CS258: Information Theory

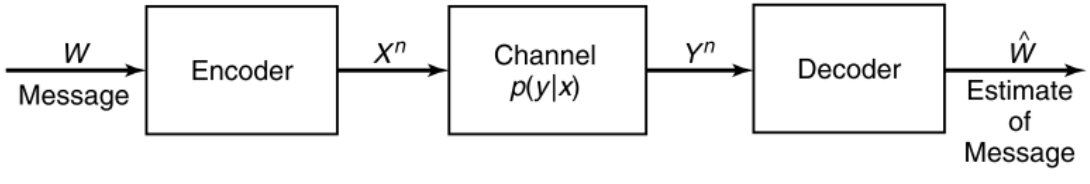
Fan Cheng
Shanghai Jiao Tong University

http://www.cs.sjtu.edu.cn/~chengfan/
chengfan@sjtu.edu.cn
Spring, 2020

Outline

- Channel Model
- Channel Capacity
- Channel Coding Theorem: Achievability
- Channel Coding Theorem: Converse
- Feedback Capacity
- Source-Channel Separation Theorem
- Hamming Code

Converse: Zero-Error Codes



$$W \to X^n \to Y^n \to \widehat{W}$$
$$X^n \to Y^n \text{ is memoryless}$$

The outline of the proof of the converse is most clearly motivated by going through the argument when **absolutely no errors are allowed**.

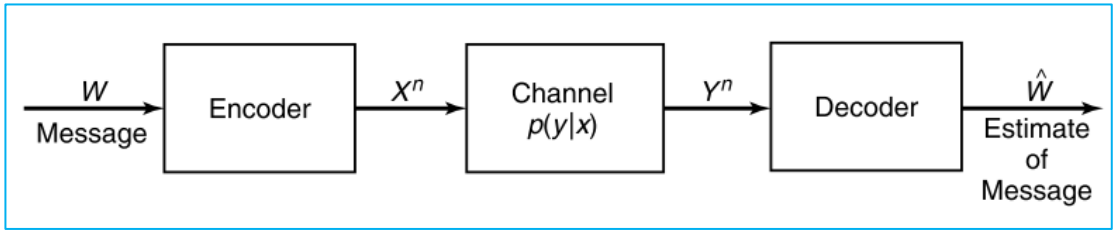
$$\begin{aligned} nR = H(W) &= H(W|Y^n)_{=0} + I(W;Y^n) \\ &= I(W;Y^n) \\ &\le I(X^n;Y^n) \quad (W \to X^n \to Y^n) \\ &\le \sum_i I(X_i;Y_i) \\ &\le nC \end{aligned}$$

$$R \le C$$

In general, $H(W|Y^n) > 0$: Fano's inequality

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum H(Y_i|X_i) \\ &\le \sum H(Y_i) - \sum H(Y_i|X_i) = \sum I(X_i;Y_i) \end{aligned}$$

Converse: Channel Coding Theorem



Fano: $H(W|\widehat{W}) \leq 1 + P_e^{(n)} nR$

By Fano's inequality

$$\begin{aligned} nR &= H(W) \\ &= H(W|\widehat{W}) + I(W;\widehat{W}) \\ &\leq \mathbf{1 + P_\epsilon^{(n)} nR} + I(W;\widehat{W}) \\ &\leq 1 + P_\epsilon^{(n)} nR + I(X^n;Y^n) \\ &\leq 1 + P_\epsilon^{(n)} nR + nC \end{aligned}$$

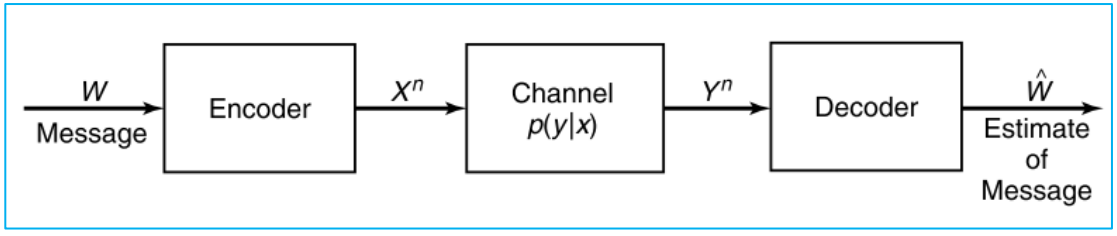
Thus

$$R \leq P_\epsilon^{(n)} R + \frac{1}{n} + C \rightarrow C$$

and

$$P_\epsilon^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR} > 0 \text{ as } R > C.$$

