check if $a_{i^*} \geq 25$. However, as we saw in Section C.2.2, this strategy is somewhat invasive. However, consider the the following implementation of the second strategy above. Say for the sake of mathematics, the instructor comes forward and volunteers the information that her age is at least 25. Since the oldest person's age has to be at least the instructor's age, this proves that $\max_i a_i \geq 25$, as desired.

Mathematically the above two strategies are the same. However, in "practice," using the strategy second turns out to be much easier. (E.g., this was true in the age example above.) Thus, here is the strategy to prove that $T(N)$ is $\Omega(f(N))$:

> For every large enough N, show that there **exists** an input $\mathbf{x}$ of size N, $t_{\mathscr{A}}(\mathbf{x})$ is $\Omega(f(N))$. Then conclude that $T(N)$ is $\Omega(f(N))$.

C.2.4 An Example

Now let us use all the strategies from Section C.2.2 and Section C.2.3 to asymptotically bound the run-time of a simple algorithm. Consider the following simple problem: given $n+1$ numbers $a_1, \ldots, a_n; v$, we should output $1 \leq i \leq n$ if $a_i = v$ (if there are multiple such i's then output any one of them) else output -1. Below is a simple algorithm to solve this problem.

Algorithm 41 Simple Search

INPUT: $a_1, \ldots, a_n; v$
OUTPUT: i if $a_i = v$; -1 otherwise

```
FOR every 1 ≤ i ≤ n DO
    IF a_i = v THEN RETURN i
RETURN −1
```

We will show the following:

Theorem C.2.1. *The Simple Search algorithm 41 has a run time of* $\Theta(n)$.

We will prove Theorem C.2.1 by proving Lemmas C.2.2 and C.2.3.

Lemma C.2.2. *$T(n)$ for Algorithm 41 is $O(n)$.*

Proof. We will use the strategy outlined in Section C.2.2. Let $a_1, \ldots, a_n; v$ be an arbitrary input. Then first note that there are at most n iterations of the for loop in Step 1. Further, each iteration of the for loop (i.e. Step 2) can be implemented in $O(1)$ time (since it involves one comparison and a potential return of the output value). Thus, by Lemma C.1.3, the total times taken overall in Steps 1 and 2 is given by

$$T_{12} \leq O(n \cdot 1) = O(n).$$

Further, since Step 3 is a simple return statement, it takes time $T_3 = O(1)$ time. Thus, we have that

$$t_{\text{Algorithm 41}}(a_1, \ldots, a_n; v) = T_{12} + T_3 \leq O(n) + O(1) \leq O(n),$$

where the last inequality follows from Lemma C.1.2 and the fact that $O(1)$ is also $O(n)$. Since the choice of $a_1, \ldots, a_n; v$ was arbitrary, the proof is complete. □

Lemma C.2.3. *$T(n)$ for Algorithm 41 is $\Omega(n)$.*

Proof. We will follow the strategy laid out in Section C.2.3. For every $n \geq 1$, consider the specific input $a'_i = n+1-i$ (for every $1 \leq i \leq n$) and $v' = 1$. For this specific input, it can be easily checked that the condition in Step 2 is only satisfied when $i = n$. In other words, the for loop runs at least (actually exactly) n times. Further, each iteration of this loop (i.e. Step 2) has to perform at least one comparison, which means that this step takes $\Omega(1)$ time. Since n is $\Omega(n)$, by Lemma C.1.3 (using notation from the proof of Lemma C.2.2), we have

$$T_{12} \geq \Omega(n \cdot 1) = \Omega(n).$$

Thus, we have

$$t_{\text{Algorithm 41}}(a'_1, \ldots, a'_n; v') \geq T_{12} \geq \Omega(n).$$

Since we have shown the existence of one input for each $n \geq 1$ for which the run-time is $\Omega(n)$, the proof is complete. □

A quick remark on the proof of Lemma C.2.3. Since by Section C.2.3, we only need to exhibit only *one* input with runtime $\Omega(n)$, the input instance in the proof of Lemma C.2.3 is only one possibility. One can choose other instances: e.g. we can choose an instance where the output has to be -1 (as a specific instance consider $a_i = i$ and $v = 0$). For this instance one can make a similar argument as in the proof of Lemma C.2.3 to show that $T(n) \geq \Omega(n)$.

C.2.5 The Best-Case Input "Trap"

We now briefly talk about a common mistake that is made when one starts trying to prove $\Omega(\cdot)$ on $T(N)$. Note that in Section C.2.3, it says that one can prove that $T(N)$ to be $\Omega(f(N))$ for every large enough N, one only needs to pick *one* input of size N for which the algorithm takes $\Omega(f(N))$ steps.

The confusing part about the strategy in Section C.2.3 is how does one get a hand on that special input that will prove the $\Omega(f(N))$ bound. There is no mechanical way of finding this input. Generally, speaking you have to look at the algorithm and get a feel for what input might force the algorithm to spend a lot of time. Sometimes, the analysis of the $O(\cdot)$ bound itself gives gives us a clue.

However, one way of picking the "special" input that **almost always never works** *in practice* is to consider (for every large enough N), the "best-case input," i.e. an input of size N on which the algorithm runs very fast. Now such an input will give you a valid lower bound but it would almost never give you a *tight* lower bound.

So for example, let us try to prove Lemma C.2.3 using the best case input. Here is one best case input: $a_i = i$ for every $i \in [n]$ and $v = 1$. Note that in this case the algorithm finds a match in the first iteration and this terminates in constant many steps. Thus, this will prove an $\Omega(1)$ lower bound but that is not tight/good enough.

Another common mistake is to make an argument for a fixed value of N (say $N = 1$). However, note that in this case one can never prove a bound better than $\Omega(1)$ and again, this trick never works in proving any meaningful lower bound.

C.3 Randomized Algorithms

So far the algorithms we have considered are deterministic, i.e. these are algorithm whose behavior is completely determined once the input is fixed. We now consider a generalization of such algorithms to algorithms that have access to random bits. In particular, even when the input is fixed, the behavior of the algorithm might change depending on the actual value of the random bits.[1] For the machine model, it is easy to modify the RAM model from Section C.2.1 to handle randomized algorithms: we can always load a register with independent and uniform random bits in one step.

Typically one considers randomized algorithms due to the following reasons:

- For some problems, it is easier to think of a randomized algorithm. Once one has designed a randomized algorithm, one could then attempt to "derandomize" the randomized algorithm to construct deterministic algorithms.

- In addition to conceptual simplicity, a randomized algorithm might run faster than all corresponding known deterministic algorithms.

- For certain problems, it might be provably impossible to design deterministic algorithms with certain guarantees but it *is* possible to design randomized algorithms with such guarantees. This is a common situation when we might be interested in algorithms that run in sub-linear time.

[1] There are many fundamental and interesting questions regarding how truly random these random bits are and how many such bits can an algorithm access. We will consider the ideal case, where an algorithm has access to as many uniform and independent random bits as it wants.

In this section, we will consider a problem where the third scenario above is applicable. For examples of the first and second scenarios, see Sections 12.3 and D.6 respectively.

Before delving into an example problem, we would like to clarify how we determine the run time and correctness of a randomized algorithm. There are multiple natural definitions but we will consider the following ones. The run time of a randomized algorithm will again be the worst-case run time as we defined for deterministic algorithms in Section C.2 for every possible choice of internal random bits that the algorithm might use (with the modification to the RAM model as discussed above). For correctness, the definition for deterministic algorithm was obvious so we did not explicitly state it: for every input, a deterministic algorithm must return the correct output. We call a randomized algorithm correct if on all its inputs, it returns the correct answer with probability bounded away from a $1/2$– to be precise let us say it has to return the correct output with probability at least $2/3$.[2]

We would like to remark on a subtle point in the definition above. In the definition of the correctness of a randomized algorithm above, the probability is taken over the random coin tosses that the algorithm might make. However, note that the guarantee is of the worst-case flavor in the sense that the algorithm has to be correct with high probability for *every* input. This should be contrasted with a scenario where the input might itself be random in which case we might be happy with an average case guarantee where the algorithm is supposed to return the correct output with high probability (over the distribution over the input). In particular, the algorithm is allowed to err on certain inputs as long as the total probability mass on the inputs on which it is incorrect is small: see Chapter 6 where this definition of correctness makes perfect sense. In such situations one can always assume that the algorithm itself is deterministic (see Exercise C.9).

C.3.1 An example problem

In the rest of the section, we will consider the following problem and will attempt to design (deterministic and randomized) algorithms with an eye to illustrate various points that were raised when we defined randomized algorithms.

> Given a vector $\mathbf{x} \in \{0,1\}^n$ determine whether $wt(\mathbf{x}) \leq \frac{n}{3}$ or $wt(\mathbf{x}) \geq \frac{2n}{3}$. For the cases where $wt(\mathbf{x}) \in (n/2, 2n/3)$ the algorithm can have arbitrary behavior.[3]

We will refer to the above as the GAPHAMMING problem.

It is easy to design an $O(n)$ time deterministic algorithm to solve GAPHAMMING: In $O(n)$ one can compute $wt(\mathbf{x})$ and then in $O(1)$ time one can verify if $wt(\mathbf{x}) \leq n/3$ or $wt(\mathbf{x}) \geq 2n/3$. In addition one can show that *any* correct deterministic algorithm will need a run time of $\Omega(n)$: see Exercise C.10.

We will now design a randomized algorithm that solves GAPHAMMING problem. Recall that we only need to determine if $wt(\mathbf{x}) \leq n/3$ or $wt(\mathbf{x}) \geq 2n/3$ (note that we assumed we do not get

[2] The choice of $2/3$ was arbitrary: see Exercise C.8.

[3] Or equivalently one can assume that the algorithm is given the *promise* that it will never encounter an input $\mathbf{x}$ with $wt(\mathbf{x}) \in (n/3, 2n/3)$.

inputs with Hamming weight in $(n/3, 2n/2)$) with high probability. We will present what is called a *sampling* algorithm for this task. To gain intuition, pick a random index $i \in [n]$. Note that then x_i is a random bit. Further, if $wt(\mathbf{x}) \le n/3$, then $\Pr_i[x_i = 1] \le 1/3$. On the other hand, if $wt(\mathbf{x}) \ge 2n/3$, then the probability is at least $2/3$. Thus, if we take s samples, with high probability in the first case we expect to see less than $s/3$ ones and in the second case we expect to see at least $2s/3$ ones. To get a constant probability of success we will invoke Chernoff bound to bound the probability of seeing more ones in the first case than the second case. Algorithm 42 for the details.

Algorithm 42 Sampling algorithm for GAPHAMMING

INPUT: $\mathbf{x} \in \{0,1\}^n$
OUTPUT: 0 if $wt(\mathbf{x}) \le n/3$ and 1 if $wt(\mathbf{x}) \ge 2n/3$ with probability at least $1-\varepsilon$

1: $s \leftarrow 98 \cdot \ln(1/\varepsilon)$
2: $C \leftarrow 0$
3: FOR $j \in [s]$ DO
4: Pick i to be a random index from $[n]$ ▷ The choice of i is independent for each j
5: $C \leftarrow C + x_i$
6: IF $C < s/2$ THEN
7: RETURN 0
8: RETURN 1

It can be checked that Algorithm 42 runs in time $O(\log(1/\varepsilon))$: see Exercise C.11. Next we argue that the algorithm is correct with probability at least $1-\varepsilon$.

Lemma C.3.1. *Algorithm 42 outputs the correct answer with probability at least* $1-\varepsilon$ *for every* $\mathbf{x}$ *(such that* $wt(\mathbf{x}) \notin (n/3, 2n/3)$*).*

Proof. We will prove the lemma for the case when $wt(\mathbf{x}) \le n/3$ and leave the other case to Exercise C.12.

Fix an arbitrary input $\mathbf{x}$ such that $wt(\mathbf{x}) \le n/3$. We will argue that at Step 6, we have

$$\Pr\left[C \ge \frac{s}{3} + \frac{s}{7}\right] \le \varepsilon. \tag{C.2}$$

Note that the above is enough to prove that the algorithm will output 0, as desired.

Towards that end for every $j \in [s]$, let Y_j be the random bit x_i that is picked. Note that $C = \sum_{j=1}^{s} Y_j$. Since each of the Y_j's are independent binary random variables, the additive Chernoff bound (Theorem 3.1.6) implies that

$$\Pr\left[C > \mathbb{E}[C] + \frac{s}{7}\right] \le e^{-\frac{s}{7^2 \cdot 2}} \le \varepsilon,$$

where the last inequality follows from our choice of s. As observed earlier for any j, $\Pr[Y_j = 1] \le 1/3$, which implies that $\mathbb{E}[C] \le s/3$, which with the above bound implies (C.2), as desired. □

Finally, we consider the average-case version of the GAPHAMMING problem. Our goal is to illustrate the difference between randomized algorithms and average-case algorithms that was alluded to earlier in this section. Recall that in the GAPHAMMING problem, we are trying to distinguish between two classes of inputs: one with Hamming weight at most $n/3$ and the other with Hamming weight at least $2n/3$. We now consider the following natural version where the inputs themselves comes from two distributions and our goal is to distinguish between the two cases.

> Let $\mathbb{D}_p$ denote the distribution on $\{0,1\}^n$, where each bit is picked independently with probability $0 \le p \le 1$. Given an $\mathbf{x} \in \{0,1\}^n$ sampled from either $\mathbb{D}_{\frac{1}{3}}$ or $\mathbb{D}_{\frac{2}{3}}$, we need to figure out which distribution $\mathbf{x}$ is sampled from.

The intuition for an algorithm that is correct with high probability (over the corresponding distributions) is same as Algorithm 42, so we directly present the the algorithm for the new version of the problem above.

