Topics in Error-Correcting Codes

Notes by Yael Kalai

MIT - 6.S953
Lecture 1 (September 9, 2026)

> ***Warning:*** This document is a rough draft, so it may contain bugs. Please feel free to email me with corrections.

Logistics

All the logistics for this class can be found in the class website: https://codes-6s953.github.io/

Error Correcting Codes

Communication is fundamentals, humans have been communicating for a very long time. Communication has intrinsic error, for example, people have different accents. Nevertheless, we tend to understand each other. This is due to the fact that languages have redundancy built it. Redundancy allows us to correct error in communication.

Errors also exist in the digital world. Error correcting codes (ECC) represent data in a dredundant way that allows for error correction. There are two equivalent ways to define error correcting codes, one due to Shannon [2], and one due to Hamming [1], the former is more algorithmic in nature, while the latter is more combinatorial.

Definition 1. [2] An error correcting code over an alphabet Σ, message length k, consists of two functions

$$\mathsf{Enc} : \Sigma^k \to \Sigma^n, \quad \mathsf{Dec} : \Sigma^n \to \Sigma^k$$

such that for every $m \in \Sigma^k$ it holds that $\mathsf{Dec}(\mathsf{Enc}(m)) = m$.

From now on we denote by $q = |\Sigma|$. The message length is denoted by k, and n is called the block length.

Definition 2. [1] A code of block length n over alphabet Σ is a subset of $\mathcal{C} \subseteq \Sigma^n$. The dimension of the code is defined to be $k := \log_q |\mathcal{C}|$.

We will use both definitions interchangably, depending on which is more convenient in terms of notation. We will study the tradeoff between the amount of redundancy used and the number of errors that can be corrected by this code.

Definition 3. The Rate of a code of dimention k and block length n is $R := k/n \leq 1$.

Intuitively, the rate is the average amount of "real" information in each of the n symbols transmitted.

We quantify the amount of errors in a codeword by the Hamming distance between the codeword and the recieved word.

Definition 4. For every $u, v \in \Sigma^n$, the Hamming distance between u and v, denoted by $\Delta(u, v)$, is defined to be

$$\Delta(u, v) := |\{i \in [n] : \ u_i \neq v_i\}|.$$

The relative Hamming distance, denoted by $\delta(u, v)$ is definited by $\delta(u, v) = \Delta(u, v)/n$.

The relative version will be useful when we study the asymptotic behavior of the code (coming up!).

Note that the Hamming distance has the following two useful properties:

- **Symmetric:** For every $u, v \in \Sigma^n$,
$$\Delta(u, v) = \Delta(v, u)$$
- **Triangle inequality:** For every $u, v, w \in \Sigma^n$,
$$\Delta(u, v) = \Delta(u, w) + \Delta(w, v).$$

Definition 5. An n-symbol t-error channel over alphabet Σ is a function $\mathsf{ch} : \Sigma^n \to \Sigma^n$ such that $\Delta(v, \mathsf{ch}(v)) \leq t$ for every $v \in \Sigma^n$.

This noise model is often called the adversarial noise model, or the Hamming model. A popular alterantive noise model is the stochastic noise model. For example, in the binary symmetric channel with probability $p \in [0, 1]$, denoted by BSC_p, first defined by Shannon, each bit is flipped with probability p.

Note that the Hamming model and the Shannon model are two extremes: Hamming assumes no knowledge about the channel except for a bound on the total number of errors, whereas Shannon assumes full knowledge. Usually, these two extremes are considered, even though in reality the situation is often somewhere in between.

Definition 6. A t-ECC is an error correcting code such that for every $m \in \Sigma^k$ and for every t-error channel ch,

$$\mathsf{Dec}(\mathsf{Enc}(m)) = m.$$

It is a t-error-detecting code if for every $m \in \Sigma^k$ and for every t-error channel ch,

$$\mathsf{Dec}(\mathsf{ch}(\mathsf{Enc}(m))) = \begin{cases} 1 & \text{if } \mathsf{ch}(\mathsf{Enc}(m)) = \mathsf{Enc}(m), \\ 0 & \text{otherwise.} \end{cases}$$

Definition 7 (t-erasure channel)**.** An n-symbol t-erasure channel over alphabet Σ is a function $\mathsf{ch} : \Sigma^n \to (\Sigma \cup \{?\})^n$ such that $\Delta(v, \mathsf{ch}(v)) \leq t$ for every $v \in \Sigma^n$ and for every $i \in [n]$, $\mathsf{ch}(v)_i \in \{v_i, ?\}$. If $\mathsf{ch}(v)_i =?$ then i is called an erasure.