Achievability: Code Construction



$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix}$$

- Fix $p(x)$. Generate a $(2^{nR}, n)$ code at random according to $p(x)$

$$p(x^n) = \prod_{i=1}^{n} p(x_i)$$

Random coding

- The probability the we generate a particular code C is

$$\Pr(C) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(w))$$

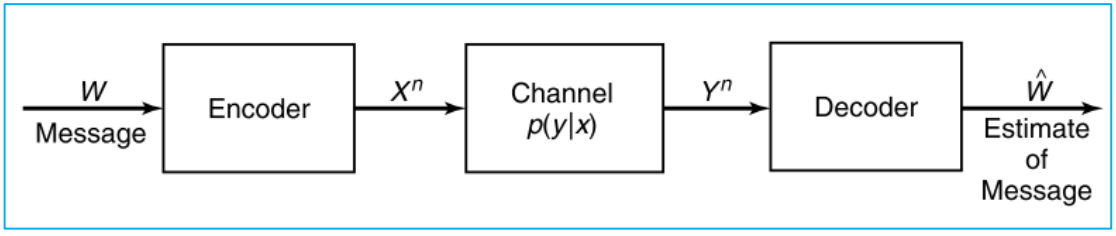
- The code C is revealed to both the sender and the receiver. Both them know $p(y|x)$
- A message W is chosen according to a **uniform distribution**

$$\Pr(W = w) = 2^{-nR}, w = 1, 2, \ldots, 2^{nR}$$

- The wth codeword $X^n(w)$ is sent over the channel
- The receiver receives a sequence Y^n according to the distribution

$$P(y^n|x^n(w)) = \prod_{i=1}^{N} p(y_i|x_i(w))$$

Achievability: Joint Decoding



The receiver guess which message was sent. In jointly typical decoding, the receiver declares that the index $\widehat{W}$ was sent if the following conditions are satisfied:

- $(X^n(\widehat{W}), Y^n)$ is jointly typical
- There is no other index $W' \neq W$, such that $(X^n(W'), Y^n) \in A_\epsilon^{(n)}$.

If no such $\widehat{W}$ exists or if there is more than one such, an error is declared. (We may assume that the receiver outputs a dummy index such as 0 in this case.)

- Let $\mathcal{E}$ be the event $\{\widehat{W} \neq W\}$
- **We need to show that**

$$\mathbf{Pr}(\mathcal{E}) \to \mathbf{0}$$

$\Pr(\mathcal{E}) \to 0$

Main idea: If we could prove that for all the codebook (all the possible C), the average $\Pr(\mathcal{E}) \le \epsilon$; then **the error probability of the best code** (one of $C's$) $\le \epsilon$

- We let W be drawn according to a uniform distribution over $\{1, 2, \ldots, 2^{nR}\}$ and use jointly typical decoding $\widehat{W}(y^n)$
- Let $\mathcal{E} = \{\widehat{W}(y^n) \neq W\}$ denote the error event
- We will calculate the average probability of error, averaged over all codewords in the codebook, and averaged over all codebooks

$$\Pr(\mathcal{E}) = \sum_C \Pr(C) P_e^{(n)}(C)$$

$$= \sum_C \Pr(C) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C)$$

$$= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_C \Pr(C)\, \lambda_w(C)$$

$$\sum_C \Pr(C)\, \lambda_1(C) = \Pr(\mathcal{E}|W=1)$$

$$\Pr(\mathcal{E}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \Pr(\mathcal{E}|W=w)$$

