Entropy

Math 4175

§3.4. Secrecy of Cryptosystems

In the previous section, we discussed the concept of perfect secrecy.

Historically, it has been the goal of cryptography to design cryptosystems where one key can be used to encrypt many messages and still maintain some measure of computational security.

We will discuss some of these systems (Data Encryption Standard and Advanced Encryption Standard) next chapter.

In this section, we will look at what happens when more and more plaintexts are encrypted using the same key, and how likely a cryptanalyst (Oscar) will be able to carry out a successful cipher text only attack.

The basic tool in studying this question is the concept of entropy introduced by Shannon (1948).

§3.4. Entropy

Entropy can be thought of as a mathematical measurement of "information", that is, how much information we gain, on average, when we learn the value of X.

More specifically, we want to know how much information is revealed about a plaintext by knowing the corresponding cipher text under some cryptosystem.

An alternative view is that the entropy of X measures the amount of uncertainty about X before we learn its value.

§3.4. Entropy

Example: Compare two coin tosses.

- A fair coin toss in which heads and tails are equally likely in each toss.
- A crooked coin toss in which the tail is weighted to appear 90% of the time.

Which has more uncertainty?

Answer: The fair coin toss does have more uncertainty, because there is more randomness in its possibilities (we will prove it rigorously later).

Example: Roll a six-sided die and a ten-sided die. Which experiment has more uncertainty? If you make a guess at the outcome of each roll, you are more likely to be wrong with the ten-sided die than with the six-sided die. Therefore, the ten sided die has more uncertainty.

§3.4. Measure of Uncertainty

We require the measure of uncertainty to satisfy the following:

- We want to make sure that two random variables X and Y that have same probability distribution have the same uncertainty.
- So the uncertainty must be a function only of the probability distributions and not of the names chosen for the outcomes.
- This serves as partial motivation for the following:

§3.4. Measure of Uncertainty

1. To each set of nonnegative numbers $p_1, \cdots, p_n$ with $\sum_i p_i = 1$, the uncertainty is given by a number $H(p_1, \cdots, p_n)$.
2. H should be a continuous function of the probability distribution. So a small change in the probability should not drastically change the uncertainty.
3. In situations where all outcomes are equally likely, the uncertainty increases when there are more possible outcomes. That is,
$$H\left(\frac{1}{n}, \cdots, \frac{1}{n}\right) \leq H\left(\frac{1}{n+1}, \cdots, \frac{1}{n+1}\right)$$
4. If $0 < q < 1$, then $H(p_1, \cdots, qp_j, (1-q)p_j, \cdots, p_n)$ can be written as
$$H(p_1, \cdots, p_j, \cdots, p_n) + p_j H(q, 1-q).$$

§3.4. Measure of Uncertainty

Last condition means that if the jth outcome is broken into two sub-outcomes, with probabilities qp_j and $(1-q)p_j$, then the total uncertainty is increased by the uncertainty caused by the choice between the two sub-outcomes, multiplied by the probability p_j of the original outcome.

Example: Roll a fair six-sided die. Observe the outcomes {even, odd}. The uncertainty in outcome should be $H(\frac{1}{2}, \frac{1}{2})$. But now suppose that if the roll is even, we further observe whether the outcome is low even (that is, $\{2\}$) or high even (that is, $\{4, 6\}$). The outcome Even, has split into two sub-outcomes with relative probabilities 1/3 and 2/3. Overall, there are now three outcomes to our experiment: low even, high even and odd. So we should have:

$$H\left(\frac{1}{6}, \frac{1}{3}, \frac{1}{2}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{1}{3}, \frac{2}{3}\right)$$

§3.4. Entropy Function

Let $S = \{x_1, \cdots, x_n\}$ be a sample space with probabilities $p_1, \cdots, p_n$. Then Shannon showed that if H is a function that satisfies properties (1)-(4) given earlier, then H must be of the form:

$$H(p_1, \cdots, p_n) = -\lambda \sum_k p_k \log_2 p_k$$

where λ is a non-negative constant and where the sum is taken over those k such that $p_k > 0$.

