CS258: Information Theory

Fan Cheng

Spring, 2018. chengfan@sjtu.edu.cn

Recap: fundamental information quantities

- Definition of KL-divergence, entropy, mutual information
- Information diagram: relationship of entropy, mutual information, etc.
- Some fundamental properties: non-negative, convexity/concativy, inequalities
- How to solve problems via information quantities

Lecture 3: Error Correcting

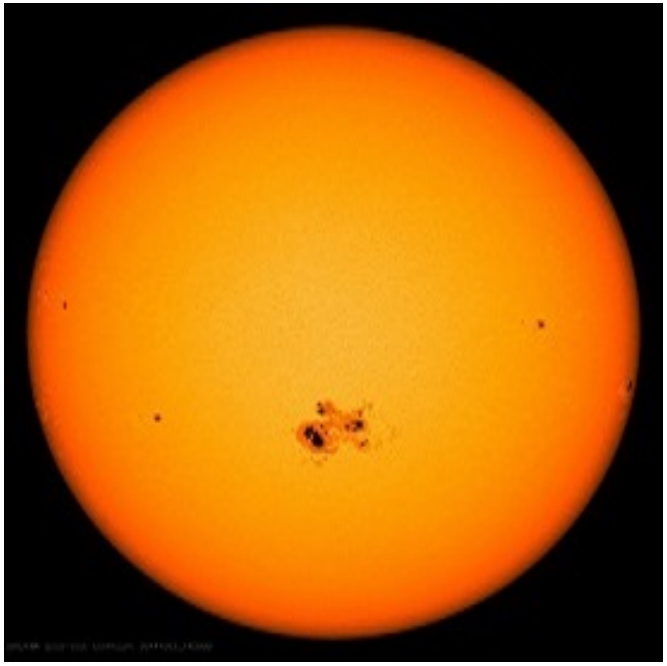
- Repetition code
- Hamming code
- Decomposability of entropy