Algorithm 43 An average-case algorithm for GAPHAMMING

INPUT: $\mathbf{x} \in \{0,1\}^n$ sampled from either $\mathbb{D}_{\frac{1}{3}}$ or $\mathbb{D}_{\frac{2}{3}}$

OUTPUT: 0 if $\mathbf{x}$ was sampled from $\mathbb{D}_{\frac{1}{3}}$ and 1 otherwise with probability at least $1-\varepsilon$

```
s ← 98 · ln(1/ε)
C ← 0
FOR j ∈ [s] DO
    C ← C + x_j
IF C < s/2 THEN
    RETURN 0
RETURN 1
```

Note that unlike Algorithm 42, Algorithm 43 is a *deterministic* algorithm and the algorithm might make an incorrect decision on certain specific inputs $\mathbf{x}$ that it receives.

Using pretty much the same analysis as in the proof of Lemma C.3.1, one can argue that:

Lemma C.3.2. *Let* $\mathbf{x}$ *be a random sample from* $\mathbb{D}_{\frac{1}{3}}$ *(*$\mathbb{D}_{\frac{2}{3}}$ *resp.). Then with probability at least* $1-\varepsilon$ *(over the choice of* $\mathbf{x}$*), Algorithm 43 outputs* 0 *(*1 *resp.)*

(See Exercise C.13 for a proof.)

C.4 Efficient Algorithms

A major focus of this book is to design algorithms that are efficient. A somewhat smaller focus is to argue that for certain problems efficient algorithms do not exists (maybe with a well accepted assumption that certain computational tasks are hard to accomplish efficiently). In this section, we first begin with the notion of efficient algorithms that will be standard for this book and then

present a peek into how one might argue that a computational task is hard. To illustrate various concepts we will focus on the following problem:

> Given n linear equations over k variables (all over $\mathbb{F}_2$: i.e. all the variables are in $\{0,1\}$ and all arithmetic operations are over the binary field [4] $\mathbb{F}_2$) and an integer $0 \le s \le n$, we want to find a solution to the systems of equations that satisfies at least s out of the n equations. We will denote this problem as MaxLinearEq(k,n,s). We will drop the arguments when we want to talk about the problem in general.

We choose the non-standard notation of k for number of variables and n for number of equations as they correspond better to problems in coding theory that we would be interested in.

An overwhelming majority of the algorithmic problems considered in this book will have the following property: there are exponentially many possible solutions and we are interested in a solution (or solutions) that satisfy a certain objective. For example, in the MaxLinearEq problem, there are 2^k possible solutions and we are interested in a solution that satisfies at least s many linear equations. Note that such problems have a very natural exponential time algorithm: generate all (the exponentially many) potential solutions and check if the current potential solution satisfy the objective. If it does, then the algorithm stops. Otherwise the algorithm continues to the next solution. For example, Algorithm 44 instantiates this general algorithm for the MaxLinearEq problem.

Algorithm 44 Exponential time algorithm for MaxLinearEq

INPUT: n linear equations over k variables and an integer $0 \le s \le n$
OUTPUT: A solution in $\{0,1\}^k$ that satisfies at least s of the equations or fail if none exists

```
FOR every x ∈ {0,1}^k DO
    Let t be the number of equations the solution x satisfies
    IF t ≥ s THEN
        RETURN x
RETURN fail
```

It is not too hard to argue that Algorithm 44 runs in time $O\left(kn2^k\right)$ (see Exercise C.14). We point out two things that will expand into more detailed discussion on what is an efficient algorithm (and what is not):

1. A run time of $\Omega(2^k)$ is not efficient for even moderate values of k: indeed for $k = 100$, the number of steps of the algorithm exceeds the number of particles in the universe.

2. In the generic exponential time algorithm mentioned earlier, we made the implicit assumption that given a potential solution we can "quickly" verify if the potential solution satisfies the objective or not.

[4] In other words, addition is XOR and multiplication is AND.

Notwithstanding the fact that an exponential run time can become infeasible for moderate input size, one might think that one cannot do better than Algorithm 44 to solve the MAXLINEAREQ problem. In particular, the lack of any extra information other than the fact that we have a system of linear equations on our hands seems to make the possibility of coming up with a faster algorithm slim. However, looks can sometimes be deceptive, as we will see shortly.

Consider the special case of the MAXLINEAREQ problem: MAXLINEAREQ(k, n, n). In other words, we want to see if there exists a solution $\mathbf{x} \in \{0,1\}^n$ that satisfies all the equations. Not only is this is an interesting special case but it is also relevant to this book since this setting corresponds to the error detection problem (Definition 1.3.4) for linear codes (Chapter 2)– see Exercise C.15. It turns out this special setting of parameters makes the problem easy: one can use Gaussian elimination to solve this problem in time $O(kn^2)$ (see Exercise C.16). For the case of $k = \Theta(n)$ (which would be the most important parameter regime for this book), this cubic run time is much faster than the exponential time Algorithm 44. In particular, any run time of the form $O(n^c)$ for some fixed constant c would be much faster than the $2^{\Omega(n)}$ run time of Algorithm 44 (for large enough n). Note that the appealing aspect of a run time of the form $O(n^c)$ is that when the input size doubles, the run time only increases by a constant (though clearly we might be pushing the boundary of a practical definition of a constant for moderately large values of c) as opposed to the exponential run time, where the run time on the new input size is quadratic in the old run time.

In theoretical computer science, the notion of an efficient algorithm is one that runs in time $O(N^c)$ on inputs of size N for some fixed constant $c \geq 0$: such algorithms are said to have *polynomial run time*. In particular, a problem is said to be in the *complexity class* P if it admits a polynomial time algorithm[5]. While one might debate the definition of P as capturing algorithms that are efficient in practice, it clearly seems to capture the difference between problems that "need" exponential time algorithms and problems that have some inherent structure that allows much faster algorithmic solutions (in this case polynomial time algorithm).

C.4.1 Computational Intractability

So far we have talked mainly about problems that admit efficient algorithms. We now consider the issue of how we talk about a problem being hard: e.g. can we somehow formally argue that a certain problem cannot admit efficient solutions? In particular, are there problems where the generic exponential time algorithm discussed earlier is the best possible? To be more precise, let us consider problems where given a potential solution one can in polynomial time determine whether the solution satisfies the objective or not. We call the class of such problems as NP.[6] For example, MAXLINEAREQ(k, n, s) is such a problem because given a potential solution $\mathbf{x}$, one can in time $O(kn)$ verify whether it satisfies at least s out of the n solutions– see Exercise C.17 and hence MAXLINEAREQ $\in$ NP. Like the earlier special case of MAXLINEAREQ(k, n, n),

[5]The technical definition of P is a bit more nuanced: in particular it only considers problems with a binary output but we will ignore this technical issue in this book.

[6]Again for the technical definition we need to only consider problems with binary output but we will ignore this technicality.

the more general problem MaxLinearEq(k, n, s) for $s < n$ is also interesting from a coding theory perspective: see Exercise C.18.

Thus, the question of whether there exists a problem where the earlier exponential time algorithm is the best possible is essentially the same as showing $\mathsf{P} \neq \mathsf{NP}$ (see Exercise C.19). While we are nowhere close to answer this fundamental question, we do know a way to identify the "core" of hard problems in NP. Such problems (called NP-complete problems) have the property that if any of them do not have a polynomial time algorithms then none of them do. (Conversely if any of them do have a polynomial time algorithm then $\mathsf{P} = \mathsf{NP}$.) Note that this implies if one assumes that $\mathsf{P} \neq \mathsf{NP}$, then these NP-complete problems are hard problems since they are not in P (which we consider to be the class of "easy" problems).

At first blush, one might wonder how one would go about proving that such problems exist. Proving the existence of such a problem is out of the scope of the book. However, we do want to give an overview of how given one such specific problem that is NP-complete one might argue that another problem is also NP-complete. The way to show such a result is to *reduce* the known NP-complete problem (let us call this problem P_1) to the other problem (let us call this problem P_2). Without going into the technical definition of a reduction, we present an informal definition, which would be sufficient for our purposes. A reduction is a polynomial time algorithm (let us call it $\mathscr{A}_1$) that given an *arbitrary* instance $\mathbf{x}_1$ of P_1 can produce in polynomial time another instance $\mathbf{x}_2$ but this time for the problem P_2 such that given the answer for problem P_2 on $\mathbf{x}_2$, one can in polynomial time exactly determine the answer of P_1 on $\mathbf{x}_1$ (by another algorithm, which let us call $\mathscr{A}_2$). There are two (equivalent) ways to think about such a reduction:

1. A reduction implies that to solve P_1 in polynomial time, it is enough to "only" solve some subset of instances for problem P_2 (in particular, those inputs for P_2 that are generated by $\mathscr{A}_1$ on all possible input instances of P_1). In other words, P_2 is "harder" to solve than P_1. Since the problem P_1 is a hard problem, P_2 is also a hard problem.

2. Let us for the sake of contradiction assume that there exists a polynomial time algorithm $\mathscr{A}_3$ that solves P_2 on all instances (i.e. P_2 is easy). Then one can construct a polynomial time algorithm to solve P_1 as follows. Given an arbitrary input $\mathbf{x}_1$ for P_1, first use $\mathscr{A}_1$ to generate an input $\mathbf{x}_2$ for P_2. Then use $\mathscr{A}_3$ to solve P_2 on $\mathbf{x}_2$ and then convert the answer of P_2 on $\mathbf{x}_2$ to the answer of P_1 on $\mathbf{x}_1$ by using $\mathscr{A}_2$. Note that this is a polynomial time algorithm and is a correct algorithm. Thus, we have proved that P_1 is easy, which contradicts our assumption that P_1 is hard.

To make the concept of reduction a bit less abstract we outline a reduction from a known NP-complete problem to our MaxLinearEq problem.[7] In particular, the following problem is known to be NP-complete

> Given a graph $G = (V, E)$ with $|V| = k$ and $|E| = n$ and an integer $0 \leq s \leq n$, does there exist a *cut* of size at least s. In other words, does there exist a subset $S \subset V$ such that the number of edges with one end point in S and the other in $V \setminus$ is at least s? We will call this the MaxCut(k, n, s) problem.

[7] We assume that the reader is familiar with the mathematical concept of graphs, where we do *not* mean graphs in the sense of plots.

Algorithm 45 is the algorithm $\mathscr{A}_1$ of the reduction from MaxCut(k, n, s) to MaxLinearEq(k, n, s).

Algorithm 45 Reduction from MaxCut to MaxLinearEq

INPUT: An instance for MaxCut(k, n, s): a graph G and an integer s
OUTPUT: An instance of MaxLinearEq(k', n', s')

```
k' ← k, n' ← n, s' ← s
FOR every vertex i ∈ V DO
    Add a variable x_i to the set of variables
FOR every edge (i, j) ∈ E DO
    Add a linear equation x_i + x_j = 1 to the system of equation
```

Further, the algorithm $\mathscr{A}_2$ is simple: given a solution $(x_1, \ldots, x_k)$ to the instance for MaxLinearEq(k, n, s) problem, consider the cut $S = \{i \in V | x_i = 1\}$. It can be checked that this algorithm and Algorithm 45 forms a valid reduction. (See Exercise C.20.)

C.5 Exercises

Exercise C.1. Prove that $f(N)$ is $O(g(N))$ (as per Definition C.1.1) if and only if

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} \le C,$$

for some absolute constant C.

Exercise C.2. Prove that $f(N)$ is $\Omega(g(N))$ (as per Definition C.1.2) if and only if

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} \ge C,$$

for some absolute constant C.

Exercise C.3. Prove that $f(N)$ is $\Theta(g(N))$ (as per Definition C.1.3) if and only if

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} = C,$$

for some absolute constant C.

Exercise C.4. Prove that $f(N)$ is $o(g(N))$ (as per Definition C.1.4) if and only if

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} = 0.$$

Exercise C.5. Prove that $f(N)$ is $\omega(g(N))$ (as per Definition C.1.5) if and only if

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} = \infty.$$

Exercise C.6. Prove Lemmas C.1.1, C.1.2 and C.1.3.

Exercise C.7. Prove or disprove the following for every $\alpha \in \{O, \Omega, \Theta, o, \omega\}$:

> Let $f(N)$ be $\alpha(h(N))$ and $g(N)$ be $\alpha(h(N))$. Then $f(N) \cdot g(N)$ is $\alpha(h(N))$.

Exercise C.8. Say there exists a randomized algorithm $\mathscr{A}$ that is correct with probability $\frac{1}{2} + \delta$ for some $\delta > 0$ with runtime $T(N)$. Then show that for every $\varepsilon > 0$, there exists another randomized algorithm that is correct with probability $1 - \varepsilon$ with runtime $O\left(\frac{\log(1/\varepsilon)}{\delta} \cdot T(N)\right)$.

Hint: Repeat $\mathscr{A}$ multiple times and pick one among the multiple outputs. For analysis use the Chernoff bound (Theorem 3.1.6).

Exercise C.9. Assume that there is a randomized algorithm $\mathscr{A}$, which one when provided with an input from a distribution $\mathbb{D}$, is correct with probability p (where the probability is taken over both $\mathbb{D}$ and internal randomness of $\mathscr{A}$). Then show that there exists a deterministic algorithm that is correct with probability at least p (where the probability is now only taken over $\mathbb{D}$) with the same run time as $\mathscr{A}$.

Exercise C.10. Argue that any correct deterministic algorithm that solves the GapHamming problem needs a run time of $\Omega(n)$.

Hint: Argue that any correct deterministic algorithm needs to read $\Omega(n)$ bits of the input.

Exercise C.11. Argue that Algorithm 42 runs in time $O(\log(1/\varepsilon))$.

Exercise C.12. Prove that for every $\mathbf{x} \in \{0,1\}^n$ such that $wt(\mathbf{x}) \geq 2n/3$, Algorithm 42 outputs 1 with probability at least $1 - \varepsilon$.

Exercise C.13. Prove Lemma C.3.2 and that Algorithm 43 runs in time $O(\log(1/\varepsilon))$.

Exercise C.14. Argue that Algorithm 44 runs in time $O\left(kn2^k\right)$. Conclude that the algorithm runs in time $2^{O(n)}$.

Exercise C.15. Show that if any MaxLinearEq(k, n, n) problem can be solved in time $T(k, n)$, then the error detection for any $[n, k]_2$ code can be solved in $T(k, n)$ time.

