INSTRUCTOR VERSION WITH ANSWERS & FEEDBACK SHOWN

May 15, 2017

Entropy & Data Compression

Center for Science of Information, A National Science Foundation Science & Technology Center `http://soihub.org`

Questions on *Entropy & Data Compression*

1. Which of the following in regard to the random coding scheme is/are true:
   - ◯ $\mathcal{X}$ and $\hat{\mathcal{X}}$ must be the same.
   - ◯ The encoder encodes the source sequence $\mathbf{X}$ into a codeword $\hat{\mathbf{X}}(K)$.
   - ◯ Given the codebook, the reproduction sequence $\hat{\mathbf{X}}$ is a function of the source sequence $\mathbf{X}$. ✓
2. Which of the following is true:
   - ◯ $d(\mathbf{X}, \hat{\mathbf{X}}) \approx Ed(X, \hat{X})$, if $\mathbf{X}$ and $\hat{\mathbf{X}}$ are jointly typical. ✓
   - ◯ $P\{d(\mathbf{X}, \hat{\mathbf{X}}) > D + \epsilon \,|\, K \neq 1\}$ can approach 0 if we choose an appropriate δ, but can never be equal to 0.
   - ◯ For each codeword which is a typical $\hat{\mathbf{X}}$ sequence, there are approximately $2^{nH(X|\hat{X})}$ typical $\mathbf{X}$ sequences that are jointly atypical with it.
3. Which of the following about Theorem 9.2 is/are true:
   - ◯ The supremum is taken over all distribution $\mathbf{r}$ on $\mathcal{X}$ such that $r(x) > 0$ for at least one $x \in \mathcal{X}$.
   - ◯ The maximum is taken over all transition matrix $\mathbf{q}$ from $\mathcal{Y}$ to $\mathcal{X}$ such that $q(x|y) = 0$ iff $p(y|x) = 0$. ✓
   - ◯ $I(X;Y)$ can be written as $I(\mathbf{r}, \mathbf{q})$, where $\mathbf{r}$ is the input distribution and $\mathbf{q}$ is the transition matrix from $\mathcal{Y}$ to $\mathcal{X}$.
4. Which of the following is true:
   - ◯ $f^* = \inf_{\mathbf{Q}\in A_1} \inf_{\mathbf{p}\in A_2} f(\mathbf{Q}, \mathbf{p})$
   - ◯ $f^* = \sup_{\mathbf{Q}\in A_1} \sup_{\mathbf{t}\in A_2} f(\mathbf{Q}, \mathbf{t})$
   - ◯ $t^*(\hat{x})$ that minimizes f when $\mathbf{Q}$ is given equals $\sum_x p(x)Q(\hat{x}|x)$. ✓
   - ◯ $Q(\hat{x}|x)$ that minimizes f when $\mathbf{t}$ is given has numerator $t(\hat{x})e^{s\sum_x d(x,\hat{x})}$.
5. Which of the following about Lemma 9.4 is/are true:
   - ◯ The lemma indicates that if f is convex, then the algorithm cannot be trapped at $\mathbf{u} if f(\mathbf{u}) < f^*$.
   - ◯ To prove that if $\delta f(\mathbf{u}) = 0$ then $\mathbf{u}_1 = c_1(\mathbf{u}_2) and \mathbf{u}_2 = c_2(\mathbf{u}_1)$, we need the fact that f is concave.
   - ◯ To prove that if $\delta f(\mathbf{u}) = 0$ then $\mathbf{u}_1 = c_1(\mathbf{u}_2) and \mathbf{u}_2 = c_2(\mathbf{u}_1)$, we need the fact that values of $c_1(\cdot)$ and $c_2(\cdot)$ are unique. ✓
   - ◯ $\nabla f \cdot \tilde{z} = 0$ and f being convex imply that f attains the maximum along the line passing through $\mathbf{u}$ and $\mathbf{v}$ at the point $\mathbf{u}$.

