

**Solution to Final Examination**

1. (10 points) **Huffman code**

Find the Huffman binary codeword lengths for the probability assignment

$$\mathbf{p} = (1/45, 1/45, 6/45, 6/45, 7/45, 8/45, 8/45, 8/45).$$

**Solution to Huffman code**

Here is one possible solution.

code									
0000	$x_1$	1	2	6	8	8	13	29	45
0001	$x_2$	1	6	7	8	8	16	16	
001	$x_3$	6	6	8	8	13	16		
100	$x_4$	6	7	8	8	16			
101	$x_5$	7	8	8	13				
110	$x_6$	8	8	8					
111	$x_7$	8	8						
01	$x_8$	8							

The codeword lengths are (4, 4, 3, 3, 3, 3, 3, 2), and the average codeword length is  $\frac{43}{15} \approx 2.8667$ .

Different solutions are possible depending on how one goes about building the tree. Another solution, with the same average codeword length, has codeword lengths of (5, 5, 4, 3, 3, 3, 2, 2).

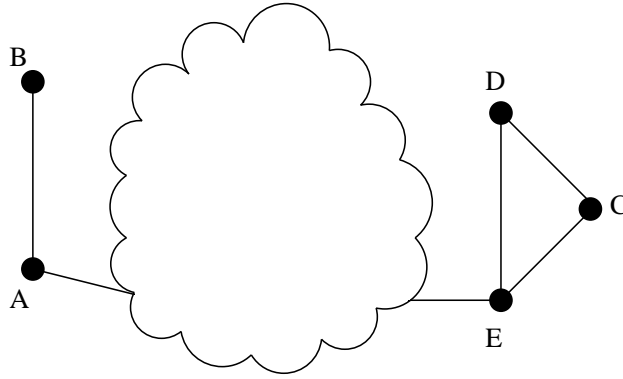
2. (20 points) **Random walk obscured by clouds**

A random walk takes place on this undirected connected graph. Part of the graph is obscured. We observe that  $\frac{1}{6}$ th of the time, in the limit as  $n \rightarrow \infty$ , is spent in  $A$ .

- (a) What fraction of the time is spent in the cloud?
- (b) What is the entropy rate of the random walk?

**Solution to Random walk obscured by clouds**

- (a) Using the formula for the stationary distribution of a random walk on an undi-



rected connected graph, the fraction of time spent at node  $A$  is  $\frac{E_A}{2E}$ , where  $E$  is the total number of edges in the graph, and  $E_A$  is the degree of  $A$ . Since  $E_A = 2$ , and  $\frac{E_A}{2E} = 1/6$ , it follows that  $E = 6$ .

From what is visible in the picture, there are either 5 or 6 edges which are not obscured, since we are not sure whether the edge that enters the cloud from  $A$  is the same as the edge that enters the cloud from  $E$ .

However, since  $E = 6$ , we know that the edges entering the cloud from  $A$  and  $E$  must be distinct, because if they were not, then the graph would only have 5 edges in total. So there is only one node in the cloud. The fraction of time spent in the cloud is thus the fraction of time spent at this hidden node, and using the formula for the stationary distribution, the answer is  $1/6$ .

(b) The entropy rate is

$$\begin{aligned} H(\mathcal{X}) &= \log_2(2E) - H\left(\frac{2}{12}, \frac{1}{12}, \frac{2}{12}, \frac{2}{12}, \frac{3}{12}, \frac{2}{12}\right) \\ &= \frac{2}{3} + \frac{1}{4} \log_2 3 \\ &= 1.06291. \end{aligned}$$

3. (20 points) **Entropy inequality**

Let  $(X, Y) \sim p(x, y)$  be discrete real valued random variables.

- (a) Compare  $H(X|X \cdot Y)$  and  $H(X, Y) - H(X \cdot Y)$ . Is the left-hand side  $\leq, \geq$ , or  $=$  to the right-hand side?
- (b) Find conditions for equality.

**Solution to Entropy inequality**

The answer for part (a) is  $\leq$ , and the answer for part (b) is twofold: either  $\Pr\{X = 0\} =$

0, or  $H(Y|X = 0) = 0$ .

Here are two different ways of arriving at these same conclusions.