Hint: The two problems are in fact *equivalent*.

Exercise C.16. Argue that the problem MaxLinearEq(k, n, n) can be solved in time $O(kn^2)$.

Exercise C.17. Show that given a system of n linear equation on k variables over $\mathbb{F}_2$, there exists a $O(kn)$ time algorithm that given a vector $\mathbf{x} \in \{0,1\}^n$ can compute the exact number of equations $\mathbf{x}$ satisfies.

Exercise C.18. Consider the following problem called the Bounded Distance Decoding problem. Given a code $C \subseteq \{0,1\}^n$, a vector $\mathbf{y} \in \{0,1\}^n$ and an integer $0 \leq e \leq n$ (called the error radius), output any codeword $\mathbf{c} \in C$ such that $\Delta(\mathbf{c}, \mathbf{y}) \leq e$ (or state that no such codeword exists).

Prove that if any MaxLinearEq(k, n, s) problem can be solved in time $T(k, n, s)$, then one can solve the Bounded Distance Decoding problem for any $[k, n]_2$ linear code with error radius $n - s$.

Exercise C.19. Argue that showing $\mathsf{P} \neq \mathsf{NP}$ is equivalent to showing that $\mathsf{NP} \setminus \mathsf{P} \neq \emptyset$.

Exercise C.20. Argue that Algorithm 45 and the algorithm $\mathscr{A}_2$ defined just below it are correct and run in polynomial time.

C.6 Bibliographic Notes

The full suite of asymptotic notation in Section C.1 was advocated for analysis of algorithms by Knuth [62]. The Big-Oh notation is credited to Bachmann [6] form a work in 1894 and the little-oh notation was first used by Landau [65] in 1909. A variant of the Big-Omega notation was defined by Hardy and Littlewood [53] in 1914 though the exact definition in Section C.1 seems to be from [62]. The Theta and little omega notation seem to have been defined by Knuth [62] in 1976: Knuth credits Tarjan and Paterson for suggesting the Theta notation to him.

The choice to use worst-case run time as measure of computational efficiency in the RAM model as well as only considering the asymptotic run time (as opposed to more fine grained analysis as advocated by Knuth) seem to have been advocated by Hopcroft and Tarjan: see Tarjan's Turing award lecture for more on this [99].

Cobham [18] and Edmonds [26] are generally credited with making the first forceful case for using P as the notion of efficiently solvable problems. Somewhat interestingly, Peterson's paper on decoding of Reed-Solomon codes [80] that predates these two work explicitly talks about why a polynomial time algorithm is better than an exponential time algorithm (though it does not explicitly define the class P). The notion of NP (along with a proof of the existence of an NP-complete problem) was defined independently by Cook [19] and Levin [67]. This notion really took off when Karp showed that 21 natural problems where NP-complete (including the MaxCut problem) [60]. For more historical context on P and NP including some relevant historical comments, see the survey by Sipser [93].

The first randomized algorithms is generally credited to Rabin [81]. However, an earlier work of Berlekamp on factoring polynomials presents a randomized algorithm (though it is not stated explicitly as such) [9].

This chapter gave a very brief overview of topics that generally span multiple classes. For further readings, please consult standard textbooks on the subjects of (introductory) algorithms and computational complexity as well as randomized algorithms.

Appendix D

Basic Algebraic Algorithms

D.1 Executive Summary

In this appendix we include some basic facts about abstract algebra that were used throughout the book. Readers who are comfortable with their background in algebra should feel free to skip it entirely. However, this background should include both aspects introduced by *finiteness* — most fields we work with are finite, and so are the vector spaces defined over them — and *computation* — the mere existence of a nice algebraic structure is not good enough for us, we need to know how to carry out basic, and not so basic operations over these structures efficiently. If you are not very comfortable with these settings you will find the appropriate sections of this appendix more useful. The opening paragraph of each section summarizes the main aspects covered in the section and the reader may use them to decide if they wish to read further.

Some of the material in this appendix appears earlier in the book (e.g. Sections 2.1, 2.2 and 5.1). Finally, this coverage of algebra in this appendix is not exhaustive and the reader is referred to the book by Lidl and Niederreiter [69] for more matetrial (and proofs) on finite fields and the book by Shoup for more details on the basic algebraic algorithms [91].

D.2 Groups, Rings, Fields

The title of this section says it all. We cover, very tersely, the definition of a group, a ring, and a field.

We begin with some terminology. We consider binary operations over some set of elements. Given a set X such a binary operator would be a function $\circ : X \times X \to X$, and we usually use $a \circ b$ to denote $\circ(a, b)$, for $a, b \in X$. We say the operator $\circ$ is *associative* if $a \circ (b \circ c) = (a \circ b) \circ c$, for every $a, b, c \in X$. For associative operations it is customary to drop the parenthesis. We say the operator $\circ$ is *commutative* if $a \circ b = b \circ a$ for every $a, b \in X$. We say an element $e \in X$ is an identity for $\circ$ if $a \circ e = e \circ a = a$ for every $a \in X$. Identities are, by definition, unique if they exist, since if $e_1, e_2 \in X$ were identities, we would have $e_1 = e_1 \circ e_2 = e_2$. Given $a \in X$ and operator $\circ$ with inverse e we say that a is is *invertible* with respect to $\circ$ if there exists an element $a^{-1} \in X$ such that $a \circ a^{-1} = a^{-1} \circ a = e$. Often $\circ$ will be clear from context in which case we will refer to a

as simply invertible.

Definition D.2.1 (Group)**.** Given a set G and a binary operation $\circ$ over G, we say that $(G, \circ)$ is a group if $\circ$ is associative, has an identity, and every element of G is invertible. A group $(G, \circ)$ is said to be an *abelian* group if $\circ$ is also commutative.

Examples of groups include the integers with addition, the non-zero rationals with multiplication and the set of permutations (one-to-one functions) on any finite set under the composition operation.

Definition D.2.2 (Ring)**.** A finite set R with two binary operations $+$ and $\cdot$ are said to form a ring if (1) $(R, +)$ form an abelian group, (2) $\cdot$ is associative and has an identity and (3) $\cdot$ *distributes* over $+$, i.e., for every $a, b, c \in R$ we have $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ and have $(b + c) \cdot a = (b \cdot a) + (c \cdot a)$. The ring $(R, +, \cdot)$ is said to be a commutative ring if $\cdot$ is commutative.

Examples include the integers over addition and multiplication (a commutative ring) and the set of $k \times k$ integer matrices (for any positive integer k) under matrix addition and matrix multiplication (which forms a non-commutative ring for $k \geq 2$).

Definition D.2.3 (Field)**.** A set $\mathbb{F}$ with operations $+$ and $\cdot$ forms a field if $(\mathbb{F}, +, \cdot)$ is a commutative ring, and $(\mathbb{F} \setminus \{0\}, \cdot)$ is a group where 0 denotes the identity for $+$.

Examples of fields include the rationals, the reals, the complexes (all under addition and multiplication) and (more interestingly to us) the integers modulo any prime number p (see Lemma 2.1.2 for the latter).

It is customary in rings and fields to let 0 denote the additive identity, 1 the multplicative identity and to let $-a$ denote the additive inverse of a and a^{-1} the multiplicative inverse of a. It is also customary to abbreviate $a + (-b)$ to $a - b$.

D.3 Polynomials

In this section we will introduce polynomial rings, mention when they satisfy the unique factorization property, describe the 'remainder algorithm', and describe the evaluation map and the polynomial distance property (where the latter is a re-statement of the degree mantra (Proposition 5.2.3)).

Definition D.3.1 (Formal Polynomials)**.** Let $(R, +, \cdot)$ be a commutative ring with identity 0. The set of formal polynomials over R in indeterminate X, denoted $R[X]$, is given by finite formal sums $R[X] = \{\sum_{i=0}^{d} f_i X^i | f_0, \ldots, f_d \in R; d \in \mathbb{Z}^{\geq 0}\}$, under the equivalence $\sum_{i=0}^{d} f_i X^i = \sum_{i=0}^{d-1} f_i X^i$ if $f_d = 0$. (The term formal refers to the fact that the summation, and the terms X^i are just formal symbols and do not have operational meaning, yet. So really polynomials are just finite sequences of elements from R under the equivalence $(f_0, \ldots, f_d, 0) \cong (f_0, \ldots, f_d)$.)

Basic terminology The elements f_i are referred to as the coefficients of f, the symbols X^i as the monomials of f and the product $f_i X^i$ as the terms of f. For $f = \sum_{i=0}^{d} f_i X^i$, its *degree*, denoted $\deg_X(f)$ or simply $\deg(f)$, is the largest integer e such that $f_e \neq 0$.

Addition The sum of two polynomials $f = \sum_{i=0}^{d} f_i X^i$ and $g = \sum_{i=0}^{d} g_i X^i$, denoted $f + g$, is the polynomial $\sum_{i=0}^{d} (f_i + g_i) X^i$. (Note that by padding the coefficients of f and g with zeroes we can always arrange it so that they have the same number of terms.)

Multiplication Finally, the product of $f = \sum_{i=0}^{d} f_i X^i$ $g = \sum_{i=0}^{e} g_i X^i$, denoted $f \cdot g$ (or sometimes simply fg), is given by $\sum_{i=0}^{d+e} \left(\sum_{j=0}^{e} f_{i-j} \cdot g_j \right) X^i$.

The following proposition follows immediately from the definitions above.

Proposition D.3.1. *For every commutative ring R, $R[X]$ is a commutative ring under the sum and product of polynomials.*

In fact R inherits many properties of R and in particular the notion of "unique factorization" which we describe next.

Definition D.3.2 (Unique Factorization Domains)**.** Let R be a commutative ring. An element $u \in R$ is said to be a *unit* if it has a multiplicative inverse in R. Elements a and b are said to be *associates* if there exists a unit u such that $a = b \cdot u$. (Note that being associates is an equivlence relationship.) Element $a \in R$ is said to be *irreducible* if $a = b \cdot c$ implies either b or c is a unit. A *factorization* of $a \in R$ is a sequence $b_1, \ldots, b_k$ such that $a = b_1 \cdot b_2 \cdots b_k$ and none of the b_i's are units. The b_i are referred to as the factors of a in this factorization. Ring R is a *factorization domain* if for non-zero every $a \in R$ that is not a unit, there is a finite bound k_a such that every factorization of a has at most k_a factors. A factorization domain R is a *unique factorization domain (UFD)* if every non-zero, non-unit element has a unique irreducible factorization, upto associates. I.e., if $a = b_1 \cdots b_k = c_1 \cdots c_\ell$ and the b_i's and c_j's are irreducible, then $k = \ell$ and there exists a bijection $\pi : [k] \to [\ell]$ such that b_i and $c_{\pi(i)}$ are associates, for every $i \in [k]$.

Since every non-zero element of a field is a unit, every field is a UFD.

Proposition D.3.2. *Every field is a UFD.*

A central result in basic commutative algebra is the following lemma of Gauss.

Lemma D.3.3 (Gauss)**.** *If R is a UFD, then so is $R[X]$.*

We omit the proof of the above lemma here, but point out its implications. It allows us to build many interesting rings from a simple base case, namely a field. Given a field $\mathbb{F}$, $\mathbb{F}[X]$ is a UFD. So is $(\mathbb{F}[X])[Y]$. Now we could have gone in the other direction and created the ring $(\mathbb{F}[Y])[X]$ and this would be a UFD too. However if X and Y commute (so $XY = YX$) then the rings $(\mathbb{F}[X])[Y]$ and $(\mathbb{F}[Y])[X]$ are isomorphic under the isomorphism that preserves $\mathbb{F}$ and sends $X \to X$ and $Y \to Y$. So we tend to compress the notation and refer to this ring as $\mathbb{F}[X, Y]$, the ring of "bivariate" polynomials over $\mathbb{F}$. Rings of univariate and mutlivariate polynomials play a central role in algebraic coding theory.

We now turn to the notion of polynomial *division* with *remainder* that lead us to some important notions associated with polynomials.

Let $f \in R[X]$ and let $f = \sum_{i=0}^{d} f_i X^i$ with $f_d \neq 0$. f is said to be *monic* if f_d is a unit in R.

Proposition D.3.4. *Given a monic polynomial f, and general polynomial p there exists a unique pair of polynomials q (for quotient) and r (for remainder) such that $p = q \cdot f + r$ and* $\deg(r) < \deg(f)$.

See Exercise D.1 for a proof.

The function $p \mapsto_f (q, r)$ described is often referred to as the 'division algorithm' (since it is the outcome of long division). A special case that is of great interest to us is when $f = X - \alpha$ for $\alpha \in R$. In this case the remainder is polynomial of degree at most 0, and so can be associated with an element of R. Denote this element $p(\alpha)$ (since it depends only on p and α) and we get the "evaluation" map which maps elements of $\mathbb{R}[X] \times R$ to R. The remainder $p(\alpha)$ can be worked out explicitly and is given by the simple form below (where the uniqueness follows from Proposition D.3.4).

Proposition D.3.5. *Given $p = \sum_{i=0}^{d} p_i X^i \in R[X]$ and $\alpha \in R$, let $p(\alpha) = \sum_{i=0}^{d} p_i \alpha^i$. Then there exists a unique $q \in R[X]$ such that $p = q \cdot (X - \alpha) + p(\alpha)$. It follows that $p(\alpha) = 0$ if and only if $X - \alpha$ divides $p(X)$.*

Finally using Proposition D.3.5 and the fact that $\mathbb{F}[X]$ is a UFD, we get the following the following central fact about univariate polynomials.