Noisy world

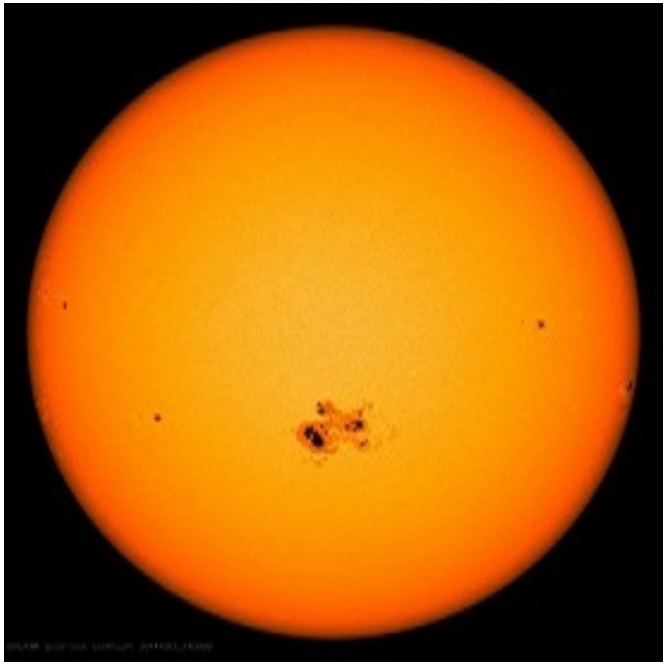
Noisy world

Noisy world

Noisy world

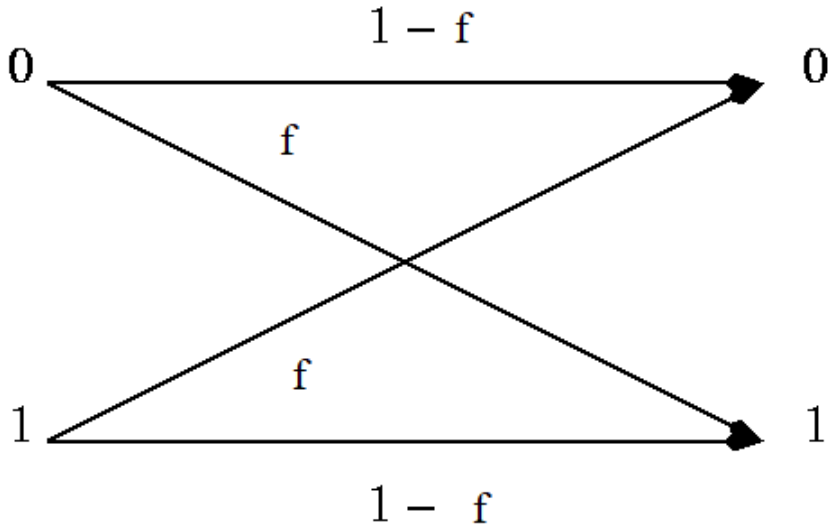


Noisy world



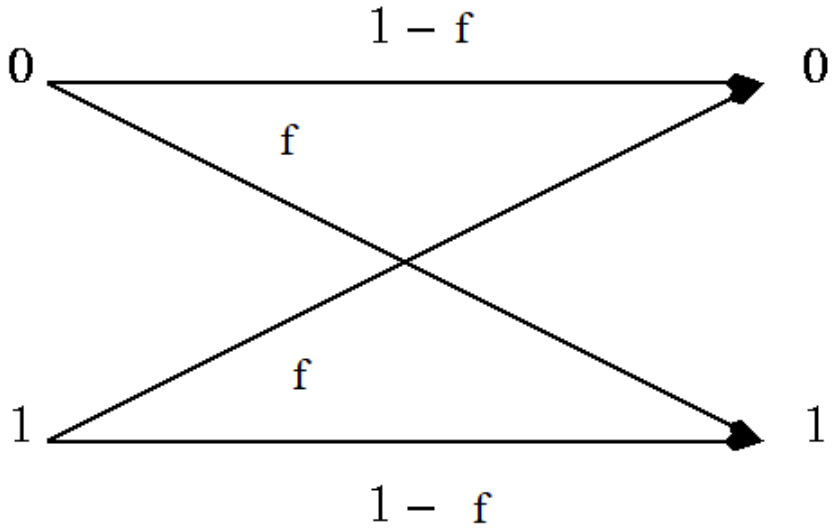
We need to correct them!

Mathematical model: binary symmetric channel



Binary symmetric channel (BSC): A message from alphabet $\{0, 1\}$ is sent through a noisy channel with flipping probability f.

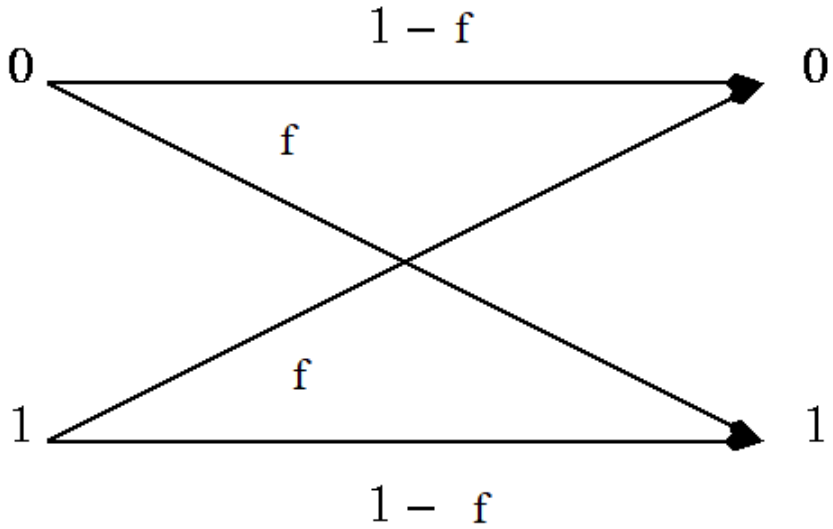
Mathematical model: binary symmetric channel



Binary symmetric channel (BSC): A message from alphabet $\{0,1\}$ is sent through a noisy channel with flipping probability f.

Equivalently, BSC can be written as: $Y = X + Z \pmod 2$, where $X, Z \in \{0,1\}$

Mathematical model: binary symmetric channel

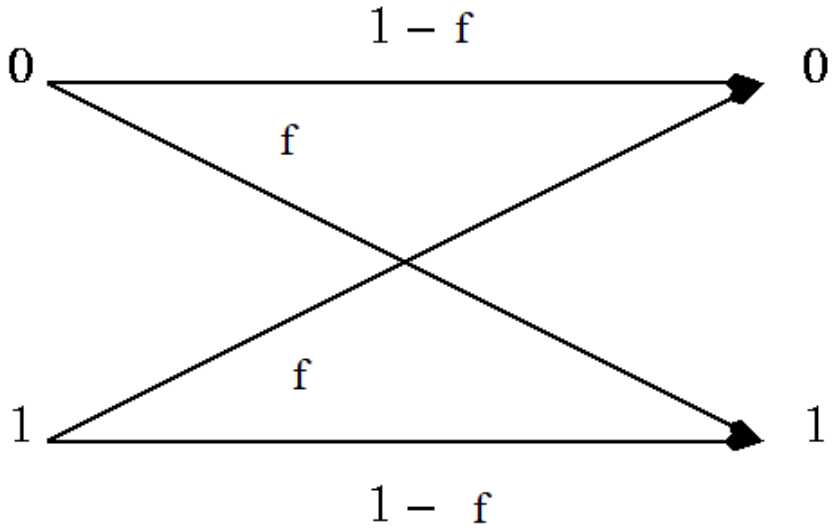


Binary symmetric channel (BSC): A message from alphabet $\{0,1\}$ is sent through a noisy channel with flipping probability f.