Definition 8 (Erasure correcting code)**.** $(\mathsf{Enc}, \mathsf{Dec})$ is a t-erasure code if for every $m \in \Sigma^k$ and for every t-erasure channel ch it holds that $\mathsf{Dec}(\mathsf{ch}(\mathsf{Enc}(m))) = m$.

Example: Repetition code. In the repetition code $\mathsf{Enc} : \Sigma^k \to \Sigma^n$ for $n = a \cdot k$, where each symbol is repeated a times. Decoding is via majority vote. The rate of this code is $R = k/n = a$ and the distance is $\Delta(\mathcal{C}) = a$. It is a t-ECC for $t = \lfloor \frac{a-1}{2} \rfloor$. It is a t-error-detecting code and a t-erasure code for $t = a - 1$.

Definition 9. Let $\mathcal{C}$ be a code. The minimum distance of $\mathcal{C}$ is $\Delta(\mathcal{C}) := \min_{c_1 \neq c_2 \in \mathcal{C}} \Delta(c_1, c_2)$. The relative minimum distance is $\delta(\mathcal{C}) := \min_{c_1 \neq c_2 \in \mathcal{C}} \delta(c_1, c_2)$.

For example, the repetition code has distance a (where a is the number each symbol is repeate).

Theorem 10. *For every code $\mathcal{C}$, the following are equivalent:*

1. *$\mathcal{C}$ has minimum distance $d \geq 2$.*
2. *If d is odd then $\mathcal{C}$ can correct $t = \frac{d-1}{2}$ errors.*[1]
3. *$\mathcal{C}$ can detect $t = d - 1$ errors*
4. *$\mathcal{C}$ can correct $t = d - 1$ erasures.*

[1] If d is even then Item 1 implies Item 2 with $t = \lfloor \frac{d-1}{2} \rfloor$ and Item 2 implies Item 1 for $t = \lceil \frac{d-1}{2} \rceil$.

Proof. Fix a code $\mathcal{C}$. We first prove that Item 1 implies Items 2, 3 and 4, starting with Item 2.

Item 1 $\to$ Item 2. Consider the **Maximum Likelyhood Decocing** (MLD) defined by

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

Let $t = \lfloor \frac{d-1}{2} \rfloor$. Let c_1 be the transmitted codeword and let y be the received codeword, where $\Delta(y, c_1) \leq t$. Suppose for contradiction that $\mathsf{MLD}(y) = c_2 \leq c_1$. Then by the definition of MLD, $\Delta(y, c_2) \leq \Delta(y, c_1)$. Therefore

$$\Delta(c_1, c_2) \leq \Delta(c_1, y) + \Delta(y, c_2) \leq 2\Delta(c_1, y) \leq 2t \leq d - 1,$$

contradiction.

Item 1 $\to$ Item 3. We construct an error detection algorithm for $t \leq d - 1$. Given $y \in \Sigma^n$ if $y \in \mathcal{C}$ then output 1 and otherwise output 0. To prove correctness, note that if no errors occured then the answer is correct. Otherwise, suppose the transmitted codeword was c_1 such that $\Delta(c_1, y) \in [1, d - 1]$, then y is not a codeword and hence the answer is correct.

Item 1 $\to$ Item 4. Let $y \in (\Sigma \cup \{?\})^n$ be the received word. We prove that there exists a unique $c \in \mathcal{C}$ that agrees with y on every index $i \in [n]$ such that $y_i \neq ?$ and such that $\Delta(y, c) \leq d - 1$. Suppose for contradiction that there exist $c_1 \neq c_2 \in \mathcal{C}$ that satisfy these two properties. This implies that for every i such that for every $i \in [n]$ such thayt $y_i \neq ?$ it holds that $c_{1,i} = c_{2,i} = y_i$. Since $|\{i \in [n] : y_i = ?\}| \leq d - 1$ this implies that $\Delta(c_1, c_2) \leq d - 1$, contradiction.
We next prove that $\neg$ Item 1 implies $\neg$ Item 2, $\neg$ Item 3 and $\neg$ Item 4. In all these cases we assume that $\mathcal{C}$ has distance $\leq d - 1$, which implies that exists codewords $c_1 \neq c_2$ such that $\Delta(c_1, c_2) = t \leq d - 1$.

