CS258: Information Theory

Fan Cheng

Spring, 2018. chengfan@sjtu.edu.cn

Recap: Huffman Coding

- Symbol code
- Unique decodeablity and suffix code
- Kraft inequality and its proof
- Huffman coding

- Noisy channel
- Useful model channels
- Channel coding theorem

Review

Review

- Entropy and mutual information

Review

- Entropy and mutual information
- Noisy world

Review

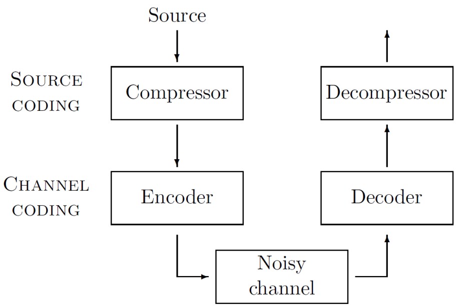
- Entropy and mutual information
- Noisy world
- Lossy source coding

Review

- Entropy and mutual information
- Noisy world
- Lossy source coding
- Channel capacity of BSC

Review

- Entropy and mutual information
- Noisy world
- Lossy source coding
- Channel capacity of BSC



system model

A simple example

The joint distribution of X and Y is

$P(x,y)$		x				$P(y)$
		1	2	3	4	
	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
y	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

A simple example

The joint distribution of X and Y is

$P(x,y)$		x				$P(y)$
		1	2	3	4	
	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
y	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

The conditional entropy of X given each y

$P(x \mid y)$		x				$H(X \mid y)$/bits
		1	2	3	4	
	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
y	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0

A simple example

The joint distribution of X and Y is

$P(x,y)$		x 1	2	3	4	$P(y)$
	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
y	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

The conditional entropy of X given each y

$P(x \mid y)$		x 1	2	3	4	$H(X \mid y)$/bits
	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
y	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0

Channel: fixed $P(Y|X)$

DMC

A **discrete** **memoryless** **channel** Q is characterized by an input alphabet $\mathcal{A}_X$, an output alphabet $\mathcal{A}_Y$, and a set of conditional probability distributions $P(y|x)$, one for each $x \in \mathcal{A}_X$. These transition probabilities may be written in a matrix

$$Q_{j|i} = P(y = b_j | x = a_i).$$

DMC

A **discrete memoryless channel** Q is characterized by an input alphabet $\mathcal{A}_X$, an output alphabet $\mathcal{A}_Y$, and a set of conditional probability distributions $P(y|x)$, one for each $x \in \mathcal{A}_X$. These transition probabilities may be written in a matrix

$$Q_{j|i} = P(y = b_j | x = a_i).$$

Channel may have memory, like tapes

Useful model channels

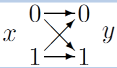
Binary symmetric channel. $\mathcal{A}_x = \{0,1\}$. $\mathcal{A}_Y = \{0,1\}$.

$$x \quad \begin{tikzcd}[row sep=small, column sep=small] 0 \arrow[r] \arrow[dr] & 0 \\ 1 \arrow[r] \arrow[ur] & 1 \end{tikzcd} \quad y$$

$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= f; \\ P(y=1 \mid x=0) &= f; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$

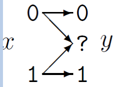
Useful model channels

Binary symmetric channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, 1\}$.



$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= f; \\ P(y=1 \mid x=0) &= f; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$

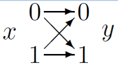
Binary erasure channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, ?, 1\}$.



$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= 0; \\ P(y=? \mid x=0) &= f; & P(y=? \mid x=1) &= f; \\ P(y=1 \mid x=0) &= 0; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$

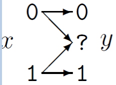
Useful model channels

Binary symmetric channel. $\mathcal{A}_X = \{0,1\}$. $\mathcal{A}_Y = \{0,1\}$.



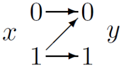
$$\begin{array}{rclrcl} P(y=0 \mid x=0) &=& 1-f; & P(y=0 \mid x=1) &=& f; \\ P(y=1 \mid x=0) &=& f; & P(y=1 \mid x=1) &=& 1-f. \end{array}$$

Binary erasure channel. $\mathcal{A}_X = \{0,1\}$. $\mathcal{A}_Y = \{0,?,1\}$.

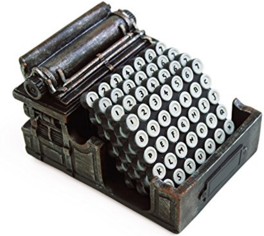


$$\begin{array}{rclrcl} P(y=0 \mid x=0) &=& 1-f; & P(y=0 \mid x=1) &=& 0; \\ P(y=? \mid x=0) &=& f; & P(y=? \mid x=1) &=& f; \\ P(y=1 \mid x=0) &=& 0; & P(y=1 \mid x=1) &=& 1-f. \end{array}$$

Z channel. $\mathcal{A}_X = \{0,1\}$. $\mathcal{A}_Y = \{0,1\}$.

