Elements of Information Theory
Second Edition
Solutions to Problems

Thomas M. Cover
Joy A. Thomas

October 17, 2006

Contents

Preface

Here we have the solutions to all the problems in the second edition of Elements of Information Theory. First a word about how the problems and solutions were generated.

The problems arose over the many years the authors taught this course. At first the homework problems and exam problems were generated each week. After a few years of this double duty, the homework problems were rolled forward from previous years and only the exam problems were fresh. So each year, the midterm and final exam problems became candidates for addition to the body of homework problems that you see in the text. The exam problems are necessarily brief, with a point, and reasonable free from time consuming calculation, so the problems in the text for the most part share these properties.

The solutions to the problems were generated by the teaching assistants and graders for the weekly homework assignments and handed back with the graded homeworks in the class immediately following the date the assignment was due. Homeworks were optional and did not enter into the course grade. Nonetheless most students did the homework. A list of the many students who contributed to the solutions is given in the book acknowledgment. In particular, we would like to thank Laura Ekroot, Will Equitz, Don Kimber, Mitchell Trott, Andrew Nobel, Jim Roche, Vittorio Castelli, Mitchell Oslick, Chien-Wen Tseng, Michael Morrell, Marc Goldberg, George Gemelos, Navid Hassanpour, Young-Han Kim, Charles Mathis, Styrmir Sigurjonsson, Jon Yard, Michael Baer, Mung Chiang, Suhas Diggavi, Elza Erkip, Paul Fahn, Garud Iyengar, David Julian, Yiannis Kontoyiannis, Amos Lapidoth, Erik Ordentlich, Sandeep Pombra, Arak Sutivong, Josh Sweetkind-Singer and Assaf Zeevi. We would like to thank Prof. John Gill and Prof. Abbas El Gamal for many interesting problems and solutions.

The solutions therefore show a wide range of personalities and styles, although some of them have been smoothed out over the years by the authors. The best way to look at the solutions is that they offer more than you need to solve the problems. And the solutions in some cases may be awkward or inefficient. We view that as a plus. An instructor can see the extent of the problem by examining the solution but can still improve his or her own version.

The solution manual comes to some 400 pages. We are making electronic copies available to course instructors in PDF. We hope that all the solutions are not put up on an insecure website—it will not be useful to use the problems in the book for homeworks and exams if the solutions can be obtained immediately with a quick Google search. Instead, we will put up a small selected subset of problem solutions on our website, http://www.elementsofinformationtheory.com, available to all. These will be problems that have particularly elegant or long solutions that would not be suitable homework or exam problems.

We have also seen some people trying to sell the solutions manual on Amazon or Ebay. Please note that the Solutions Manual for Elements of Information Theory is copyrighted and any sale or distribution without the permission of the authors is not permitted.

We would appreciate any comments, suggestions and corrections to this solutions manual.

Tom Cover
Durand 121, Information Systems Lab
Stanford University
Stanford, CA 94305.
Ph. 650-723-4505
FAX: 650-723-8473
Email: cover@stanford.edu

Joy Thomas
Stratify
701 N Shoreline Avenue
Mountain View, CA 94043.
Ph. 650-210-2722
FAX: 650-988-2159
Email: joythomas@stanfordalumni.org

Chapter 1

Introduction

Chapter 2

Entropy, Relative Entropy and Mutual Information

1. **Coin flips.** A fair coin is flipped until the first head occurs. Let X denote the number of flips required.

   (a) Find the entropy $H(X)$ in bits. The following expressions may be useful:

   $$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}, \qquad \sum_{n=0}^{\infty} nr^n = \frac{r}{(1-r)^2}.$$

   (b) A random variable X is drawn according to this distribution. Find an "efficient" sequence of yes-no questions of the form, "Is X contained in the set S?" Compare $H(X)$ to the expected number of questions required to determine X.

   Solution:

   (a) The number X of tosses till the first head appears has the geometric distribution with parameter $p = 1/2$, where $P(X = n) = pq^{n-1}$, $n \in \{1, 2, \ldots\}$. Hence the entropy of X is

   $$\begin{aligned} H(X) &= -\sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1}) \\ &= -\left[\sum_{n=0}^{\infty} pq^n \log p + \sum_{n=0}^{\infty} npq^n \log q\right] \\ &= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2} \\ &= \frac{-p \log p - q \log q}{p} \\ &= H(p)/p \text{ bits.} \end{aligned}$$

   If $p = 1/2$, then $H(X) = 2$ bits.

(b) Intuitively, it seems clear that the best questions are those that have equally likely chances of receiving a yes or a no answer. Consequently, one possible guess is that the most "efficient" series of questions is: Is $X = 1$? If not, is $X = 2$? If not, is $X = 3$? ... with a resulting expected number of questions equal to $\sum_{n=1}^{\infty} n(1/2^n) = 2$. This should reinforce the intuition that $H(X)$ is a measure of the uncertainty of X. Indeed in this case, the entropy is exactly the same as the average number of questions needed to define X, and in general $E(\#\text{ of questions}) \geq H(X)$. This problem has an interpretation as a source coding problem. Let $0 =$no, $1 =$yes, $X =$Source, and $Y =$Encoded Source. Then the set of questions in the above procedure can be written as a collection of (X, Y) pairs: $(1, 1)$, $(2, 01)$, $(3, 001)$, etc. . In fact, this intuitively derived code is the optimal (Huffman) code minimizing the expected number of questions.

2. **Entropy of functions.** Let X be a random variable taking on a finite number of values. What is the (general) inequality relationship of $H(X)$ and $H(Y)$ if

(a) $Y = 2^X$?

(b) $Y = \cos X$?

Solution: Let $y = g(x)$. Then

$$p(y) = \sum_{x:\, y=g(x)} p(x).$$

Consider any set of x's that map onto a single y. For this set

$$\sum_{x:\, y=g(x)} p(x) \log p(x) \leq \sum_{x:\, y=g(x)} p(x) \log p(y) = p(y) \log p(y),$$

since log is a monotone increasing function and $p(x) \leq \sum_{x:\, y=g(x)} p(x) = p(y)$. Extending this argument to the entire range of X (and Y), we obtain

$$\begin{aligned} H(X) &= -\sum_x p(x) \log p(x) \\ &= -\sum_y \sum_{x:\, y=g(x)} p(x) \log p(x) \\ &\geq -\sum_y p(y) \log p(y) \\ &= H(Y), \end{aligned}$$

with equality iff g is one-to-one with probability one.

(a) $Y = 2^X$ is one-to-one and hence the entropy, which is just a function of the probabilities (and not the values of a random variable) does not change, i.e., $H(X) = H(Y)$.

(b) $Y = \cos(X)$ is not necessarily one-to-one. Hence all that we can say is that $H(X) \geq H(Y)$, with equality if cosine is one-to-one on the range of X.

3. **Minimum entropy.** What is the minimum value of $H(p_1, ..., p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges over the set of n-dimensional probability vectors? Find all $\mathbf{p}$'s which achieve this minimum.

   Solution: We wish to find *all* probability vectors $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ which minimize

   $$H(\mathbf{p}) = -\sum_i p_i \log p_i.$$

   Now $-p_i \log p_i \geq 0$, with equality iff $p_i = 0$ or 1. Hence the only possible probability vectors which minimize $H(\mathbf{p})$ are those with $p_i = 1$ for some i and $p_j = 0, j \neq i$. There are n such vectors, i.e., $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$, and the minimum value of $H(\mathbf{p})$ is 0.

4. **Entropy of functions of a random variable.** Let X be a discrete random variable. Show that the entropy of a function of X is less than or equal to the entropy of X by justifying the following steps:

   $$\begin{aligned} H(X, g(X)) &\overset{(a)}{=} H(X) + H(g(X) \mid X) && (2.1)\\ &\overset{(b)}{=} H(X); && (2.2)\\ H(X, g(X)) &\overset{(c)}{=} H(g(X)) + H(X \mid g(X)) && (2.3)\\ &\overset{(d)}{\geq} H(g(X)). && (2.4) \end{aligned}$$

   Thus $H(g(X)) \leq H(X)$.

   Solution: *Entropy of functions of a random variable.*

   (a) $H(X, g(X)) = H(X) + H(g(X)|X)$ by the chain rule for entropies.

   (b) $H(g(X)|X) = 0$ since for any particular value of X, g(X) is fixed, and hence $H(g(X)|X) = \sum_x p(x) H(g(X)|X = x) = \sum_x 0 = 0$.

   (c) $H(X, g(X)) = H(g(X)) + H(X|g(X))$ again by the chain rule.

   (d) $H(X|g(X)) \geq 0$, with equality iff X is a function of $g(X)$, i.e., $g(.)$ is one-to-one. Hence $H(X, g(X)) \geq H(g(X))$.

   Combining parts (b) and (d), we obtain $H(X) \geq H(g(X))$.

5. **Zero conditional entropy.** 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 $p(x, y) > 0$.

   Solution: *Zero Conditional Entropy.* Assume that there exists an x, say x_0 and two different values of y, say y_1 and y_2 such that $p(x_0, y_1) > 0$ and $p(x_0, y_2) > 0$. Then $p(x_0) \geq p(x_0, y_1) + p(x_0, y_2) > 0$, and $p(y_1|x_0)$ and $p(y_2|x_0)$ are not equal to 0 or 1. Thus

   $$\begin{aligned} H(Y|X) &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) && (2.5)\\ &\geq p(x_0)(-p(y_1|x_0) \log p(y_1|x_0) - p(y_2|x_0) \log p(y_2|x_0)) && (2.6)\\ &> > 0, && (2.7) \end{aligned}$$

since $-t\log t \geq 0$ for $0 \leq t \leq 1$, and is strictly positive for t not equal to 0 or 1. Therefore the conditional entropy $H(Y|X)$ is 0 if and only if Y is a function of X.

6. **Conditional mutual information vs. unconditional mutual information.** Give examples of joint random variables X, Y and Z such that

   (a) $I(X;Y \mid Z) < I(X;Y)$,

   (b) $I(X;Y \mid Z) > I(X;Y)$.

   Solution: *Conditional mutual information vs. unconditional mutual information.*

   (a) The last corollary to Theorem 2.8.1 in the text states that if $X \to Y \to Z$ that is, if $p(x,y \mid z) = p(x \mid z)p(y \mid z)$ then, $I(X;Y) \geq I(X;Y \mid Z)$. Equality holds if and only if $I(X;Z) = 0$ or X and Z are independent.

   A simple example of random variables satisfying the inequality conditions above is, X is a fair binary random variable and $Y = X$ and $Z = Y$. In this case,

   $$I(X;Y) = H(X) - H(X \mid Y) = H(X) = 1$$

   and,

   $$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z) = 0.$$

   So that $I(X;Y) > I(X;Y \mid Z)$.

   (b) This example is also given in the text. Let X,Y be independent fair binary random variables and let $Z = X + Y$. In this case we have that,

   $$I(X;Y) = 0$$

   and,

   $$I(X;Y \mid Z) = H(X \mid Z) = 1/2.$$

   So $I(X;Y) < I(X;Y \mid Z)$. Note that in this case X,Y,Z are not markov.

7. **Coin weighing.** Suppose one has n coins, among which there may or may not be one counterfeit coin. If there is a counterfeit coin, it may be either heavier or lighter than the other coins. The coins are to be weighed by a balance.

   (a) Find an upper bound on the number of coins n so that k weighings will find the counterfeit coin (if any) and correctly declare it to be heavier or lighter.

   (b) *(Difficult)* What is the coin weighing strategy for $k = 3$ weighings and 12 coins?

   Solution: *Coin weighing.*

   (a) For n coins, there are $2n + 1$ possible situations or "states".

   - One of the n coins is heavier.
   - One of the n coins is lighter.
   - They are all of equal weight.

Each weighing has three possible outcomes - equal, left pan heavier or right pan heavier. Hence with k weighings, there are 3^k possible outcomes and hence we can distinguish between at most 3^k different "states". Hence $2n+1 \leq 3^k$ or $n \leq (3^k-1)/2$.

Looking at it from an information theoretic viewpoint, each weighing gives at most $\log_2 3$ bits of information. There are $2n+1$ possible "states", with a maximum entropy of $\log_2(2n+1)$ bits. Hence in this situation, one would require at least $\log_2(2n+1)/\log_2 3$ weighings to extract enough information for determination of the odd coin, which gives the same result as above.

