Name:

Problem	Points	Score
1a	10	
1b	10	
1c	10	
1d	10	
2a	10	
2b	10	
2c	10	
3a	10	
3b	10	
3c	10	
Total	100	

Notes:

1. The exam is closed books/closed notes - except for one page (double-sided) of notes.

2. Please show ALL work. Answers with no supporting explanations or work will be given no credit.

3. Please indicate clearly your answer to the problem. If I can't read it (and I am the judge of legibility), it is wrong. If I can't follow your solution (and I get lost easily), it is wrong. All things being equal, neat and legible work will get the higher grade:)

Problem No. 1: General Concepts

(a) Explain the relationship between mutual information and relative entropy. Do not simply write the equations or restate them in words — show some insight. (Hint: recall the discussion in Chap. 1.)

We can view mutual information as the reduction in uncertainty of one random variable due to the knowledge of the other. Also relative entropy is a distance measure between two random variables. Hence we can intuitively say that the mutual information will be minimum when the distance measure between the two distributions is a maximum. If the two distributions had been highly correlated, then the distance measure would have been very less and we can also say that the decrease in uncertainty is large because we are almost certain of one given the other i.e. the mutual information is more. This argument also leads to the definition of mutual information as being the relative entropy between the joint distribution and the product distribution. We can visualize mutual information as a distance measure between the joint distribution and the product distribution (which is he same as the joint distribution of two independent random variables).

(b) Explain the impact of information theory on the data compression and data transmission problems (Hint: recall the discussion in Chap. 1.)

From concepts learned so far in Information Theory, we know that data can be transmitted over a channel without any loss of information up to a particular rate known as the Shannon limit.

Regarding data compression, we know that we can transmit the data without error using a particular number of bits, determined from the entropy of the typical set.

Thus we can say that Information theory helps us to understand data transmission and compression better.

(c) Prove the chain rule for relative entropy.

The chain rule of relative entropy is
$D(p(x,y) \| q(x,y)) = D(p(x) \| q(x)) + D(p(y|x\rangle \| p(y|x\rangle)$

Proof:
$D(p(x,y) \| q(x,y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y))}$

$= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x\rangle}{q(x)q(y|x\rangle}$

$= \left(\sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x\rangle}{q(y|x\rangle} \right)$

$= (D(p(x) \| q(x)) + D(p(y|x\rangle \| q(y|x\rangle))$

This proves the chain rule for relative entropy.

(d) Prove the independence bound on entropy:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, X_{i-2}, \ldots, X_1)$$

The chain rule for entropy states that

$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1\rangle$ Also we know that conditioning reduces entropy.So

$\sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1\rangle \le \sum_{i=1}^{n} H(X_i)$ which implies that

$H(X_1, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i)$ where the equality follows if and only if all X_i are independent.This proves the independence bound on entropy.

Problem No. 2: Calculations

(a) **What is the relationship between $H(Y)$ and $H(X)$ when $Y = \cos X$ and X is a random variable that takes on a finite number of values.**

We know that conditioning reduces entropy. Hence we can write that $H(Y) \le H(X)$.

If X takes on values between -π to + π only then Y will be a complete statistic with respect to X and $H(Y) = H(X)$.

(b) **Show that if $H(Y|X) = 0$, then Y is a function of X, i.e., for all x with $p(x) > 0$, there is only one possible value of y with $p(x, y) > 0$.**

$$H(Y|X\rangle = \sum p(x) H(Y|(X = x)\rangle$$

$$= -\sum_x p(x) \sum_y p(y|x\rangle \log p(y|x\rangle$$

$$= -\sum_x \sum_y p(x, y) \log p(y|x\rangle$$

$$= -E_{p(x, y)} \log p(Y|X\rangle$$

If $H(Y|X\rangle$ =0 then $\log p(Y|X\rangle = 0$ which implies that Y is a function of X.($P(Y|X\rangle = 1$)

Also $p(X, Y) = p(\dot{X}) p(Y|X\rangle = p(X)$. So for all x with $p(x) > 0$, there is only one possible value of y with $p(x, y) > 0$.

(c) X **and** Y **are random variables, and let** $Z = X + Y$**. Show that** $H(Z|X) = H(Y|X)$**.**

$Z = X + Y$ implies that $p(Z = z|(X = x)\rangle = p(Y = z - x|(X = x)\rangle$

$H(Z|X\rangle = \sum p(x)H(Z|(X = x)\rangle$

$= -\sum_x p(x)\sum_z p(Z = z|(X = x)\rangle \log p(Z = z|(X = x)\rangle$

$= \sum_x p(x)\sum_y p(Y = z - x|(X = x)\rangle \log p(Y = z - x|(X = x)\rangle$

$= \sum p(x)H(Y|(X = x)\rangle$

$= H(Y|X\rangle$

Hence if $Z = X + Y$ then $H(Z|X\rangle = H(Y|X\rangle$.

Problem No. 3: A Deck of Cards

(a) A deck of cards contains 52 cards, evenly distributed between hearts, diamonds, spades, and clubs. Which has higher entropy: drawing two red cards with or without replacement.

Case (i) Drawing two cards with replacement

$$H(X) = -[p(A) \bullet \log(p(A)) + (1 - p(A)) \bullet \log(1 - p(A))]$$
$$= -[0.25 \log(0.25) + 0.75\log(0.75)] = 0.81 \text{ bits}$$

Case(ii) Drawing two cards without replacement

$$H(X) = -[0.5 \log(0.5) + (25/51)\log(25/51)] = 0.80 \text{ bits}$$

So drawing two cards with replacement has a higher entropy. This can also be thought of as an example of conditioning reducing entropy. In the second case, by drawing out the first card and not replacing it, we are conditioning the data.

(b) Suppose the first card drawn is a king. What are the chances that the next card drawn will result in a sum under 21. Prove that you are better off knowing that the first card drawn was a king.

If the sum should result under 21 and the first card drawn is a king, then the next card cannot be an ace. So our chances of not drawing an ace are (1-4/51)=0.92

In case we did know that the first card drawn was a king then the probability of drawing two cards whose sum is less than 21 is the same as the probability of not drawing an ace and a figure card together = (1 - (13/52)(15/51)=0.93

Hence the chances of the sum resulting under 21 have decreased once we know that the first card drawn is a king. So we are better off not to bid knowing that the first card is a king.

(c) You are playing blackjack with a dealer who is using one or two decks of cards. The player to your right has a king exposed on the table, and tells you he has reached a score of 20. Can you use information learned in this class to determine what your safest betting strategy might be? Or is this simply a straightforward probability calculation?

Once we know that he has a K exposed, his chances of reaching a score of 20 are equivalent to his chances of drawing either a figure card or an ace out of the remaining 51 cards which is equivalent to (15/51) = 0.29.

The above calculation was based on the assumption that there is just one deck of cards. Suppose there were two decks of cards, the probability would have been (15+16)/(103)=0.30.

We notice that the probability increases with the number of decks used.

Another argument that we can make is based on information theory. As the number of decks increase, so does the entropy of the cards. So the higher the number of decks, the better are my chances of betting on the player reaching a score of 20.