This motivates the following formal definition:

§3.4. Entropy Function

Definition: Let X be a random variable on a finite sample space S with probability distribution p. Then the entropy of the random variable X is defined by the quantity

$$H(X) = -\sum_{x \in S} p[x] \log_2 p[x] = -\sum_{x \in S} p[x] \frac{\ln p[x]}{\ln 2}$$

Remark:

- Notice that if $p[x] = 0$, then $\log_2 p[x]$ is undefined. However, $\lim_{y \to 0^+} y \log_2 y = 0$ and so there is no problem in allowing $p[x] = 0$.
- Notice also that $H(X)$ is a positive valued function, because $\log_2 p[x] \leq 0$.

§3.4. Entropy Function

- The entropy $H(X)$ is a measure of uncertainty in the outcome of X.
- Note that the choice of two as the base of the logarithms is not necessary. Another base choice would only change the value by a constant factor.
- However, it is the typical convention that the logarithm is taken with base 2, in which case the entropy is measured in bits.
- Notice that if S has exactly n outcomes with $p[x] = 1/n$ for each $x \in S$, then $H(X) = \log_2 n$. (For a fair die, $H(X) = \log_2 6$.)
- In particular, consider a fair coin toss. Then $S = \{H, T\}$, exactly two outcomes with each probability $1/2$. So in this case, $H(X) = 1$.
- Notice also that for any sample space S, $H(X) = 0$ if and only if there exists $x_0 \in S$ such that $p[x_0] = 1$ and $p[x] = 0$ for all $x \neq x_0$.

§3.4. Entropy Function

Now we will show that a fair coin toss has more uncertainty than any other coin toss. Suppose that a coin toss has the probabilities

$$p[head] = t \text{ and } p[tail] = 1 - t.$$

$$\text{Then} \quad H(X) = -t \log_2 t - (1-t) \log_2 (1-t) = f(t)$$

By calculus, one can now show that $f(t)$ has maximum value at $t = 1/2$ and $f(1/2) = 1$.

Now we will consider the entropy of the various components of a cryptosystem such as $H(P)$, $H(C)$ and $H(K)$.

§3.4. Entropy Function

For illustration, let us consider the example* from previous section where $\mathcal{P} = \{a, b\}$ with $p[a] = 1/4$ and $p[b] = 3/4$.

Recall: $\mathcal{K} = \{k_1, k_2, k_3\}$ with $p[k_1] = 1/2$, $p[k_2] = 1/4$, and $p[k_3] = 1/4$.

$\mathcal{C} = \{1, 2, 3, 4\}$ with $p[1] = 1/8$, $p[2] = 7/16$, $p[3] = 1/4$ and $p[4] = 3/16$.

Now $H(P) = -(1/4)\log_2(1/4) - (3/4)\log_2(3/4) \approx 0.81$.

$H(K) = -(0.5)\log_2(0.5) - (0.25)\log_2(0.25) - (0.25)\log_2(0.25) = 1.5$

and similarly
$H(C) = -(1/8)\log_2(1/8) - (7/16)\log_2(7/16) - (1/4)\log_2(1/4) - (3/16)\log_2(3/16) \approx 1.85$.

§3.4. Entropy Function

Entropy and yes-no questions:

There is a relationship between entropy and the number of yes-no questions needed to determine accurately the outcome of a random experiment.

Suppose we randomly draw a number from the set $\{0, 1, 2, 3\}$ at random.

Then the entropy is $H(X) = 2$.

Interpretation: How many yes-no questions are necessary on average to determine the outcome? We can do so with 2 questions:

First question : is the number even?
Second question: is the number greater than 1?

§3.4. Entropy Function

We flip two fair coins. Let X be the number of heads, where possible outcomes are $\{0, 1, 2\}$ with probabilities $1/4,\ 1/2,\ 1/4$ respectively.