6. Which of the following is/are true:
   - ◯ If X has a pdf, then $F_X(x)$ is differentiable, implying that $F_X(x)$ is continuous and hence X is continuous. ✓
   - ◯ If X is continuous, then $F_X(x)$ is continuous and hence differentiable.
   - ◯ Conditional CDF $F_{Y|X}(y|x) = Pr\{Y \leq y \mid X \leq x\} = \frac{F_{XY}(x,y)}{F_X(x)}$.
   - ◯ $\text{var}(\sum_{i=0}^{n} X_i) = \sum_{i=0}^{n} \text{var}X_i + \sum_{i=0}^{n} \sum_{j=i+1}^{n} \text{cov}(X_i, X_j)$.

7. Consider an arbitrary covariance matrix K, which of the following is/are always true:
   - ◯ K is symmetric positive definite.
   - ◯ K is invertible.
   - ◯ K is diagonalizable and can be written as $Q\lambda Q^{\mathsf{T}}$, where Q is orthogonal and λ is a full rank diagonal matrix.
   - ◯ K is diagonalizable and can be written as $Q\lambda Q^{\mathsf{T}}$, where Q is orthogonal and λ is nonnegative. ✓

8. Which of the following about decorrelation is/are true:
   - ◯ Let $\mathbf{Y} = Q\mathbf{X}, where K_{\mathbf{X}} = Q\lambda Q^{\mathsf{T}}, then K_{\mathbf{Y}} = \lambda$.
   - ◯ Let $\mathbf{Y} = Q\mathbf{X}, where K_{\mathbf{X}} = Q^{\mathsf{T}}\lambda Q, then K_{\mathbf{Y}} = \lambda$. ✓
   - ◯ If $K_{\mathbf{Y}} = \lambda$, then the random variables in $\mathbf{Y}$ are pairwise independent.
   - ◯ The total energy of a random vector is preserved under correlation.

9. Which of the following is/are true:(Only for continuous X):
   - ◯ (Differential) entropy represents the uncertainty of r.v. X ✓
   - ◯ Chain rule of (differential) entropy
   - ◯ Conditioning does not increase (differential) entropy
   - ◯ none

10. Which of the following is/are true: (Only for continuous X):
    - ◯ Independence bound for (differential) entropy
    - ◯ $I(X;Y|T) \geq 0$
    - ◯ $I(X;Y|T) = 0$ if and only if X and Y are independent given T
    - ◯ none ✓

11. Which of the following is/are true: (For general X, Y and T):
    - ◯ (Differential) entropy represents the uncertainty of r.v. X
    - ◯ Chain rule of (differential) entropy
    - ◯ Conditioning does increase (differential) entropy
    - ◯ none ✓

12. Which of the following is/are true (Only for continuous X):
    - ◯ Independence bound for (differential) entropy
    - ◯ $I(X;Y|T) \geq 0$
    - ◯ $I(X;Y|T) = 0$ if and only if X and Y are independent given T
    - ◯ none ✓

13. Which of the following is/are true (For general X, Y and T):
    - ◯ (Differential) entropy represents the uncertainty of r.v. X
    - ◯ Chain rule of (differential) entropy ✓

- ◯ Conditioning does increase (differential) entropy
- ◯ none

14. Which of the following is/are true: (For general X, Y and T):
    - ◯ Dependence bound for (differential) entropy
    - ◯ $I(X;Y|T) \geq 0$ ✓
    - ◯ $I(X;Y|T) = 0$ if and only if X and Y are dependent given T
    - ◯ none
15. Which of the following is/are true:
    - ◯ The volume of a set A in $\mathfrak{R}^n$ is its cardinality.
    - ◯ The squences in the typical sets have empirical differential entropies very close to the amount of information the generic random variable carries.
    - ◯ The larger the differential entropy is, the larger the volume of the typical set is. ✓
    - ◯ $h(X) < 0$ implies that $f(\mathbf{x}) > 1$ for any sequence $\mathbf{x}$.
16. Which of the following is true:
    - ◯ On the alphabet of natural numbers, if the mean is specified, the distribution to obtain the maximum entropy is the geometric distribution. ✓
    - ◯ $h(X) \leq \frac{1}{4}\log(2\pi e\sigma^2)$ for any continuous random variable X with variance σ^2, and the equality is obtained when X is normally distributed with arbitrary mean.
    - ◯ $h(X) \leq \frac{1}{2}\log(2\pi e\kappa)$ for any continuous random variable X with $EX^2 = \kappa$, and the equality is obtained when X is normally distributed with arbitrary mean.