(b) There are many solutions to this problem. We will give one which is based on the ternary number system.

We may express the numbers $\{-12,-11,\ldots,-1,0,1,\ldots,12\}$ in a ternary number system with alphabet $\{-1,0,1\}$. For example, the number 8 is (-1,0,1) where $-1\times 3^0+0\times 3^1+1\times 3^2=8$. We form the matrix with the representation of the positive numbers as its columns.

	1	2	3	4	5	6	7	8	9	10	11	12	
3^0	1	-1	0	1	-1	0	1	-1	0	1	-1	0	$\Sigma_1 = 0$
3^1	0	1	1	1	-1	-1	-1	0	0	0	1	1	$\Sigma_2 = 2$
3^2	0	0	0	0	1	1	1	1	1	1	1	1	$\Sigma_3 = 8$

Note that the row sums are not all zero. We can negate some columns to make the row sums zero. For example, negating columns 7,9,11 and 12, we obtain

	1	2	3	4	5	6	7	8	9	10	11	12	
3^0	1	-1	0	1	-1	0	-1	-1	0	1	1	0	$\Sigma_1 = 0$
3^1	0	1	1	1	-1	-1	1	0	0	0	-1	-1	$\Sigma_2 = 0$
3^2	0	0	0	0	1	1	-1	1	-1	1	-1	-1	$\Sigma_3 = 0$

Now place the coins on the balance according to the following rule: For weighing $\#i$, place coin n

- On left pan, if $n_i = -1$.
- Aside, if $n_i = 0$.
- On right pan, if $n_i = 1$.

The outcome of the three weighings will find the odd coin if any and tell whether it is heavy or light. The result of each weighing is 0 if both pans are equal, -1 if the left pan is heavier, and 1 if the right pan is heavier. Then the three weighings give the ternary expansion of the index of the odd coin. If the expansion is the same as the expansion in the matrix, it indicates that the coin is heavier. If the expansion is of the opposite sign, the coin is lighter. For example, (0,-1,-1) indicates $(0)3^0+(-1)3+(-1)3^2=-12$, hence coin #12 is heavy, (1,0,-1) indicates #8 is light, (0,0,0) indicates no odd coin.

Why does this scheme work? It is a single error correcting Hamming code for the ternary alphabet (discussed in Section 8.11 in the book). Here are some details.

First note a few properties of the matrix above that was used for the scheme. All the columns are distinct and no two columns add to (0,0,0). Also if any coin

is heavier, it will produce the sequence of weighings that matches its column in the matrix. If it is lighter, it produces the negative of its column as a sequence of weighings. Combining all these facts, we can see that any single odd coin will produce a unique sequence of weighings, and that the coin can be determined from the sequence.

One of the questions that many of you had whether the bound derived in part (a) was actually achievable. For example, can one distinguish 13 coins in 3 weighings? No, not with a scheme like the one above. Yes, under the assumptions under which the bound was derived. The bound did not prohibit the division of coins into halves, neither did it disallow the existence of another coin known to be normal. Under both these conditions, it is possible to find the odd coin of 13 coins in 3 weighings. You could try modifying the above scheme to these cases.

8. **Drawing with and without replacement.** An urn contains r red, w white, and b black balls. Which has higher entropy, drawing $k \geq 2$ balls from the urn with replacement or without replacement? Set it up and show why. (There is both a hard way and a relatively simple way to do this.)

   Solution: *Drawing with and without replacement.* Intuitively, it is clear that if the balls are drawn with replacement, the number of possible choices for the i-th ball is larger, and therefore the conditional entropy is larger. But computing the conditional distributions is slightly involved. It is easier to compute the unconditional entropy.

   - With replacement. In this case the conditional distribution of each draw is the same for every draw. Thus

   $$X_i = \begin{cases} \text{red} & \text{with prob.} \frac{r}{r+w+b} \\ \text{white} & \text{with prob.} \frac{w}{r+w+b} \\ \text{black} & \text{with prob.} \frac{b}{r+w+b} \end{cases} \tag{2.8}$$

   and therefore

   $$H(X_i|X_{i-1}, \ldots, X_1) = H(X_i) \tag{2.9}$$

   $$= \log(r+w+b) - \frac{r}{r+w+b}\log r - \frac{w}{r+w+b}\log w - \frac{b}{r+w+b}\log b \tag{2.10}$$

   - Without replacement. The unconditional probability of the i-th ball being red is still $r/(r+w+b)$, etc. Thus the unconditional entropy $H(X_i)$ is still the same as with replacement. The conditional entropy $H(X_i|X_{i-1}, \ldots, X_1)$ is less than the unconditional entropy, and therefore the entropy of drawing without replacement is lower.

9. **A metric.** A function $\rho(x, y)$ is a metric if for all x, y,

   - $\rho(x, y) \geq 0$
   - $\rho(x, y) = \rho(y, x)$

- $\rho(x,y) = 0$ if and only if $x = y$
- $\rho(x,y) + \rho(y,z) \geq \rho(x,z)$.

(a) Show that $\rho(X,Y) = H(X|Y) + H(Y|X)$ satisfies the first, second and fourth properties above. If we say that $X = Y$ if there is a one-to-one function mapping from X to Y, then the third property is also satisfied, and $\rho(X,Y)$ is a metric.

(b) Verify that $\rho(X,Y)$ can also be expressed as

$$\begin{aligned} \rho(X,Y) &= H(X) + H(Y) - 2I(X;Y) && (2.11)\\ &= H(X,Y) - I(X;Y) && (2.12)\\ &= 2H(X,Y) - H(X) - H(Y). && (2.13) \end{aligned}$$

Solution: *A metric*

(a) Let

$$\rho(X,Y) = H(X|Y) + H(Y|X). \tag{2.14}$$

Then

- Since conditional entropy is always ≥ 0, $\rho(X,Y) \geq 0$.
- The symmetry of the definition implies that $\rho(X,Y) = \rho(Y,X)$.
- By problem 2.6, it follows that $H(Y|X)$ is 0 iff Y is a function of X and $H(X|Y)$ is 0 iff X is a function of Y. Thus $\rho(X,Y)$ is 0 iff X and Y are functions of each other - and therefore are equivalent up to a reversible transformation.
- Consider three random variables X, Y and Z. Then

$$\begin{aligned} H(X|Y) + H(Y|Z) &\geq H(X|Y,Z) + H(Y|Z) && (2.15)\\ &= H(X,Y|Z) && (2.16)\\ &= H(X|Z) + H(Y|X,Z) && (2.17)\\ &\geq H(X|Z), && (2.18) \end{aligned}$$

from which it follows that

$$\rho(X,Y) + \rho(Y,Z) \geq \rho(X,Z). \tag{2.19}$$

Note that the inequality is strict unless $X \to Y \to Z$ forms a Markov Chain and Y is a function of X and Z.

(b) Since $H(X|Y) = H(X) - I(X;Y)$, the first equation follows. The second relation follows from the first equation and the fact that $H(X,Y) = H(X) + H(Y) - I(X;Y)$. The third follows on substituting $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

10. **Entropy of a disjoint mixture.** Let X_1 and X_2 be discrete random variables drawn according to probability mass functions $p_1(\cdot)$ and $p_2(\cdot)$ over the respective alphabets $\mathcal{X}_1 = \{1, 2, \ldots, m\}$ and $\mathcal{X}_2 = \{m+1, \ldots, n\}$. Let

$$X = \begin{cases} X_1, & \text{with probability } \alpha, \\ X_2, & \text{with probability } 1-\alpha. \end{cases}$$

(a) Find $H(X)$ in terms of $H(X_1)$ and $H(X_2)$ and α.

(b) Maximize over α to show that $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$ and interpret using the notion that $2^{H(X)}$ is the effective alphabet size.

Solution: *Entropy.* We can do this problem by writing down the definition of entropy and expanding the various terms. Instead, we will use the algebra of entropies for a simpler proof.

Since X_1 and X_2 have disjoint support sets, we can write

$$X = \begin{cases} X_1 & \text{with probability} \quad \alpha \\ X_2 & \text{with probability} \quad 1-\alpha \end{cases}$$

Define a function of X,

$$\theta = f(X) = \begin{cases} 1 & \text{when } X = X_1 \\ 2 & \text{when } X = X_2 \end{cases}$$

Then as in problem 1, we have

$$\begin{aligned} H(X) &= H(X, f(X)) = H(\theta) + H(X|\theta) \\ &= H(\theta) + p(\theta=1)H(X|\theta=1) + p(\theta=2)H(X|\theta=2) \\ &= H(\alpha) + \alpha H(X_1) + (1-\alpha)H(X_2) \end{aligned}$$

where $H(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$.

11. **A measure of correlation**. Let X_1 and X_2 be identically distributed, but not necessarily independent. Let

$$\rho = 1 - \frac{H(X_2 \mid X_1)}{H(X_1)}.$$

(a) Show $\rho = \frac{I(X_1;X_2)}{H(X_1)}$.

(b) Show $0 \leq \rho \leq 1$.

(c) When is $\rho = 0$?

(d) When is $\rho = 1$?

Solution: *A measure of correlation.* X_1 and X_2 are identically distributed and

$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}$$

(a)

$$\begin{aligned} \rho &= \frac{H(X_1) - H(X_2|X_1)}{H(X_1)} \\ &= \frac{H(X_2) - H(X_2|X_1)}{H(X_1)} \text{ (since } H(X_1) = H(X_2)) \\ &= \frac{I(X_1;X_2)}{H(X_1)}. \end{aligned}$$

(b) Since $0 \le H(X_2|X_1) \le H(X_2) = H(X_1)$, we have

$$0 \le \frac{H(X_2|X_1)}{H(X_1)} \le 1$$

$$0 \le \rho \le 1.$$

(c) $\rho = 0$ iff $I(X_1; X_2) = 0$ iff X_1 and X_2 are independent.

(d) $\rho = 1$ iff $H(X_2|X_1) = 0$ iff X_2 is a function of X_1. By symmetry, X_1 is a function of X_2, i.e., X_1 and X_2 have a one-to-one relationship.

12. **Example of joint entropy.** Let $p(x, y)$ be given by

X \ Y	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

Find

(a) $H(X), H(Y)$.

(b) $H(X \mid Y), H(Y \mid X)$.

(c) $H(X, Y)$.

(d) $H(Y) - H(Y \mid X)$.

(e) $I(X; Y)$.

(f) Draw a Venn diagram for the quantities in (a) through (e).

Solution: *Example of joint entropy*

(a) $H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 = 0.918$ bits $= H(Y)$.

(b) $H(X|Y) = \frac{1}{3}H(X|Y=0) + \frac{2}{3}H(X|Y=1) = 0.667$ bits $= H(Y|X)$.

(c) $H(X, Y) = 3 \times \frac{1}{3}\log 3 = 1.585$ bits.

(d) $H(Y) - H(Y|X) = 0.251$ bits.

(e) $I(X; Y) = H(Y) - H(Y|X) = 0.251$ bits.

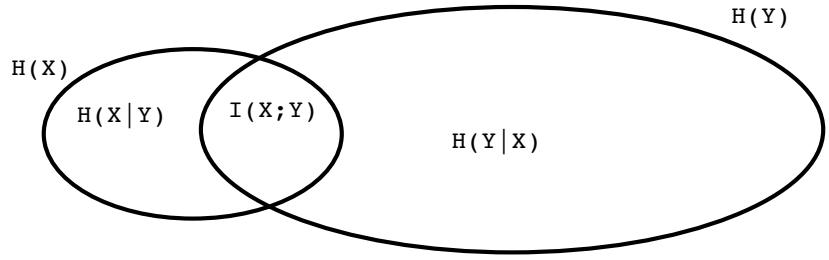
(f) See Figure 1.

13. **Inequality.** Show $\ln x \ge 1 - \frac{1}{x}$ for $x > 0$.

Solution: *Inequality.* Using the Remainder form of the Taylor expansion of $\ln(x)$ about $x = 1$, we have for some c between 1 and x

$$\ln(x) = \ln(1) + \left(\frac{1}{t}\right)_{t=1}(x-1) + \left(\frac{-1}{t^2}\right)_{t=c}\frac{(x-1)^2}{2} \le x - 1$$

Figure 2.1: Venn diagram to illustrate the relationships of entropy and relative entropy



since the second term is always negative. Hence letting $y = 1/x$, we obtain

$$-\ln y \leq \frac{1}{y} - 1$$

or

$$\ln y \geq 1 - \frac{1}{y}$$

with equality iff $y = 1$.