Now $H(X) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4}\right) = \frac{3}{2}$.

Let's try to interpret this in terms of yes-no questions. Note that in the worst case, we'll need two questions (a single yes-no question can only distinguish two outcomes). The goal is to ask the first question such that it is as likely as possible not to need a second question.

First question: is there exactly one head? If so, we're done and we don't need another question. This happens half the time.

Second question: is there a head? This determines whether there are two or no heads.
The average number of questions asked is 1.5.

§3.4. Entropy Function

What if we flip three coins and record the number of heads?

The set of outcomes is the same as in the first example, but note that the entropy is:

$$-\left(\frac{1}{8}\log_2\frac{1}{8}+\frac{1}{8}\log_2\frac{1}{8}+\frac{3}{8}\log_2\frac{3}{8}+\frac{3}{8}\log_2\frac{3}{8}\right)\approx 1.81$$

Indicating that maybe we can get away with fewer than 2 questions on average to determine an outcome:

1. Is there exactly one head?
2. Are there exactly two?
3. Are there exactly three?

Average number of questions: 1.875. Not exactly the entropy, but close.

§3.4.0. Coding Theory

In this section, we will briefly discuss the connection between entropy and Huffman encoding.

A code is simply a system to convert information from one system of representation into some other system of representation.

There are various reasons we may want to do so. We have already seen an example of a code:

- Representing letters as integers in $\mathbb{Z}_{26}$ in order to facilitate definitions and computations in cryptosystems.
- ASCII (American Standard Code for Information Interchange) is a character encoding for electronic communication. It uses seven-bit binary code for each character. Though, it became common to use 8-bit byte in extended ASCII later.

§3.4.0. Coding Theory

More generally, we may want to look at other binary codes, with an eye to efficiency of storage or transmission.

Definition: Let X be a random variable on a finite set S with a probability distribution p. A binary encoding is a function $f : S \to \{0,1\}^*$ where $\{0,1\}^*$ is the set of all finite strings of 0's and 1's.

Example: Suppose we have $S = \{a, b, c, d\}$. We wish to represent each with a binary codeword. We'll need at least four codewords, so let's use all four binary strings of length 2. So we have:

letter	a	b	c	d
codeword	00	11	10	01

§3.4.0. Coding Theory

What if were concerned about efficiency? Then we want to use shorter codewords whenever possible. In the previous example, if we have any string of length n with the alphabet a, b, c, d, we'll use a binary string of length exactly $2n$ to encode it.

What if we try the following in order to get away with using fewer than $2n$ symbols?

letter	a	b	c	d
codeword	0	1	00	01

This is a valid code, but it presents a problem while decoding. What if we see the string 001?

It can be decoded in three possible ways: aab, ad, or cb. Note that this is not an issue for the previous coding; since all codewords have length 2. So we can simply separate them into blocks of two and then decode.

§3.4.0. Coding Theory

Now, given that we are going to encode strings using a mapping f, it is important that we are able to decode in an unambiguous fashion. Thus it should be the case that the encoding function f is injective. In that case, a code string is uniquely decipherable.

A sufficient, but not necessary, condition to guarantee the unique decipherable is the following:

Definition: A code is prefix-free if no codeword is an initial segment of any other; i.e., there are no triples $(\mathbf{c}, \mathbf{d}, \mathbf{y})$ such that $\mathbf{cy} = \mathbf{d}$ where $\mathbf{c}, \mathbf{d}$ are codewords and $\mathbf{y}$ is a binary string.

Exercise: If $a = 0$, $b = 10$, $c = 110$, check that they are prefix-free and then decode 1000110.

