Information, Codes and Ciphers

By Jeremy Le for MATH3411 24T3

1 Introduction

1.1 Mathematical Model

To give a mathematical framework for digital data transmission, define

- a **source alphabet** $S = \{s_1, s_2, \ldots, s_q\}$ of q symbols
- a **code alphabet** A of r symbols probabilities $p_i = P(s_i)$
- a **code** that encodes each symbol s_i by a codeword which is a **string** of code symbols.

1.2 Assumed Knowledge

- Modular Arithmetic and the Division Algorithm
- Probability (Binomial Distribution and Bayes' Rule)
- Linear Algebra (Linear combination, independence, etc...)

1.3 Morse Code

Morse code is a **ternary** code (radix 3). Its alphabet is

1. $\bullet$ called **dot**
2. — called **dash**
3. p a **pause**

The codewords are strings of $\bullet$ and — **terminated** by p.

1.4 ASCII

American National **S**tandard **C**ode for **I**nformation **I**nterchange.

Binary code of fixed codeword length, namely 7, with $2^7 = 128$ encoded symbols.

The extended ASCII is a code like the 7-bit ASCII but with an extra bit in the front used as a check bit, requiring the number of 1's to be even.

1.5 ISBN

International **S**tandard **B**ook **N**umber.

They have 10 bits, with it's last bit being a check bit, requiring

$$\sum_{i=1}^{10} i x_i \equiv 0 \pmod{11}.$$

2 Error Detection and Correction Codes

We say that $\mathbf{x}$ **corrupted** to $\mathbf{y}$ is denoted by $\mathbf{x} \rightsquigarrow \mathbf{y}$.

2.1 ISBN-10 Error Capability

ISBN-10 numbers are capable of detecting the two types of errors:

1. getting a digit wrong,
2. interchanging two (unequal) digits.

2.2 Types of Codes

- **variable length code**: codewords have different lengths
- **block code**: codewords have the same lengths
- **t-error correcting code**: code can always correct up to t errors
- **systematic code**: code with **information digits** and **check digits** distinct

2.3 Binary Repetition Codes

A **binary r-repetition code** encodes $0 \to \overbrace{0 \cdots 0}^{r}$ and $1 \to \overbrace{1 \cdots 1}^{r}$.

The binary $(2t+1)$-repetition code is t-error correcting.
The binary $2t$-repetition code is $(t-1)$-error correcting and t-error detecting.

2.4 Information Rate and Redundancy

The **information rate** R is given by,

- For a code C of radix r and length n, $R = \dfrac{\log_r |C|}{n}$
- For a **systematic code**, $R = \dfrac{\text{\# information digits}}{\text{length of code}}$

We then define **redundancy** $= \frac{1}{R}$.

2.5 Binary Hamming Error-Correcting Codes

A Binary Hamming (n, k) code is a code of length n with k information bits, such that it is a single error correcting and has a parity check matrix, H, of size $n-k$ by n.

2.6 Hamming Distance, Weights

The **weight** of an n-bit word $\mathbf{x}$ is defined to be

$$w(\mathbf{x}) = \#\{i : 1 \le i \le n, x_i \ne 0\}.$$

Given two n-bit words, the **Hamming distance** between them is

$$d(\mathbf{x}, \mathbf{y}) = \#\{i : 1 \le i \le n, x_i \ne y_i\}.$$

Given some code with set of codewords C, we define **(minimum) weight** of C to

$$w = w(C) = \min\{w(\mathbf{x}) : \mathbf{x} \in C, \mathbf{x} \ne \mathbf{0}\}.$$

Similarly, the **(minimum) distance of** C is defined by

$$d = d(C) = \min\{d(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in C, \mathbf{x} \ne \mathbf{y}\}.$$

If $\mathbf{x} \rightsquigarrow \mathbf{y}$, then $d(\mathbf{x}, \mathbf{y})$ is the number of errors in $\mathbf{y}$.

2.7 Decoding Strategies

Minimum Distance Decoding Strategy Given a received word y, decode to *closest* codeword x.

Standard Strategy If received word y is distance at most t from a codeword x, then decode y to x; otherwise flag an error.

Pure Error Detection If received word y is not a codeword x, then flag an error.

2.8 Sphere Packing

The **sphere** of radius r around $\mathbf{c}$:

$$S_r(\mathbf{c}) = \{\mathbf{x} \in \mathbb{Z}_2^n : d(\mathbf{x}, \mathbf{c}) \le r\}.$$

The volume of this sphere is its size $|S_r(\mathbf{c})|$.

Sphere-Packing Condition Theorem A t-error correcting binary code C of length n has minimum distance $d = 2y + 1$ or $2t + 2$, and

$$|C| \sum_{i=0}^{t} \binom{n}{i} \le 2^n.$$

If we have equality in the bound, then we say that the code is perfect. This means that codewords are evenly spread around in $\mathbb{Z}_2^n$ space.

2.9 Binary Linear Codes

A linear code C is a vector space over some field $\mathbb{F}$. Equivalently it is the null-space of

$$C = \{\mathbf{x} \in \mathbb{F}^n : H\mathbf{x}^T = \mathbf{0}\}$$

of an $m \times n$ parity check matrix H with $m = \operatorname{rank}(H)$.

- $\dim C = k = n - m$ by the Rank-Nullity Theorem.
- If C is binary, then $|C| = 2^k$.
- C is systematic.
- If H is **reduced echelon form**, then we can choose the non-leading columns of H to be **information bits** and the leading columns of H to be **check bits**.

Minimum Distance for Linear Codes If C is a linear code with parity check matrix H, then

- $w(C) = d(C)$,
- $d(C) = \min\{r : H \text{ has } r \text{ linearly dependent columns }\}$.

For a linear code C, the **row space** (or **row span**) of a $k \times n$ **generator matrix** G over $\mathbb{F}$ generates C, in the sense that C is a set of linear combinations of G.

2.10 Standard Form Matrices

For a linear code C of dimension k and length $n = k + m$,

- $H = (I_m \mid B)$ is a parity check matrix for C *if and only if*
- $G = (-B^T \mid I_k)$ is a generator matrix C.

Linear codes $C$and C' are **equivalent** if C' is obtained by permuting the codeword entries of C by a fixed permutation:

$$C' = CP = \{\mathbf{x}P : \mathbf{x} \in C\} \text{ for some permutation matrix } P$$

Note that $G' = GP$ and $H' = HP$.

2.11 Extending Linear Codes

The **extension** of a linear code C:

$$\hat{C} = \{x_0 x_1 \cdots x_n : x_1 \cdots x_n \in C, x_0 = -(x_1 + \cdots + x_n)\}.$$

The **extension** $\hat{C}$ is a linear code with minimum distance $d(C)$ or $d(C) + 1$.

2.12 Radix r Hamming Codes

- Let r be a prime number and $m \ge 1$ some integer.
- Write the numbers $1, \ldots, r^m - 1$ in base r, as length m column vectors.
- Of each set of $r - 1$ parallel columns, delete all whose *first nonzero entry* is not 1.
- This gives the **radix r Hamming code** of length $n = \frac{r^m - 1}{r - 1}$.

3 Compression Coding