Equivalently, BSC can be written as: $Y = X + Z \pmod 2$, where $X, Z \in \{0,1\}$

For example: $01010101 \rightarrow 01110110$

Mathematical model: binary symmetric channel



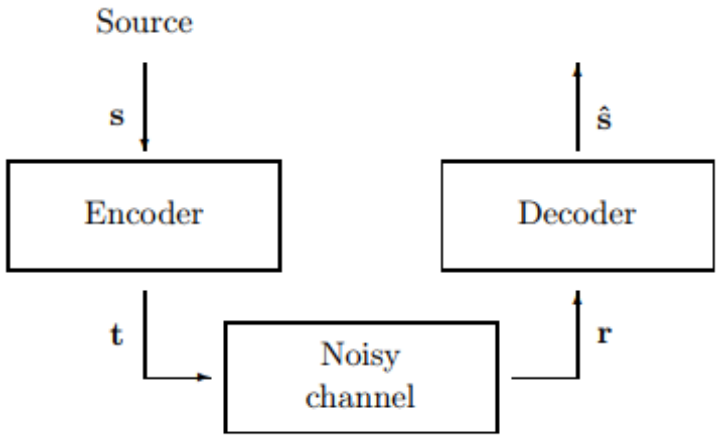
Binary symmetric channel (BSC): A message from alphabet $\{0,1\}$ is sent through a noisy channel with flipping probability f.

Equivalently, BSC can be written as: $Y = X + Z \pmod 2$, where $X, Z \in \{0,1\}$

For example: $01010101 \to 01110110$

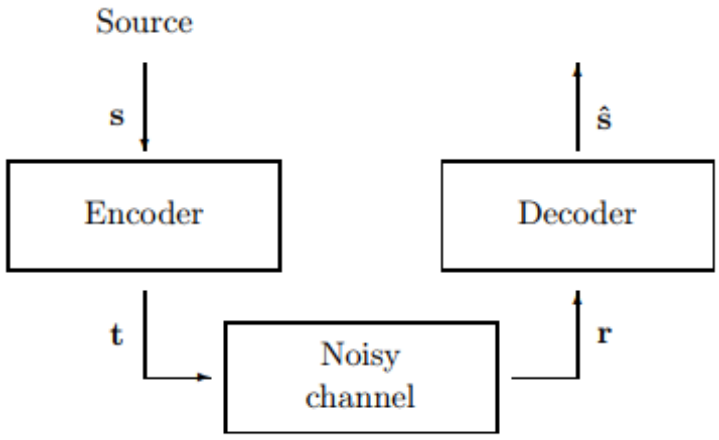
Question: how to transmit a message with very low error probability (e.g. 10^{-15})?

A system solution



System solution can turn noisy channels into reliable communication channels with the only cost being a *computational* requirement at the encoder and decoder.

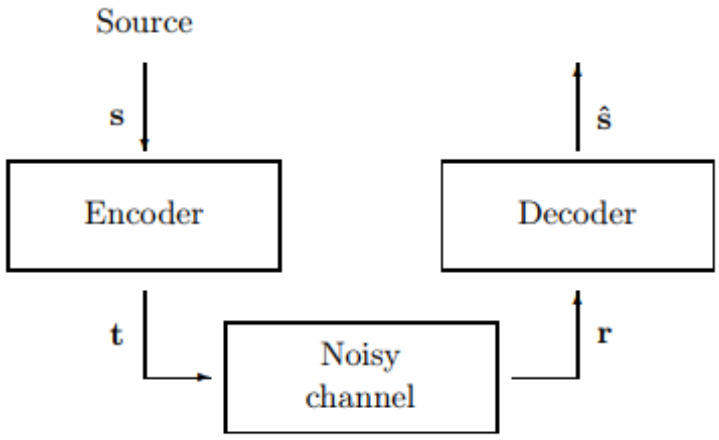
A system solution



System solution can turn noisy channels into reliable communication channels with the only cost being a *computational* requirement at the encoder and decoder.

Information theory: What is the best error-correcting performance we could achieve?

A system solution



System solution can turn noisy channels into reliable communication channels with the only cost being a *computational* requirement at the encoder and decoder.

Information theory: What is the best error-correcting performance we could achieve?
Coding theory: The creation of practical encoding and decoding systems.