Lemma D.3.6 (Polynomial Distance Lemma)**.** *Let $f \neq g \in \mathbb{F}[X]$ be polynomials of degree at most d. Then there exist at most d elements $\alpha \in \mathbb{F}$ such that $f(\alpha) = g(\alpha)$.*

Proof. Let $h = f - g$. We have h is non-zero and of degree at most d. Let $S = \{\alpha | f(\alpha) = g(\alpha)\}$. Then we have $(X - \alpha)$ divides h for every $\alpha \in S$. Furthermore, by the unique factorization property we have $\tilde{h} = \prod_{\alpha \in S}(X - \alpha)$ divides h. But if $\tilde{h}$ divides h, then $\deg(\tilde{h}) \leq \deg(h)$ and $\deg(\tilde{h}) = |S|$. We conclude $|S| \leq d$. □

D.4 Vector Spaces

In this section we introduce vector spaces over fields and describe two basic views of describing a finite dimensional vector space: first via its generators (and the generator matrix) and next via constraints on the vector space (and its parity check matrix). We first start with a quick overview of matrices and the corresponding notation and then move on to vector spaces.

D.4.1 Matrices and Vectors

In this book, a vector of length n over the field $\mathbb{F}$ (i.e. $\mathbf{x} \in \mathbb{F}^n$) is a row vector[1]. E.g., we have $\mathbf{x} = \begin{pmatrix} 0 & 1 & 3 & 4 & 0 \end{pmatrix} \in \mathbb{F}_5^4$. Given two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{F}^n$, their *inner product* is defined as

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{n} u_i \cdot v_i,$$

[1]We acknowledge that this is different from the usual assumption in linear algebra that all vectors are column vectors. We are assuming row vectors to be consistent with how message vectors are assumed to be row vectors in coding theory.

where the multiplication and addition are over $\mathbb{F}$.

A matrix $M \in \mathbb{F}^{k\times n}$ is a two-dimensional array/vector, where we refer to the (i,j)'th entry (for $(i,j) \in [k]\times[n]$) as $M_{i,j}$ (or M_{ij} if the two indices are clear without being separated by a comma). We will use $M_{i,\cdot}$ as the i'th row and $M_{\cdot,j}$ as the jth column of M respectively.

So e.g. consider $G \in \mathbb{F}_3^{2\times 3}$ as follows:

$$G = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \end{pmatrix}.$$

In the above $G_{1,2} = 2$, $G_{1,\cdot} = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}$ and $G_{\cdot,2} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$.

The *transpose* of a matrix $M \in \mathbb{F}^{k\times n}$, denoted by M^T is an $n\times k$ matrix over $\mathbb{F}$ such that for any $(j,i) \in [n]\times[k]$, we have

$$M^T_{j,i} = M_{ij}.$$

Note that if $k = 1$, then the above says that for a row vector $\mathbf{x} \in \mathbb{F}^n$, its transpose $\mathbf{x}^T$ is a column vector.

The product of two matrices $A \in \mathbb{F}^{k\times n}$ and $B \in \mathbb{F}^{n\times m}$ is a matrix $C \in \mathbb{F}^{k\times m}$ such that for any $(i,j) \in [k]\times[m]$, we have

$$C_{i,j} = \langle A_{i,\cdot}, B_{\cdot,j} \rangle.$$

D.4.2 Definition and Properties of Vector Spaces

This section will repeat some of the material from Section 2.2. We begin with the definition of a vector space:

Definition D.4.1 (Vector Space)**.** Over a field $\mathbb{F}$, a vector space is given by a triple $(V,+,\cdot)$ where $(V,+)$ is a commutative group and $\cdot : \mathbb{F}\times V \to V$ distributes over addition, so that $\alpha\cdot(\mathbf{u}+\mathbf{v}) = \alpha\cdot\mathbf{u}+\alpha\cdot\mathbf{v}$ for every $\alpha\in\mathbb{F}$ and $\mathbf{u},\mathbf{v}\in V$. It is customary to denote the identity of the group $(V,+)$ by $\mathbf{0}$ and to refer to V as an $\mathbb{F}$-vector space.

The simplest example of an $\mathbb{F}$-vector space is $\mathbb{F}^n$, whose elements are sequences of n elements of $\mathbb{F}$. The sum is coordinate-wise summation and product is "scalar" product, so if $\mathbf{u} = (u_1,\dots,u_n)$, $\mathbf{v} = (v_1,\dots,v_n)$ and $\alpha\in\mathbb{F}$ then $\mathbf{u}+\mathbf{v} = (u_1+v_1,\dots,u_n+v_n)$ and $\alpha\cdot\mathbf{u} = (\alpha\cdot u_1,\dots,\alpha\cdot u_n)$. Essentially these are the only vector spaces (as we will make precise soon), but representations of the vectors is important to us, and will make a difference.

Definition D.4.2 (Dimension of a vector space)**.** A sequence of vectors $\mathbf{v}_1,\dots,\mathbf{v}_k \in V$ are said to be *linearly independent* if $\sum_{i=1}^k \beta_i\cdot\mathbf{v}_i = \mathbf{0}$ implies that $\beta_1 = \cdots = \beta_k = 0$. $\mathbf{v}_1,\dots,\mathbf{v}_k \in V$ are said to linearly dependent otherwise.

V is said to be finite dimensional of dimension k if every sequence of $k+1$ vectors from V is linearly dependent and there exists a sequence of length k that is linearly independent.

A linearly independent set $\mathbf{v}_1,\dots,\mathbf{v}_k \in V$ is said to form a *basis* for V if V has dimension k.

Every $\mathbb{F}$-vector space of dimension k is isomorphic to $\mathbb{F}^k$ as described by the following proposition.

Proposition D.4.1. *If* $\mathbf{v}_1, \ldots, \mathbf{v}_k$ *for a basis for an* $\mathbb{F}$*-vector space* V*, then* $V = \{\sum_{i=a}^{k} \beta_i \cdot \mathbf{v}_i | \beta_1, \ldots, \beta_k \in \mathbb{F}\}$ *and the map* $(\beta_1, \ldots, \beta_k) \mapsto \sum_{i=1}^{k} \beta_i \cdot \mathbf{v}_i$ *is an isomorphism from* $\mathbb{F}^k$ *to* V*.*

The point we wish to stress now is that even though all vector spaces are isomorphic, different spaces do lead to different codes with different error-correction properties, and these properties are not preserved by such isomorphisms. So not all k-dimensional vector spaces are identical for our purposes. We will specially be interested in k-dimensional vector spaces contained in $\mathbb{F}^n$, and how these can be represented succinctly in matrix form.

Definition D.4.3 (Generator Matrix, Parity Check Matrix)**.** A matrix $G \in \mathbb{F}^{k \times n}$ is said to be a *generator matrix* of an $\mathbb{F}$-vector space $V \subseteq \mathbb{F}^n$ if the rows of G are linearly independent in $\mathbb{F}^n$ and $V = \{\mathbf{x} \cdot G | \mathbf{x} \in \mathbb{F}^k\}$. The rows of G form a basis of V. A matrix $H \in \mathbb{F}^{n \times (n-k)}$ is said to be a *parity check matrix* of an $\mathbb{F}$-vector space $V \subseteq \mathbb{F}^n$ if the columns of H are linearly independent and $V = \{\mathbf{y} \in \mathbb{F}^n | H \cdot \mathbf{y}^T = \mathbf{0}\}$. Given a vector space V with parity check matrix H, its dual space, denoted $V^\perp$, is the vector space generated by the transpose of H, i.e., $V^\perp = \{\mathbf{x} \cdot H^T | \mathbf{x} \in \mathbb{F}^{n-k}\}$.

Our goal below is to show that every space has a generator matrix and a parity check matrix. The former is obvious from definitions. If $V \subseteq \mathbb{F}^n$ is a k-dimensional vector space, then it has a basis $\mathbf{v}_1, \ldots, \mathbf{v}_k$ and if we build a matrix G with these vectors as its rows, then G satisfies the conditions of the generator matrix.

We sketch the idea for construction of a parity check matrix. We say that a $k \times k$ matrix R forms a row operation if $R_{ii} = 1$, and $R_{ij} = 0$ for all must at most one pair $i \neq j \in [k]$. We say that $\tilde{G}$ is obtained from G by row operations, denoted $G \rightsquigarrow \tilde{G}$, if $\tilde{G} = R_m \cdot R_{m-1} \cdots R_1 \cdot G$ where the R_i's are row operations. Note that if G is a generator matrix for V then so is $\tilde{G}$. Gaussian elimination allows us to "simplify" G till its columns are special, and in particular after permuting the columns $\tilde{G}$ would look like $[I_k | A]$ where I_k denotes the $k \times k$ identity matrix. Assume for simplicity that $\tilde{G} = [I_k | A]$ (without permuting columns). Now let H be given by $H^T = [-A^T | I_{n-k}]$. It can be verified that $\tilde{G} \cdot H = 0$ and so $G \cdot H = 0$. Furthermore all columns of H are linearly independent and so H satisfies the conditions of the parity check matrix of V. We conclude with the following.

Proposition D.4.2. *If* $V \subseteq \mathbb{F}^n$ *is a* k*-dimensional vector space then it has a generator matrix* $G \in \mathbb{F}^{k \times n}$ *and a parity check matrix* $H \in \mathbb{F}^{n \times (n-k)}$*. Furthermore its dual* $V^\perp$ *is generated by* H^T*, has dimension* $n - k$*, and has* G^T *as its parity check matrix. Finally* $(V^\perp)^\perp = V$*.*

Before concluding we mention one important difference from the case of *orthoganility* of real vectors. For vector spaces over finite fields it is possible that there are non-zero vectors in $V \cap V^\perp$ and indeed even have $V = V^\perp$. (See Exercise 2.29 for more on this.)

D.5 Finite Fields

In this section we describe the existence and uniqueness of finite fields. We also describe the basic maps going from prime fields to extensions and vice versa. Parts of this section will repeat material from Section 2.1 and 5.1.

D.5.1 Prime Fields

We start by describing a field of size p, for any prime number p. Let $\mathbb{Z}_p$ be the set of integers $\{0,\dots,p-1\}$. For integer a and postive integer b, let $a \mod b$ denote the unique integer c in $\mathbb{Z}_p$ such that b divides $a-c$. Let $+_p$ be the binary operation on $\mathbb{Z}_p$ that maps a and b to $(a+b) \mod p$. Let $\cdot_p$ map a and b to $(ab) \mod p$. We have the following (see Section 2.1 for a proof).

Proposition D.5.1. $(\mathbb{Z}_p, +_p, \cdot_p)$ *form a field of cardinality* p.

Given a finite field $\mathbb{F}$, its characteristic, denoted $\mathrm{char}(\mathbb{F})$, is the smallest positive integer p such that $p \cdot 1 = 1+1+\cdots+1 = 0$. (See Exercise D.2 for why such a finite characteristic exists.)

Proposition D.5.2. *For every finite field* $\mathbb{F}$, $\mathrm{char}(\mathbb{F})$ *is a prime. Furthermore,* $\mathbb{F}$ *is a* $\mathbb{Z}_p$*-vector space, where* $p = \mathrm{char}(\mathbb{F})$. *Thus* $\mathbb{F}$ *has cardinality* p^n *for prime* p *and integer* n.

Proof. Let $p = \mathrm{char}(\mathbb{F})$. We first note that p is the smallest integer such that $p \cdot a = 0$ for any non-zero element of $\mathbb{F}$. This is so since $p \cdot a = p \cdot 1 \cdot a = 0$, and if $p \cdot a = 0$ then so is $p \cdot a \cdot a^{-1} = p \cdot 1$. Next we note that if $p = qr$ then for the element $w = q \cdot 1 \in \mathbb{F}$, we have $w \cdot r = 0$ which contradicts the minimality of p.

Next we note that $(\mathbb{F}, +, \circ)$ satisfy the conditions of a $\mathbb{Z}_p$-vector space where $i \circ a = (a + \cdots + a)$ (i times), for $i \in \mathbb{Z}_p$ and $a \in \mathbb{F}$. We conclude that $|\mathbb{F}| = p^n$ where n is the dimension of the vector space $(\mathbb{F}, +, \circ)$ and $p = \mathrm{char}(\mathbb{F})$. □

We conclude now by claiming that $\mathbb{Z}_p$ is the unique field of cardinality p.

Proposition D.5.3. *For any prime* p*, there is a unique field of cardinality* p *upto isomorphism.*

Proof. Let $\mathbb{F}$ be a field of cardinality p. By Proposition D.5.2 we have that $\mathrm{char}(\mathbb{F}) = p$. It can be verified that the map $1_{\mathbb{F}} \to 1$ extends to an isomorphism (see Exercise D.3). □

The uniqueness of the field of cardinality p allows us to call it $\mathbb{F}_p$ in the future.

D.5.2 Extension fields and subfields

We now move towards determining when non-prime fields exists. While the answer is simple (they exist for every number of the form p^n for prime p and positive integer n), proving when they exist requires some structural understanding of how fields behave.

We will first present a basic property of all finite fields that is crucial when working with fields.

We recall a basic result about finite groups (see Exercise D.4 for a proof for the abelian case).

Proposition D.5.4. *If* $(G, \cdot)$ *is a finite group with identity* 1*, then for every* $a \in G$*, we have* $a^{|G|} = 1$.

Proposition D.5.5. *Let* $\mathbb{F}$ *be a field of cardinality* q. *The every element* $\alpha \in \mathbb{F}$ *is a root of the polynomial* $X^q - X$ *and so* $X^q - X = \prod_{\alpha \in \mathbb{F}} (X - \alpha)$.

Proof. If $\alpha = 0$, then it is trivial to see that is a root of $X^q - X$. If $\alpha \neq 0$, then it is a member of a group $(\mathbb{F} \setminus \{0\}, \cdot)$ and so by Proposition D.5.4, satisfies $\alpha^{|\mathbb{F}\setminus\{0\}|} = 1$. Thus, $\alpha^{q-1} = 1$, and finally $\alpha^q = \alpha$, as desired. □

Let $\mathcal{K}$ be a field and $\mathbb{F} \subseteq \mathcal{K}$ be a set that is closed under addition and multiplication. Then $\mathbb{F}$ is itself a field and we denote if $\mathbb{F} \lhd \mathcal{K}$ to denote that it is a subfield of $\mathcal{K}$. We say $\mathcal{K} \rhd \mathbb{F}$ to denote that $\mathcal{K}$ extends $\mathbb{F}$.