- Solution I (Concise):

$$\begin{aligned} H(X|XY) + H(XY) &\stackrel{(a)}{=} H(X, XY) \\ &\stackrel{(b)}{\leq} H(X, Y) \end{aligned}$$

where each step is justified as follows:

- (a) chain rule
- (b)  $(X, XY)$  is a function of  $(X, Y)$

Thus, moving  $H(XY)$  to the other side,

$$H(X|XY) \leq H(X, Y) - H(XY).$$

Equality holds if and only if

- either  $\Pr\{X = 0\} = 0$ , or
  - $Y$  is a constant (degenerate random variable) when  $X = 0$ .
- Solution II (Less Concise, but More Explicit):

$$\begin{aligned} H(X|XY) &\stackrel{(a)}{=} H(X, XY) - H(XY) \\ &\stackrel{(b)}{=} H(X, Y, XY) - H(Y|X, XY) - H(XY) \\ &\stackrel{(c)}{=} H(X, Y) - H(Y|X, XY) - H(XY) \\ &\stackrel{(d)}{=} H(X, Y) - H(XY) \\ &\quad - \Pr\{X = 0\} H(Y|X = 0, XY) - \Pr\{X \neq 0\} H(Y|X \neq 0, XY) \\ &\stackrel{(e)}{=} H(X, Y) - H(XY) \\ &\quad - \Pr\{X = 0\} H(Y|X = 0) - \Pr\{X \neq 0\} H(Y|X \neq 0, XY) \\ &\stackrel{(f)}{=} H(X, Y) - H(XY) - \Pr\{X = 0\} H(Y|X = 0) \end{aligned}$$

where each step is justified as follows:

- (a) chain rule
- (b) chain rule
- (c)  $XY$  is a function of  $X$  and  $Y$

- (d) conditioning on whether  $X = 0$  or  $X \neq 0$
- (e) if  $X = 0$ , then  $XY = 0$  regardless of  $Y$ , so the  $XY$  term can be dropped from the conditioning in  $H(Y|X = 0, XY)$
- (f) if  $X \neq 0$ , then we can determine  $Y$  given  $X$  and  $XY$ , so  $H(Y|X \neq 0, XY) = 0$ .

Thus we conclude that

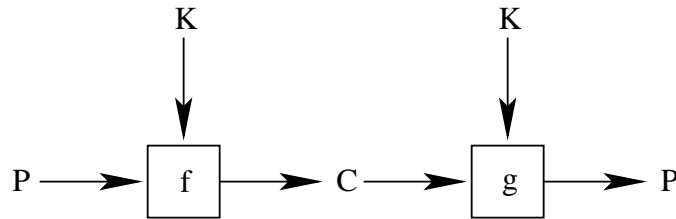
$$H(X|X \cdot Y) \leq H(X, Y) - H(X \cdot Y).$$

We see that equality holds if and only if  $\Pr\{X = 0\}H(Y|X = 0)$  equals zero.

Thus,

- either  $\Pr\{X = 0\} = 0$ , or
- $H(Y|X = 0) = 0$ .

#### 4. (20 points) Shannon cryptography



In cryptography, we want to know how large the random key must be in order to obscure the text. Consider the following relations between random variables  $P$ ,  $K$ , and  $C$ :

- i.* The plaintext  $P$  and the key  $K$  are independent of each other.
- ii.* The ciphertext is a deterministic function of the plaintext and key:  $C = f(P, K)$ .
- iii.* The plaintext is a deterministic function of the ciphertext and key:  $P = g(C, K)$ .

Note we do not assume that  $f$  and  $g$  are invertible functions.

Show the following:

$$\text{If } P \text{ and } C \text{ are independent, then } H(K) \geq H(P).$$

That is, if we wish the ciphertext and plaintext to be independent, and we wish to be able to decode the plaintext from the ciphertext, then the entropy of the secret key  $K$  must be greater than the entropy of the plaintext  $P$  it is encrypting.

## Solution to Shannon cryptography

There are many solutions. Below are some of them, including some rather unconventional approaches.