Take $\Pr(\mathcal{E}|W=1)$ for example

$$E_i = \left\{ \left((X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)} \right), i \in \{1,2,\ldots,2^{nR}\} \right\}$$

$$\Pr(\mathcal{E}|W=1) = P(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}|W=1)$$

$$\le P(E_1^c|W=1) + \sum_{i=2}^{2^{nR}} P(E_i|W=1)$$

$\Pr(\mathcal{E}) \to 0$ (cont'd)

$$\boldsymbol{\Pr(\mathcal{E}|W=1) \le P(E_1^c|W=1) + \sum_{i=2}^{2^{nR}} P(E_i|W=1)}$$

- By Joint AEP, $P(E_1^c|W=1) \to 0$, and hence $\mathrm{P}(\mathrm{E}_1^{\mathrm{c}}|\mathrm{W}=1) \le \epsilon$, for n sufficiently large
- For $i \ge 2$, $(E_i|W=1)$: Since by the code generation process, $X^n(1)$ and $X^n(i)$ are independent for $i \ne 1$, so are Y^n and $X^n(i)$. Hence, the probability that $X^n(i)$ and Y^n are jointly typical is $\le 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP

$$\begin{aligned}\Pr(\mathcal{E}|W=1) &\le \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + \left(2^{nR}-1\right)2^{-n(I(X;Y)-3\epsilon)} \\ &\le \epsilon + 2^{nR}2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)}\end{aligned}$$

If n is sufficiently large and $R < I(X;Y) - 3\epsilon$,

$$\Pr(\mathcal{E}|W=1) \le 2\epsilon$$

$$\Pr(\mathcal{E}) \le 2\epsilon$$

Choose $p(x)$ in the proof to be $p^*(x)$, the distribution on X that achieving capacity. Then

$$\boldsymbol{R \le I(X^*;Y) = C}$$

$$\boldsymbol{\lambda^{(n)} \le 4\epsilon}$$

$\Pr(\mathcal{E}) \to 0 \Rightarrow \lambda^{(n)} \to 0$

There exists a best codebook C^* such that

$$\Pr(\mathcal{E}|C^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(C^*) \leq 2\epsilon$$

By the definition of $(n, 2^{nR})$ code, we need to further show that

$$\lambda^{(n)} \to 0$$

Without loss of generality, assume

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{2^{nR}}$$

By $Pr(\mathcal{E}|C^*) \leq 2\epsilon$, we have

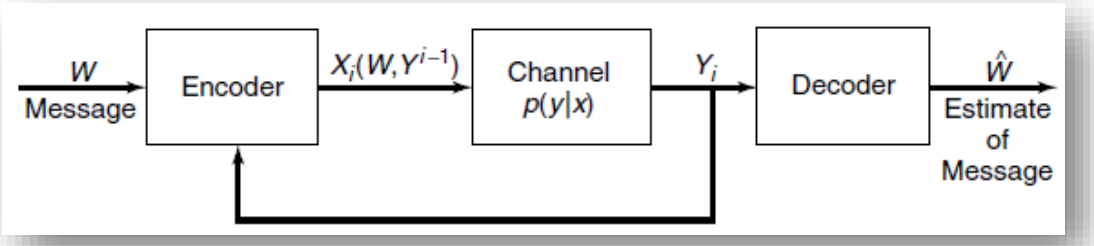
$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{2^{nR-1}} \leq 4\epsilon$$

(Or $\lambda_{2^{nR-1}} > 4\epsilon$, $\frac{1}{2^{nR}} \sum_{i=1+2^{nR-1}}^{2^{nR}} \lambda_i(C^*) > \frac{1}{2} 4\epsilon = 2\epsilon$, contradiction!)