Proposition D.5.6. *If $\mathcal{K} \rhd \mathbb{F}$ then $\mathcal{K}$ is an $\mathbb{F}$-vector space and so $|\mathcal{K}| = |\mathbb{F}|^n$ where n is the dimension of $\mathcal{K}$ as an $\mathbb{F}$-vector space. Furthermore there is a unique copy of $\mathbb{F}$ in $\mathcal{K}$.*

Proof. The fact that $\mathcal{K}$ is a vector space follows from the definitions, and thus the claim about its cardinality. The fact that there is a unique copy of $\mathbb{F}$ follows from the fact that the elements of $\mathbb{F}$ satisfy $X^q - X = 0$, where $q = |\mathbb{F}|$ and there can be at most q roots of this polynomial. □

D.5.3 Existence of Finite Fields

In what follows we will rely heavily on the modular reduction of polynomials. Following the notation of previous sections, for field $\mathbb{F}$ and $f, g \in \mathbb{F}[X]$, we let $f \mod g$ be the remainder when f is divided by g - so $\deg(f \mod g) < \deg(g)$ and g divides $f - (f \mod g)$. Let $f +_g h = (f + h) \mod g$ and let $f \cdot_g h = (fh) \mod g$. Recall that an irreducible polynomial in $\mathbb{F}_q[X]$ is one that does not have any non-trivial factor (recall Definition 5.1.4).

Proposition D.5.7. *Let $\mathbb{F}$ be a finite field of cardinality q and let $g \in \mathbb{F}[X]$ be an irreducible polynomial of degree n. Then $(\mathbb{F}[X]/g, +_g, \cdot_g)$ form a field of cardinality q^n.*

Essentially all fields can be obtained in the above manner, but to prove this fact, we need to prove that there is an irreducible polynomial of degree n over $\mathbb{F}_p$ for every p and unfortunately this proof is not much simpler than proving the existence of a field of cardinality p^n. So we prove the existence directly, or rather sketch a proof of this fact.

The rough idea of the proof is as follows: First we establish that every polynomial $f \in \mathbb{F}[X]$ splits completely (into linear factors) over some extension $\mathcal{K}$ of $\mathbb{F}$. To do this we work slowly, working away at one irreducible factor of f at a time. If g is such an irreducible factor, we consider the field[2] $\mathbb{L} = \mathbb{F}[Z]/g(Z)$ and note that Z is a root of g,[3] and hence of f, in $\mathbb{L}$ and so f splits more in $\mathbb{L}$. We continue this process till f splits completely in some field $\mathcal{K}$.

Now we work with a very special polynomial f, namely $f(X) = X^{p^n} - X$ in the ring $\mathbb{F}_p[X]$ and let $\mathcal{K}$ be a field in which f splits completely. Now let $S \subseteq \mathcal{K}$ be the set $S = \{\alpha \in \mathcal{K} | f(\alpha) = 0\}$. We note that this set, miraculously, is closed under addition and multiplication. The latter is easy: $f(\alpha) = 0$ if and only if $\alpha^{p^n} = \alpha$. So if $f(\alpha) = f(\beta) = 0$ then $\alpha^{p^n} = \alpha$ and $\beta^{p^n} = \beta$ and so $(\alpha\beta)^{p^n} = \alpha^{p^n}\beta^{p^n} = \alpha\beta$ and so $\alpha\beta \in S$. For the former we explicitly highlight another crucial fact in finite fields.

[2] Recall Theorem 5.1.1.

[3] This is because $g(Z) \equiv 0$ in $\mathbb{L}$.

Proposition D.5.8. *Let $\mathcal{K}$ be a field of characteristic p and let $A, B \in \mathcal{K}[X, Y]$. Then for all positive integers n we have $(A+B)^{p^n} = A^{p^n} + B^{p^n}$.*

The proof of the lemma above follows immediately from the fact that $\binom{p}{i} \mod p$ is 0 unless p divides i (see Exercise D.5). And while the lemma is stated for very general A and B, we only need it for $A, B \in \mathcal{K}$ itself. However we state it generally since it is fundamental to working over extension fields and indeed we will see a few applications later.

Returning to our quest to prove that S is closed under addition, let us apply the above proposition to $\alpha, \beta \in S$. We get that $(\alpha+\beta)^{p^n} = \alpha^{p^n} + \beta^{p^n} = \alpha + \beta$ and so S is closed under addition as well. What we will show next is that S has exactly p^n elements and so is a field of size p^n (it is closed under addition and multiplication and the rest of the properties follow from the fact that S is a subseteq of a field $\mathcal{K}$).

First note that S has all roots of f. We note further that f has no multiple roots. In general this is proved by looking at derivatives etc., but in this case we can do it by inspection. We wish to show that $(X-\alpha)^2$ does not divide $X^{p^n} - X$, but this is the same as showing that Z^2 does not divide $(Z+\alpha)^{p^n} - (Z+\alpha) = Z^{p^n} - Z + \alpha^{p^n} - \alpha$, but the latter polynomial has a coefficient of $-1 \neq 0$ for Z and so is not divisible by Z^2. We conclude that since S has all roots of $X^{p^n} - X$ and this polynomial has p^n distinct roots, and so $|S| \geq p^n$. On the other hand since every element of S is a root of $X^{p^n} - X$ and this polynomial has at most p^n roots, we conclude that $|S| = p^n$ and so there exists a field of cardinality p^n. Thus we get the following theorem (the first part follows from Proposition D.5.2 and the second part follows from Proposition D.5.7).

Theorem D.5.9. *If $\mathbb{F}$ is a finite field, then it has charatestic p for some prime p and its cardinality is p^n for positive integer n. Conversely, for every prime p and positive integer n, there is a field of cardinality p^n.*

D.5.4 Uniqueness of finite fields

We start by proving that every finite field has a multiplicative generator. To do so we need to understand cyclic groups a bit better.

The cyclic group of order n is the group $\mathbb{Z}_n = \{0, \ldots, n-1\}$ with addition modulo n being the binary operation. This group clearly has an element of order n (namely the number 1). Let $N^=(G, m)$ denote the number of elements of order exactly m in G and let $N(G, m)$ denote the number of elements of order dividing m in G. We have $N(G.m) = \sum_{k|m} N^=(G, k)$. For the cyclic group, we have for every $k|n$, $N(\mathbb{Z}_n, k) = k$ and $N^=(\mathbb{Z}_n, k) \geq 0$. (The latter is trivial and for the former see Exercise D.6.)

We now turn to understanding $(\mathbb{F}^*, \cdot)$ the group of non-zero elements of $\mathbb{F}$ under multiplication.

Lemma D.5.10. *Let $q = |\mathbb{F}|$ and $n = q-1$. We claim that for every k dividing n, $N(\mathbb{F}^*, k) = N(\mathbb{Z}_n, k)$ and $N^=(\mathbb{F}^*, k) = N^=(\mathbb{Z}_n, k)$.*

Proof. The claim is straightforward for $N(\mathbb{F}^*, k)$. We have that every $\alpha \in \mathbb{F}^*$ is a root of of the

polynomial $X^n - 1$ and since $X^k - 1$ divides[4] $X^n - 1$, k elements of $\mathbb{F}^*$ must be roots of this polynomial also. We thus have $N(\mathbb{F}^*, k) = k = N(\mathbb{Z}_n, k)$.

For the claim about $N^=(\mathbb{F}^*, k)$, we use induction and the inductive formula. We have $\sum_{\ell|k} N^=(\mathbb{F}^*, \ell) = N(\mathbb{F}^*, k) = k = N(\mathbb{Z}_n, k) = \sum_{\ell|k} N^=(\mathbb{Z}_n, \ell)$. But since by induction we have $N^=(\mathbb{F}^*, \ell) = N^=(\mathbb{Z}_n, \ell)$ for $\ell < k$, we may conclude that the remaining term $N^=(\mathbb{F}^*, k) = N^=(\mathbb{Z}_n, k)$. □

We say that an element $\omega \in \mathbb{F}$ is primitive if $\omega^i \neq 1$ for $i < |\mathbb{F}|-1$ and $\omega^{|\mathbb{F}|-1} = 1$. Since $N^=(\mathbb{F}^*, n)$ counts the number of primitive elements, Lemma D.5.10 implies that the number of primitive elements is at least one. Indeed, if p is the smallest prime divisor of n, then we have that $N^=(\mathbb{F}^*, n) = N(\mathbb{F}^*, n) - N(\mathbb{F}^*, n/p) - N(\mathbb{F}^*, p) = n - n/p - p > 0$, assuming $p < n/p$. Otherwise if $n = p^2$, then we have $N^=(\mathbb{F}^*, n) = N(\mathbb{F}^*, n) - N(\mathbb{F}^*, p) = n - p > 0$. If n itself is a prime then we have $N^=(\mathbb{F}^*, n) = N(\mathbb{F}^*, n) = n > 0$.

Proposition D.5.11. *Every finite field $\mathbb{F}$ has a primitive element. Consequently the multiplicative group is cyclic.*

We now describe a weaker form of special element in $\mathbb{F}$. Let $\mathcal{K}$ extend $\mathbb{F}$. We say that $\alpha \in \mathcal{K}$ is an $\mathbb{F}$-generator for $\mathcal{K}$ if for every element $\beta \in \mathcal{K}$ there is a polynomial $p \in \mathbb{F}[X]$ such that $\beta = p(\alpha)$.

Proposition D.5.12. *Let $\mathcal{K}$ be a finite field and let ω be a primitive element in $\mathcal{K}$. Then for every subfield $\mathbb{F} \lhd \mathcal{K}$ we have that ω is an $\mathbb{F}$-generator of $\mathcal{K}$. As a consequence, for every $\mathcal{K} \rhd \mathbb{F}$ there is an $\mathbb{F}$-generator in $\mathcal{K}$.*

Proof. Consider the lowest degree polynomial $p \in \mathbb{F}[X]$ such that $p(\omega) = 0$. Let $|\mathbb{F}| = q$ and $|\mathcal{K}| = q^n$.

We claim that $\deg(p) = n$. If $\deg(p) > n$, we have that $1, \omega, \omega^2, \ldots, \omega^n$ are linearly independent over $\mathbb{F}$ and so $\mathcal{K}$ has size strictly larger than q^n. Now if $\deg(p) < n$, then consider the polynomials $X, X^2, X^3, \ldots, X^{q^n-1}$ modulo $p \in \mathbb{F}[X]$. Since we have only $q^{\deg(p)}$ options for the residues, two of these must be equal modulo p and so there exist $i \neq j$ and $f \in \mathbb{F}[X]$ such that $X^i = X^j + p \cdot f$. Substituting $X = \omega$ yields $\omega^i = \omega^j + p(\omega) f(\omega) = \omega^j$. But this contradicts the assumption that ω is a primitive element.

Finally, note that any element $\beta \in \mathcal{K}$ can be written as the polynomial $q^j \mod p(X)$ evaluated at $X = \omega$ for some $0 \leq j < q^n$. □

Generators are useful in that they show that the only way to construct field extensions is via irreducible polynomials.

Proposition D.5.13. *Let $\mathcal{K} \rhd \mathbb{F}$ and let α be an $\mathbb{F}$-generator of $\mathcal{K}$. Then, if p is the minimal polynomial in $\mathbb{F}[X]$ such that $p(\alpha) = 0$, we have p is irreducible and $\mathcal{K}$ is isomorphic to $\mathbb{F}[X]/p$.*

Proof. Irreducibility of p follows from its minimality (see Exercise D.8). The isomorphism is obtained by fixing $\mathbb{F} \lhd \mathcal{K}$ and letting $\alpha \mapsto X$. We leave it to the reader to verify that this extends to an isomorphism (uniquely)– see Exercise D.9. □

[4] See Exercise D.7.

We are almost ready to prove uniqueness of finite fields. We need one more fact about irreducible polynomials to do so.

Proposition D.5.14. *If $f \in \mathbb{F}_p[X]$ is irreducible of degree n, then f divides $X^{p^n} - X$.*

Proof. Consider the field $\mathcal{K} = \mathbb{F}_p[X]/(f)$. This is a field of cardinality p^n and so every element $\alpha \in \mathcal{K}$ satisfies $\alpha^{p^n} = \alpha$. In particular $X \in \mathcal{K}$ also satisfies this, implying that $X^{p^n} - X = 0($ mod $f)$ and so f divides $X^{p^n} - X$. □

We now turn to proving uniqueness of finite fields.

Theorem D.5.15. *For every prime p and integer n, there is a unique field of cardinality p^n up to isomorphism.*

Proof. Suppose $\mathcal{K}, \mathbb{L}$ are both fields of cardinality p^n. Both fields contain a unique copy of $\mathbb{F}_p$, and by mapping $1_{\mathcal{K}}$ to $1_{\mathbb{L}}$ and extending additively, we get a partial isomorphism between these copies of $\mathbb{F}_p$. Now we show how to extend it. Let $\alpha \in \mathcal{K}$ be a $\mathbb{F}_p$-generator and let $f \in \mathbb{F}_p[X]$ be its minimal polynomial. Since f is irreducible of degree n, we have that $f(X)$ divides the polynomial $X^{p^n} - X$ (see Proposition D.5.14).

Using the fact that $X^{p^n} - X = \prod_{\beta \in \mathbb{L}}(X - \beta)$, we conclude that $\mathbb{L}$ contains a root β of f. We assert (see Exercise D.10) that the map that sends $\alpha \mapsto \beta$ is an isomorphism from $\mathcal{K}$ to $\mathbb{L}$. □

D.5.5 The Trace and Norm maps

We conclude this section with two basic polynomials that have some very nice regularity property when dealing with finite fields and their extensions.

Definition D.5.1. Let $\mathbb{F} = \mathbb{F}_q$ and let $\mathcal{K} = \mathbb{F}_{q^n}$. Then the Trace function $\mathrm{Tr} = \mathrm{Tr}_{\mathcal{K} \to \mathbb{F}}$ is the function obtained by the evaluation of the polynomial $\mathrm{Tr}(X) = X + X^q + X^{q^2} + \cdots + X^{q^{n-1}}$. The Norm function is obtained by evaluation of $N(X) = X^{1+q+q^2+\cdots+q^{n-1}}$.