Section 1: Bits as a Measure Part 1

1. Suppose an octopus labels its eight legs A through H and selects one at random. How many bits of information do we obtain from knowing which leg it chooses?

   Solution: Each leg has probability $\frac{1}{8}$ of being selected, so the entropy contributed by each leg is

   $$-\frac{1}{8}\log_2\left(\frac{1}{8}\right) = \frac{3}{8}.$$

   Since there are eight legs capable of being selected, there are eight terms in the sum, and the total entropy is 3.

2. Suppose we have a biased coin which lands heads $\frac{2}{3}$ of the time, and tails otherwise. How many bits of information are obtained from knowing that a flip of the coin landed heads? What about tails? How many bits of information are obtained on average from knowing the result of one flip of the coin?

   Solution: The number of bits obtained from knowing that a flip landed heads is $-\log_2\left(\frac{2}{3}\right)$ and the number of bits obtained from knowing that a flip landed tails is $-\log_2\left(\frac{1}{3}\right)$. Hence, the average number of bits, which is also the entropy is

   $$-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right).$$

Section 2: Bits as a Measure Part 2

3. Suppose we draw all 52 cards again, from a deck of cards. However, if the first card drawn is the Ace of Spades, we draw repeatedly draw all 52 cards again, until the first card drawn is NOT the Ace of Spades. How many bits of information are obtained from knowing the order of the 52 cards that are drawn?

   Solution: The Ace of Spades cannot be possibly be drawn, but all other 51 cards may be drawn. Hence, the entropy is $\log_2(51)$.

4. Suppose there are n outcomes to a random process. If this random process occurs once, and the outcome is given, what are possible conditions so that the entropy of the process is NOT $\log_2(n)$?

   Solution: If the probability of all outcomes are uniform, then the entropy for each outcome will be $\frac{1}{n}\log_2(n)$, and the entropy of entire process is then $\log_2(n)$. Thus, one possible condition for the entropy of the process to not be $\log_2(n)$ is that the probability of all outcomes are not equivalent, so that some outcomes are more likely than others.

Section 3: The Kraft Inequality and Code Length Part 1

5. Prove the following lengths in bits for the encoding of each of the following 9 possible outcomes as instantaneous codes is not uniquely decodable given a binary alphabet.

Outcome	Length in Bits of Encoding
A	3
B	4
C	2
D	5
E	6
F	3
G	3
H	2
I	5

   Solution: In a binary alphabet, we use the base 2 in Kraft's Inequality:

$$2^{-3} + 2^{-4} + 2^{-2} + 2^{-5} + 2^{-6} + 2^{-3} + 2^{-3} + 2^{-2} + 2^{-5} = \frac{65}{64} \geq 1.$$

   Hence, the code is not uniquely decodable.

6. Suppose we use the English alphabet of 26 letters to encode some words so that the encoded words have length $a_1, a_2, \ldots, a_n$ to send messages, without spaces. Determine Kraft's Inequality to ensure the code is uniquely decodable.

   Solution: In an alphabet with 26 letters, we require

$$\sum_{i=1}^{n} 26^{-a_i} \leq 1$$

   to ensure the code is uniquely decodable.

Section 4: Calculating Average Code Length - Shannon's First Theorem Part 1

Consider two different encryption schemes with the same distribution. In the first scheme, all messages have fixed length 2. Message 00 appears with probability $p = 0.7$, message 01 appears with probability $p = 0.1$ and message 10 appears with probability $p = 0.1$ and message 11 appears with probability $p = 0.1$.

In the second scheme, message 0 appears with probability $p = 0.7$, message 10 appears with probability $p = 0.1$, message 110 appears with probability $p = 0.1$ and message 111 appears with probability $p = 0.1$

7. What is the average length of a message in each scheme? Is this surprising, given that the second scheme has messages of length three, while the first scheme does not?

   Solution: The average length of a message in the first scheme is 2, since all messages have fixed length. The average length of a message in the second scheme is

   $$1 \cdot 0.7 + 2 \cdot 0.1 + 3 \cdot 0.1 + 4 \cdot 0.1 = 1.6.$$

   The reason the second scheme has shorter average length is because the message of length one appears frequently, while the longer messages appear very infrequently.