§3.4.0. Coding Theory

Example: Suppose again that $S = \{a, b, c, d\}$ and consider the following three encodings:

	a	b	c	d
f	1	10	100	1000
g	0	10	110	111
h	0	01	10	11

1. Notice that the first encoding f is injective, but the code is not prefix-free [because $f(a)0 = f(b)$].
2. Second encoding g is injective and the code is prefix-free.
3. Third encoding h is not injective and the code is not prefix-free. Notice that $h(ac) = h(ba) = 010$ and $h(b) = h(a)1$.

§3.4.0. Coding Theory

One prefers the encoding g from the point of view of ease of decoding, because the decoding can be done sequentially from the beginning to end if g is used and so no memory is required.

Among various encodings, we are interested in more efficient encoding and so we need to measure the efficiency of encoding.

A string $x_1 x_2 \cdots x_n$ is said to be produced by a memoryless source if the probability $p[x_i]$ does not depend on the probability of the previous characters and this means that

$$p[x_1 \cdots x_n] = p[x_1] \times \cdots \times p[x_n]$$

Example: Consider n tosses of a fair coin.

§3.4.0. Efficiency of the Coding

Define $|\mathbf{x}|$ to be the length of the binary string $\mathbf{x}$; e.g., $|1011| = 4$.

Suppose that a string of symbols is produced by a memoryless source, which is an ordered pair (S, p), where $S = \{s_1, \ldots, s_q\}$ is a source alphabet, and p is probability distribution on S.

To measure the efficiency of an encoding scheme f, we use the weighted average codeword length:

$$\ell(f) = \sum_{i=1}^{q} |f(s_i)| p[s_i].$$

Now, our fundamental problem is to find an injective coding f with minimum $\ell(f)$.

§3.4.0. Efficiency of the Coding

Example: Consider $S = \{a, b, c, d\}$ where $p(a) = 0.5$, $p(b) = 0.3$, $p(c) = 0.1$ and $p(d) = 0.1$ with encoding:

letter	a	b	c	d
codeword	00	11	10	01

Notice that the weighted average codeword length is

$$\ell(f) = (0.5)(2) + (0.3)(2) + (0.1)(2) + (0.1)(2) = 2.0$$

Can we improve on the average word-length of 2?

§3.4.0. Efficiency of the Coding

Let's try to do so by using a prefix-free code.

letter	a	b	c	d
codeword	1	01	001	000

Now, the average codeword length is

$$\ell(f) = (0.5)(1) + (0.3)(2) + (0.1)(3) + (0.1)(3) = 1.7$$

Compare with entropy of the probability distribution: $H(X) = 1.685$

§3.4.0. Huffman encoding

There is a well-known algorithm, known as Huffman encoding which yields a pre-fix free encoding f with minimum $\ell(f)$.

Moreover, for Huffman encoding, we get:

$$H(X) \leq \ell(f) \leq H(X) + 1$$

Remark: Huffman Encoding is a technique for compressing data. It (provably) achieves maximum compression among single-character encodings, although there are block-level compression algorithms that achieve a higher compression rate

§3.4.0. Huffman encoding

Example: Suppose we want to compress a text in which these (and only these) letters appeared with the given frequencies:

letter	Z	K	O	R	S	A
frequency	5	7	10	15	20	45

In order to determine the Huffman encoding for this text, we first build a **Huffman tree** as follows:

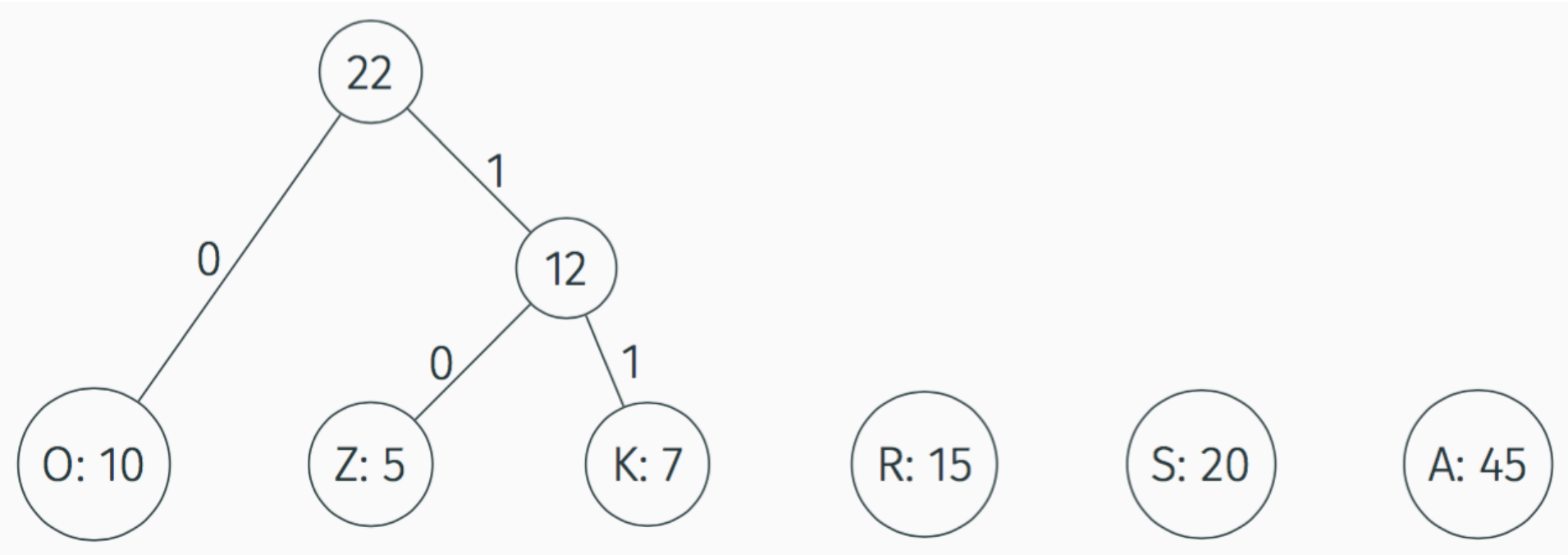
Step 1: We build the tree bottom-up by first constructing a leaf node for every letter in the text, and assigning a weight to that node corresponding to its frequency.

§3.4.0. Huffman encoding

Step 2: We look for the two parentless nodes with the smallest weights and join them together, assigning their combined weight to their parent. We will assign a 0 to the edge leading to the smaller-weighted child and a 1 to the other edge. Since Z and K have the smallest weights, we will join them first. (Ties may be broken arbitrary.)