Repetition codes

R_k: Repeat each bit k times; e.g., $R_3 :=$

s	t
0	000
1	111

Repeat every bit of the message a prearranged number of times

Repetition codes

R_k: Repeat each bit k times; e.g., $R_3 :=$

s	t
0	000
1	111

Repeat every bit of the message a prearranged number of times

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000

Repetition codes

R_k: Repeat each bit k times; e.g., $R_3 :=$

s	t
0	000
1	111

Repeat every bit of the message a prearranged number of times

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000

Encoding is obvious. How to decoding? Majority vote? Is it optimal?

Decoding

Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given $\mathbf{r}$; i.e., $\max_s P(s|\mathbf{r})$.

Decoding

Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given $\mathbf{r}$; i.e., $\max_s P(s|\mathbf{r})$.

Recall: $s \to t \to r \to \hat{s}$

Decoding

Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given $\mathbf{r}$; i.e., $\max_s P(s|\mathbf{r})$.

Recall: $s \to t \to r \to \hat{s}$

- [1.] If $P(s=0) = P(s=1) = 0.5$, $\max_s P(s|\mathbf{r}) = \max_s P(\mathbf{r}|s)$.

Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given $\mathbf{r}$; i.e., $\max_s P(s|\mathbf{r})$.

Recall: $s \to t \to r \to \hat{s}$

- [1.] If $P(s=0) = P(s=1) = 0.5$, $\max_s P(s|\mathbf{r}) = \max_s P(\mathbf{r}|s)$.

Proof.

By Bayes' theorem, $P(s|\mathbf{r}) = \frac{P(r,s)}{P(\mathbf{r})} = \frac{P(s)P(\mathbf{r}|s)}{P(\mathbf{r})}$ □

Decoding

Optimal decoding

The optimal decoding decision is to find which value of s is most probable, given $\mathbf{r}$; i.e., $\max_s P(s|\mathbf{r})$.

Recall: $s \to t \to r \to \hat{s}$

- [1.] If $P(s=0) = P(s=1) = 0.5$, $\max_s P(s|\mathbf{r}) = \max_s P(\mathbf{r}|s)$.

Proof.

By Bayes' theorem, $P(s|\mathbf{r}) = \frac{P(r,s)}{P(\mathbf{r})} = \frac{P(s)P(\mathbf{r}|s)}{P(\mathbf{r})}$ □

- [2.] Assume $f < 0.5$, the winning hypothesis is the one with the most 'votes'.

Proof.

$$P(\mathbf{r}|s) = P(\mathbf{r}|t(s)) = \prod_{n=1}^{N} P(r_n|t_n(s))$$

Proof.

$$P(\mathbf{r}|s) = P(\mathbf{r}|t(s)) = \prod_{n=1}^{N} P(r_n|t_n(s))$$

$$P(r_n|t_n(s)) = \begin{cases} 1-f, & \text{if } r_n = t_n; \\ f, & \text{if } r_n \neq t_n. \end{cases}$$

□

Proof.

$$P(\mathbf{r}|s) = P(\mathbf{r}|t(s)) = \prod_{n=1}^{N} P(r_n|t_n(s))$$

$$P(r_n|t_n(s)) = \begin{cases} 1-f, & \text{if } r_n = t_n; \\ f, & \text{if } r_n \neq t_n. \end{cases}$$

The likelihood ratio (Since $s \in \{0,1\}$, likelihood ratio test can tell us the optimal solution) for the two hypotheses is

$$\frac{P(\mathbf{r}|s=1)}{P(\mathbf{r}|s=0)} = \prod_{n=1}^{N} \frac{P(r_n|t_n(1))}{P(r_n|t_n(0))}$$

Proof.

$$P(\mathbf{r}|s) = P(\mathbf{r}|t(s)) = \prod_{n=1}^{N} P(r_n|t_n(s))$$

$$P(r_n|t_n(s)) = \begin{cases} 1-f, & \text{if } r_n = t_n; \\ f, & \text{if } r_n \neq t_n. \end{cases}$$

The likelihood ratio (Since $s \in \{0,1\}$, likelihood ratio test can tell us the optimal solution) for the two hypotheses is

$$\frac{P(\mathbf{r}|s=1)}{P(\mathbf{r}|s=0)} = \prod_{n=1}^{N} \frac{P(r_n|t_n(1))}{P(r_n|t_n(0))}$$

Each factor $\frac{P(r_n|t_n(1))}{P(r_n|t_n(0))}$ equals $\frac{1-f}{f}$ if $r_n = 1$ and $\frac{f}{1-f}$ if $r_n = 0$. Thus majority vote is optimal if $f < 0.5$ □

Majority vote decoder

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{s}$	0	0	1	0	0	1	0

Majority vote decoder

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{s}$	0	0	1	0	0	1	0

Not all the errors can be detected.

Majority vote decoder

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{s}$	0	0	1	0	0	1	0

Not all the errors can be detected.