The Norm and Trace functions are important because they map elements of $\mathcal{K}$ to the subfield $\mathbb{F}$ and they do so in a nice uniform way. We mention some properties below.

Proposition D.5.16.

1. *Trace is a $\mathbb{F}$-linear map, i.e., for every $\alpha \in \mathbb{F}$ and $\beta, \gamma \in \mathcal{K}$, we have* $\mathrm{Tr}(\alpha \cdot \beta + \gamma) = \alpha \cdot \mathrm{Tr}(\beta) + \mathrm{Tr}(\gamma)$.
2. *Norm is multiplicative, i.e., $N(\beta \cdot \gamma) = N(\beta)N(\gamma)$*
3. *Trace is a q^{n-1}-to-one map from $\mathcal{K}$ to $\mathbb{F}$.*
4. *Norm is a $(q^n - 1)/(q - 1)$-to-one map from $\mathcal{K}^*$ to $\mathbb{F}^*$.*

Proof.

1. The $\mathbb{F}$-linearity follows from the facts that $(\alpha\beta + \gamma)^{q^i} = \alpha^{q^i}\beta^{q^i} + \gamma^{q^i}$ and $\alpha^{q^i} = \alpha$.
2. The multiplicativity is obvious from definition.

3. For $\beta \in \mathcal{K}$ we have $\mathrm{Tr}(\beta)^q = \beta^q + \cdots \beta^{q^n} = \beta^q + \cdots + \beta^{q^{n-1}} + 1 = \mathrm{Tr}(\beta)$ and so the range of Tr is $\mathbb{F}$. Since Tr is a polynomial of degree q^{n-1} is can take on any value in the range at most q^{n-1} times. But it has a domain of size q^n and range of size q, so it must take on every value exactly q^{n-1} times.

4. The is similar to Part (3) above. By Exercise D.11, we have that $N(\beta)^q = N(\beta)$ and furthermore note that by definition, $N(\beta)$ is non-zero iff $\beta \neq 0$. We then use the degree of N and counting to determine that it is a regular function on non-zero values.

□

The Trace function from $\mathcal{K} \to \mathbb{F}$ is especially important since it captures all $\mathbb{F}$-linear maps from $\mathcal{K} \to \mathbb{F}$, as explained below.

Proposition D.5.17. *A function $L : \mathcal{K} \to \mathbb{F}$ is $\mathbb{F}$-linear if and only if there exists $\lambda \in \mathcal{K}$ such that $L(\beta) = \mathrm{Tr}(\lambda\beta)$ for every $\beta \in \mathcal{K}$.*

Proof. First note that $f(\beta) = \mathrm{Tr}(\lambda\beta)$ is obviously $\mathbb{F}$-linear since

$$f(\alpha\beta + \gamma) = \mathrm{Tr}(\lambda(\alpha\beta + \gamma)) = \mathrm{Tr}(\lambda\alpha\beta) + \mathrm{Tr}(\lambda\gamma) = \alpha\,\mathrm{Tr}(\lambda\beta) + \mathrm{Tr}(\lambda\gamma) = \alpha f(\beta) + f(\gamma)$$

for every $\alpha \in \mathbb{F}$ and $\beta, \gamma \in \mathcal{K}$ (where in the above we used Propositon D.5.16). This concludes one direction of the proposition.

To see the converse we employ a *counting argument*. First note that if $\lambda \neq 0$ then the function $f_\lambda(\beta) = \mathrm{Tr}(\lambda\beta)$ is not identically zero. (To see this, note that $f_\lambda(Z)$, viewed as a polynomial in Z has degree $|\mathcal{K}|/|\mathbb{F}|$ and it is a non-zero polynomial since the coefficient of Z is non-zero.) By linearity this implies that $f_\lambda \neq f_\tau$ if $\lambda \neq \tau$ since $f_\lambda - f_\tau = f_{\lambda-\tau} \neq 0$. So, including $\lambda = 0$, we have at least $|\mathcal{K}|$ distinct linear functions of the form $f_\lambda(\cdot)$. We now note there are also at most $|\mathcal{K}|$ such functions. To see this let $\beta_1, \ldots, \beta_n \in \mathcal{K}$ be $\mathbb{F}$-linearly independent elements of $\mathcal{K}$ (i.e., $\sum_{i=1}^n \alpha_i\beta_i \neq 0$ if $(\alpha_1, \ldots, \alpha_n) \in \mathbb{F}^n \setminus \{0\}$). Since $\mathcal{K}$ is a degree n extension of $\mathbb{F}$ we know such a sequence exists and furthermore the β_i's generate $\mathcal{K}$ in that for every $\beta \in \mathcal{K}$ there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$ such that $\beta = \sum_i \alpha_i\beta_i$. We note that a linear function $L : \mathcal{K} \to \mathbb{F}$ is completely determined by its values at $\beta_1, \ldots, \beta_n$, since for every $\beta = \sum_i \alpha_i\beta_i \in \mathcal{K}$ with $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$, we have $L(\beta) = L(\sum_i \alpha_i\beta_i) = \sum_i \alpha_i L(\beta_i)$. Thus the number of linear functions is upper bounded by $|\{(L(\beta_1), \ldots, L(\beta_n)) \in \mathbb{F}^n\}| \leq |\mathbb{F}|^n = |\mathcal{K}|$. We conclude that these are exactly $|\mathcal{K}|$ functions that are $\mathbb{F}$-linear from $\mathcal{K} \to \mathbb{F}$, and these are exactly the Trace functions. □

D.6 Algorithmic aspects of Finite Fields

In this section we show how finite fields may be represented and field operations computed efficiently.

Let $q = p^t$ for prime p and positive integer t. We consider how to work with $\mathbb{F}_q$ — the field on q elements.

We start by noticing that if $O(q^2)$ space is not imposing, then four tables — one for addition, and one for multiplication, and one each for additive and multiplicative inverses would suffice

for working with fields, with each field operation now requiring a single table look up. In what follows we give more succinct descriptions that still allow moderately fast (some polynomial in $\log q$) operations.

D.6.1 Prime Fields

We start with the case of $t = 1$. Here there is not much to do. The most natural representation of the field is by specifying the prime p which takes $\log_2 p + 1 = \log q + 1$ bits to specify. The most complex part of addition and multiplication is the computation of the remainder of the operation modulo p, and this takes at most $O((\log p)^2)$ steps by the naive method. More sophisiticated algorithms can bring this compexity down to $O((\log p)(\log\log p)^2)$.

D.6.2 General fields as vectors

For general fields, we can adopt one of two approaches. The first of these uses less of the knowledge that we have about finite fields, but helps abstract away many issues. In this view we use the isomorphism between $\mathbb{F}_{p^t}$ and $\mathbb{F}_p^t$ to represent elements of the former as vectors in $\mathbb{F}_p^t$. This, thus represents elements of $\mathbb{F}_q$ by $O(\log q)$ bits which is nice. This also tells us how to add in $\mathbb{F}_q$ since it is simply coordinatewise $\mathbb{F}_p$-addition. However this representation by itself is not sufficient to do $\mathbb{F}_q$-multiplication. To multiply elements we enhance this representation by maintaining t^2 vectors $\mathbf{w}_{ij} \in \mathbb{F}_p^t$ with $\mathbf{w}_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$ where the $\mathbf{e}_i$'s are the unit vectors (so $\mathbf{e}_i = (0,\ldots,0,1,0,\ldots,0)$ is 1 in the ith coordinate and zero elsewhere). Now given $\mathbf{u} = (u_1,\ldots,u_t)$ and $\mathbf{v} = (v_1,\ldots,v_t)$ we can compute $\mathbf{u}\cdot\mathbf{v} = \sum_{i=1}^t \sum_{j=1}^t u_i v_j \mathbf{w}_{ij}$. This leads to a roughly $O(t^3(\log p)^2) = O((\log q)^3)$ time algorithm for multiplying in $\mathbb{F}_q$ while requiring $O(t^3 \log p)$ bits to store p and the vectors $\mathbf{w}_{ij}$. While not the most efficient representation, this may afford a clean representation that may be sufficient in several settings.

D.6.3 General fields as polynomial rings

Our final representation uses the fact that the field $\mathbb{F}_{p^t}$ is isomorphic to $\mathbb{F}_p[X]/(g)$ for any irreducible polynomial g of degree t. Here, a field element is simply a polynomial in $\mathbb{F}_p[X]$ of degree strictly less than t which is maintained as a vector of coefficients. Addition is just coordinate-wise addition, whereas multiplication is polynomial multiplication followed by a remainder computation modulo g. Thus addition takes time $O(t(\log p)^2)$ while mutliplication naively takes time $O(t^2(\log p)^2)$. The only ingredients that need to be remembered to do field operations are the integer p and the polynomial $g \in \mathbb{F}_p[X]$, all of which take $O(t \log p)$ bits. So this representation definitely outperforms the generic representation via vector spaces in almost all senses (though we might find the vector space view helpful when discussing certain operations with codes).

D.6.4 Finding primes and irreducible polynomials

The final question that remains to be discussed is how hard is to find the ingredients that describe a field. Of course, this depends on how the field is described, and the most natural one may be by giving the cardinality q of the field.

Given $q = p^t$, it is straightforward to enumerate all candidate (p, t) pairs such that $q = p^t$ — there are only $\log q$ possible values of t and thus $\log q$ such integers. Only one of these, the one with the largest t could correspond to prime p. Testing if an integer is prime can be done efficiently with randomization, and thanks to a recent breakthrough [1] even deterministically in time polynomial in $\log q$.

When $t = 1$ no further works needs to be done. If $t > 1$ one needs to find an irreducible polynomial g of degree t and this can be a challenge. There are several possible solutions here:

Randomized It is known (and can actually be proved with a little effort, given the ingredients of this chapter) that a random polynomial $g \in \mathbb{F}_p[X]$ of degree t is irreducible with probability at least $1/t$. Furthermore, irreducibility can be tested (see next section) in time $\text{poly}(\log q)$. Thus repeatedly sampling random polynomials till an irreducible polynomial is found takes expected time $\text{poly}(\log q)$. (See Algorithm 8.)

Deterministic Shoup [90] gave an algorithm to deterministically find an irreducible polynomial of degree t in $\mathbb{F}_p[X]$ in time $\text{poly}(t, p)$. Notice that this dependence is slower than one may hope for in terms of p, but works well when p is small (say, smaller than t).

Explicit In some rare cases, i.e., a few choices of p and t, explicit polynomials are known that are irreducible. These may be used when the field size seems appropriate. One such family of irreducible polynomials is given in the following proposition.

Proposition D.6.1 ([69])**.** *Let $p = 2$ and $t = 2 \cdot 3^\ell$ for any non-negative integer ℓ. Then the polynomial $X^t + X^{t/2} + 1$ is irreducble in $\mathbb{F}_2[X]$.*

D.7 Algorithmic aspects of Polynomials

In this section we review basic facts about algorithmic aspects of manipulating polynomials. We start with basic tasks and move to more complex tasks ending with factoring and root-finding.

D.7.1 Adding, Multiplying, Dividing

Given two polynomials $f, g \in \mathbb{F}_q[X]$ of degree at most n, they can be added with $O(n)$ operations in $\mathbb{F}_q$ and no more needs to be said. f and g can also be multiplied with $O(n^2)$ operations by the standard *long multiplication*. Similarly the quotient and remainder obtained when dividing f by g can be computed wih $O(n^2)$ operations using the *long division* algorithm. More efficient algorithms do exist for both these tasks making $O(n(\log n)^c)$ field operations, for some constant c. (See [103] for this and other references for this section.)

D.7.2 Greatest Common Divisor

Perhaps the most surprising algorithm in algebra is that of finding greatest common divisors (of integers or polynomials), and would be even more so, if it were not for over 2000 years of exposure. To explain, let us look at the definition of the problem.

Definition D.7.1 (Greatest Common Divisor)**.** Given polynomials $f, g \in \mathbb{F}[X]$, their greatest common divisor, denoted $\gcd(f, g)$, is the maximal degree polynomial $h(X)$ with leading coefficient being 1 such that h divides f and g.

The natural algorithm for finding $\gcd(f, g)$ would be to factor f and g into irreducible factors, and then to take all common factors (with multiplicity) and take their product to get h. Unfortunately this reduces gcd to factoring which goes in the wrong direction. (As we will see below, factoring can also be solved efficiently for polynomials, but by reduction to gcd computation.)

But fortunately, we can employ Euclid's algorithm which uses the following algorithmic reduction: If $\deg(g) < \deg(f)$ and g does not divide f, then $\gcd(f, g) = \gcd(g, r)$ where $f = q \cdot g + r$ with $\deg(r) < \deg(g)$ is as given by the division algorithm. This simple fact turns out to be algorithmically effective reducing the (sum of the) degree of the polynomials in a single step of polynomial division, and thus leading to a polynomial time algorithm for finding the greatest common divisor.

Once again the steps of this algorithm can be combined in clever ways to get an implementation in $O(n(\log n)^c)$ time.

D.7.3 Factoring and Root-Finding

Finally, one of the most striking tasks related to polynomials that turns out to have a polynomial time algorithm is the factorization of polynomials. Polynomials, even multivariate ones, can be factored extremely efficiently with randomization and this is a consequence of many years of research in algebraic computing. We won't give the strongest results here, since even stating the result is non-trivial. For our purposes it will suffice to know that polynomials in $\mathbb{F}_q[X]$ of degree n can be factored in time $\mathrm{poly}(n, \log q)$. We state this general result, and prove a very special case of it, which will suffice for the algorithms in this book.

Theorem D.7.1. *There exists a constant c and a randomized algorithm running in expected time $O((n \log q)^c)$ that factors polynomials of degree n in $\mathbb{F}_q[X]$. Furthermore, if $q = p^t$ for prime t, then there is a deterministic algorithm with running time $O((npt)^c)$ for factoring.*

To give an idea behind this powerful algorithm, we consider a simple special case of root-finding.