14. **Entropy of a sum.** Let X and Y be random variables that take on values $x_1, x_2, \ldots, x_r$ and $y_1, y_2, \ldots, y_s$, respectively. Let $Z = X + Y$.

    (a) Show that $H(Z|X) = H(Y|X)$. Argue that if X, Y are independent, then $H(Y) \leq H(Z)$ and $H(X) \leq H(Z)$. Thus the addition of *independent* random variables adds uncertainty.

    (b) Give an example of (necessarily dependent) random variables in which $H(X) > H(Z)$ and $H(Y) > H(Z)$.

    (c) Under what conditions does $H(Z) = H(X) + H(Y)$?

    Solution: *Entropy of a sum.*

    (a) $Z = X + Y$. Hence $p(Z = z|X = x) = p(Y = z - x|X = x)$.

$$\begin{aligned} H(Z|X) &= \sum p(x) H(Z|X = x) \\ &= -\sum_x p(x) \sum_z p(Z = z|X = x) \log p(Z = z|X = x) \\ &= \sum_x p(x) \sum_y p(Y = z - x|X = x) \log p(Y = z - x|X = x) \\ &= \sum p(x) H(Y|X = x) \\ &= H(Y|X). \end{aligned}$$

If X and Y are independent, then $H(Y|X) = H(Y)$. Since $I(X;Z) \geq 0$, we have $H(Z) \geq H(Z|X) = H(Y|X) = H(Y)$. Similarly we can show that $H(Z) \geq H(X)$.

(b) Consider the following joint distribution for X and Y Let

$$X = -Y = \begin{cases} 1 & \text{with probability } 1/2 \\ 0 & \text{with probability } 1/2 \end{cases}$$

Then $H(X) = H(Y) = 1$, but $Z = 0$ with prob. 1 and hence $H(Z) = 0$.

(c) We have

$$H(Z) \leq H(X,Y) \leq H(X) + H(Y)$$

because Z is a function of (X,Y) and $H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$. We have equality iff (X,Y) is a function of Z and $H(Y) = H(Y|X)$, i.e., X and Y are independent.

15. **Data processing.** Let $X_1 \to X_2 \to X_3 \to \cdots \to X_n$ form a Markov chain in this order; i.e., let

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)\cdots p(x_n|x_{n-1}).$$

Reduce $I(X_1; X_2, \ldots, X_n)$ to its simplest form.

Solution: *Data Processing.* By the chain rule for mutual information,

$$I(X_1; X_2, \ldots, X_n) = I(X_1; X_2) + I(X_1; X_3|X_2) + \cdots + I(X_1; X_n|X_2, \ldots, X_{n-2}). \quad (2.20)$$

By the Markov property, the past and the future are conditionally independent given the present and hence all terms except the first are zero. Therefore

$$I(X_1; X_2, \ldots, X_n) = I(X_1; X_2). \quad (2.21)$$

16. **Bottleneck.** Suppose a (non-stationary) Markov chain starts in one of n states, necks down to $k < n$ states, and then fans back to $m > k$ states. Thus $X_1 \to X_2 \to X_3$, i.e., $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$, for all $x_1 \in \{1, 2, \ldots, n\}$, $x_2 \in \{1, 2, \ldots, k\}$, $x_3 \in \{1, 2, \ldots, m\}$.

(a) Show that the dependence of X_1 and X_3 is limited by the bottleneck by proving that $I(X_1; X_3) \leq \log k$.

(b) Evaluate $I(X_1; X_3)$ for $k = 1$, and conclude that no dependence can survive such a bottleneck.

Solution:

Bottleneck.

(a) From the data processing inequality, and the fact that entropy is maximum for a uniform distribution, we get

$$\begin{aligned} I(X_1; X_3) &\leq I(X_1; X_2) \\ &= H(X_2) - H(X_2 \mid X_1) \\ &\leq H(X_2) \\ &\leq \log k. \end{aligned}$$

Thus, the dependence between X_1 and X_3 is limited by the size of the bottleneck. That is $I(X_1; X_3) \leq \log k$.

(b) For $k = 1$, $I(X_1; X_3) \leq \log 1 = 0$ and since $I(X_1, X_3) \geq 0$, $I(X_1, X_3) = 0$. Thus, for $k = 1$, X_1 and X_3 are independent.

17. **Pure randomness and bent coins.** Let $X_1, X_2, \ldots, X_n$ denote the outcomes of independent flips of a *bent* coin. Thus $\Pr\{X_i = 1\} = p$, $\Pr\{X_i = 0\} = 1 - p$, where p is unknown. We wish to obtain a sequence $Z_1, Z_2, \ldots, Z_K$ of *fair* coin flips from $X_1, X_2, \ldots, X_n$. Toward this end let $f : \mathcal{X}^n \to \{0,1\}^*$, (where $\{0,1\}^* = \{\Lambda, 0, 1, 00, 01, \ldots\}$ is the set of all finite length binary sequences), be a mapping $f(X_1, X_2, \ldots, X_n) = (Z_1, Z_2, \ldots, Z_K)$, where $Z_i \sim$ Bernoulli $(\frac{1}{2})$, and K may depend on $(X_1, \ldots, X_n)$. In order that the sequence $Z_1, Z_2, \ldots$ appear to be fair coin flips, the map f from bent coin flips to fair flips must have the property that all 2^k sequences $(Z_1, Z_2, \ldots, Z_k)$ of a given length k have equal probability (possibly 0), for $k = 1, 2, \ldots$. For example, for $n = 2$, the map $f(01) = 0$, $f(10) = 1$, $f(00) = f(11) = \Lambda$ (the null string), has the property that $\Pr\{Z_1 = 1 | K = 1\} = \Pr\{Z_1 = 0 | K = 1\} = \frac{1}{2}$.

Give reasons for the following inequalities:

$$\begin{aligned} nH(p) &\overset{(a)}{=} H(X_1, \ldots, X_n) \\ &\overset{(b)}{\geq} H(Z_1, Z_2, \ldots, Z_K, K) \\ &\overset{(c)}{=} H(K) + H(Z_1, \ldots, Z_K | K) \\ &\overset{(d)}{=} H(K) + E(K) \\ &\overset{(e)}{\geq} EK. \end{aligned}$$

Thus no more than $nH(p)$ fair coin tosses can be derived from $(X_1, \ldots, X_n)$, on the average. Exhibit a good map f on sequences of length 4.

Solution: *Pure randomness and bent coins.*

$$\begin{aligned} nH(p) &\overset{(a)}{=} H(X_1, \ldots, X_n) \\ &\overset{(b)}{\geq} H(Z_1, Z_2, \ldots, Z_K) \end{aligned}$$

$$\begin{aligned} &\overset{(c)}{=} H(Z_1, Z_2, \ldots, Z_K, K) \\ &\overset{(d)}{=} H(K) + H(Z_1, \ldots, Z_K|K) \\ &\overset{(e)}{=} H(K) + E(K) \\ &\overset{(f)}{\geq} EK \ . \end{aligned}$$

(a) Since $X_1, X_2, \ldots, X_n$ are i.i.d. with probability of $X_i = 1$ being p, the entropy $H(X_1, X_2, \ldots, X_n)$ is $nH(p)$.

(b) $Z_1, \ldots, Z_K$ is a function of $X_1, X_2, \ldots, X_n$, and since the entropy of a function of a random variable is less than the entropy of the random variable, $H(Z_1, \ldots, Z_K) \leq H(X_1, X_2, \ldots, X_n)$.

(c) K is a function of $Z_1, Z_2, \ldots, Z_K$, so its conditional entropy given $Z_1, Z_2, \ldots, Z_K$ is 0. Hence $H(Z_1, Z_2, \ldots, Z_K, K) = H(Z_1, \ldots, Z_K) + H(K|Z_1, Z_2, \ldots, Z_K) = H(Z_1, Z_2, \ldots, Z_K)$.

(d) Follows from the chain rule for entropy.

(e) By assumption, $Z_1, Z_2, \ldots, Z_K$ are pure random bits (given K), with entropy 1 bit per symbol. Hence

$$\begin{aligned} H(Z_1, Z_2, \ldots, Z_K|K) &= \sum_k p(K=k) H(Z_1, Z_2, \ldots, Z_k|K=k) && (2.22) \\ &= \sum_k p(k)k && (2.23) \\ &= EK. && (2.24) \end{aligned}$$

(f) Follows from the non-negativity of discrete entropy.

(g) Since we do not know p, the only way to generate pure random bits is to use the fact that all sequences with the same number of ones are equally likely. For example, the sequences 0001,0010,0100 and 1000 are equally likely and can be used to generate 2 pure random bits. An example of a mapping to generate random bits is

$$\begin{array}{cccc} 0000 \to \Lambda & & & \\ 0001 \to 00 & 0010 \to 01 & 0100 \to 10 & 1000 \to 11 \\ 0011 \to 00 & 0110 \to 01 & 1100 \to 10 & 1001 \to 11 \\ 1010 \to 0 & 0101 \to 1 & & \\ 1110 \to 11 & 1101 \to 10 & 1011 \to 01 & 0111 \to 00 \\ 1111 \to \Lambda & & & \end{array} \qquad (2.25)$$

The resulting expected number of bits is

$$\begin{aligned} EK &= 4pq^3 \times 2 + 4p^2q^2 \times 2 + 2p^2q^2 \times 1 + 4p^3q \times 2 && (2.26) \\ &= 8pq^3 + 10p^2q^2 + 8p^3q. && (2.27) \end{aligned}$$

For example, for $p \approx \frac{1}{2}$, the expected number of pure random bits is close to 1.625. This is substantially less then the 4 pure random bits that could be generated if p were exactly $\frac{1}{2}$.

We will now analyze the efficiency of this scheme of generating random bits for long sequences of bent coin flips. Let n be the number of bent coin flips. The algorithm that we will use is the obvious extension of the above method of generating pure bits using the fact that all sequences with the same number of ones are equally likely.

Consider all sequences with k ones. There are $\binom{n}{k}$ such sequences, which are all equally likely. If $\binom{n}{k}$ were a power of 2, then we could generate $\log \binom{n}{k}$ pure random bits from such a set. However, in the general case, $\binom{n}{k}$ is not a power of 2 and the best we can to is the divide the set of $\binom{n}{k}$ elements into subset of sizes which are powers of 2. The largest set would have a size $2^{\lfloor \log \binom{n}{k} \rfloor}$ and could be used to generate $\lfloor \log \binom{n}{k} \rfloor$ random bits. We could divide the remaining elements into the largest set which is a power of 2, etc. The worst case would occur when $\binom{n}{k} = 2^{l+1} - 1$, in which case the subsets would be of sizes $2^l, 2^{l-1}, 2^{l-2}, \ldots, 1$.

Instead of analyzing the scheme exactly, we will just find a lower bound on number of random bits generated from a set of size $\binom{n}{k}$. Let $l = \lfloor \log \binom{n}{k} \rfloor$. Then at least half of the elements belong to a set of size 2^l and would generate l random bits, at least $\frac{1}{4}$th belong to a set of size 2^{l-1} and generate $l-1$ random bits, etc. On the average, the number of bits generated is

$$\begin{aligned} E[K|k \text{ 1's in sequence}] &\geq \frac{1}{2}l + \frac{1}{4}(l-1) + \cdots + \frac{1}{2^l}1 && (2.28)\\ &= l - \frac{1}{4}\left(1 + \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \cdots + \frac{l-1}{2^{l-2}}\right) && (2.29)\\ &\geq l - 1, && (2.30) \end{aligned}$$

since the infinite series sums to 1.

Hence the fact that $\binom{n}{k}$ is not a power of 2 will cost at most 1 bit on the average in the number of random bits that are produced.