- Error probability can be reduced by R_k
- Rate of information transfer has fallen by a factor of k ($k \to \infty$, :( )

Majority vote decoder

An example transmission using R_3

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{s}$	0	0	1	0	0	1	0

Not all the errors can be detected.

- Error probability can be reduced by R_k
- Rate of information transfer has fallen by a factor of k ($k \to \infty$, :( )

Proper redundancy is needed to reliable communication. Tradeoff exists between rate of information and error probability.

Hamming code

A **block code** is a rule for converting a sequence of source bits s, of length K, into a transmitted sequence of length N bits. The extra $N - K$ bits are linear functions of the original K bits, called **parity-check bits**.

Hamming code

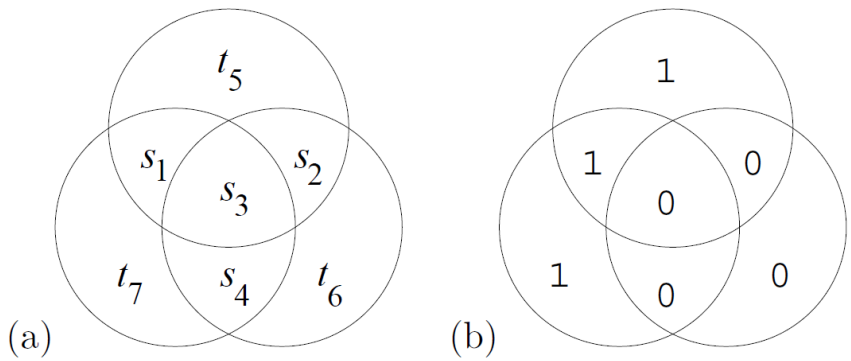
A **block code** is a rule for converting a sequence of source bits s, of length K, into a transmitted sequence of length N bits. The extra $N-K$ bits are linear functions of the original K bits, called **parity-check bits**.

Rate of information: K/N

Hamming code

A **block code** is a rule for converting a sequence of source bits s, of length K, into a transmitted sequence of length N bits. The extra $N-K$ bits are linear functions of the original K bits, called **parity-check bits**.

Rate of information: K/N

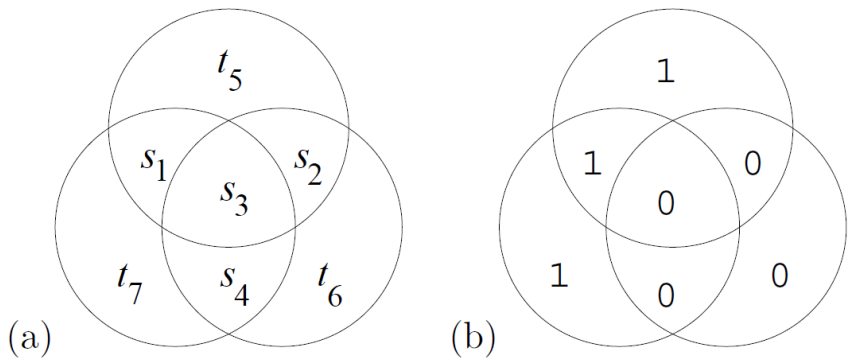


Encoder of (7,4) Hamming code:

Hamming code

A **block code** is a rule for converting a sequence of source bits s, of length K, into a transmitted sequence of length N bits. The extra $N-K$ bits are linear functions of the original K bits, called **parity-check bits**.

Rate of information: K/N



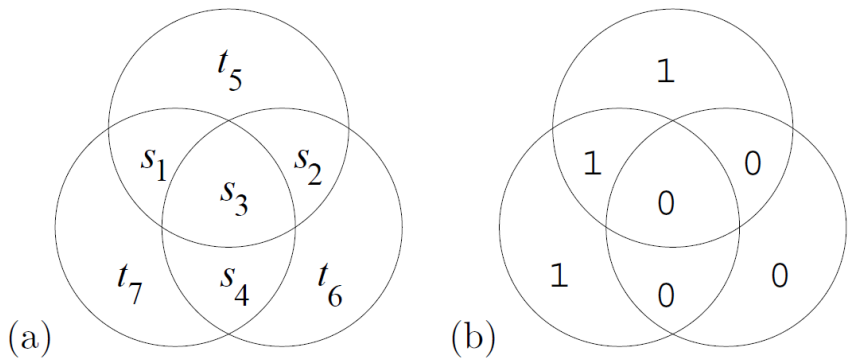
Encoder of (7,4) Hamming code:

- $t_1 t_2 t_3 t_4$ are set equal to $s_1 s_2 s_3 s_4$

Hamming code

A **block code** is a rule for converting a sequence of source bits s, of length K, into a transmitted sequence of length N bits. The extra $N-K$ bits are linear functions of the original K bits, called **parity-check bits**.