- Solution I (Ming-Yang Chen, Student), Most Concise:

$$\begin{aligned} H(P) &\stackrel{(a)}{=} H(P|C) \\ &\stackrel{(b)}{=} H(P|C) - H(P|K, C) \\ &\stackrel{(c)}{=} I(P; K|C) \\ &\stackrel{(d)}{\leq} H(K|C) \\ &\stackrel{(e)}{\leq} H(K) \end{aligned}$$

where each step is justified as follows:

- (a) premise that  $P$  is independent of  $C$
  - (b) entropy is non-negative
  - (c) definition of mutual information
  - (d) definition of mutual information
  - (e) conditioning reduces entropy
- Solution II (Jia Shuo Yue, Student), Most Resourceful:  
We will expand  $H(P, K, C)$  in two different ways. Firstly,

$$\begin{aligned} H(P, K, C) &= H(P, K) + H(C|P, K) \\ &= H(P, K) \\ &= H(P) + H(K). \end{aligned}$$

Secondly,

$$\begin{aligned} H(P, K, C) &= H(P, C) + H(K|P, C) \\ &= H(P) + H(C) + H(K|P, C). \end{aligned}$$

Setting these two expansions equal to each other and cancelling out  $H(P)$ , we have  $H(C) + H(K|P, C) = H(K)$ , and thus,

$$H(C) \leq H(K).$$

Now recall that from the midterm, we proved the relation  $H(K|C) = H(K) + H(P) - H(C)$ , which can be rewritten as  $H(C) - H(P) = H(K) - H(K|C)$ . Since conditioning reduces entropy, this relation implies that

$$H(P) \leq H(C).$$

Combining our deductions, we have the result:

$$H(P) \leq H(C) \leq H(K).$$

Equality holds when  $K$  and  $C$  are independent, and when  $K$  is a deterministic function of  $P$  and  $C$ .

- Solution III (William Wu, TA + Aaron Lehmann, Student), Most Unconventional

$$\begin{aligned} H(P) &\stackrel{(a)}{=} H(P|C) \\ &\stackrel{(b)}{=} \sum \Pr\{C=c\} H(P|C=c) \\ &\stackrel{(c)}{=} \sum \Pr\{C=c\} H(g(C,K)|C=c) \\ &\stackrel{(d)}{=} \sum \Pr\{C=c\} H(g(C=c,K)|C=c) \\ &\stackrel{(e)}{\leq} \sum \Pr\{C=c\} H(K|C=c) \\ &\stackrel{(f)}{=} H(K|C) \\ &\stackrel{(g)}{\leq} H(K) \end{aligned}$$

where each step is justified as follows:

- (a)  $P$  and  $C$  are independent
  - (b) definition of conditional entropy  $H(P|C)$
  - (c)  $P = g(C, K)$
  - (d) plugging in the condition  $C = c$
  - (e)  $g(C = c, K)$  is a deterministic function of  $K$
  - (f) definition of conditional entropy  $H(K|C)$
  - (g) conditioning reduces entropy
- Solution IV (Lei Zhao, TA), Second Most Boring

We can expand  $H(K, P, C)$  in two different ways:

$$\begin{aligned}H(K, P, C) &= H(C) + H(P|C) + H(K|P, C) \\H(K, P, C) &= H(C) + H(K|C) + H(P|K, C)\end{aligned}$$

Now we equate our expressions, cancelling out the  $H(C)$ :

$$H(P|C) + H(K|P, C) = H(K|C) + H(P|K, C)$$

Since we require  $P$  and  $C$  to be independent,  $H(P|C) = H(P)$ . Furthermore, since  $P$  is a function of  $K$  and  $C$ ,  $H(P|K, C) = 0$ . Thus,

$$H(P) + H(K|P, C) = H(K|C)$$

which implies that

$$H(P) \leq H(K|C).$$

Lastly, since conditioning reduces entropy,

$$H(K) \geq H(K|C) \geq H(P).$$

- Solution V (William Wu, TA), Most Boring

$$\begin{aligned}H(C) + H(P|C) &\stackrel{(a)}{=} H(P, C) \\&\stackrel{(b)}{\leq} H(P, C) + H(K|P, C) \\&\stackrel{(c)}{=} H(P, C, K) \\&\stackrel{(d)}{=} H(C, K) + H(P|C, K) \\&\stackrel{(e)}{=} H(C, K) \\&\stackrel{(f)}{=} H(C) + H(K|C) \\&\stackrel{(g)}{\leq} H(C) + H(K)\end{aligned}$$

where each step is justified as follows:

- (a) chain rule
- (b) entropy is non-negative
- (c) chain rule
- (d) chain rule
- (e)  $P$  is a function of  $C$  and  $K$