Hence, the expected number of pure random bits produced by this algorithm is

$$\begin{aligned} EK &\geq \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log \binom{n}{k} - 1 \rfloor && (2.31)\\ &\geq \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left(\log \binom{n}{k} - 2\right) && (2.32)\\ &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 2 && (2.33)\\ &\geq \sum_{n(p-\epsilon) \leq k \leq n(p+\epsilon)} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 2. && (2.34) \end{aligned}$$

Now for sufficiently large n, the probability that the number of 1's in the sequence is close to np is near 1 (by the weak law of large numbers). For such sequences, $\frac{k}{n}$ is close to p and hence there exists a δ such that

$$\binom{n}{k} \geq 2^{n(H(\frac{k}{n})-\delta)} \geq 2^{n(H(p)-2\delta)} \tag{2.35}$$

using Stirling's approximation for the binomial coefficients and the continuity of the entropy function. If we assume that n is large enough so that the probability that $n(p-\epsilon) \leq k \leq n(p+\epsilon)$ is greater than $1-\epsilon$, then we see that $EK \geq (1-\epsilon)n(H(p)-2\delta)-2$, which is very good since $nH(p)$ is an upper bound on the number of pure random bits that can be produced from the bent coin sequence.

18. **World Series.** The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; possible values of X are AAAA, BABABAB, and BBBAAAA. Let Y be the number of games played, which ranges from 4 to 7. Assuming that A and B are equally matched and that the games are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, and $H(X|Y)$.

Solution:

World Series. Two teams play until one of them has won 4 games.

There are 2 (AAAA, BBBB) World Series with 4 games. Each happens with probability $(1/2)^4$.

There are $8 = 2\binom{4}{3}$ World Series with 5 games. Each happens with probability $(1/2)^5$.

There are $20 = 2\binom{5}{3}$ World Series with 6 games. Each happens with probability $(1/2)^6$.

There are $40 = 2\binom{6}{3}$ World Series with 7 games. Each happens with probability $(1/2)^7$.

The probability of a 4 game series $(Y=4)$ is $2(1/2)^4 = 1/8$.

The probability of a 5 game series $(Y=5)$ is $8(1/2)^5 = 1/4$.

The probability of a 6 game series $(Y=6)$ is $20(1/2)^6 = 5/16$.

The probability of a 7 game series $(Y=7)$ is $40(1/2)^7 = 5/16$.

$$\begin{aligned} H(X) &= \sum p(x) log \frac{1}{p(x)} \\ &= 2(1/16)\log 16 + 8(1/32)\log 32 + 20(1/64)\log 64 + 40(1/128)\log 128 \\ &= 5.8125 \end{aligned}$$

$$\begin{aligned} H(Y) &= \sum p(y) log \frac{1}{p(y)} \\ &= 1/8\log 8 + 1/4\log 4 + 5/16\log(16/5) + 5/16\log(16/5) \\ &= 1.924 \end{aligned}$$

Y is a deterministic function of X, so if you know X there is no randomness in Y. Or, $H(Y|X) = 0$.

Since $H(X) + H(Y|X) = H(X,Y) = H(Y) + H(X|Y)$, it is easy to determine $H(X|Y) = H(X) + H(Y|X) - H(Y) = 3.889$

19. **Infinite entropy.** This problem shows that the entropy of a discrete random variable can be infinite. Let $A = \sum_{n=2}^{\infty} (n \log^2 n)^{-1}$. (It is easy to show that A is finite by bounding the infinite sum by the integral of $(x \log^2 x)^{-1}$.) Show that the integer-valued random variable X defined by $\Pr(X = n) = (An \log^2 n)^{-1}$ for $n = 2, 3, \ldots$, has $H(X) = +\infty$.

Solution: *Infinite entropy.* By definition, $p_n = \Pr(X = n) = 1/An \log^2 n$ for $n \geq 2$. Therefore

$$\begin{aligned} H(X) &= -\sum_{n=2}^{\infty} p(n) \log p(n) \\ &= -\sum_{n=2}^{\infty} \left(1/An \log^2 n\right) \log \left(1/An \log^2 n\right) \\ &= \sum_{n=2}^{\infty} \frac{\log(An \log^2 n)}{An \log^2 n} \\ &= \sum_{n=2}^{\infty} \frac{\log A + \log n + 2 \log\log n}{An \log^2 n} \\ &= \log A + \sum_{n=2}^{\infty} \frac{1}{An \log n} + \sum_{n=2}^{\infty} \frac{2 \log\log n}{An \log^2 n}. \end{aligned}$$

The first term is finite. For base 2 logarithms, all the elements in the sum in the last term are nonnegative. (For any other base, the terms of the last sum eventually all become positive.) So all we have to do is bound the middle sum, which we do by comparing with an integral.

$$\sum_{n=2}^{\infty} \frac{1}{An \log n} > \int_2^{\infty} \frac{1}{Ax \log x} dx = K \ln\ln x \Big|_2^{\infty} = +\infty.$$

We conclude that $H(X) = +\infty$.

20. **Run length coding.** Let $X_1, X_2, \ldots, X_n$ be (possibly dependent) binary random variables. Suppose one calculates the run lengths $\mathbf{R} = (R_1, R_2, \ldots)$ of this sequence (in order as they occur). For example, the sequence $\mathbf{X} = 0001100100$ yields run lengths $\mathbf{R} = (3, 2, 2, 1, 2)$. Compare $H(X_1, X_2, \ldots, X_n)$, $H(\mathbf{R})$ and $H(X_n, \mathbf{R})$. Show all equalities and inequalities, and bound all the differences.

Solution: *Run length coding.* Since the run lengths are a function of $X_1, X_2, \ldots, X_n$, $H(\mathbf{R}) \leq H(\mathbf{X})$. Any X_i together with the run lengths determine the entire sequence

$X_1, X_2, \ldots, X_n$. Hence

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_i, \mathbf{R}) && (2.36)\\ &= H(\mathbf{R}) + H(X_i|\mathbf{R}) && (2.37)\\ &\leq H(\mathbf{R}) + H(X_i) && (2.38)\\ &\leq H(\mathbf{R}) + 1. && (2.39) \end{aligned}$$

21. **Markov's inequality for probabilities.** Let $p(x)$ be a probability mass function. Prove, for all $d \geq 0$,

$$\Pr\{p(X) \leq d\} \log\left(\frac{1}{d}\right) \leq H(X). \tag{2.40}$$

Solution: *Markov inequality applied to entropy.*

$$\begin{aligned} P(p(X) < d) \log \frac{1}{d} &= \sum_{x:p(x)<d} p(x) \log \frac{1}{d} && (2.41)\\ &\leq \sum_{x:p(x)<d} p(x) \log \frac{1}{p(x)} && (2.42)\\ &\leq \sum_{x} p(x) \log \frac{1}{p(x)} && (2.43)\\ &= H(X) && (2.44) \end{aligned}$$

22. **Logical order of ideas.** Ideas have been developed in order of need, and then generalized if necessary. Reorder the following ideas, strongest first, implications following:

    (a) Chain rule for $I(X_1, \ldots, X_n; Y)$, chain rule for $D(p(x_1, \ldots, x_n)||q(x_1, x_2, \ldots, x_n))$, and chain rule for $H(X_1, X_2, \ldots, X_n)$.

    (b) $D(f||g) \geq 0$, Jensen's inequality, $I(X;Y) \geq 0$.

Solution: *Logical ordering of ideas.*

(a) The following orderings are subjective. Since $I(X;Y) = D(p(x,y)||p(x)p(y))$ is a special case of relative entropy, it is possible to derive the chain rule for I from the chain rule for D.

Since $H(X) = I(X;X)$, it is possible to derive the chain rule for H from the chain rule for I.

It is also possible to derive the chain rule for I from the chain rule for H as was done in the notes.

(b) In class, Jensen's inequality was used to prove the non-negativity of D. The inequality $I(X;Y) \geq 0$ followed as a special case of the non-negativity of D.

23. **Conditional mutual information.** Consider a sequence of n binary random variables $X_1, X_2, \ldots, X_n$. Each sequence with an even number of 1's has probability $2^{-(n-1)}$ and each sequence with an odd number of 1's has probability 0. Find the mutual informations

$$I(X_1; X_2), \quad I(X_2; X_3|X_1), \ldots, \ I(X_{n-1}; X_n|X_1, \ldots, X_{n-2}).$$

Solution: *Conditional mutual information.*

Consider a sequence of n binary random variables $X_1, X_2, \ldots, X_n$. Each sequence of length n with an even number of 1's is equally likely and has probability $2^{-(n-1)}$.

Any $n-1$ or fewer of these are independent. Thus, for $k \leq n-1$,

$$I(X_{k-1}; X_k|X_1, X_2, \ldots, X_{k-2}) = 0.$$

However, given $X_1, X_2, \ldots, X_{n-2}$, we know that once we know either X_{n-1} or X_n we know the other.

$$\begin{aligned} I(X_{n-1}; X_n|X_1, X_2, \ldots, X_{n-2}) &= H(X_n|X_1, X_2, \ldots, X_{n-2}) - H(X_n|X_1, X_2, \ldots, X_{n-1}) \\ &= 1 - 0 = 1 \text{ bit.} \end{aligned}$$

24. **Average entropy.** Let $H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$ be the binary entropy function.

(a) Evaluate $H(1/4)$ using the fact that $\log_2 3 \approx 1.584$. *Hint:* You may wish to consider an experiment with four equally likely outcomes, one of which is more interesting than the others.

(b) Calculate the average entropy $H(p)$ when the probability p is chosen uniformly in the range $0 \leq p \leq 1$.

(c) *(Optional)* Calculate the average entropy $H(p_1, p_2, p_3)$ where (p_1, p_2, p_3) is a uniformly distributed probability vector. Generalize to dimension n.

Solution: *Average Entropy.*

(a) We can generate two bits of information by picking one of four equally likely alternatives. This selection can be made in two steps. First we decide whether the first outcome occurs. Since this has probability $1/4$, the information generated is $H(1/4)$. If not the first outcome, then we select one of the three remaining outcomes; with probability $3/4$, this produces $\log_2 3$ bits of information. Thus

$$H(1/4) + (3/4) \log_2 3 = 2$$

and so $H(1/4) = 2 - (3/4) \log_2 3 = 2 - (.75)(1.585) = 0.811$ bits.

(b) If p is chosen uniformly in the range $0 \le p \le 1$, then the average entropy (in nats) is

$$-\int_0^1 p\ln p + (1-p)\ln(1-p)dp = -2\int_0^1 x\ln x\,dx = -2\left(\frac{x^2}{2}\ln x + \frac{x^2}{4}\right)\Big|_0^1 = \tfrac{1}{2}\,.$$

Therefore the average entropy is $\frac{1}{2}\log_2 e = 1/(2\ln 2) = .721$ bits.

(c) Choosing a uniformly distributed probability vector (p_1, p_2, p_3) is equivalent to choosing a point (p_1, p_2) uniformly from the triangle $0 \le p_1 \le 1$, $p_1 \le p_2 \le 1$. The probability density function has the constant value 2 because the area of the triangle is 1/2. So the average entropy $H(p_1, p_2, p_3)$ is

$$-2\int_0^1\int_{p_1}^1 p_1\ln p_1 + p_2\ln p_2 + (1-p_1-p_2)\ln(1-p_1-p_2)dp_2dp_1\,.$$

After some enjoyable calculus, we obtain the final result $5/(6\ln 2) = 1.202$ bits.

25. **Venn diagrams.** There isn't realy a notion of mutual information common to three random variables. Here is one attempt at a definition: Using Venn diagrams, we can see that the mutual information common to three random variables X, Y and Z can be defined by

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z)\,.$$

This quantity is symmetric in X, Y and Z, despite the preceding asymmetric definition. Unfortunately, $I(X;Y;Z)$ is not necessarily nonnegative. Find X, Y and Z such that $I(X;Y;Z) < 0$, and prove the following two identities:

(a) $I(X;Y;Z) = H(X,Y,Z) - H(X) - H(Y) - H(Z) + I(X;Y) + I(Y;Z) + I(Z;X)$

(b) $I(X;Y;Z) = H(X,Y,Z) - H(X,Y) - H(Y,Z) - H(Z,X) + H(X) + H(Y) + H(Z)$

The first identity can be understood using the Venn diagram analogy for entropy and mutual information. The second identity follows easily from the first.