$\neg$ *Item 1* $\rightarrow$ $\neg$ *Item 2*. Let y be a word such that agrees with c_1 and c_2 on all coordinates $i \in [n]$ for which $c_{1,i} = c_{2,i}$, and on the remaining the t coordinates, it agrees with c_1 on $t/2$ of them and with $t/2$ of them (if t is odd then break the tie arbitrarily). Note that

$$\Delta(y, c_1), \Delta(y, c_2) \leq \lceil \frac{t}{2} \rceil,$$

and thus the word y cannot be decoded since it could have originated from either c_1 or c_2

$\neg$ *Item 1* $\rightarrow$ $\neg$ *Item 3*. Let $y = c_2$. It cannot be detected if y was originated from c_1 or c_2.

$\neg$ *Item 1*$\rightarrow$ $\neg$ *Item 4*. Let y be the word that agrees with c_1 and c_2 on every coordinate $i \in [n]$ for which $c_{1,i} = c_{2,i}$, and on the remaining coordinates $y_i =?$. As above, it cannot be detected if y was originated from c_1 or c_2. □

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

We saw that the repetition code $\mathcal{C}_{\mathsf{rep},a}$ has distance a and rate $1/a$. Can we do better?

Definition 11 (Hamming ball)**.** : For every $x \in \Sigma^n$,

$$B(x, e) = \{y \in \Sigma^n : \Delta(x, y) \leq e\}.$$

Definition 12. A code $\mathcal{C} \in \Sigma^n$ with dimension k and distance d is called a $(n, k, d)_\Sigma$ code, and is also called a $(n, k, d)_q$ code.

Theorem 13. *For every $(n, k, d)_q$ code,*

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

Proof. Let $e = \lfloor \frac{d-1}{2} \rfloor$. By Theorem 10, for every $c_1 \neq c_2 \in \mathcal{C}$,

$$B(c_1, e) \cap B(c_2, e) = \emptyset. \tag{1}$$

Note that for every $x \in [q]^n$,

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

This, together with Equation (1), implies that

$$q^n \geq |\bigcup_{c \in \mathcal{C}} B(C, e)| = \sum_{c \in \mathcal{C}} |B(C, e)| \leq \sum_{c \in \mathcal{C}} \sum_{i=0}^{e} \binom{n}{i} (q-1)^i = q^k \cdot \sum_{i=0}^{e} \binom{n}{i} (q-1)^i.$$

Applying $\log_q$ to both sides of the equation we get our desired result. □

Corollary 14. *For every code of distance d,*

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

Definition 15. Codes that meet the Hamming bound are called **perfect codes**.

Perfect codes lead to "perfect packing." If one constructs Hamming ball of radius $\lfloor \frac{d-1}{2} \rfloor$ around all codewords, then in covers the entire space Σ^n.

Hamming Codes

Consider the following code, which is called the Hamming code:

$$\mathsf{Enc}_H(x_1, x_2, x_3, x_4) = (x_1, x_2, x_3, x_4, x_2 \oplus x_3 \oplus x_4, x_1 \oplus x_3 \oplus x_4, x_1 \oplus x_2 \oplus x_4)$$

We denote this code by $\mathcal{C}_H$.

Claim 1. $\mathcal{C}_H$ is a $(7,4,3)_2$ code, and hence it is a perfect code.[2]

[2] One can generalize this code to a perfect $(2^r - 1, 2^r - r - 1, 3)_2$ for every $r \in \mathbb{N}$.

Proof. We need to prove that this code has distance $d = 3$, since $\frac{k}{n} = 4/7$, and

$$\begin{aligned}
&1 - \frac{\log_q \left(\sum_{i=0}^{e} \binom{n}{i}(q-1)^i\right)}{n} = \\
&1 - \frac{\log_2 \left(\sum_{i=0}^{1} \binom{n}{i}\right)}{n} = \\
&1 - \frac{\log_2 (1+n)}{n} = \\
&1 - 3/7,
\end{aligned}$$

where the frist equation follows from the fact that $e = \lfloor \frac{d-1}{2} \rfloor = 1$ and $q = 2$, and the last equation follows from the fact that $n = 7$.[3]

[3] Note that for general $n = 2^r - 1$ the last equation is replaced by $1 - \frac{r}{2^r - 1} = \frac{k}{n}$, and thus is a perfect code for every $r \in \mathbb{N}$, assuming the distance is $d = 3$.