Rate of information: K/N



Encoder of (7,4) Hamming code:

- $t_1 t_2 t_3 t_4$ are set equal to $s_1 s_2 s_3 s_4$
- The parity-check bits $t_5 t_6 t_7$ are set so that the parity within each circle is even

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{\mathbf{t}\}$ of the (7, 4) Hamming code.

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{\mathbf{t}\}$ of the (7, 4) Hamming code.

Any pair of codewords differ from each other in at least three bits.

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{\mathbf{t}\}$ of the (7, 4) Hamming code.

Any pair of codewords differ from each other in at least three bits.

Matrix form: $t = G^t s$ (or $t = sG$),
where G is the generator matrix of the code.

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{\mathbf{t}\}$ of the (7, 4) Hamming code.

Any pair of codewords differ from each other in at least three bits.

Matrix form: $t = G^t s$ (or $t = sG$),
where G is the generator matrix of the code.

- Higher information rate

Codewords of (7,4) Hamming code

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

The sixteen codewords $\{\mathbf{t}\}$ of the (7, 4) Hamming code.

Any pair of codewords differ from each other in at least three bits.

Matrix form: $t = G^t s$ (or $t = sG$),
where G is the generator matrix of the code.

- Higher information rate
- More complicated encoder

Matrix form of Hamming (7,4) code

$$t = sG,$$

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Decoding the (7,4) Hamming code

Facts:

Decoding the (7,4) Hamming code

Facts:

- $s \to t \to r$, any of the bits may have been flipped, including the parity bits

Decoding the (7,4) Hamming code

Facts:

- $s \to t \to r$, any of the bits may have been flipped, including the parity bits
- If we assume that the channel is BSC and all the source vector s are equiprobable, then the optimal decoder identifies the source vector s whose encoding $t(s)$ differs from the received vector r in the fewest bits.

Decoding the (7,4) Hamming code

Facts:

- $s \to t \to r$, any of the bits may have been flipped, including the parity bits
- If we assume that the channel is BSC and all the source vector s are equiprobable, then the optimal decoder identifies the source vector s whose encoding $t(s)$ differs from the received vector r in the fewest bits.
- We could solve the decoding problem by measuring how far r is from each of the sixteen codewords, then picking the closest.

Decoding the (7,4) Hamming code

Facts:

- $s \to t \to r$, any of the bits may have been flipped, including the parity bits
- If we assume that the channel is BSC and all the source vector s are equiprobable, then the optimal decoder identifies the source vector s whose encoding $t(s)$ differs from the received vector r in the fewest bits.
- We could solve the decoding problem by measuring how far r is from each of the sixteen codewords, then picking the closest.

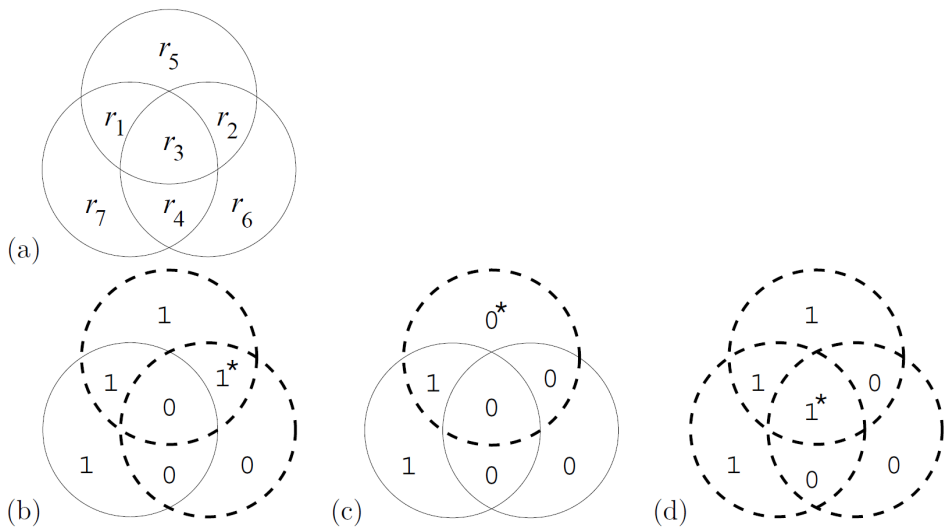
Not efficient!

Syndrome decoding for the Hamming code

The pattern of violations of the parity checks is called the syndrome, and can be written as a binary vector (In Fig. b, the syndrome is $z = (1, 1, 0)$). (Syndrome: Happy (parity 0) and unhappy (parity 1))

Syndrome decoding

Find a unique bit that lies inside all the 'unhappy' circles and outside all the 'happy' circles. Flip this bit for correction.