Solution: *Venn Diagrams.* To show the first identity,

$$\begin{aligned}
I(X;Y;Z) &= I(X;Y) - I(X;Y|Z) \quad \text{by definition}\\
&= I(X;Y) - (I(X;Y,Z) - I(X;Z)) \quad \text{by chain rule}\\
&= I(X;Y) + I(X;Z) - I(X;Y,Z)\\
&= I(X;Y) + I(X;Z) - (H(X) + H(Y,Z) - H(X,Y,Z))\\
&= I(X;Y) + I(X;Z) - H(X) + H(X,Y,Z) - H(Y,Z)\\
&= I(X;Y) + I(X;Z) - H(X) + H(X,Y,Z) - (H(Y) + H(Z) - I(Y;Z))\\
&= I(X;Y) + I(X;Z) + I(Y;Z) + H(X,Y,Z) - H(X) - H(Y) - H(Z).
\end{aligned}$$

To show the second identity, simply substitute for $I(X;Y)$, $I(X;Z)$, and $I(Y;Z)$ using equations like

$$I(X;Y) = H(X) + H(Y) - H(X,Y)\,.$$

These two identities show that $I(X;Y;Z)$ is a symmetric (but not necessarily nonnegative) function of three random variables.

26. **Another proof of non-negativity of relative entropy.** In view of the fundamental nature of the result $D(p||q) \geq 0$, we will give another proof.

(a) Show that $\ln x \leq x - 1$ for $0 < x < \infty$.

(b) Justify the following steps:

$$\begin{aligned} -D(p||q) &= \sum_x p(x) \ln \frac{q(x)}{p(x)} && (2.45) \\ &\leq \sum_x p(x) \left(\frac{q(x)}{p(x)} - 1 \right) && (2.46) \\ &\leq 0 && (2.47) \end{aligned}$$

(c) What are the conditions for equality?

Solution: *Another proof of non-negativity of relative entropy.* In view of the fundamental nature of the result $D(p||q) \geq 0$, we will give another proof.

(a) Show that $\ln x \leq x - 1$ for $0 < x < \infty$.
There are many ways to prove this. The easiest is using calculus. Let

$$f(x) = x - 1 - \ln x \tag{2.48}$$

for $0 < x < \infty$. Then $f'(x) = 1 - \frac{1}{x}$ and $f''(x) = \frac{1}{x^2} > 0$, and therefore $f(x)$ is strictly convex. Therefore a local minimum of the function is also a global minimum. The function has a local minimum at the point where $f'(x) = 0$, i.e., when $x = 1$. Therefore $f(x) \geq f(1)$, i.e.,

$$x - 1 - \ln x \geq 1 - 1 - \ln 1 = 0 \tag{2.49}$$

which gives us the desired inequality. Equality occurs only if $x = 1$.

(b) We let A be the set of x such that $p(x) > 0$.

$$\begin{aligned} -D_e(p||q) &= \sum_{x \in A} p(x) ln \frac{q(x)}{p(x)} && (2.50) \\ &\leq \sum_{x \in A} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) && (2.51) \\ &= \sum_{x \in A} q(x) - \sum_{x \in A} p(x) && (2.52) \\ &\leq 0 && (2.53) \end{aligned}$$

The first step follows from the definition of D, the second step follows from the inequality $\ln t \leq t - 1$, the third step from expanding the sum, and the last step from the fact that the $q(A) \leq 1$ and $p(A) = 1$.

(c) What are the conditions for equality?

We have equality in the inequality $\ln t \le t-1$ if and only if $t=1$. Therefore we have equality in step 2 of the chain iff $q(x)/p(x)=1$ for all $x \in A$. This implies that $p(x)=q(x)$ for all x, and we have equality in the last step as well. Thus the condition for equality is that $p(x)=q(x)$ for all x.

27. **Grouping rule for entropy:** Let $\mathbf{p}=(p_1,p_2,\ldots,p_m)$ be a probability distribution on m elements, i.e, $p_i \ge 0$, and $\sum_{i=1}^{m} p_i = 1$. Define a new distribution $\mathbf{q}$ on $m-1$ elements as $q_1=p_1$, $q_2=p_2,\ldots$, $q_{m-2}=p_{m-2}$, and $q_{m-1}=p_{m-1}+p_m$, i.e., the distribution $\mathbf{q}$ is the same as $\mathbf{p}$ on $\{1,2,\ldots,m-2\}$, and the probability of the last element in $\mathbf{q}$ is the sum of the last two probabilities of $\mathbf{p}$. Show that

$$H(\mathbf{p}) = H(\mathbf{q}) + (p_{m-1}+p_m)H\left(\frac{p_{m-1}}{p_{m-1}+p_m}, \frac{p_m}{p_{m-1}+p_m}\right). \tag{2.54}$$

Solution:

$$\begin{aligned}
H(\mathbf{p}) &= -\sum_{i=1}^{m} p_i \log p_i && (2.55)\\
&= -\sum_{i=1}^{m-2} p_i \log p_i - p_{m-1}\log p_{m-1} - p_m \log p_m && (2.56)\\
&= -\sum_{i=1}^{m-2} p_i \log p_i - p_{m-1}\log \frac{p_{m-1}}{p_{m-1}+p_m} - p_m \log \frac{p_m}{p_{m-1}+p_m} && (2.57)\\
&\quad -(p_{m-1}+p_m)\log(p_{m-1}+p_m) && (2.58)\\
&= H(\mathbf{q}) - p_{m-1}\log \frac{p_{m-1}}{p_{m-1}+p_m} - p_m \log \frac{p_m}{p_{m-1}+p_m} && (2.59)\\
&= H(\mathbf{q}) - (p_{m-1}+p_m)\left(\frac{p_{m-1}}{p_{m-1}+p_m}\log\frac{p_{m-1}}{p_{m-1}+p_m} - \frac{p_m}{p_{m-1}+p_m}\log\frac{p_m}{p_{m-1}+p_m}\right) && (2.60)\\
&= H(\mathbf{q}) + (p_{m-1}+p_m)H_2\left(\frac{p_{m-1}}{p_{m-1}+p_m}, \frac{p_m}{p_{m-1}+p_m}\right), && (2.61)
\end{aligned}$$

where $H_2(a,b) = -a\log a - b\log b$.

28. **Mixing increases entropy.** Show that the entropy of the probability distribution, $(p_1,\ldots,p_i,\ldots,p_j,\ldots,p_m)$, is less than the entropy of the distribution $(p_1,\ldots,\frac{p_i+p_j}{2},\ldots,\frac{p_i+p_j}{2},\ldots,p_m)$. Show that in general any transfer of probability that makes the distribution more uniform increases the entropy.

Solution:

Mixing increases entropy.

This problem depends on the convexity of the log function. Let

$$\begin{aligned}
P_1 &= (p_1,\ldots,p_i,\ldots,p_j,\ldots,p_m)\\
P_2 &= (p_1,\ldots,\frac{p_i+p_j}{2},\ldots,\frac{p_j+p_i}{2},\ldots,p_m)
\end{aligned}$$

Then, by the log sum inequality,

$$\begin{aligned} H(P_2) - H(P_1) &= -2(\frac{p_i + p_j}{2}) \log(\frac{p_i + p_j}{2}) + p_i \log p_i + p_j \log p_j \\ &= -(p_i + p_j) \log(\frac{p_i + p_j}{2}) + p_i \log p_i + p_j \log p_j \\ &\geq 0. \end{aligned}$$

Thus,

$$H(P_2) \geq H(P_1).$$

29. **Inequalities.** Let X, Y and Z be joint random variables. Prove the following inequalities and find conditions for equality.

(a) $H(X, Y|Z) \geq H(X|Z)$.

(b) $I(X, Y; Z) \geq I(X; Z)$.

(c) $H(X, Y, Z) - H(X, Y) \leq H(X, Z) - H(X)$.

(d) $I(X; Z|Y) \geq I(Z; Y|X) - I(Z; Y) + I(X; Z)$.

Solution: *Inequalities.*

(a) Using the chain rule for conditional entropy,

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z) \geq H(X|Z),$$

with equality iff $H(Y|X, Z) = 0$, that is, when Y is a function of X and Z.

(b) Using the chain rule for mutual information,

$$I(X, Y; Z) = I(X; Z) + I(Y; Z|X) \geq I(X; Z),$$

with equality iff $I(Y; Z|X) = 0$, that is, when Y and Z are conditionally independent given X.

(c) Using first the chain rule for entropy and then the definition of conditional mutual information,

$$\begin{aligned} H(X, Y, Z) - H(X, Y) &= H(Z|X, Y) = H(Z|X) - I(Y; Z|X) \\ &\leq H(Z|X) = H(X, Z) - H(X), \end{aligned}$$

with equality iff $I(Y; Z|X) = 0$, that is, when Y and Z are conditionally independent given X.

(d) Using the chain rule for mutual information,

$$I(X; Z|Y) + I(Z; Y) = I(X, Y; Z) = I(Z; Y|X) + I(X; Z),$$

and therefore

$$I(X; Z|Y) = I(Z; Y|X) - I(Z; Y) + I(X; Z).$$

We see that this inequality is actually an equality in all cases.

30. **Maximum entropy.** Find the probability mass function $p(x)$ that maximizes the entropy $H(X)$ of a non-negative integer-valued random variable X subject to the constraint

$$EX = \sum_{n=0}^{\infty} np(n) = A$$

for a fixed value $A > 0$. Evaluate this maximum $H(X)$.

Solution: *Maximum entropy*

Recall that,

$$-\sum_{i=0}^{\infty} p_i \log p_i \leq -\sum_{i=0}^{\infty} p_i \log q_i.$$

Let $q_i = \alpha(\beta)^i$. Then we have that,

$$\begin{aligned} -\sum_{i=0}^{\infty} p_i \log p_i &\leq -\sum_{i=0}^{\infty} p_i \log q_i \\ &= -\left(\log(\alpha)\sum_{i=0}^{\infty} p_i + \log(\beta)\sum_{i=0}^{\infty} ip_i\right) \\ &= -\log\alpha - A\log\beta \end{aligned}$$

Notice that the final right hand side expression is independent of $\{p_i\}$, and that the inequality,

$$-\sum_{i=0}^{\infty} p_i \log p_i \leq -\log\alpha - A\log\beta$$

holds for all α, β such that,

$$\sum_{i=0}^{\infty} \alpha\beta^i = 1 = \alpha\frac{1}{1-\beta}.$$

The constraint on the expected value also requires that,

$$\sum_{i=0}^{\infty} i\alpha\beta^i = A = \alpha\frac{\beta}{(1-\beta)^2}.$$

Combining the two constraints we have,

$$\begin{aligned} \alpha\frac{\beta}{(1-\beta)^2} &= \left(\frac{\alpha}{1-\beta}\right)\left(\frac{\beta}{1-\beta}\right) \\ &= \frac{\beta}{1-\beta} \\ &= A, \end{aligned}$$

which implies that,

$$\begin{aligned}\beta &= \frac{A}{A+1} \\ \alpha &= \frac{1}{A+1}.\end{aligned}$$

So the entropy maximizing distribution is,

$$p_i = \frac{1}{A+1}\left(\frac{A}{A+1}\right)^i.$$

Plugging these values into the expression for the maximum entropy,

$$-\log\alpha - A\log\beta = (A+1)\log(A+1) - A\log A.$$

The general form of the distribution,

$$p_i = \alpha\beta^i$$

can be obtained either by guessing or by Lagrange multipliers where,

$$F(p_i, \lambda_1, \lambda_2) = -\sum_{i=0}^{\infty} p_i \log p_i + \lambda_1(\sum_{i=0}^{\infty} p_i - 1) + \lambda_2(\sum_{i=0}^{\infty} i p_i - A)$$

is the function whose gradient we set to 0.

To complete the argument with Lagrange multipliers, it is necessary to show that the local maximum is the global maximum. One possible argument is based on the fact that $-H(p)$ is convex, it has only one local minima, no local maxima and therefore Lagrange multiplier actually gives the global maximum for $H(p)$.

31. **Conditional entropy.** Under what conditions does $H(X \mid g(Y)) = H(X \mid Y)$?

Solution: (*Conditional Entropy*). If $H(X|g(Y)) = H(X|Y)$, then $H(X) - H(X|g(Y)) = H(X) - H(X|Y)$, i.e., $I(X; g(Y)) = I(X; Y)$. This is the condition for equality in the data processing inequality. From the derivation of the inequality, we have equality iff $X \to g(Y) \to Y$ forms a Markov chain. Hence $H(X|g(Y)) = H(X|Y)$ iff $X \to g(Y) \to Y$. This condition includes many special cases, such as g being one-to-one, and X and Y being independent. However, these two special cases do not exhaust all the possibilities.