- (f) chain rule
- (g) conditioning reduces entropy

Comparing the first and last expressions and cancelling  $H(C)$ , we see that

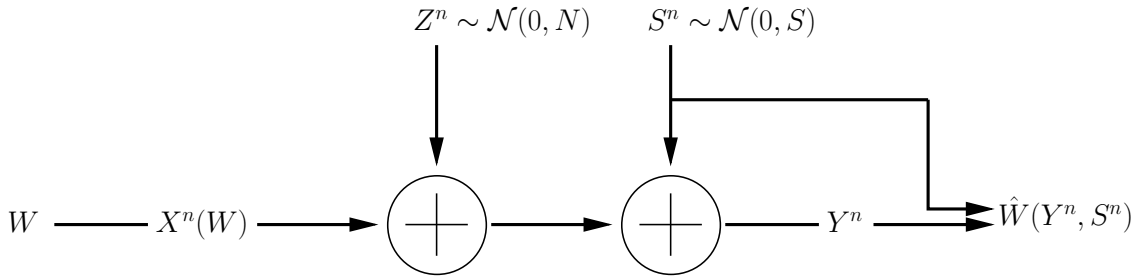
$$H(P|C) \leq H(K)$$

Lastly, by the premise that  $P$  and  $C$  are independent,  $H(P|C) = H(P)$ . Thus,

$$H(P) \leq H(K).$$

5. (20 points) **Noise partially known**

What is the capacity of the Gaussian channel with part of the noise known at the receiver? *You are not required to give proofs of achievability and converse.*



Here  $Z, S$  are i.i.d. Gaussian, independent of  $X$  and each other.  $X, S$ , and  $Z$  are also jointly independent.  $Y^n = X^n + Z^n + S^n$ . There is a power  $P$  constraint on the input  $X$ . That is,  $\frac{1}{n} \sum_{i=1}^n x_i^2(w) \leq P$  for all  $w \in \{1, 2, \dots, 2^{nR}\}$ .

**Solution to Noise partially known**

The received signal is  $Y^n = X^n + Z^n + S^n$ . However,  $S^n$  is known at the receiver, so we can subtract it from the sum, resulting in  $X^n + Z^n$ . Thus, we have the standard Gaussian channel, and the capacity is  $C = \frac{1}{2} \log(1 + \frac{P}{N})$ .

More formally, we can think of the channel as having input  $X$  and outputs  $Y$  and  $S$ . Then we wish to maximize the mutual information between  $X$  and  $(Y, S)$ .

$$\begin{aligned} I(X; Y, S) &\stackrel{(a)}{=} I(X; S) + I(X; Y|S) \\ &\stackrel{(b)}{=} I(X; X + Z + S|S) \\ &\stackrel{(c)}{=} I(X; X + Z|S) \\ &\stackrel{(d)}{=} I(X; X + Z) \end{aligned}$$

where each line is justified as follows:



Recall that  $H(p)$  is concave and maximized for  $p = 1/2$ . Thus, there are three regions:

$$C = \begin{cases} 1 - \alpha & 1/2 \leq E - 1 \\ (1 - \alpha)H(E - 1) & 0 \leq E - 1 \leq 1/2 \\ 0 & E - 1 < 0 \end{cases}$$

7. (20 points) **Side information**

Consider a horse race  $X \sim p(x)$  with  $m$ -for-1 fair odds. Let  $(X, Y, Z) \sim p(x, y, z)$ . Gambler  $A$  has side information  $Y$ , and gambler  $B$  has side information  $Z$ .

- (a) Suppose  $X \rightarrow Z \rightarrow Y$  forms a Markov chain. Which gambler does best?
- (b) Now, gambler  $A$  is given side information  $Z$  in addition to his information  $Y$ , where  $p(x, y, z)$  is arbitrary. How much does this combined information  $(Z, Y)$  increase his growth rate of wealth beyond his rate from  $Y$  alone?

**Solution to Side information**

- (a) Let  $W^*(X)$  denote the unconditional optimal doubling rate, let  $W^*(X|Y)$  denote the optimal doubling rate conditioned on side information  $Y$ , and let  $W^*(X|Z)$  denote the optimal doubling rate conditioned on side information  $Z$ . Then as shown on page 165 of the text,

$$W^*(X|Y) = I(X; Y) + W^*(X),$$

and

$$W^*(X|Z) = I(X; Z) + W^*(X).$$

Thus the difference in doubling rate between Gambler B and Gambler A is difference of the mutual informations  $I(X; Z)$  and  $I(X; Y)$ .