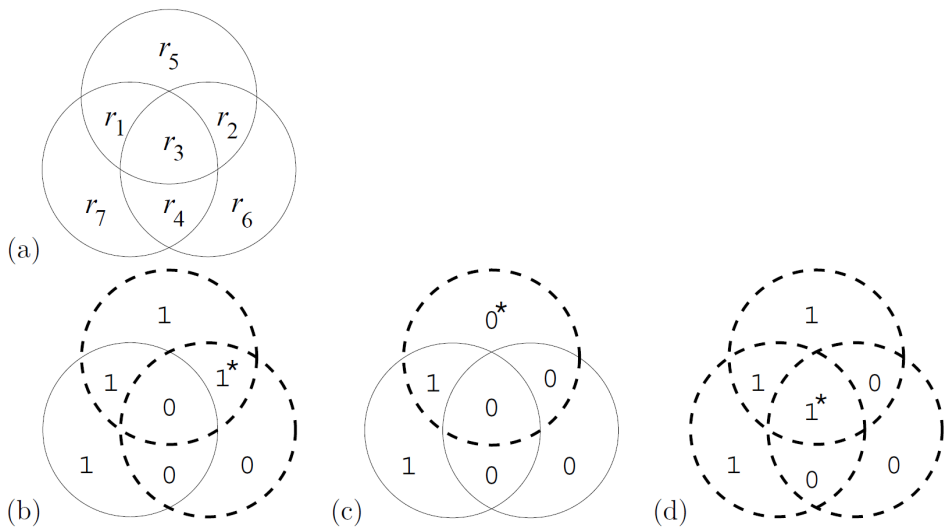


Syndrome decoding for the Hamming code

The pattern of violations of the parity checks is called the syndrome, and can be written as a binary vector (In Fig. b, the syndrome is $z = (1, 1, 0)$). (Syndrome: Happy (parity 0) and unhappy (parity 1))

Syndrome decoding

Find a unique bit that lies inside all the 'unhappy' circles and outside all the 'happy' circles. Flip this bit for correction.



Syndrome z	000	001	010	011	100	101	110	111
Unflip this bit	none	r_7	r_6	r_4	r_5	r_1	r_2	r_3

Decoding: matrix form of (7, 4) Hamming code

Denote

$$\mathbf{t} = \mathbf{sG}, \mathbf{G}^T = \begin{pmatrix} \mathbf{I_4} \\ \mathbf{P} \end{pmatrix},$$

where $\mathbf{I_4}$ is the identity matrix, then the syndrome vector $\mathbf{z} = \mathbf{Hr}$, where the parity-check matrix $\mathbf{H}$ is given by $\mathbf{H} = [-\mathbf{P}\ \mathbf{I_3}]$. Since $-1 \equiv 1 \pmod 2$, $\mathbf{H} = [\mathbf{P}\ \mathbf{I_3}]$.

Decoding: matrix form of (7, 4) Hamming code

Denote

$$\mathbf{t} = \mathbf{sG}, \mathbf{G}^T = \begin{pmatrix} \mathbf{I_4} \\ \mathbf{P} \end{pmatrix},$$

where $\mathbf{I_4}$ is the identity matrix, then the syndrome vector $\mathbf{z} = \mathbf{Hr}$, where the parity-check matrix $\mathbf{H}$ is given by $\mathbf{H} = [-\mathbf{P}\ \mathbf{I_3}]$. Since $-1 \equiv 1 \pmod 2$, $\mathbf{H} = [\mathbf{P}\ \mathbf{I_3}]$.

G and H

All the codewords $\mathbf{t} = \mathbf{G^t s}$ of the code satisfy $\mathbf{Ht} = \mathbf{0}$, i.e., $\mathbf{HG^t} = \mathbf{0}$
In general, $\mathbf{G} = \left[\mathbf{I}_k | \mathbf{P}\right], \mathbf{H} = \left[-\mathbf{P}^\top | \mathbf{I}_{n-k}\right]$.

Decoding: matrix form of (7, 4) Hamming code

Denote

$$\mathbf{t} = \mathbf{sG}, \mathbf{G}^T = \begin{pmatrix} \mathbf{I_4} \\ \mathbf{P} \end{pmatrix},$$

where $\mathbf{I_4}$ is the identity matrix, then the syndrome vector $\mathbf{z} = \mathbf{Hr}$, where the parity-check matrix $\mathbf{H}$ is given by $\mathbf{H} = [-\mathbf{P}\ \mathbf{I_3}]$. Since $-1 \equiv 1 \pmod 2$, $\mathbf{H} = [\mathbf{P}\ \mathbf{I_3}]$.