Further refine the codebook C^*

- Throw away the worst half of the codewords in the best codebook C^*
- The best half of the codewords have a maximal probability of error less than 4ϵ
- If we reindex these codewords, we have 2^{nR-1} codewords. Throwing out half the codewords has changed the rate from R to $R - \frac{1}{n}$, which is negligible for large n

Feedback Capacity



- DMC: X_i is determined by $W, X_1, X_2, \ldots, X_{i-1}$
- $Y_1, Y_2, \ldots Y_{i-1}$ could be used to encode with $W, X_1, X_2, \ldots, X_{i-1}$ to determine X_i

- We assume that all the received symbols are sent back immediately and noiselessly to the transmitter, which can then use them to decide which symbol to send next
- We define a $(2^{nR}, n)$ *feedback code* as a sequence of mappings $x_i(W, Y^{i-1})$, where each x_i is a function only of the message $W \in 2^{nR}$ and the previous received values, $Y_1, Y_2, \ldots, Y_{i-1}$, and a sequence of decoding functions $g: \mathcal{Y}^n \rightarrow \{1,2, \ldots, 2^{nR}\}$. Thus,

$$P_e^{(n)} = \Pr(g(Y^n) \neq W)$$

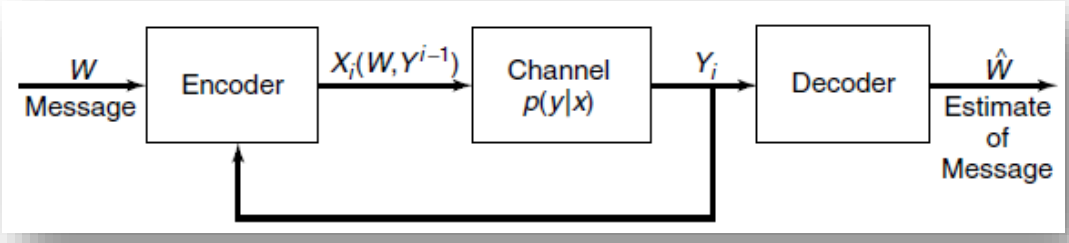
when W is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$.

- Feedback capacity

$$\boldsymbol{C_{FB} = C = \max_{p(x)} I(X;Y)}$$

Feedback cannot increase capacity

Feedback Can't Increase Capacity



$$nR = H(W) = H(W|\hat{W}) + I(W;\hat{W})$$
$$\leq 1 + P_e^{(n)} nR + I(W;\hat{W})$$
$$\leq 1 + P_e^{(n)} nR + I(W;Y^n)$$

$$I(W;Y^n) = H(Y^n) - H(Y^n|W)$$
$$= H(Y^n) - \sum_i H(Y_i|Y_1, Y_2, \dots, Y_{i-1}, W)$$
$$= H(Y^n) - \sum_i H(Y_i|Y_1, Y_2, \dots, Y_{i-1}, W, X_i)$$
$$= H(Y^n) - \sum_i H(Y_i|X_i)$$

since X_i is a function of $Y_1, Y_2, \dots, Y_{i-1}$ and W; and conditional on X_i, Y_i is independent of W and past samples of Y.

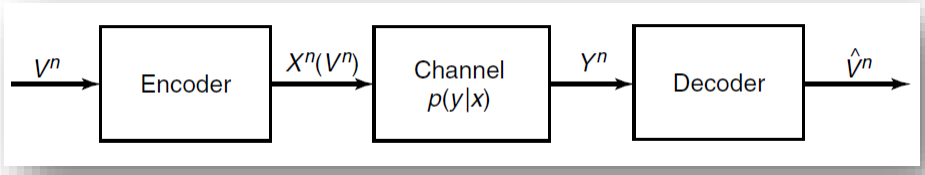
$$I(W;Y^n) = H(Y^n) - \sum_i H(Y_i|X_i)$$
$$\leq \sum_i H(Y_i) - \sum_i H(Y_i|X_i)$$
$$= \sum_i I(X_i;Y_i)$$
$$\leq nC$$

$$nR \leq P_e^{(n)} nR + 1 + nC$$

$$\boldsymbol{R \leq C}$$

In DMC: $\boldsymbol{I(X^n;Y^n) \leq \sum I(X_i;Y_i)}$

Source-Channel Separation



- In data compression: $R > H$
- In data transmission: $R < C$

⇒ Is the condition $H < C$ sufficient and necessary?