32. **Fano.** We are given the following joint distribution on (X, Y)

X \ Y	a	b	c
1	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$
2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
3	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$

Let $\hat{X}(Y)$ be an estimator for X (based on Y) and let $P_e = \Pr\{\hat{X}(Y) \neq X\}$.

(a) Find the minimum probability of error estimator $\hat{X}(Y)$ and the associated P_e.

(b) Evaluate Fano's inequality for this problem and compare.

Solution:

(a) From inspection we see that

$$\hat{X}(y) = \begin{cases} 1 & y = a \\ 2 & y = b \\ 3 & y = c \end{cases}$$

Hence the associated P_e is the sum of $P(1,b)$, $P(1,c)$, $P(2,a)$, $P(2,c)$, $P(3,a)$ and $P(3,b)$. Therefore, $P_e = 1/2$.

(b) From Fano's inequality we know

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$

Here,

$$\begin{aligned} H(X|Y) &= H(X|Y=a)\Pr\{y=a\} + H(X|Y=b)\Pr\{y=b\} + H(X|Y=c)\Pr\{y=c\} \\ &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)\Pr\{y=a\} + H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)\Pr\{y=b\} + H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)\Pr\{y=c\} \\ &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)(\Pr\{y=a\} + \Pr\{y=b\} + \Pr\{y=c\}) \\ &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right) \\ &= 1.5 \text{ bits.} \end{aligned}$$

Hence

$$P_e \geq \frac{1.5 - 1}{\log 3} = .316.$$

Hence our estimator $\hat{X}(Y)$ is not very close to Fano's bound in this form. If $\hat{X} \in \mathcal{X}$, as it does here, we can use the stronger form of Fano's inequality to get

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}|\text{-}1)}.$$

and

$$P_e \geq \frac{1.5 - 1}{\log 2} = \frac{1}{2}.$$

Therefore our estimator $\hat{X}(Y)$ is actually quite good.

33. **Fano's inequality.** Let $\Pr(X = i) = p_i$, $i = 1, 2, \ldots, m$ and let $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_m$. The minimal probability of error predictor of X is $\hat{X} = 1$, with resulting probability of error $P_e = 1 - p_1$. Maximize $H(\mathbf{p})$ subject to the constraint $1 - p_1 = P_e$

to find a bound on P_e in terms of H. This is Fano's inequality in the absence of conditioning.

Solution: (*Fano's Inequality.*) The minimal probability of error predictor when there is no information is $\hat{X} = 1$, the most probable value of X. The probability of error in this case is $P_e = 1 - p_1$. Hence if we fix P_e, we fix p_1. We maximize the entropy of X for a given P_e to obtain an upper bound on the entropy for a given P_e. The entropy,

$$\begin{aligned} H(\mathbf{p}) &= -p_1 \log p_1 - \sum_{i=2}^{m} p_i \log p_i & (2.62)\\ &= -p_1 \log p_1 - \sum_{i=2}^{m} P_e \frac{p_i}{P_e} \log \frac{p_i}{P_e} - P_e \log P_e & (2.63)\\ &= H(P_e) + P_e H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \ldots, \frac{p_m}{P_e}\right) & (2.64)\\ &\leq H(P_e) + P_e \log(m-1), & (2.65) \end{aligned}$$

since the maximum of $H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \ldots, \frac{p_m}{P_e}\right)$ is attained by an uniform distribution. Hence any X that can be predicted with a probability of error P_e must satisfy

$$H(X) \leq H(P_e) + P_e \log(m-1), \tag{2.66}$$

which is the unconditional form of Fano's inequality. We can weaken this inequality to obtain an explicit lower bound for P_e,

$$P_e \geq \frac{H(X) - 1}{\log(m-1)}. \tag{2.67}$$

34. **Entropy of initial conditions.** Prove that $H(X_0|X_n)$ is non-decreasing with n for any Markov chain.

Solution: *Entropy of initial conditions.* For a Markov chain, by the data processing theorem, we have

$$I(X_0; X_{n-1}) \geq I(X_0; X_n). \tag{2.68}$$

Therefore

$$H(X_0) - H(X_0|X_{n-1}) \geq H(X_0) - H(X_0|X_n) \tag{2.69}$$

or $H(X_0|X_n)$ increases with n.

35. **Relative entropy is not symmetric:** Let the random variable X have three possible outcomes $\{a, b, c\}$. Consider two distributions on this random variable

Symbol	$p(x)$	$q(x)$
a	1/2	1/3
b	1/4	1/3
c	1/4	1/3

Calculate $H(p)$, $H(q)$, $D(p||q)$ and $D(q||p)$. Verify that in this case $D(p||q) \neq D(q||p)$.

Solution:

$$H(p) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = 1.5 \text{ bits.} \tag{2.70}$$

$$H(q) = \frac{1}{3}\log 3 + \frac{1}{3}\log 3 + \frac{1}{3}\log 3 = \log 3 = 1.58496 \text{ bits.} \tag{2.71}$$

$$D(p||q) = \frac{1}{2}\log\frac{3}{2} + \frac{1}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{3}{4} = \log(3) - 1.5 = 1.58496 - 1.5 = 0.08496 \tag{2.72}$$

$$D(q||p) = \frac{1}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{4}{3} + \frac{1}{3}\log\frac{4}{3} = \frac{5}{3} - \log(3) = 1.66666 - 1.58496 = 0.08170 \tag{2.73}$$

36. **Symmetric relative entropy:** Though, as the previous example shows, $D(p||q) \neq D(q||p)$ in general, there could be distributions for which equality holds. Give an example of two distributions p and q on a binary alphabet such that $D(p||q) = D(q||p)$ (other than the trivial case $p = q$).

Solution:

A simple case for $D((p, 1-p)||(q, 1-q)) = D((q, 1-q)||(p, 1-p))$, i.e., for

$$p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p} \tag{2.74}$$

is when $q = 1 - p$.

37. **Relative entropy:** Let X, Y, Z be three random variables with a joint probability mass function $p(x, y, z)$. The relative entropy between the joint distribution and the product of the marginals is

$$D(p(x,y,z)||p(x)p(y)p(z)) = E\left[\log\frac{p(x,y,z)}{p(x)p(y)p(z)}\right] \tag{2.75}$$

Expand this in terms of entropies. When is this quantity zero?

Solution:

$$\begin{aligned} D(p(x,y,z)||p(x)p(y)p(z)) &= E\left[\log\frac{p(x,y,z)}{p(x)p(y)p(z)}\right] && (2.76)\\ &= E[\log p(x,y,z)] - E[\log p(x)] - E[\log p(y)] - E[\log p(z)] && (2.77)\\ &= -H(X,Y,Z) + H(X) + H(Y) + H(Z) && (2.78) \end{aligned}$$

We have $D(p(x,y,z)||p(x)p(y)p(z)) = 0$ if and only $p(x,y,z) = p(x)p(y)p(z)$ for all (x, y, z), i.e., if X and Y and Z are independent.

38. **The value of a question** Let $X \sim p(x)$, $x = 1, 2, \ldots, m$. We are given a set $S \subseteq \{1, 2, \ldots, m\}$. We ask whether $X \in S$ and receive the answer

$$Y = \begin{cases} 1, & \text{if } X \in S \\ 0, & \text{if } X \notin S. \end{cases}$$

Suppose $\Pr\{X \in S\} = \alpha$. Find the decrease in uncertainty $H(X) - H(X|Y)$.

Apparently any set S with a given α is as good as any other.

Solution: *The value of a question.*

$$\begin{aligned} H(X) - H(X|Y) &= I(X;Y) \\ &= H(Y) - H(Y|X) \\ &= H(\alpha) - H(Y|X) \\ &= H(\alpha) \end{aligned}$$

since $H(Y|X) = 0$.

39. **Entropy and pairwise independence.**
Let X, Y, Z be three binary Bernoulli $(\frac{1}{2})$ random variables that are pairwise independent, that is, $I(X;Y) = I(X;Z) = I(Y;Z) = 0$.

(a) Under this constraint, what is the minimum value for $H(X,Y,Z)$?

(b) Give an example achieving this minimum.

Solution:

(a)

$$\begin{aligned} H(X,Y,Z) &= H(X,Y) + H(Z|X,Y) && (2.79) \\ &\geq H(X,Y) && (2.80) \\ &= 2. && (2.81) \end{aligned}$$

So the minimum value for $H(X,Y,Z)$ is at least 2. To show that is is actually equal to 2, we show in part (b) that this bound is attainable.

(b) Let X and Y be iid Bernoulli($\frac{1}{2}$) and let $Z = X \oplus Y$, where $\oplus$ denotes addition mod 2 (xor).

40. **Discrete entropies**

Let X and Y be two independent integer-valued random variables. Let X be uniformly distributed over $\{1, 2, \ldots, 8\}$, and let $\Pr\{Y = k\} = 2^{-k}$, $\quad k = 1, 2, 3, \ldots$

(a) Find $H(X)$

(b) Find $H(Y)$

(c) Find $H(X + Y, X - Y)$.

Solution:

(a) For a uniform distribution, $H(X) = \log m = \log 8 = 3$.

(b) For a geometric distribution, $H(Y) = \sum_k k2^{-k} = 2$. (See solution to problem 2.1

(c) Since $(X,Y) \to (X+Y, X-Y)$ is a one to one transformation, $H(X+Y, X-Y) = H(X,Y) = H(X) + H(Y) = 3 + 2 = 5$.

41. **Random questions**

One wishes to identify a random object $X \sim p(x)$. A question $Q \sim r(q)$ is asked at random according to $r(q)$. This results in a deterministic answer $A = A(x,q) \in \{a_1, a_2, \ldots\}$. Suppose X and Q are independent. Then $I(X;Q,A)$ is the uncertainty in X removed by the question-answer (Q,A).

(a) Show $I(X;Q,A) = H(A|Q)$. Interpret.

(b) Now suppose that two i.i.d. questions $Q_1, Q_2, \sim r(q)$ are asked, eliciting answers A_1 and A_2. Show that two questions are less valuable than twice a single question in the sense that $I(X;Q_1,A_1,Q_2,A_2) \leq 2I(X;Q_1,A_1)$.

Solution: *Random questions.*

(a)

$$\begin{aligned} I(X;Q,A) &= H(Q,A) - H(Q,A,|X) \\ &= H(Q) + H(A|Q) - H(Q|X) - H(A|Q,X) \\ &= H(Q) + H(A|Q) - H(Q) \\ &= H(A|Q) \end{aligned}$$

The interpretation is as follows. The uncertainty removed in X by (Q,A) is the same as the uncertainty in the answer given the question.

(b) Using the result from part a and the fact that questions are independent, we can easily obtain the desired relationship.

$$\begin{aligned} I(X;Q_1,A_1,Q_2,A_2) &\overset{(a)}{=} I(X;Q_1) + I(X;A_1|Q_1) + I(X;Q_2|A_1,Q_1) + I(X;A_2|A_1,Q_1,Q_2) \\ &\overset{(b)}{=} I(X;A_1|Q_1) + H(Q_2|A_1,Q_1) - H(Q_2|X,A_1,Q_1) + I(X;A_2|A_1,Q_1,Q_2) \\ &\overset{(c)}{=} I(X;A_1|Q_1) + I(X;A_2|A_1,Q_1,Q_2) \\ &= I(X;A_1|Q_1) + H(A_2|A_1,Q_1,Q_2) - H(A_2|X,A_1,Q_1,Q_2) \\ &\overset{(d)}{=} I(X;A_1|Q_1) + H(A_2|A_1,Q_1,Q_2) \\ &\overset{(e)}{\leq} I(X;A_1|Q_1) + H(A_2|Q_2) \\ &\overset{(f)}{=} 2I(X;A_1|Q_1) \end{aligned}$$

(a) Chain rule.
(b) X and Q_1 are independent.

(c) Q_2 are independent of X, Q_1, and A_1.
(d) A_2 is completely determined given Q_2 and X.
(e) Conditioning decreases entropy.
(f) Result from part a.

42. **Inequalities**. Which of the following inequalities are generally $\geq, =, \leq$? Label each with $\geq, =,$ or $\leq$.