$\mathbf{G}$ and $\mathbf{H}$

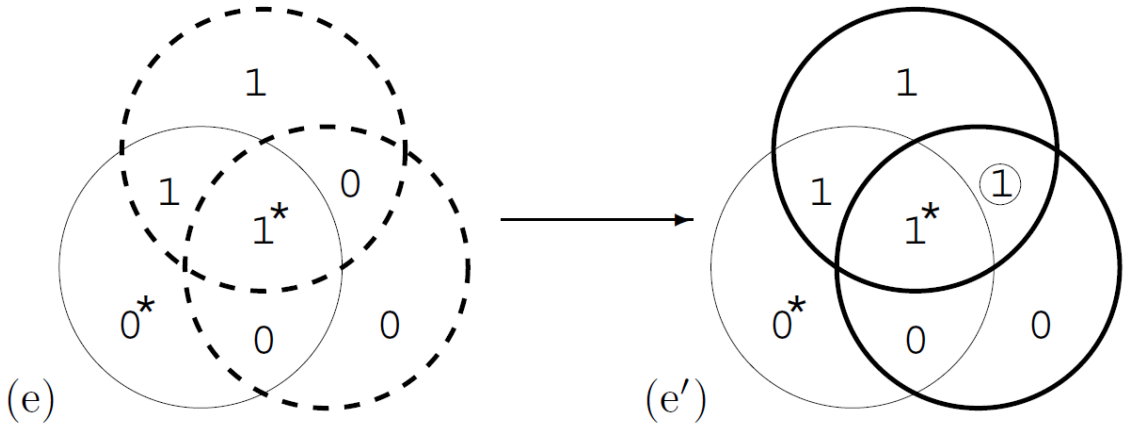
All the codewords $\mathbf{t} = \mathbf{G^t s}$ of the code satisfy $\mathbf{Ht} = \mathbf{0}$, i.e., $\mathbf{HG^t} = \mathbf{0}$
In general, $\mathbf{G} = \left[\mathbf{I}_k | \mathbf{P}\right], \mathbf{H} = \left[-\mathbf{P}^\top | \mathbf{I}_{n-k}\right]$.

Since the received vector $\mathbf{r} = \mathbf{G^t s} + \mathbf{n}$, the syndrome-decoding problem is to find the most probable noise vector n satisfying the equation

$$\mathbf{Hn} = \mathbf{z}$$

A decoding algorithm that solves this problem is called a *maximum-likelihood decoder*.

More than two bits are flipped



$$1000101 \rightarrow 1010100$$

When two bits, r_3 and r_7, are received flipped. The syndrome, 110, makes us suspect the single bit r_2; so our optimal decoding algorithm flips this bit. If we use the optimal decoding algorithm, any two-bit error pattern will lead to a decoded seven-bit vector that contains three errors.

$$1000101 \rightarrow 1010100 \rightarrow 1110100$$

Performance of best code

There is a tradeoff between p_b and R. The feasible region of (p_b, R) is solved by C. E. Shannon in 1948.

Performance of best code

There is a tradeoff between p_b and R. The feasible region of (p_b, R) is solved by C. E. Shannon in 1948.
The maximum rate at which communication is possible with arbitrarily small p_b is called the capacity of the channel.

Performance of best code

There is a tradeoff between p_b and R. The feasible region of (p_b, R) is solved by C. E. Shannon in 1948.
The maximum rate at which communication is possible with arbitrarily small p_b is called the capacity of the channel.

Capacity of BSC

$$C(f) = 1 - H_2(f)$$

Performance of best code

There is a tradeoff between p_b and R. The feasible region of (p_b, R) is solved by C. E. Shannon in 1948.
The maximum rate at which communication is possible with arbitrarily small p_b is called the capacity of the channel.

Capacity of BSC

$$C(f) = 1 - H_2(f)$$

Exercise

1.3, 1.6, 2.28, 2.29

Decomposability of the entropy

The entropy of any probability distribution $p = \{p_1, p_2, ..., p_I\}$ is that

$$H(p) = H(p_1, 1 - p_1) + (1 - p_1)H(p_2/(1 - p_1), ..., p_I/(1 - p_1))$$

Generalizing further,

$$\begin{aligned} H(p) = & H(p_1 + ... + p_m, p_{m+1} + ... + p_I) \\ & + (p_1 + ... + p_m)H(p_1/(p_1 + ... + p_m), ..., p_m/(p_1 + ... + p_m)) \\ & + (p_{m+1} + ... + p_I)H(p_{m+1}/(p_{m+1} + ... + p_I), ..., p_I/(p_{m+1} + ... + p_I)) \end{aligned}$$

An unbiased coin is flipped until one head is thrown. What is the entropy of the random variable $x \in \{1, 2, 3, ...\}$, the number of flips?

An unbiased coin is flipped until one head is thrown. What is the entropy of the random variable $x \in \{1, 2, 3, ...\}$, the number of flips?

$$H(X) = H_2(f) + (1 - f)H(X)$$