7.2 Some Coding Theory and the proof of Theorem 7.3

In this section we (very) briefly introduce error-correcting codes and use Reed-Solomon codes to prove Theorem 7.3. We direct the reader to [GRS15] for more on the subject.

Let’s say Alice wants to send a message to Bob but they can only communicate through a channel that erases or replaces some of the letters in Alice’s message. If Alice and Bob are communicating with an alphabet $\Sigma$ and can send messages with length $N$ they can pre-decide a set of allowed messages (or codewords) such that even if a certain number of elements of the codeword gets erased or replaced there is no risk for the codeword sent to be confused with another codeword. The set $C$ of codewords (which is a subset of $\Sigma^N$) is called the codebook and $N$ is the blocklength.

If every two codewords in the codebook differs in at least $d$ coordinates, then there is no risk of confusion with either up to $d-1$ erasures or up to $\lfloor \frac{d-1}{2} \rfloor$ replacements. We will be interested in codebooks that are a subset of a finite field, meaning that we will take $\Sigma$ to be $\mathbb{F}_q$ for $q$ a prime power and $C$ to be a linear subspace of $\mathbb{F}_q$.

The dimension of the code is given by

$$m = \log_q |C|,$$

and the rate of the code by

$$R = \frac{m}{N}.$$

Given two code words $c_1, c_2$ the Hamming distance $\Delta(c_1, c_2)$ is the number of entries where they differ. The distance of a code is defined as

$$d = \min_{c_1 \neq c_2} \Delta(c_1, c_2).$$

We say that a linear code $C$ is a $[N, m, d]_q$ code (where $N$ is the blocklength, $m$ the dimension, $d$ the distance, and $\mathbb{F}_q$ the alphabet.

One of the main goals of the theory of error-correcting codes is to understand the possible values of rates, distance, and $q$ for which codes exist. We simply briefly mention a few of the bounds and refer the reader to [GRS15]. An important parameter is given by the entropy function:

$$H_q(x) = x \frac{\log(q - 1)}{\log q} - x \frac{\log x}{\log q} - (1 - x) \frac{\log(1 - x)}{\log q}.$$

- Hamming bound follows essentially by noting that if a code has distance $d$ then balls of radius $\lfloor \frac{d-1}{2} \rfloor$ centered at codewords cannot intersect. It says that

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

- Another particularly simple bound is Singleton bound (it can be easily proven by noting that the first $n + d + 2$ of two codewords need to differ in at least 2 coordinates)

$$R \leq 1 - \frac{d}{N} + o(1).$$
There are probabilistic constructions of codes that, for any $\epsilon > 0$, satisfy
\[
R \geq 1 - H_q\left(\frac{d}{N}\right) - \epsilon.
\]
This means that $R^*$, the best rate achievable, satisfies
\[
R^* \geq 1 - H_q\left(\frac{d}{N}\right),
\] (65)
known as the Gilbert-Varshamov (GV) bound [Gil52, Var57]. Even for $q = 2$ (corresponding to binary codes) it is not known whether this bound is tight or not, nor are there deterministic constructions achieving this rate. This motivates the following problem.

**Open Problem 7.1**

1. Construct an explicit (deterministic) binary code ($q = 2$) satisfying the GV bound (65).

2. Is the GV bound tight for binary codes ($q = 2$)?

**References**