(a) $H(5X)$ vs. $H(X)$
(b) $I(g(X); Y)$ vs. $I(X; Y)$
(c) $H(X_0|X_{-1})$ vs. $H(X_0|X_{-1}, X_1)$
(d) $H(X_1, X_2, \ldots, X_n)$ vs. $H(c(X_1, X_2, \ldots, X_n))$, where $c(x_1, x_2, \ldots, x_n)$ is the Huffman codeword assigned to $(x_1, x_2, \ldots, x_n)$.
(e) $H(X, Y)/(H(X) + H(Y))$ vs. 1

Solution:

(a) $X \to 5X$ is a one to one mapping, and hence $H(X) = H(5X)$.
(b) By data processing inequality, $I(g(X); Y) \leq I(X; Y)$.
(c) Because conditioning reduces entropy, $H(X_0|X_{-1}) \geq H(X_0|X_{-1}, X_1)$.
(d) $H(X, Y) \leq H(X) + H(Y)$, so $H(X, Y)/(H(X) + H(Y)) \leq 1$.

43. **Mutual information of heads and tails.**

(a) Consider a fair coin flip. What is the mutual information between the top side and the bottom side of the coin?
(b) A 6-sided fair die is rolled. What is the mutual information between the top side and the front face (the side most facing you)?

Solution:

Mutual information of heads and tails.

To prove (a) observe that

$$\begin{aligned} I(T; B) &= H(B) - H(B|T) \\ &= \log 2 = 1 \end{aligned}$$

since $B \sim Ber(1/2)$, and $B = f(T)$. Here B, T stand for Bottom and Top respectively.

To prove (b) note that having observed a side of the cube facing us F, there are four possibilities for the top T, which are equally probable. Thus,

$$\begin{aligned} I(T; F) &= H(T) - H(T|F) \\ &= \log 6 - \log 4 \\ &= \log 3 - 1 \end{aligned}$$

since T has uniform distribution on $\{1, 2, \ldots, 6\}$.

44. **Pure randomness**

We wish to use a 3-sided coin to generate a fair coin toss. Let the coin X have probability mass function

$$X = \begin{cases} A, & p_A \\ B, & p_B \\ C, & p_C \end{cases}$$

where p_A, p_B, p_C are unknown.

(a) How would you use 2 independent flips X_1, X_2 to generate (if possible) a Bernoulli($\frac{1}{2}$) random variable Z?

(b) What is the resulting maximum expected number of fair bits generated?

Solution:

(a) The trick here is to notice that for any two letters Y and Z produced by two independent tosses of our bent three-sided coin, YZ has the same probability as ZY. So we can produce $B \sim$Bernoulli($\frac{1}{2}$) coin flips by letting $B = 0$ when we get AB, BC or AC, and $B = 1$ when we get BA, CB or CA (if we get AA, BB or CC we don't assign a value to B.)

(b) The expected number of bits generated by the above scheme is as follows. We get one bit, except when the two flips of the 3-sided coin produce the same symbol. So the expected number of fair bits generated is

$$0 * [P(AA) + P(BB) + P(CC)] + 1 * [1 - P(AA) - P(BB) - P(CC)], \quad (2.82)$$

$$\text{or,} \quad 1 - p_A^2 - p_B^2 - p_C^2. \quad (2.83)$$

45. **Finite entropy.** Show that for a discrete random variable $X \in \{1, 2, \ldots\}$, if $E \log X < \infty$, then $H(X) < \infty$.

Solution: Let the distribution on the integers be $p_1, p_2, \ldots$. Then $H(p) = -\sum p_i log p_i$ and $E \log X = \sum p_i log i = c < \infty$.

We will now find the maximum entropy distribution subject to the constraint on the expected logarithm. Using Lagrange multipliers or the results of Chapter 12, we have the following functional to optimize

$$J(p) = -\sum p_i \log p_i - \lambda_1 \sum p_i - \lambda_2 \sum p_i \log i \quad (2.84)$$

Differentiating with respect to p_i and setting to zero, we find that the p_i that maximizes the entropy set $p_i = a i^{\lambda}$, where $a = 1/(\sum i^{\lambda})$ and λ chosed to meet the expected log constraint, i.e.

$$\sum i^{\lambda} \log i = c \sum i^{\lambda} \quad (2.85)$$

Using this value of p_i, we can see that the entropy is finite.

46. **Axiomatic definition of entropy.** If we assume certain axioms for our measure of information, then we will be forced to use a logarithmic measure like entropy. Shannon used this to justify his initial definition of entropy. In this book, we will rely more on the other properties of entropy rather than its axiomatic derivation to justify its use. The following problem is considerably more difficult than the other problems in this section.

If a sequence of symmetric functions $H_m(p_1, p_2, \ldots, p_m)$ satisfies the following properties,

- Normalization: $H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$,
- Continuity: $H_2(p, 1-p)$ is a continuous function of p,
- Grouping: $H_m(p_1, p_2, \ldots, p_m) = H_{m-1}(p_1+p_2, p_3, \ldots, p_m)+(p_1+p_2)H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$,

prove that H_m must be of the form

$$H_m(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i, \qquad m = 2, 3, \ldots. \tag{2.86}$$

There are various other axiomatic formulations which also result in the same definition of entropy. See, for example, the book by Csiszár and Körner[3].

Solution: *Axiomatic definition of entropy.* This is a long solution, so we will first outline what we plan to do. First we will extend the grouping axiom by induction and prove that

$$\begin{aligned} H_m(p_1, p_2, \ldots, p_m) &= H_{m-k}(p_1 + p_2 + \cdots + p_k, p_{k+1}, \ldots, p_m) \\ &\quad + (p_1 + p_2 + \cdots + p_k) H_k\left(\frac{p_1}{p_1 + p_2 + \cdots + p_k}, \ldots, \frac{p_k}{p_1 + p_2 + \cdots + p_k}\right) \end{aligned} \tag{2.87}$$

Let $f(m)$ be the entropy of a uniform distribution on m symbols, i.e.,

$$f(m) = H_m\left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right). \tag{2.88}$$

We will then show that for any two integers r and s, that $f(rs) = f(r) + f(s)$. We use this to show that $f(m) = \log m$. We then show for rational $p = r/s$, that $H_2(p, 1-p) = -p \log p - (1-p) \log(1-p)$. By continuity, we will extend it to irrational p and finally by induction and grouping, we will extend the result to H_m for $m \geq 2$.

To begin, we extend the grouping axiom. For convenience in notation, we will let

$$S_k = \sum_{i=1}^{k} p_i \tag{2.89}$$

and we will denote $H_2(q, 1-q)$ as $h(q)$. Then we can write the grouping axiom as

$$H_m(p_1, \ldots, p_m) = H_{m-1}(S_2, p_3, \ldots, p_m) + S_2 h\left(\frac{p_2}{S_2}\right). \tag{2.90}$$

Applying the grouping axiom again, we have

$$\begin{aligned} H_m(p_1,\ldots,p_m) &= H_{m-1}(S_2,p_3,\ldots,p_m) + S_2 h\left(\frac{p_2}{S_2}\right) && (2.91)\\ &= H_{m-2}(S_3,p_4,\ldots,p_m) + S_3 h\left(\frac{p_3}{S_3}\right) + S_2 h\left(\frac{p_2}{S_2}\right) && (2.92)\\ &\vdots && (2.93)\\ &= H_{m-(k-1)}(S_k,p_{k+1},\ldots,p_m) + \sum_{i=2}^{k} S_i h\left(\frac{p_i}{S_i}\right). && (2.94) \end{aligned}$$

Now, we apply the same grouping axiom repeatedly to $H_k(p_1/S_k,\ldots,p_k/S_k)$, to obtain

$$\begin{aligned} H_k\left(\frac{p_1}{S_k},\ldots,\frac{p_k}{S_k}\right) &= H_2\left(\frac{S_{k-1}}{S_k},\frac{p_k}{S_k}\right) + \sum_{i=2}^{k-1}\frac{S_i}{S_k} h\left(\frac{p_i/S_k}{S_i/S_k}\right) && (2.95)\\ &= \frac{1}{S_k}\sum_{i=2}^{k} S_i h\left(\frac{p_i}{S_i}\right). && (2.96) \end{aligned}$$

From (2.94) and (2.96), it follows that

$$H_m(p_1,\ldots,p_m) = H_{m-k}(S_k,p_{k+1},\ldots,p_m) + S_k H_k\left(\frac{p_1}{S_k},\ldots,\frac{p_k}{S_k}\right), \qquad (2.97)$$

which is the extended grouping axiom.

Now we need to use an axiom that is not explicitly stated in the text, namely that the function H_m is symmetric with respect to its arguments. Using this, we can combine any set of arguments of H_m using the extended grouping axiom.

Let $f(m)$ denote $H_m(\frac{1}{m},\frac{1}{m},\ldots,\frac{1}{m})$.

Consider

$$f(mn) = H_{mn}(\frac{1}{mn},\frac{1}{mn},\ldots,\frac{1}{mn}). \qquad (2.98)$$

By repeatedly applying the extended grouping axiom, we have

$$\begin{aligned} f(mn) &= H_{mn}(\frac{1}{mn},\frac{1}{mn},\ldots,\frac{1}{mn}) && (2.99)\\ &= H_{mn-n}(\frac{1}{m},\frac{1}{mn},\ldots,\frac{1}{mn}) + \frac{1}{m}H_n(\frac{1}{n},\ldots,\frac{1}{n}) && (2.100)\\ &= H_{mn-2n}(\frac{1}{m},\frac{1}{m},\frac{1}{mn},\ldots,\frac{1}{mn}) + \frac{2}{m}H_n(\frac{1}{n},\ldots,\frac{1}{n}) && (2.101)\\ &\vdots && (2.102)\\ &= H_m(\frac{1}{m},\ldots.\frac{1}{m}) + H(\frac{1}{n},\ldots,\frac{1}{n}) && (2.103)\\ &= f(m) + f(n). && (2.104) \end{aligned}$$

We can immediately use this to conclude that $f(m^k) = kf(m)$.

Now, we will argue that $H_2(1,0) = h(1) = 0$. We do this by expanding $H_3(p_1, p_2, 0)$ $(p_1 + p_2 = 1)$ in two different ways using the grouping axiom

$$\begin{aligned} H_3(p_1, p_2, 0) &= H_2(p_1, p_2) + p_2 H_2(1,0) && (2.105)\\ &= H_2(1,0) + (p_1 + p_2) H_2(p_1, p_2) && (2.106) \end{aligned}$$

Thus $p_2 H_2(1,0) = H_2(1,0)$ for all p_2, and therefore $H(1,0) = 0$.

We will also need to show that $f(m+1) - f(m) \to 0$ as $m \to \infty$. To prove this, we use the extended grouping axiom and write

$$\begin{aligned} f(m+1) &= H_{m+1}(\frac{1}{m+1}, \ldots, \frac{1}{m+1}) && (2.107)\\ &= h(\frac{1}{m+1}) + \frac{m}{m+1} H_m(\frac{1}{m}, \ldots, \frac{1}{m}) && (2.108)\\ &= h(\frac{1}{m+1}) + \frac{m}{m+1} f(m) && (2.109) \end{aligned}$$

and therefore

$$f(m+1) - \frac{m}{m+1} f(m) = h(\frac{1}{m+1}). \quad (2.110)$$

Thus $\lim f(m+1) - \frac{m}{m+1} f(m) = \lim h(\frac{1}{m+1})$. But by the continuity of H_2, it follows that the limit on the right is $h(0) = 0$. Thus $\lim h(\frac{1}{m+1}) = 0$.