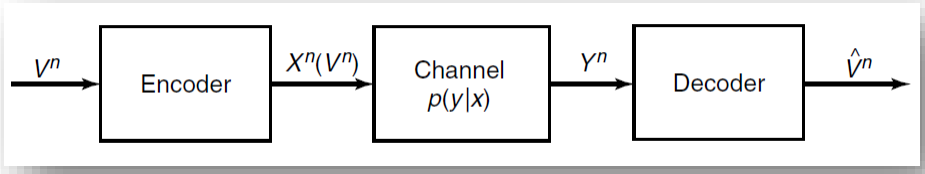
- We want to send the sequence of symbols $V^n = V_1, V_2, \ldots, V_n$ over the channel so that the receiver can reconstruct the sequence
- To do this, we map the sequence onto a codeword $X^n(V^n)$ and send the codeword over the channel
- The receiver looks at his received sequence Y^n and makes an estimate $\hat{V}^n$ of the sequence V^n that was sent. The receiver makes an error if $V^n \neq \hat{V}^n$. We define the probability of error as

$$\Pr(V^n \neq \hat{V}^n) = \sum_{y^n} \sum_{v^n} p(v^n) p(y^n | x^n(v^n)) I(g(y^n) \neq v^n)$$

Where I is the indicator function and $g(y^n)$ is the decoding function

$H < C$ is sufficient and necessary

Source-Channel Separation Theorem



Theorem (Source–channel coding theorem). If $V_1, V_2, \ldots, V_n$ is a finite alphabet stochastic process that satisfies the AEP and $H(\mathcal{V}) < C$, there exists a source–channel code with probability of error $\Pr(\hat{V}^n \neq V^n) \to 0$. Conversely, for any stationary stochastic process, if $H(\mathcal{V}) > C$, the probability of error is bounded away from zero, and it is not possible to send the process over the channel with arbitrarily low probability of error.

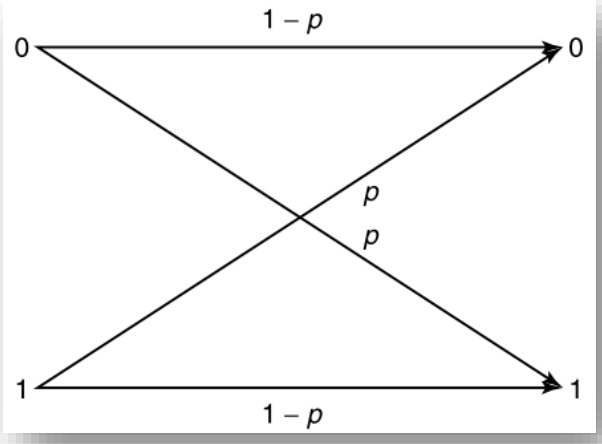
Converse: We wish to show that $\Pr(\hat{V}^n \neq V^n) \to 0$ implies that $H(\mathcal{V}) \leq C$ for any sequence of source-channel codes

$$X^n(V^n): \mathcal{V}^n \to \mathcal{X}^n$$
$$g_n(Y^n): \mathcal{Y}^n \to \mathcal{V}^n$$
$$H\left(V^n \middle| \hat{V}^n\right) \leq 1 + \Pr\left(\hat{V}^n \neq V^n\right) \log|\mathcal{V}^n| = 1 + \Pr\left(\hat{V}^n \neq V^n\right) n \log|\mathcal{V}|$$
$$H(\mathcal{V}) \leq \frac{H(V_1, V_2, \ldots, V_n)}{n} = \frac{H(V^n)}{n} = \frac{1}{n} H\left(V^n \middle| \hat{V}^n\right) + \frac{1}{n} I\left(V^n; \hat{V}^n\right)$$
$$\leq \frac{1}{n}\left(1 + \Pr\left(\hat{V}^n \neq V^n\right) n \log|\mathcal{V}|\right) + \frac{1}{n} I\left(V^n; \hat{V}^n\right)$$
$$\leq \frac{1}{n} + \Pr\left(\hat{V}^n \neq V^n\right) \log|\mathcal{V}| + \frac{1}{n} I(X^n; Y^n) \to C$$

Error Correction Code

The object of coding is to introduce redundancy so that even if some of the information is lost or corrupted, it will still be possible to recover the message at the receiver.