By the data processing inequality,  $I(X; Z) \geq I(X; Y)$ , so gambler B does best.

- (b) When gambler  $A$  has both  $Y$  and  $Z$ , his doubling rate is

$$W^*(X|Y, Z) = I(X; Y, Z) + W^*(X).$$

Thus,

$$\begin{aligned} W^*(X|Y, Z) - W^*(X|Y) &= I(X; Y, Z) - I(X; Y) \\ &= I(X; Y) + I(X; Z|Y) - I(X; Y) \\ &= I(X; Z|Y). \end{aligned}$$

8. (20 points) **Poisoned wine**

You have eight wine bottles, and exactly one is poisoned, according to the following probabilities:

Bottle:	1	2	3	4	5	6	7	8
Probability of poison:	1/45	1/45	6/45	6/45	7/45	8/45	8/45	8/45

You intend to determine which bottle is poisoned by testing the wines on lab rats. If a rat imbibes poison, it instantly dies. Since we don't want to deal with drunken rats, there is one constraint:

**No rat can be used more than once.**

*The following two problems should be considered separately.*

- (a) Determine which bottle is poisoned in a way that ensures *no more than one rat is killed*.
- (b) Determine which bottle is poisoned in a way that minimizes the average number of rats used.

**Solution to Poisoned wine**

- (a) Simplest solution: take eight rats, and have each sample a different bottle. Exactly one rat will die, thereby identifying which bottle is poisoned.

Note that we cannot use only one rat and have it sip a bottle, sober up, sip a different bottle, sober up, and so on. The problem statement says that no rat can be used more than once.

- (b) The solution is to use Huffman coding. Let  $X$  be a random variable corresponding to which bottle was poisoned. Determining each bit of the codeword for  $X$  corresponds to having a rat taste some mixture of wines. Minimizing the average codeword length corresponds to minimizing the average number of rats used. Since the probability distribution is the same as that given in Problem 1, the Huffman codewords are:

code									
0000	$x_1$	1	2	6	8	8	13	29	45
0001	$x_2$	1	6	7	8	8	16	16	
001	$x_3$	6	6	8	8	13	16		
100	$x_4$	6	7	8	8	16			
101	$x_5$	7	8	8	13				
110	$x_6$	8	8	8					
111	$x_7$	8	8						
01	$x_8$	8							

The tastings should be conducted in the following order, terminating as soon as all bits of the codeword, revealed from most significant bit (leftmost) to least significant bit (rightmost), are fully determined:

- i. Rat 1: Taste a mixture of bottles 4,5,6,7.
- ii. Rat 2: Taste a mixture of bottles 6,7,8.
- iii. Rat 3: Taste a mixture of bottles 3,5,7.
- iv. Rat 4: Taste bottle 2.

For example, if  $X = 1$ , then we will require four rats.

The average number of rats killed is  $\frac{43}{15} \cong 2.86667$ .

9. (30 points) **Memory chip**

A memory chip has one million memory cells, each storing one bit. However, due to the manufacturing process, each cell is defective with probability  $p = .01$ . Cell defects are independent.

We have to live with the defective chips because otherwise, we would have to wait a long time to get a chip that doesn't have any defects.

So one option is to measure each cell before the chip is used, identify the defects, and not use those cells.

- (a) What is the average usable memory size of a chip?

Now suppose we don't make these measurements ahead of time, and instead decide to simply use error correction codes to combat the defects, under several different assumptions.

- (b) First, assume that defective cells give illegitimate voltage readings (erasures) and therefore can be identified as soon as the memory is read. What is the effective memory size of the chip?

- (c) Now assume that the defective cells give the opposite reading from that which was written in memory. What is the effective memory size of the chip?

**Solution to Memory chip**

- (a)  $0.99 \times 1000000 = 990,000$  memory cells.
- (b) This is an erasure channel. Since the capacity of the BEC is  $1 - \alpha$ , the effective memory size is again 990,000 memory cells.
- (c) This is a binary symmetric channel. Since the capacity of the BSC is  $1 - H(\alpha)$ , the effective memory size is  $1000000(1 - H(0.01)) \approx 919,207$  memory cells.