It thus suffices to prove that $d = 3$. To this end, we first note that it is a linear code; namely for every $x, y \in \{0,1\}^4$,

$$\mathsf{Enc}_H(x \oplus y) = \mathsf{Enc}_H(x) \oplus \mathsf{Enc}_H(y).$$

We use the following definition:

Definition 16. The Hamming weight of a vector $v \in \{0, 1, \ldots, q-1\}$ is defined to be

$$\mathsf{wt}(v) = \{i \in [n] : \ v_i \neq 0\}$$

We note that by case analysis (by considering the weight of the messages) one can argue that

$$\min_{0 \neq c \in \mathcal{C}_H} \mathsf{wt}(c) = 3.$$

Thus, the proof of Claim 1 follows from Claim 2 below. □

Claim 2. For every binary linear code $\mathcal{C}$, its distance is equal to the minimum Hamming weight of every non-zero codeword. Namely,

$$\min_{0 \neq c \in \mathcal{C}} \mathsf{wt}(c) = \min_{c_1 \neq c_2 \in \mathcal{C}} \Delta(c_1, c_2)$$

Proof. Fix a binary linear code $(\mathsf{Enc}, \mathsf{Dec})$. The fact that the code is linear implies that there exists a matrix $G \in \{0,1\}^{k \times n}$ such that for every message $x \in \{0,1\}^k$,

$$\mathsf{Enc}(x) = x \cdot G$$

G is called the *generator matrix* of the code. Note that

$$\min_{0 \neq c \in \mathcal{C}} \mathsf{wt}(c) = \min_{x \in \{0,1\}^k \setminus \{0^k\}} \mathsf{wt}(x \cdot G)$$

and that

$$\min_{c_1 \neq c_2 \in \mathcal{C}} \Delta(c_1, c_2) = \min_{x_1 \neq x_2 \in \{0,1\}^k} \Delta(x_1 \cdot G, x_2 \cdot G)$$

It remains to notice that

$$\Delta(x_1 \cdot G, x_2 \cdot G) = \mathsf{wt}(x_1 \cdot G - x_2 \cdot G, 0) = \mathsf{wt}((x_1 - x_2) \cdot G, 0)$$

which implies that

$$\min_{c_1 \neq c_2 \in \mathcal{C}} \Delta(c_1, c_2) = \min_{x_1 \neq x_2 \in \{0,1\}^k} \mathsf{wt}((x_1 - x_2) \cdot G, 0) = \min_{x \in \{0,1\}^k \setminus \{0^k\}} \mathsf{wt}(x \cdot G, 0)$$

□

Remark. We note that a binary linear code $\mathcal{C}$ is not only associated with a generator matrix $G \in \{0,1\}^{k \times n}$, but is also with a parity-check matrix $H \in \{0,1\}^{n \times (n-k)}$ such that for every $x \in \{0,1\}^n$

$$x \cdot H = 0 \quad \text{if and only if} \quad x \in \mathcal{C},$$

where H is a matrix such that $G \cdot H \equiv 0$.

Family of Codes

When we perform an asymptotic study of codes it makes sense to consider a family of codes, and study their asymptotic rate and distance.

Definition 17. Let $\mathcal{C} = \{\mathcal{C}_i\}_{i \in \mathbb{N}}$ be a family of codes where each $\mathcal{C}_i$ is a $(n_i, k_i, d_i)_{q_i}$ code. We define the rate of this family by

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

and we define the relative distance of this family by

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

Question: Does there exist a constant q and a family of codes $\mathcal{C}$ such that $q_i = q$ for every $i \in \mathbb{N}$ such that $R(\mathcal{C}) > 0$ and $\delta(\mathcal{C}) > 0$?

Such codes are called **asymptotically good codes**.

Spoiler: We will present several asymptotically code codes.

References

[1] Richard W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, 1950.

[2] Claude E. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656–715, 1949.