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** $S = \{s_1, s_2, \dots, s_q\}$ 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. • called **dot**
2. — called **dash**
3. `p` a **pause**

The codewords are strings of • 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

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

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

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

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

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

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

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

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