Definition D.7.2 (Root-Finding Problem)**.** The input to the root finding problem is a polynomial $f \in \mathbb{F}_q[X]$ of degree at most n (given as a list of coefficients $f_0, \ldots, f_n \in \mathbb{F}_q$). The task is to find all $\alpha \in \mathbb{F}_q$ that are roots of f, i.e., to output the set $\{\alpha \in \mathbb{F}_q | f(\alpha) = 0\}$.

We now turn towards the root-finding algorithm. The algorithm relies crucially on the algorithm for computing greatest common divisors (mentioned in the previous section) and two additional facts. First we use the fact $X^q - X = \prod_{\alpha \in \mathbb{F}_q}(X - \alpha)$ to algorithmic advantage as follows.

Lemma D.7.2. *A polynomial $f \in \mathbb{F}_q[X]$ has a root in $\mathbb{F}_q$ if and only if* $\gcd(f, X^q - X) \neq 1$.

Proof. The proof is immediate. If f has a root α, then $X - \alpha$ divides $\gcd(f, X^q - X)$ and so their gcd can't be trivial. Conversely a factor of $X^q - X$ is of the form $\prod_{\alpha \in S}(X - \alpha)$ for some $S \subseteq \mathbb{F}_q$, and so $\gcd(f, X^q - X)$ must be of this form. If the gcd is non-trivial, then S must be non-empty implying that for every $\alpha \in S$ we have $X - \alpha$ divides f and thus f has a root in $S \subseteq \mathbb{F}_q$. □

The step above is almost algorithmic, but to verify this, we need to stress that the gcd of f and $X^q - X$ can be computed in time polynomial in $\deg(f)$ and $\log q$. We explain how this can be done in the next few paragraphs, taking a detour on sparse polynomials. But assuming this can be done, this provides a natural starting point for a root finding algorithm. Given f we compute $g = \gcd(f, X^q - X)$. If $g \neq 1$, then we take the set S_1 of roots of g and the set S_2 of roots of f/g and output $S_1 \cup S_2$. The set S_2 can be computed recursively (since f/g has smaller degree than f), but for S_1 we need some new ideas to determine how to compute the roots of g, when g splits into linear and distinct factors over $\mathbb{F}_q$. To get to this point we will use the fact that $X^q - X$ splits into some high-degree sparse factors and this will turn out to be crucial to finding S. Indeed the sparsity of $X^q - X$ and its factors are heavily used concepts and we now take a detour to explain these effects.

Sparse high degree polynomials

We start with some terminology. We say that a polynomial $h \in \mathbb{F}[X]$ is t-sparse if at most t of its coefficients are non-zero. Every polynomial h is thus $(\deg(h) + 1)$-sparse, but often the sparsity can be smaller, and this will be useful. One sparse polynomial that is already motivated by the earlier discussion is $X^q - X$, which is 2-sparse. We will see a few more below.

Lemma D.7.3. *Let $f \in \mathbb{F}[X]$ be a polynomial of degree n and let $h \in F[X]$ be a t-sparse polynomial of degree D. Then $h \mod f$ and* $\gcd(f, h)$ *can be computed in time* $\mathrm{poly}(n, t, \log D)$.

Proof. It obviously suffices to compute $h \mod f$ in time $\mathrm{poly}(n, t, \log D)$ and then one can use Euclid's algorithm to compute $\gcd(f, h) = \gcd(f, h \mod f)$ in time $\mathrm{poly}(n)$. In turns out that if $h = \sum_{i=1}^{t} h_i X^{d_i}$ and we can compute $h_i X^{d_i} \mod f$ in time $\mathrm{poly}(n, \log d_i)$ for every i, then we can add the results in time $\mathrm{poly}(n, t)$ to get $h \mod f$. Finally, we note that $X^d \mod f$ can be computed by repeated squaring. Let $d = \sum_{j=0}^{\log_2 d} d_j 2^j$. We can first compute the sequence of polynomials $g_j = X^{2^j} \mod f = g_{j-1}^2 \mod f$ by repeatedly squaring the output of the previous step. Then we can compute $X^d \mod f = \prod_{j=0}^{\log_2 d} (g_j)^{d_j}$ by using $\log d$ more multiplications, yielding the desired result. □

The lemma above shows that sparse polynomials can be used effectively. The following lemma shows that the central sparse polynomial also has sparse factorizations, and this will be useful later.

Proposition D.7.4. *1. Let $\mathbb{F}_q$ be a field of odd characteristic (and so q is odd). Then $X^q - X = X \cdot (X^{(q-1)/2} - 1) \cdot (X^{(q-1)/2} + 1)$. In particular $X^q - X$ factors into three 2-sparse polynomials of degree at most $q/2$.*

2. Let $q = 2^t$ for integer $t \geq 2$. Then $(X^q - X) = \mathrm{Tr}(X) \cdot (\mathrm{Tr}(X) - 1)$ where $\mathrm{Tr}(X) = Tr_{\mathbb{F}_q \to \mathbb{F}_2}(X) = X + X^2 + X^4 + \cdots + X^{2^{r-1}}$ is the Trace map from $\mathbb{F}_q$ to $\mathbb{F}_2$. In particular $X^q - X$ factors into two $(2 + \log_2 q)$-sparse polynomials of degree $q/2$.

Proof. The case of odd q is obvious by inspection. Only aspect to be stressed is that $(q-1)/2$ is an integer.

For the case of even q, we use the fact that the trace map is a map from $\mathbb{F}_q$ to $\mathbb{F}_2$. So every $\alpha \in \mathbb{F}_q$ satisfies $\mathrm{Tr}(\alpha) = 0$ or $\mathrm{Tr}(\alpha) = 1$. It follows that $X - \alpha$ divides $\mathrm{Tr}(X) \cdot (\mathrm{Tr}(X) - 1)$ for every α. Consequently $X^q - X$ divides $\mathrm{Tr}(X) \cdot (\mathrm{Tr}(X) - 1)$. The identity $X^q - X = \mathrm{Tr}(X) \cdot (\mathrm{Tr}(X) - 1)$ now follows from the fact that both polynomials have the same degree and have leading coefficient 1. □

The existence of such sparse polynomials with many roots is one of the remarkable aspects of finite fields and leads to many algorithmic effects. We demonstrate this by showing how this is utilized in root-finding.

Root finding algorithm

We now complete the root-finding algorithm. Recall that by letting $g = \gcd(f, X^q - X)$ we can reduce to the case of polynomials that split into distinct linear factors in $\mathbb{F}_q$. We now focus on this case. We will also focus on the case of odd q for simplicity, though all we will use is the fact that $X^q - X$ splits into sparse factors of degree at most $q/2$.

If we were lucky, then g would have two roots α and β with $X - \alpha$ dividing $(X^{(q-1)/2} - 1)$ and $X - \beta$ not dividing it. Then we would have that $g_1 = \gcd(g, X^{(q-1)/2} - 1)$ would be a non-trivial factor of g and we could recurse on g_1 and $g_2 = g/g_1$. The key to the randomized root-finding is that by an appropriate affine change of variables, we can try to arrange to be "lucky".

Specifically, fix $a \in \mathbb{F}_q^*$ and $b \in \mathbb{F}_q$ and let $g_{a,b}(X) = g((X - a)/b)$. We have the following proposition.

Proposition D.7.5. *Let $g \in \mathbb{F}_q[X]$ have $\alpha \neq \beta$ as its roots. Then we have:*

1. *The coefficients of $g_{a,b}$ can be computed efficiently given a, b and the coefficients of g.*

2. *$g_{a,b}$ has $a\alpha + b$ and $a\beta + b$ as its roots.*

3. *If $a \in \mathbb{F}_q^*$ and $b \in \mathbb{F}_q$ are chosen uniformly at random independently, then the probability that exactly one of $a\alpha + b$ and $a\beta + b$ is a root of $X^{(q-1)/2-1}$ is at least $1/2$.*

Proof. Parts (1) and (2) are straightforward to verify. For part (3) we note that for any pair of distinct elements $\gamma, \delta \in \mathbb{F}_q$ there is exactly one pair $a \in \mathbb{F}_q^*$ and $b \in \mathbb{F}_q$ such that $a\alpha + b = \gamma$ and $a\beta + b = \delta$. Since the fraction of distinct pairs $\gamma, \delta \in \mathbb{F}_q$ such that exacty one of them comes

from a set of size $(q-1)/2$ (the set of roots of $X^{(q-1)/2}-1$) is at least $1/2$ (the exact formula is $1/2+1/(2q)$) we have that the probability that exactly one of $a\alpha+b$ and $a\beta+b$ is a root of $X^{(q-1)/2-1}$ is at least $1/2$. □

We conclude by giving the full root-finding algorithm and summary of analysis of its runtime.

Algorithm 46 ROOT-FIND($\mathbb{F}_q, f$)

INPUT: $\mathbb{F}_q, f(X) \in \mathbb{F}_q[X]$
OUTPUT: $\mathbb{F}_q$ roots of $f(X)$

```
g ← gcd(f, X^q − X)
IF g = 1 THEN
    RETURN ∅
RETURN LINEAR-ROOT-FIND(F_q, g) ∪ ROOT-FIND(F_q, (f/g))
```

Algorithm 47 LINEAR-ROOT-FIND($\mathbb{F}_q, g$)

INPUT: $\mathbb{F}_q, g(X) \in \mathbb{F}_q(X)$
OUTPUT: $\mathbb{F}_q$ roots of $g(X)$ if $g(X)$ divides $X^q - X$

```
IF deg(g) = 1 THEN
    RETURN {α} where g = X − α
REPEAT
    Pick a ∈ F_q^* and b ∈ F_q uniformly independently
    g_{a,b} ← g((X − b)/a)
    h_1 ← gcd(g_{a,b}, X^{(q−1)/2−1})
    g_1 ← h_1(aX + b)
UNTIL 0 < deg(g_1) < deg(g)
RETURN LINEAR-ROOT-FIND(F_q, g_1) ∪ LINEAR-ROOT-FIND(F_q, (g/g_1))
```

Lemma D.7.6. ROOT-FIND($\mathbb{F}_q, f$) *outputs the multiset of roots of f in expected time* $\mathrm{poly}(n, \log q)$.

Proof. Let $n = \deg(f)$. It is straightforward to see that ROOT-FIND makes at most n calls to LINEAR-ROOT-FIND. (This is a very weak estimate, but we leave out optimizations here, in favor of simplicity.) By Proposition D.7.5, Part (3), we have that the loop in LINEAR-ROOT-FIND will be executed an expected constant number of times before a non-trivial split is found. Finally, the degrees of the polynomials in the two recursive calls add up to $\deg(g)$ and so this leads to a tree of recurvie calls of size at most n with each internal node and leaf performing $\mathrm{poly}(n, \log q)$ work (to compute the various gcds, and the transformation of variables). Thus the overall expected running time is $\mathrm{poly}(n, \log q)$. □

D.8 Exercises

Exercise D.1. Let R be a commutative ring. Then prove the following:

1. If a is a unit in R, then $b \cdot a = 0$ if and only if $b = 0$.
2. Using the previous part or otherwise, prove Proposition D.3.4.

Exercise D.2. Argue that every finite field $\mathbb{F}$ has a finite characteristic $\text{char}(\mathbb{F})$.

Exercise D.3. Let $\mathbb{F}$ be a field with p elements for prime p. Argue that the map $1_{\mathbb{F}} \to 1$ can be extended to an isomorphism between $\mathbb{F}$ and $\mathbb{Z}_p$.

Exercise D.4. Let G be an abelian group with identity 1 and let $a \in G$.

1. Argue that the map $x \to a \cdot x$ for $x \in G$ is a bijection.
2. Argue that
$$\prod_{x \in G} x = a^n \cdot \prod_{x \in G} x.$$
3. Using the previous part or otherwise prove Proposition D.5.4 for abelian groups.

Exercise D.5. Let p be a prime and let $0 \le i \le p$. The show that

$$\binom{p}{i} \mod p = \begin{cases} 1 & \text{if } i = p \text{ or } i = 0 \\ 0 & \text{otherwise} \end{cases}.$$

Exercise D.6. In this exercise we will argue that for every k that divides n, we have $N(\mathbb{Z}_n, k) = k$, i.e. show that the number of elements of $\mathbb{Z}_n$ that have an order that divides k is exactly k. Consider the following:

1. Prove that
$$S_k = \left\{ a \cdot \frac{n}{k} | 0 \le a < k \right\}$$
is a sub-group of $\mathbb{Z}_n$.
2. Argue that any $b \in \mathbb{Z}_n$ that has an order that divides k satisfies, $k \cdot b \mod n = 0$.
3. Argue that any $b \in \mathbb{Z}_n \setminus S_k$ it must be the case that $k \cdot b \mod n \neq 0$.
4. Argue that any $b \in S_k$ has an order that divides k.
5. Using the above parts or otherwise, argue that S_k contains all elements of $\mathbb{Z}_n$ with an order that divides k. Conclude that $N(\mathbb{Z}_n, k) = k$.

Exercise D.7. If k divides n, then show that $X^k - 1$ divides $X^n - 1$.

Exercise D.8. Let $\mathscr{K} \rhd \mathbb{F}$ and let α be an $\mathbb{F}$-generator of $\mathscr{K}$. Let p be the minimal polynomial in $\mathbb{F}[X]$ such that $p(\alpha) = 0$. Argue that have p is irreducible.

Exercise D.9. Let $\mathscr{K} \rhd \mathbb{F}$ and let α be an $\mathbb{F}$-generator of $\mathscr{K}$. Let p be the minimal polynomial in $\mathbb{F}[X]$ such that $p(\alpha) = 0$. Argue that there is an isomorphism between $\mathscr{K}$ and $\mathbb{F}[x]/p$ obtained by fixing $\mathbb{F} \lhd \mathscr{K}$ and letting $\alpha \mapsto X$, which can be extended to all other elements.

Exercise D.10. Using notation in proof of Theorem D.5.15, prove that the map $\alpha \mapsto \beta$ can be extended to an isomorphism between $\mathscr{K}$ and $\mathbb{L}$.