8. What is the entropy of the distribution? How does it compare to the average length of a message in each scheme?

   Solution: The entropy of the distribution is

   $$-0.7\log_2(0.7) - 0.1\log_2(0.1) - 0.1\log_2(0.1) - 0.1\log_2(0.1) \approx 1.357,$$

   which is less than the average length of a message in each above scheme.

Section 5: Calculating Average Code Length - Shannon's First Theorem Part 2

9. Verify the entropy inequality discussed in lecture for the following case:

   $$-0.6\log_2(0.6) - 0.4\log_2(0.4) \leq -0.6\log_2(0.2) - 0.4\log_2(0.8),$$

   $$-0.2\log_2(0.2) - 0.8\log_2(0.8) \leq -0.6\log_2(0.2) - 0.4\log_2(0.8).$$

   Solution: We have

   $$-0.6\log_2(0.2) - 0.4\log_2(0.8) \approx 1.522,$$

   $$-0.2\log_2(0.2) - 0.8\log_2(0.8) \approx 0.722,$$

   $$-0.6\log_2(0.6) - 0.4\log_2(0.4) \approx 0.971,$$

   and the desired inequalities follow.

10. Prove the specific case of the entropy inequality discussed in lecture, for only two variables. That is, given $0 \leq x, y \leq 1$, prove

    $$-x\log_2(x) - (1-x)\log_2(1-x) \leq -x\log_2(y) - (1-x)\log_2(1-y).$$

Solution: Write $y = x + c$ and consider the function $f(c) = -x\log_2(x+c) - (1-x)\log_2(1-x-c)$ and take the derivative with respect to c,

$$\frac{df}{dc} = -\frac{x}{(x+c)\ln(2)} + \frac{1-x}{(1-x-c)\ln(2)}.$$

Note that when $c = 0$, $\frac{df}{dc} = 0$, and in fact the value at $c = 0$ is a local minimum. Hence, the inequality follows.

Section 6: Calculating Average Code Length - Shannon's First Theorem Part 3

11. How does Shannon's idea shed light on why can there be equality in the entropy inequality if the probability of all messages are powers of $\frac{1}{2}$ in a binary alphabet?

Solution: Shannon's idea says that encoded lengths for message m_i should be $\log_2(p_i)$ in a binary alphabet, where p_i is the probability that the message occurs. For probabilities that are not powers of $\frac{1}{2}$, the message lengths must be rounded up to the next integer, but for powers of $\frac{1}{2}$, the logarithm values are already integers, and so equality holds.

Section 7: Huffman Coding Part 1

12. Determine a Huffman encoding scheme for the following distribution on 4 outcomes:

Outcome	Probability
A	$\frac{1}{12}$
B	$\frac{1}{3}$
C	$\frac{1}{12}$
D	$\frac{1}{2}$

Solution: One possible scheme is:

Outcome	Message
A	111
B	10
C	110
D	0

Regardless, the encoded message for D must have length 1, the encoded message for B must have length 2, and the encoded messages for A and C must have length 3 each, for optimality.

Section 8: Huffman Coding Part 2

13. Suppose we had a fixed length encoding scheme for 7 outcomes. How many bits would be needed? What is the average length of the encoding scheme?

Solution: Because 2 bits can only decipher $2^2 = 4$ messages, we need 3 bits to encode and decode 7 messages, as $2^3 = 8 \geq 7$. Moreover, the encoding scheme is fixed lengths, so each message, and therefore the average length, is 3 bits.

14. Now suppose we are given the following distribution for 7 outcomes. Determine a Huffman encoding scheme for the outcomes:

Outcome	Probability
A	$\frac{1}{8}$
B	$\frac{1}{4}$
C	$\frac{1}{16}$
D	$\frac{1}{16}$
E	$\frac{1}{8}$
F	$\frac{1}{8}$
G	$\frac{1}{4}$

Solution: One possible encoding appears below. There are many possibilities, but the average length of the schemes are the same.

Outcome	Message
A	111
B	01
C	1101
D	1100
E	101
F	100
G	00

15. What is the average length of the Huffman encoding scheme? How does it compare to the fixed length encoding scheme?

Solution: The average length is $\frac{42}{16} = \frac{21}{8}$, which is less than 3, so the Huffman encoding scheme performs better than the fixed length encoding scheme.

