MAD 4471: Cryptography and Coding Theory **Fall 2020**

Lecture 9: Introduction to Coding Theory

Lecturer: Jean-François Biasse *TA: William Youmans*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Here is the problem: messages are transmitted through a noisy channel. We want to:

- Detect the presence of transmission errors.
- Correct transmission errors.

$$m \quad \longrightarrow \quad Noise \quad \longrightarrow \quad m + c$$

Example 1 (Repetition code) *Here is a simple example: repetition code $\mathcal{C}_{rep}$ where $0 \leftrightarrow (0,0,0) \quad 1 \leftrightarrow (1,1,1)$*

- *If the noise induces no more than 2 errors, then if we do not receive $(0,0,0)$ or $(1,1,1)$, there has been a transmission error.*
- *If the noise induces no more than 1 error, then we can correct the error by choosing to repeat 3 times the coordinate that occurs 2 or 3 times.*

For example (1,0,0) $\longrightarrow$ (0,0,0)

9.1 Basic concepts in coding theory

Definition 9.1 (Code) *A code $\mathcal{C}$ is a set of codewords.*

Example 2 *the code $\mathcal{C}_{rep}$ is the set $\{(0,0,0),(1,1,1)\}$.*

Definition 9.2 *The codewords of a code $\mathcal{C}$ are strings of m symbols from an alphabet $\mathcal{A}$ of size q.*

- *We say that m is the length of $\mathcal{C}$.*
- *We say that $\mathcal{C}$ is a q-ary code.*

Example 3 *The main parameters of $\mathcal{C}_{rep}$ are:*

- *The length of $\mathcal{C}_{rep}$ is 3.*
- *$\mathcal{C}_{rep}$ is a binary (2-ary) code.*
- *The alphabet of $\mathcal{C}_{rep}$ is $\{0,1\}$.*

Question: How many transmission errors can we tolerate ?

The two tasks we want to perform are:

- Detecting the presence of an error.
- Correcting an error.

If we receive a vector r in $\mathcal{A}^m$ that does not belong to $\mathcal{C}$, then clearly there has been a transmission error. If $r = c' \in \mathcal{C}$, then maybe r is the original message, or maybe there has been so many errors that we went from one codeword to the other. The important parameter of $\mathcal{C}$ that allows us to determine how many errors we can detect/correct is its distance.

Definition 9.3 (Distance of $\mathcal{C}$) *The Hamming distance between $u, v \in \mathcal{A}^m$ is the number of indices on which their symbols differ. We denote it by $d(u, v)$. The distance of $\mathcal{C}$ is denoted by $d(\mathcal{C})$ and is by definition*

$$d(\mathcal{C}) = \min\{d(u, v) \mid u \neq v \in \mathcal{C}\}$$

Proposition 9.4 *We can detect the presence of a transmission error if the number of errors s satisfies $d(\mathcal{C}) \geq s + 1$. In this case, if the message received r is in $\mathcal{C}$, then we know there was no error.*

Proof: Suppose the message r we receive is in $\mathcal{C}$. Let $c \in \mathcal{C}$ be the original codeword that was sent and let $c' = r \in \mathcal{C}$. There has been at most s errors, so $d(c, c') \leq s$. But $d(\mathcal{C}) \geq s + 1$, so either $d(c, c') \geq s + 1$ or $c = c'$. In our case, $c = c'$, and there was no transmission error if the received message r is in $\mathcal{C}$. Of course if $r \notin \mathcal{C}$, there was necessarily an error. ■

Proposition 9.5 *We can correct the transmission of up to t errors if $d(\mathcal{C}) \geq 2t+1$. In this case, the original codeword sent is the closet codeword of $\mathcal{C}$ to the received message r.*

Proof: We prove that there can be only one codeword c at distance less or equal to t from the received message r. In this case, since the number of errors is bounded by t, c has to be the original codeword. Suppose that there is another $c' \in \mathcal{C}$ such that $c \neq c'$ and $d(c', r) \leq t$. Then $d(c', c) \leq d(c', r) + d(c, r) \leq t + t = 2t$, but $d(c', c) \geq d(\mathcal{C}) \geq 2t + 1$ which is a contradiction. Therefore $c \in \mathcal{C}$ such that $d(c, r) \leq t$ is unique and is the original codeword. ■

Example 4 *For $\mathcal{C}_{rep}$, $d(\mathcal{C}) = d(\{(0, 0, 0), (1, 1, 1)\}) = 3$. So*

- *We can detect the presence of an error if the number of errors does not exceed 2.*
- *We can correct an error if there is no more than 1 error.*

Question: How do we quantify the efficiency?

One way of looking at it is to measure how much redundancy we need to detect/correct errors. Indeed, with enough redundancy, $\mathcal{C}_{rep}$ allows the correction of any number of errors, but at the price of an overload of the bandwidth.

Definition 9.6 (Code rate) *An (m, M, d) code is a code of length m, with M codewords and of distance d. The code rate is the value $\frac{\log_q M}{m}$ ($\log_q M$ is the necessary length to represent M codwords over an alphabet of size q).*

Example 5 *For $\mathcal{C}_{rep}$, $q = 2, M = 2, \log_q(M) = 1, m = 3, Rate(\mathcal{C}_{rep}) = 1/3$.*

The smaller the ratio is, the more redundancy we have. Given m, d, we can give an upper bound on the rate (which means that we show that there is a limit to its efficiency). First, we need to find a bound on M.

Proposition 9.7 *Let $\mathcal{C}$ be a q-ary (m, M, d) code, then $M \leq q$.*

Proof: Let c be a codeword, $c = (a_1, ..., a_m)$. We define $c' = (a_d, ..., a_m)$ by cutting the first d coordinates. If $c_1 \neq c_2$, they have to differ in at least d coordinates. This means that c'_1 and c'_2 must differ in at least 1 coordinate. So M is less than the number of different possible c'. They are words of $m - d + 1$ symbols over an alphabet of size q. Their number is less than q^{m-d+1}. Therefore $M \leq q^{m-d+1}$.

$\blacksquare$