Exercise D.11. Argue that for any $\beta \in \mathbb{F}_{q^n}$, the norm function satisfies $N(\beta)^q = N(\beta)$.

Appendix E

Some Information Theory Essentials

We used the notions of Shannon entropy and conditional entropy in Chapter 6. This appendix collects some basic facts relating to entropy and mutual information as a quick refresher. It is very elementary and can be skipped by readers with basic familiarity with information theory.

E.1 Entropy

Let X be a discrete random variable, say,

$$X \sim \begin{cases} a & \frac{2}{3} \\ b & \frac{1}{6} \\ c & \frac{1}{6} \end{cases}$$

and define $H(X)$ to be the entropy of the random variable X. We'd like the entropy to measure the *amount of information conveyed* by the value of the random variable X, where the amount of information is, roughly speaking, the amount of surprise we get from this random variable.

Suppose we had such a function $S : [0,1] \to \mathbb{R}^+$ that takes a probability and maps it to a non-negative real. Let's think about some natural properties of S.

Suppose we told you that X is a, b, or c. Are you surprised? No—because this occurs with probability 1 and we already know this. So therefore we'd like $S(1) = 0$: flipping a two-headed coin doesn't give us any surprise.

Now suppose we told you that $X = a$. This is a little bit surprising, but $X = b$ is more surprising because this is a rarer event. So we'd like S to satisfy $S(p) > S(q)$ if $p < q$.

We'd also like $S(x)$ to be a continuous function of x, because we'd like the surprise to not exhibit jumps (if we had instead

$$X \sim \begin{cases} a & \frac{2}{3} + 10^{-6} \\ b & \frac{1}{6} - 10^{-6} \\ c & \frac{1}{6} \end{cases}$$

our impression of the "surprise" of a should not be much different than the original X).

Now suppose X_1 and X_2 are two independent instantiations of X. It's natural for our surprise from $X_1 = a$ and $X_2 = b$ to be additive (as each event "adds" to our surprise):

$$S\left(\frac{2}{3}\cdot\frac{1}{3}\right) = S\left(\frac{2}{3}\right) + S\left(\frac{1}{3}\right)$$

which means we must have $S(pq) = S(p) + S(q)$.

Exercise E.1.1. *The only S that satisfies the above requirements is $S(p) = c\log_2\left(\frac{1}{p}\right), c > 0$. A convenient normalization constant is that we are unit surprised by a fair coin flip ($S(1/2) = 1$), which sets $c = 1$.*

Definition E.1.1 (Entropy)**.** The entropy of a discrete random variable X, denoted $H(X)$, is the average surprise for an outcome of X:

$$\mathbb{E}_x\left(\log_2\frac{1}{P(x=x)}\right) = \sum_{x\in\mathrm{supp}(()X)} p(x)\log_2\frac{1}{p(x)},$$

where we define $0\log_2\frac{1}{0} = 0$.

We have the following obvious facts:

1. $H(X) \geq 0$; $H(X) = 0$ if and only if X is deterministic.
2. Let X be uniform on $\{a_1, \ldots, a_n\}$. Then $H(X) = \log_2 n$.

Lemma E.1.2. $|\mathrm{supp}(()X)| = n$ *implies that* $H(X) \leq \log_2 n$.

Proof. We use Jensen's inequality: for convex functions f, we have $\mathbb{E}[f(x)] \leq f(\mathbb{E}[x])$ (this implies $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$ and $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$.) Since we have

$$\begin{aligned} H(X) &= \mathbb{E}_{x\leftarrow X}\left[\log\frac{1}{p(x)}\right] \\ &\leq \log\left(\mathbb{E}_{x\leftarrow X}\frac{1}{p(x)}\right) = \log n \end{aligned}$$

where we have applied Jensen's inequality to the expression with $f(x) = \log x$ while using the random variable $Z \sim 1/p(x)$ w.p. $p(x)$:

$$H(X) = \mathbb{E}_Z\left[\log Z\right] .$$

□

Communication. Let's see if we can derive the binary entropy function in a different setting. Suppose $X \leftarrow \mathrm{Bernoulli}(1/2)$, then $H(X) = 1$. And if $X \leftarrow \mathrm{Bernoulli}(1)$, then $H(X) = 0$. Now suppose $X \leftarrow \mathrm{Bernoulli}(p)$, and we have $X_1, X_2, \ldots, X_n$ iid samples from this distribution, represented as a bitstring $\{0,1\}^n$. A way to communicate this in an efficient way is that we first send

k, the number of 1 bits in the string, and then use $\lceil \log_2 \binom{n}{k} \rceil$ bits to reveal which subset those 1s go in. The expected number of bits communicated with this scheme is

$$\mathbb{E}(\text{number of bits}) = \underbrace{\lceil \log_2 n \rceil}_{\text{sending } k} + \underbrace{\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \left\lceil \log_2 \binom{n}{k} \right\rceil}_{\text{number of bits required for each } k}.$$

What we are interested is the average fraction of bits we need as the number of bits as n grows large: $\lim_{n\to\infty} \frac{1}{n}\mathbb{E}(\text{number of bits})$. Dividing our expression by n, we can omit the first term ($\log n/n \to 0$ as $n \to \infty$) and we can use Stirling's approximation (or the intuitive concept that if k is far from pn the term in the sum is close to zero) to obtain that

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}(\text{number of bits}) = h(p) = p \log_2 \frac{1}{p} + (1-p)\log_2 \frac{1}{1-p},$$

which tells us that we can take $H(\text{Bernoulli}(p))$ bits, on average, to communicate a Bernoulli(p) random variable.

Non-binary random variable. We can extend this definition of entropy in the communications context to non-binary random variables. Suppose $X \leftarrow \{a_1, \ldots, a_n\}$. Suppose we have related variables

$$Z_1 = \begin{cases} 1 & X = a_1 \\ 0 & \text{else} \end{cases}$$

and

$$Z_2 = \begin{cases} a_2 & \frac{p_2}{1-p_1} \\ \vdots & \\ a_n & \frac{p_n}{1-p_1} \end{cases}.$$

We can break

$$H(X) = H(Z_1) + \Pr(Z_1 = 0) H(Z_2),$$

as we need to communicate Z_1 and then, if Z_1 is 0, we need to communicate Z_2. Since we know how to compute $H(Z_1)$ from the binary case, we can recursively use this to obtain $H(X) = \sum_{i=1}^{n} p_i \log 1/p_i$, which is the same expression as we got before.

E.2 Joint and conditional entropy

Definition E.2.1 (Joint entropy)**.** Let X and Y be two possibly correlated random variables. The joint entropy of X and Y, denoted $H(X, Y)$, is

$$H(X, Y) = \sum_{x,y} p(x, y) \log \frac{1}{p(x, y)},$$

where $p(x, y)$ is defined to be $\Pr(X = x \wedge Y = y)$.

If X and Y are independent, $p(x,y) = p(x)p(y)$ and

$$H(X,Y) = \sum_{x,y} p(x)p(y)\left(\log\frac{1}{p(x)} + \log\frac{1}{p(y)}\right) = H(X) + H(Y).$$

In general,

$$H(X,Y) = \sum_{x,y} p(x,y)\log\frac{1}{p(x)p(y|x)},$$

where $p(y|x) = \Pr(Y = y|X = x)$.

We can then do the following calculation:

$$\begin{aligned}
H(X,Y) &= \sum_{x,y} p(x,y)\log\frac{1}{p(x)p(y|x)} \\
&= \sum_{x,y} p(x,y)\log\frac{1}{p(x)} + \sum_{x,y} p(x,y)\log\frac{1}{p(y|x)} \\
&= \sum_{x} p(x)\log\frac{1}{p(x)} + \sum_{x} p(x)\sum_{y} p(y|x)\log\frac{1}{p(y|x)} \\
&= H(X) + \sum_{x} p(x)H(Y|X = x) \\
&= H(X) + \mathbb{E}_x[H(Y|X = x)].
\end{aligned}$$

This motivates the definition of conditional entropy:

Definition E.2.2 (Conditional entropy)**.** The conditional entropy of Y given X is

$$H(Y|X) = \mathbb{E}_x[H(Y|X = x)].$$

Our calculation then shows this lemma:

Lemma E.2.1. $H(X,Y) = H(X) + H(Y|X)$.

Intuitively, this says that how surprised we are by drawing from the joint distribution of X and Y is how surprised we are by X plus how surprised we are by Y given that we know X already.

Note that if X and Y are independent, $H(Y|X) = H(Y)$ and $H(X,Y) = H(X) + H(Y)$.

Recall the chain rule for probability: $p(x,y) = p(x)p(y|x)$, or, more generally,

$$p(x_1,\ldots,x_n) = p(x_1)p(x_2|x_1)\ldots p(x_n|x_1,\ldots,x_n).$$

There is a similar chain rule for entropy:

Theorem E.2.2 (Chain rule)**.** *For random variables X, Y, and Z,*

$$H(X,Y,Z) = H(X) + H(Y|X) + H(Z|X,Y).$$

For n random variables $X_1,\ldots,X_n$,

$$H(X_1,X_2,\ldots,X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1,X_2,\ldots,X_{n-1}).$$

The log in the definition of entropy changes the multiplication in the probability chain rule to addition. Also, the order of the random variables does not matter. For example, it also holds that

$$H(X,Y) = H(Y) + H(X|Y).$$

Note that $H(X|X) = 0$.

Example E.2.3. *Let X be a random variable that is uniform on $\{0,1,2,3\}$. Let $Y = X \mod 2$.*
Clearly, $H(X) = 2$.
$H(Y) = 1$ since Y is uniform on $\{0,1\}$.
$H(X|Y) = 1$ because knowing Y tells us if X is odd or even.
$H(Y|X) = 0$ since knowing X tell us the exact value of Y.
$H(X,Y) = 2$ because X tells us everything about X and Y.

Intuitively, it seems like conditioning should never increase entropy: knowing more should never increase our surprise. This is indeed the case:

Lemma E.2.4 (Conditioning cannot increase entropy)**.** $H(Y|X) \le H(Y)$.

Proof. First, we have that

$$\begin{aligned} H(Y|X) - H(Y) &= H(X,Y) - H(X) - H(Y) \\ &= \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)} - \sum_x p(x) \log \frac{1}{p(x)} - \sum_y p(y) \log \frac{1}{p(y)}. \end{aligned}$$

Since $p(x) = \sum_y p(x,y)$ and $p(y) = \sum_x p(x,y)$,

$$H(Y|X) - H(Y) = \sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)}.$$

We now define Z to be a random variable taking value $\frac{p(x)p(y)}{p(x,y)}$ with probability $p(x,y)$, so

$$\begin{aligned} H(Y|X) - H(Y) &= \mathbb{E}_{x,y}[\log Z] \\ &\le \log \mathbb{E}[Z] \quad \text{by Jensen's Inequality} \\ &= \log \left(\sum_{x,y} p(x,y) \frac{p(x)p(y)}{p(x,y)} \right) \\ &= \log \left(\left(\sum_x p(x) \right) \left(\sum_y p(y) \right) \right) \\ &= \log 1 \\ &= 0. \end{aligned}$$

□

As a corollary, we have a statement similar to the union bound:

Corollary E.2.5. *For random variables X and Y,*

$$H(X,Y) \le H(X) + H(Y).$$

More generally, for random variables $X_1, \dots, X_n$,

$$H(X_1, \dots, X_n) \le \sum_{i=1}^{n} H(X_i).$$

Exercise E.2.6. *For a random variable X with support size n, we can think of entropy as a function from $[0,1]^n$ to $\mathbb{R}_{\ge 0}$. If X takes on n different values with probabilities $p_1, \cdots, p_n$, then for $\mathbf{p} = (p_1, \cdots, p_n)$, $H(\mathbf{p}) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$. Show that $H(\mathbf{p})$ is a concave function, i.e., show*

$$H(\lambda \mathbf{p} + (1-\lambda)\mathbf{q}) \ge \lambda H(\mathbf{p}) + (1-\lambda) H(\mathbf{q})$$

for all $\lambda \in [0,1]$, $\mathbf{p}, \mathbf{q} \in [0,1]^n$.

E.3 Mutual information

Definition E.3.1 (Mutual information)**.** The mutual information between random variables X and Y, denoted $I(X;Y)$, is

$$I(X;Y) = H(X) - H(X|Y).$$

Intuitively, mutual information is the reduction in the uncertainty of X that comes from knowing Y.

We can write $I(X;Y)$ in several other equivalent ways:

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

Note that $I(X;Y) = I(Y;X)$.

The next lemma follows from the fact that conditioning cannot increase entropy.

Lemma E.3.1. $I(X;Y) \ge 0$.

Also, if X and Y are independent, $I(X;Y) = 0$.

Example E.3.2. *Consider X and Y as defined in Example E.2.3. Then*

$$I(X;Y) = H(X) - H(X|Y) = 2 - 1 = 1.$$

Example E.3.3. *Let the Z_i's be i.i.d. random variables that are uniform over $\{0,1\}$. Let $X = Z_1 Z_2 Z_3 Z_4 Z_5$ and $Y = Z_4 Z_5 Z_6 Z_7$. Then $I(X;Y) = 2$ since X and Y have 2 bits in common.*

Figure E.1: Relationship between entropy, joint entropy, conditional entropy, and mutual information for two random variables.

We show the relationship between entropy, joint entropy, conditional entropy, and mutual information for two random variables X and Y in Figure 5.1.

We can also define a conditional version of mutual information.

Definition E.3.2 (Conditional mutual information)**.** The conditional mutual information between X and Y given Z is

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= H(Y|Z) - H(Y|X,Z). \end{aligned}$$

Exercise E.3.4. *Prove the chain rule for mutual information:*

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, X_2, \ldots, X_{i-1}).$$

Note that the order of the X_i's does not matter.

Exercise E.3.5. *It is not true that conditioning never inreases mutual information. Give an exampleof random variables X, Y, X where $I(X;Y|Z) > I(X;Y)$.*