16. What is the entropy of the distribution? How does it compare to the average length of the Huffman encoding scheme? Is it possible for the Huffman encoding scheme to do better?

Solution: The entropy of the distribution is

$$\frac{1}{8}\log_2\left(\frac{1}{8}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{16}\log_2\left(\frac{1}{16}\right) + \cdots + \frac{1}{4}\log_2\left(\frac{1}{4}\right) = \frac{21}{8},$$

which is the same as the average length of the Huffman encoding scheme. Equality occurs because the probability of all outcomes are powers of 2. It is not possible for Huffman scheme to do better, because any encoding scheme must have average length at least as great as the entropy.

Section 9: Huffman Coding Part 3

17. Suppose we are given a distribution in which 0 shows up with probability $p = 0.8$ and 1 shows up with probability $p = 0.2$. Determine a Huffman encoding scheme for the outcomes and compute the difference between the average length of the encoding scheme and the entropy of the distribution.

Solution: If we just assign the messages 0 and 1 to the corresponding outcomes, then the average length of the encoding scheme is 1. The entropy of the distribution is

$$-0.2\log_2(0.2) - 0.8\log_2(0.8) \approx 0.7219,$$

and the difference is approximately 0.2781.

18. Now, group outcomes in series of two, so that there are now 4 possible outcomes. Determine a Huffman encoding scheme for the outcomes and compute the difference between the average length of the encoding scheme and the entropy of the distribution.

Solution:

Outcome	Probability	Message
00	0.64	0
01	0.16	10
10	0.16	110
11	0.04	111

The average length is now 1.5 while the entropy is now 1.4439, so the difference is approximately 0.0561.

19. How does the difference in the first scheme compare to the difference in the second scheme? This is the idea behind Shannon's First Theorem!

Solution: The difference in the second scheme is smaller than the difference in the first scheme even though no new information was gained!

Section 10: Universal Coding

20. The idea behind universal coding is to create an encoding method that asymptotically approaches the entropy of the data. As hinted in the video, what are some disadvantages to having such codes?

Solution: The disadvantages is that the look-up tables will become very long and so not only will the tables require a lot of space, but the look-up process will start becoming time-consuming as well.

Section 11: Lempel-Ziv

21. Don't forget to do the exercise at

`www.soihub.org/wiki/images/5/5d/LZ77_example.pdf`

22. Using the Lempel-Ziv 77 compression algorithm, compress the following data:

01110101001101110001010011010110011101 1

Solution:

23. Confirm the correctness of the compression process by decompressing the Lempel-Ziv 77 compression algorithm from the previous part.

Solution:

24. Using the Lempel-Ziv 77 compression algorithm, compress the phrase "ban nana anna banana" (without spaces).

Solution:

25. Confirm the compression of the data by decompressing the Lempel-Ziv 77 compression algorithm from the previous part.

Solution:

Section 12: Binary Symmetric Channel Part 1

26. Consider a binary symmetric channel with error probability 0.05 and a second binary symmetric channel with error probability 0.03. If the output of the first channel is used as the input to the second channel, what is the probability that the output of the second channel is correct for the input to the first channel? Does the error probability changed if the channel with error probability 0.03 is used first and the channel with error probability 0.05 is used second instead?

Solution: The output is correct if either both channels function properly or both channels function improperly. The probability of this is

$$0.95 \cdot 0.97 + 0.05 \cdot 0.03 = 0.9221.$$

The error probability does not depend on the order of the channels, and so it does not change if the order of the channels is changed.

Section 13: Binary Symmetric Channel Part 2

27. *[Bayes' Theorem]* Recall that

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

How do $P(A|B)$ and $P(B|A)$ relate?

Solution: Since $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

which is Bayes' Theorem.

Section 14: Binary Symmetric Channel Part 3

28. Alice is an avid wine-taster who tries red wines 60% of the time and white wines the other 40%. She enjoys red wines 75% of the time and white wines 50% of the time. Given that she does not enjoy a bottle of wine, what is the probability the bottle of wine contains white wine?