Let us define

$$a_{n+1} = f(n+1) - f(n) \quad (2.111)$$

and

$$b_n = h(\frac{1}{n}). \quad (2.112)$$

Then

$$\begin{aligned} a_{n+1} &= -\frac{1}{n+1} f(n) + b_{n+1} && (2.113)\\ &= -\frac{1}{n+1} \sum_{i=2}^{n} a_i + b_{n+1} && (2.114) \end{aligned}$$

and therefore

$$(n+1) b_{n+1} = (n+1) a_{n+1} + \sum_{i=2}^{n} a_i. \quad (2.115)$$

Therefore summing over n, we have

$$\sum_{n=2}^{N} n b_n = \sum_{n=2}^{N} (n a_n + a_{n-1} + \ldots + a_2) = N \sum_{n=2}^{N} a_i. \quad (2.116)$$

Dividing both sides by $\sum_{n=1}^{N} n = N(N+1)/2$, we obtain

$$\frac{2}{N+1}\sum_{n=2}^{N} a_n = \frac{\sum_{n=2}^{N} n b_n}{\sum_{n=2}^{N} n} \tag{2.117}$$

Now by continuity of H_2 and the definition of b_n, it follows that $b_n \to 0$ as $n \to \infty$. Since the right hand side is essentially an average of the b_n's, it also goes to 0 (This can be proved more precisely using ϵ's and δ's). Thus the left hand side goes to 0. We can then see that

$$a_{N+1} = b_{N+1} - \frac{1}{N+1}\sum_{n=2}^{N} a_n \tag{2.118}$$

also goes to 0 as $N \to \infty$. Thus

$$f(n+1) - f(n) \to 0 \quad \text{as} n \to \infty. \tag{2.119}$$

We will now prove the following lemma

Lemma 2.0.1 *Let the function $f(m)$ satisfy the following assumptions:*

- $f(mn) = f(m) + f(n)$ *for all integers* m, n.
- $\lim_{n\to\infty}(f(n+1) - f(n)) = 0$
- $f(2) = 1$,

then the function $f(m) = \log_2 m$.

Proof of the lemma: Let P be an arbitrary prime number and let

$$g(n) = f(n) - \frac{f(P)\log_2 n}{\log_2 P} \tag{2.120}$$

Then $g(n)$ satisfies the first assumption of the lemma. Also $g(P) = 0$.

Also if we let

$$\alpha_n = g(n+1) - g(n) = f(n+1) - f(n) + \frac{f(P)}{\log_2 P}\log_2\frac{n}{n+1} \tag{2.121}$$

then the second assumption in the lemma implies that $\lim \alpha_n = 0$.

For an integer n, define

$$n^{(1)} = \left\lfloor \frac{n}{P} \right\rfloor. \tag{2.122}$$

Then it follows that $n^{(1)} < n/P$, and

$$n = n^{(1)}P + l \tag{2.123}$$

where $0 \leq l < P$. From the fact that $g(P) = 0$, it follows that $g(Pn^{(1)}) = g(n^{(1)})$, and

$$g(n) = g(n^{(1)}) + g(n) - g(Pn^{(1)}) = g(n^{(1)}) + \sum_{i=Pn^{(1)}}^{n-1} \alpha_i \tag{2.124}$$

Just as we have defined $n^{(1)}$ from n, we can define $n^{(2)}$ from $n^{(1)}$. Continuing this process, we can then write

$$g(n) = g(n^{(k)}) + \sum_{j=1}^{k} \left(\sum_{i=Pn^{(i)}}^{n^{(i-1)}} \alpha_i \right). \tag{2.125}$$

Since $n^{(k)} \leq n/P^k$, after

$$k = \left\lfloor \frac{\log n}{\log P} \right\rfloor + 1 \tag{2.126}$$

terms, we have $n^{(k)} = 0$, and $g(0) = 0$ (this follows directly from the additive property of g). Thus we can write

$$g(n) = \sum_{i=1}^{t_n} \alpha_i \tag{2.127}$$

the sum of b_n terms, where

$$b_n \leq P \left(\frac{\log n}{\log P} + 1 \right). \tag{2.128}$$

Since $\alpha_n \to 0$, it follows that $\frac{g(n)}{\log_2 n} \to 0$, since $g(n)$ has at most $o(\log_2 n)$ terms α_i. Thus it follows that

$$\lim_{n \to \infty} \frac{f(n)}{\log_2 n} = \frac{f(P)}{\log_2 P} \tag{2.129}$$

Since P was arbitrary, it follows that $f(P)/\log_2 P = c$ for every prime number P. Applying the third axiom in the lemma, it follows that the constant is 1, and $f(P) = \log_2 P$.

For composite numbers $N = P_1 P_2 \ldots P_l$, we can apply the first property of f and the prime number factorization of N to show that

$$f(N) = \sum f(P_i) = \sum \log_2 P_i = \log_2 N. \tag{2.130}$$

Thus the lemma is proved.

The lemma can be simplified considerably, if instead of the second assumption, we replace it by the assumption that $f(n)$ is monotone in n. We will now argue that the only function $f(m)$ such that $f(mn) = f(m) + f(n)$ for all integers m, n is of the form $f(m) = \log_a m$ for some base a.

Let $c = f(2)$. Now $f(4) = f(2 \times 2) = f(2) + f(2) = 2c$. Similarly, it is easy to see that $f(2^k) = kc = c \log_2 2^k$. We will extend this to integers that are not powers of 2.

For any integer m, let $r > 0$, be another integer and let $2^k \le m^r < 2^{k+1}$. Then by the monotonicity assumption on f, we have

$$kc \le rf(m) < (k+1)c \tag{2.131}$$

or

$$c\frac{k}{r} \le f(m) < c\frac{k+1}{r} \tag{2.132}$$

Now by the monotonicity of $\log$, we have

$$\frac{k}{r} \le \log_2 m < \frac{k+1}{r} \tag{2.133}$$

Combining these two equations, we obtain

$$\left| f(m) - \frac{\log_2 m}{c} \right| < \frac{1}{r} \tag{2.134}$$

Since r was arbitrary, we must have

$$f(m) = \frac{\log_2 m}{c} \tag{2.135}$$

and we can identify $c = 1$ from the last assumption of the lemma.

Now we are almost done. We have shown that for any uniform distribution on m outcomes, $f(m) = H_m(1/m, \ldots, 1/m) = \log_2 m$.

We will now show that

$$H_2(p, 1-p) = -p \log p - (1-p)\log(1-p). \tag{2.136}$$

To begin, let p be a rational number, r/s, say. Consider the extended grouping axiom for H_s

$$f(s) = H_s(\frac{1}{s}, \ldots, \frac{1}{s}) = H(\underbrace{\frac{1}{s}, \ldots, \frac{1}{s}}_{r}, \frac{s-r}{s}) + \frac{s-r}{s} f(s-r) \tag{2.137}$$

$$= H_2(\frac{r}{s}, \frac{s-r}{s}) + \frac{s}{r} f(s) + \frac{s-r}{s} f(s-r) \tag{2.138}$$

Substituting $f(s) = \log_2 s$, etc, we obtain

$$H_2(\frac{r}{s}, \frac{s-r}{s}) = -\frac{r}{s}\log_2 \frac{r}{s} - \left(1 - \frac{s-r}{s}\right) \log_2 \left(1 - \frac{s-r}{s}\right). \tag{2.139}$$

Thus (2.136) is true for rational p. By the continuity assumption, (2.136) is also true at irrational p.

To complete the proof, we have to extend the definition from H_2 to H_m, i.e., we have to show that

$$H_m(p_1, \ldots, p_m) = -\sum p_i \log p_i \tag{2.140}$$

for all m. This is a straightforward induction. We have just shown that this is true for $m = 2$. Now assume that it is true for $m = n - 1$. By the grouping axiom,

$$\begin{aligned} H_n(p_1, \ldots, p_n) &= H_{n-1}(p_1 + p_2, p_3, \ldots, p_n) && (2.141)\\ &\quad + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) && (2.142)\\ &= -(p_1 + p_2)\log(p_1 + p_2) - \sum_{i=3}^{n} p_i \log p_i && (2.143)\\ &\quad - \frac{p_1}{p_1 + p_2}\log\frac{p_1}{p_1 + p_2} - \frac{p_2}{p_1 + p_2}\log\frac{p_2}{p_1 + p_2} && (2.144)\\ &= -\sum_{i=1}^{n} p_i \log p_i. && (2.145) \end{aligned}$$

Thus the statement is true for $m = n$, and by induction, it is true for all m. Thus we have finally proved that the only symmetric function that satisfies the axioms is

$$H_m(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i. \tag{2.146}$$

The proof above is due to Rényi[10]

47. **The entropy of a missorted file.**
A deck of n cards in order $1, 2, \ldots, n$ is provided. One card is removed at random then replaced at random. What is the entropy of the resulting deck?

Solution: *The entropy of a missorted file.*

The heart of this problem is simply carefully counting the possible outcome states. There are n ways to choose which card gets mis-sorted, and, once the card is chosen, there are again n ways to choose where the card is replaced in the deck. Each of these shuffling actions has probability $1/n^2$. Unfortunately, not all of these n^2 actions results in a unique mis-sorted file. So we need to carefully count the number of distinguishable outcome states. The resulting deck can only take on one of the following three cases.

- The selected card is at its original location after a replacement.
- The selected card is at most one location away from its original location after a replacement.
- The selected card is at least two locations away from its original location after a replacement.

To compute the entropy of the resulting deck, we need to know the probability of each case.

Case 1 (resulting deck is the same as the original): There are n ways to achieve this outcome state, one for each of the n cards in the deck. Thus, the probability associated with case 1 is $n/n^2 = 1/n$.

Case 2 (adjacent pair swapping): There are $n-1$ adjacent pairs, each of which will have a probability of $2/n^2$, since for each pair, there are two ways to achieve the swap, either by selecting the left-hand card and moving it one to the right, or by selecting the right-hand card and moving it one to the left.

Case 3 (typical situation): None of the remaining actions "collapses". They all result in unique outcome states, each with probability $1/n^2$. Of the n^2 possible shuffling actions, $n^2 - n - 2(n-1)$ of them result in this third case (we've simply subtracted the case 1 and case 2 situations above).

The entropy of the resulting deck can be computed as follows.

$$\begin{aligned} H(X) &= \frac{1}{n}log(n) + (n-1)\frac{2}{n^2}log(\frac{n^2}{2}) + (n^2-3n+2)\frac{1}{n^2}log(n^2) \\ &= \frac{2n-1}{n}log(n) - \frac{2(n-1)}{n^2} \end{aligned}$$

48. **Sequence length.**
How much information does the length of a sequence give about the content of a sequence? Suppose we consider a Bernoulli $(1/2)$ process $\{X_i\}$.

Stop the process when the first 1 appears. Let N designate this stopping time. Thus X^N is an element of the set of all finite length binary sequences $\{0,1\}^* = \{0, 1, 00, 01, 10, 11, 000, \ldots\}$.

(a) Find $I(N; X^N)$.

(b) Find $H(X^N|N)$.

(c) Find $H(X^N)$.

Let's now consider a different stopping time. For this part, again assume $X_i \sim$ Bernoulli $(1/2)$ but stop at time $N = 6$, with probability $1/3$ and stop at time $N = 12$ with probability $2/3$. Let this stopping time be independent of the sequence $X_1X_2 \ldots X_{12}$.

(d) Find $I(N; X^N)$.

(e) Find $H(X^N|N)$.

(f) Find $H(X^N)$.

Solution:

(a)

$$\begin{aligned} I(X^N; N) &= H(N) - H(N|X^N) \\ &= H(N) - 0 \end{aligned}$$

$$\begin{aligned} & \stackrel{(a)}{=} E(N) \\ I(X^N;N) &= 2 \end{aligned}$$

where (a) comes from the fact that the entropy of a geometric random variable is just the mean.

(b) Since given N we know that $X_i = 0$ for all $i < N$ and $X_N = 1$,

$$H(X^N|N) = 0.$$

(c)

$$\begin{aligned} H(X^N) &= I(X^N;N) + H(X^N|N) \\ &= I(X^N;N) + 0 \\ H(X^N) &= 2. \end{aligned}$$

(d)

$$\begin{aligned} I(X^N;N) &= H(N) - H(N|X^N) \\ &= H(N) - 0 \\ I(X^N;N) &= H_B(1/3) \end{aligned}$$

(e)

$$\begin{aligned} H(X^N|N) &= \frac{1}{3}H(X^6|N=6) + \frac{2}{3}H(X^{12}|N=12) \\ &= \frac{1}{3}H(X^6) + \frac{2}{3}H(X^{12}) \\ &= \frac{1}{3}6 + \frac{2}{3}12 \\ H(X^N|N) &= 10. \end{aligned}$$

(f)

$$\begin{aligned} H(X^N) &= I(X^N;N) + H(X^N|N) \\ &= I(X^N;N) + 10 \\ H(X^N) &= H(1/3) + 10. \end{aligned}$$