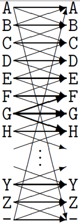


$$\begin{array}{rclrcl} P(y=0 \mid x=0) &=& 1; & P(y=0 \mid x=1) &=& f; \\ P(y=1 \mid x=0) &=& 0; & P(y=1 \mid x=1) &=& 1-f. \end{array}$$



Noisy typewriter. $\mathcal{A}_X = \mathcal{A}_Y =$ the 27 letters $\{$ A, B, . . . , Z, - $\}$. The letters are arranged in a circle, and when the typist attempts to type B, what comes out is either A, B or C, with probability $1/3$ each; when the input is C, the output is B, C or D; and so forth, with the final letter '-' adjacent to the first letter A.

Noisy typewriter. $\mathcal{A}_X = \mathcal{A}_Y =$ the 27 letters { A, B, . . . , Z, - }. The letters are arranged in a circle, and when the typist attempts to type B, what comes out is either A, B or C, with probability 1/3 each; when the input is C, the output is B, C or D; and so forth, with the final letter '-' adjacent to the first letter A.



$$
\begin{array}{rcl}
 & \vdots & \\
P(y = \mathtt{F} \mid x = \mathtt{G}) & = & 1/3; \\
P(y = \mathtt{G} \mid x = \mathtt{G}) & = & 1/3; \\
P(y = \mathtt{H} \mid x = \mathtt{G}) & = & 1/3; \\
 & \vdots &
\end{array}
$$

The capacity of a channel Q is:

$$C(Q) = \max_{\mathcal{P}_X} I(X;Y).$$

The distribution $\mathcal{P}_X$ that achieves the maximum is called the optimal input distribution, denoted by $\mathcal{P}_X^*$. [There may be multiple optimal input distributions achieving the same value of $I(X;Y)$.]

Real channels and Gaussian Channel

Consider a physical (electrical, say) channel with inputs and outputs that are continuous in time. We put in $x(t)$, and out comes $y(t) = x(t) + n(t)$. Our transmission has a power cost. The average power of a transmission of length T may be constrained thus

$$\int_0^T [x(t)]^2 / T dt \leq P.$$

Real channels and Gaussian Channel

Consider a physical (electrical, say) channel with inputs and outputs that are continuous in time. We put in $x(t)$, and out comes $y(t) = x(t) + n(t)$. Our transmission has a power cost. The average power of a transmission of length T may be constrained thus

$$\int_0^T [x(t)]^2 / T dt \leq P.$$

Gaussian Channel

The Gaussian channel has a real input x and a real output y. The conditional distribution of y given x is a Gaussian distribution:

$$P(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-(y-x)^2/2\sigma^2)$$

This channel is sometimes called the additive white Gaussian noise (AWGN) channel.

For a Gaussian channel with power constraint v and variance of noise σ^2, its channel capacity is

$$C = \max_{X:\mathrm{Var}(X)\leq v} I(X; X+Z) = \frac{1}{2}\log(1+\frac{v}{\sigma^2})$$

signal to noise ration: $\frac{v}{\sigma^2}$.

For a Gaussian channel with power constraint v and variance of noise σ^2, its channel capacity is

$$C = \max_{X:\mathrm{Var}(X)\leq v} I(X; X+Z) = \frac{1}{2}\log(1+\frac{v}{\sigma^2})$$

signal to noise ration: $\frac{v}{\sigma^2}$.
If we fixed v, C(Z) is minimized iff Z is gaussian.

For a Gaussian channel with power constraint v and variance of noise σ^2, its channel capacity is

$$C = \max_{X:\mathrm{Var}(X)\leq v} I(X;X+Z) = \frac{1}{2}\log(1+\frac{v}{\sigma^2})$$

signal to noise ration: $\frac{v}{\sigma^2}$.
If we fixed v, C(Z) is minimized iff Z is gaussian.
Entropy Power inequality (Shannon 1948) For any independent random variables X and Y

$$e^{2h(X+Y)} \geq e^{2h(X)} + e^{2h(Y)}$$

Equality holds iff X, Y are gaussian.

Reading and Exercise

Reading: Ch. 8, Ch. 9, Ch. 10, Ch. 11 (David MacKay)

Exercise

8.5, 8.7, 9.2, 9.4, 9.15, 9.17, 10.8, 10.9, 10.12