Solution: The probability Alice does not enjoy a red wine is $0.6 \cdot 0.25 = 0.15$ and the probability she does not enjoy a white wine is $0.4 \cdot 0.5 = 0.2$. Hence the probability that Alice does not enjoy a white wine, given that she does not enjoy a bottle of wine is

$$\frac{0.2}{0.2 + 0.15} = \frac{4}{7}.$$

Section 15: Conditional and Joint Entropies

29. What could cause the difference between the conditional entropies $H(X|Y)$ and $H(X)$ to be hugely different?

Solution: Entropy is heavily dependent on likelihood and could wildly fluctuate if the probability changes. Hence, one situation in which the two conditional entropies will differ hugely is when an outcome of one variable greatly changes the distribution of results in the other variable, and that outcome is reasonably likely to occur.

Section 16: Joint Entropy and Mutual Information

A cruel and unusual teacher distributes grades according to the flip of a coin. Suppose X is a random variable which represents the outcome of the coin. That is, $X = 1$ if the coin is heads and $X = 0$ if the coin is tails. Now, suppose Y is a random variable representing the grade given by the teacher. If the coin is heads, $Y = 100$ with probability $p = 0.5$, $Y = 80$ with probability $p = 0.25$ and $Y = 60$ with probability $p = 0.25$. On the other hand, if the coin is tails, $Y = 100$ with probability 0, $Y = 80$ with probability $p = \frac{1}{3}$ and $Y = 60$ with probability $p = \frac{2}{3}$.

(a) Determine the conditional entropy of Y given X, $H(Y|X)$.

Solution:

(b) Determine the conditional entropy of X given Y, $H(X|Y)$.

Solution:

(c) Determine the joint entropy $H(X, Y)$ and confirm that

$$H(X, Y) = H(Y|X) + H(X) = H(X|Y) + H(Y).$$

Solution:

Section 17: Mutual Information

30. Prove that the mutual information, $I(X; Y)$ satisfies

$$I(X; Y) \geq 0.$$

Solution: Mathematically, this can be proven using Jensen's inequality. From a logical standpoint, mutual information is the quantity $H(X) - H(X|Y)$, where entropy is the idea of randomness. If we know the outcome of the random variable Y, we certainly can't know less about the outcome of X. At worst, we gain no new information, in which case the mutual information is zero. If we gain information, then $H(X|Y) < H(X)$, and so mutual information is always non-negative.

31. Given a random variable X, prove that $H(X|X) = 0$ and hence $I(X; X) = H(X)$.

Solution: For any term in the sum, $p(X = x_1|X = x_2) = 0$ if $x_1 \neq x_2$ or $\log_2(p(X = x_1|X = x_2)) = 0$ if $x_1 = x_2$. Thus, the sum is zero, and $H(X|X) = 0$. Therefore, $I(X; X) = H(X)$.

Section 18: Mutual Information Cont.

X and Y are two random variables such that $X = 0$ with $p = 0.3$ and $X = 1$ with $p = 0.7$. If $X = 0$, then $Y = 0$ with $p = 0.5$ and $Y = 1$ with $p = 0.5$. On other hand, if $X = 1$, then $Y = 0$ with $p = 0.1$ and $Y = 1$ with $p = 0.9$.

32. Compute the entropy of X, $H(X)$.

33. Compute the entropy of Y, $H(Y)$.

34. Compute the conditional entropy of Y given X, $H(Y|X)$.

35. Compute the conditional entropy of X given Y, $H(X|Y)$.

36. Compute $H(Y|X) - H(Y)$ and $H(X|Y) - H(X)$ and verify the two values are the same. This is the mutual information.

Section 19: Asymmetric Channels

Every afternoon, John makes a cup of tea. He makes green tea with probability $p = 0.1$, after which he chooses to work with probability $p = 0.2$, take a nap with probability $p = 0.3$, or eat a snack with probability $p = 0.5$. John makes black tea with probability $p = 0.3$ after which he chooses to work with probability $p = 0.4$, take a nap with probability $p = 0.2$, or eat a snack with probability $p = 0.4$. Finally, John makes yellow tea with probability $p = 0.6$ after which he chooses to work with probability $p = 0.7$, take a nap with probability $p = 0.1$, or eat a snack with probability $p = 0.2$.

37. What is the entropy for T, the type of tea which John makes?

38. What is the entropy for A, the activity that John chooses?

39. What is $H(A|T)$?

40. What is $H(T|A)$?

41. What is the mutual information?