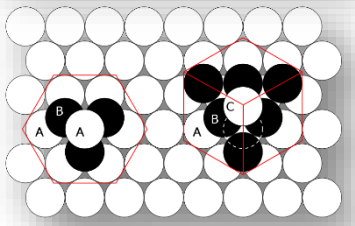
- Repetition code: For example, to send a 1, we send 11111, and to send a 0, we send 00000. The decoding scheme is to take the majority vote.
- Parity check code: Starting with a block of $n-1$ information bits, we choose the nth bit so that the parity of the entire block is 0.
- The code does not detect an even number of errors and does not give any information about how to correct the errors that occur.

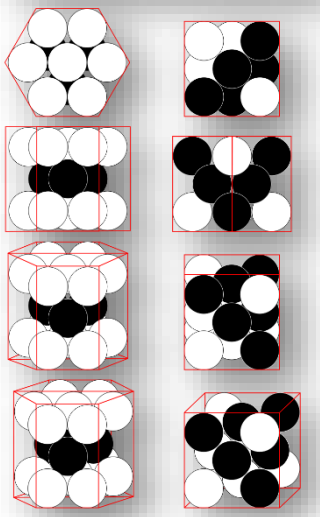


- Suppose that we use a length n 0-1 string x to denote message.
- Then after x is transmitted through the BSC, by the law of large number, about np bits will be changed. Denote the symbol received by y
- The **distance** between x and y is $d(x,y) \leq np$

All the points y are within the sphere with center x and radius np

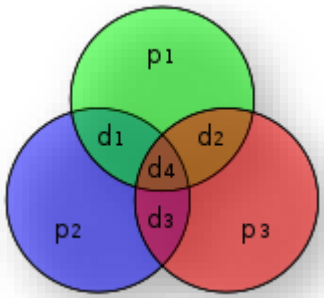
Hamming Code

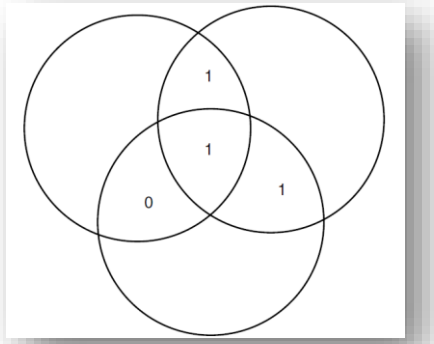


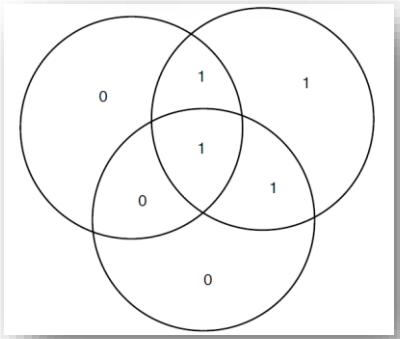


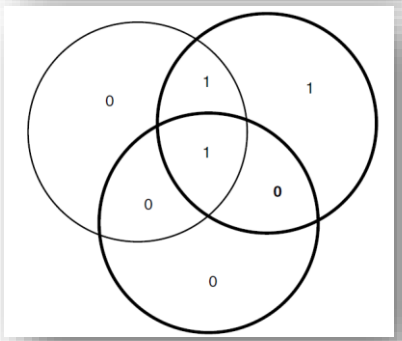
- Deonte the codeword by x, then the noisy version y of x stays inside the sphere with center x and radius r
- Sphere packing: the art of error correction code
- https://en.wikipedia.org/wiki/Sphere_packing

- (n, k, d) Hamming code: the first k bits in each codeword represent the message, and the last $n - k$ bits are parity check bits
- $n = 2^l - 1, k = 2^l - l - 1, d = 3$: (7,4,3) example









- BCH code
- Convolutional code
- Turbo code
- LDPC code
- Polar code

Ref:

1. Algebraic Coding Theory, E. Berlekamp
2. Sphere Packings, Lattices and Groups, J. H. Conway, N. J.A. Sloane
3. Modern Coding Theory, R. Urbanke and T. Richardson

Summary

Cover: 7.4, 7.5, 7.6, 7.7, 7.8, 7.9, 7.10