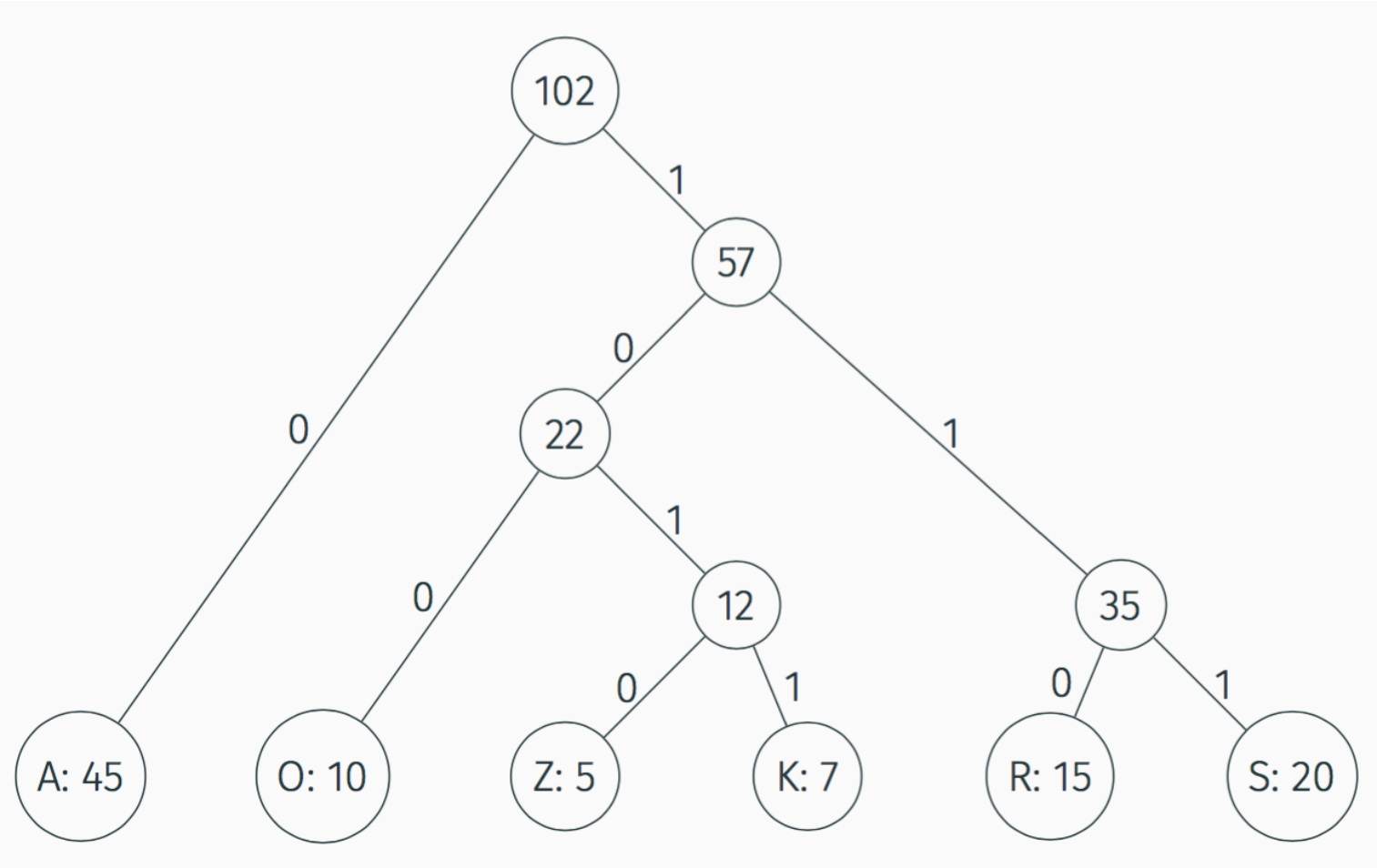
§3.4.0. Huffman encoding

Step 3: Now the two smallest weights are 10 and 12, so we join them together. (We have rearranged some nodes now).



§3.4.0. Huffman encoding

The completed tree:



§3.4.0. Huffman encoding

To read the encoding, we simply traverse the tree to each leaf:

letter	Z	K	O	R	S	A
probability	5/102	7/102	10/102	15/102	20/102	45/102
codeword	1010	1011	100	110	111	0

Note: Instead of using a tree, one may use a table as indicated in some text books to write Huffman encoding.

§3.4.0. Huffman encoding

Notice the prefix-free property. Now the average codeword length is:

$$\ell = \frac{45}{102} \times 1 + \frac{10}{102} \times 3 + \frac{5}{102} \times 4 + \frac{7}{102} \times 4 + \frac{15}{102} \times 3 + \frac{20}{102} \times 3 \approx 2.24$$

The entropy of this text is :

$$H(X) = -\left[\frac{45}{102} \log_2 \frac{45}{102} + \frac{10}{102} \log_2 \frac{10}{102} + \cdots + \frac{20}{102} \log_2 \frac{20}{102}\right] \approx 2.19$$

Notice the entropy is very close to the length of the Huffman encoding. In fact, it satisfies:

$$H(X) \leq \ell \leq H(X) + 1.$$

§3.4.1. Properties of Entropy - Concave Function

In this section, we will discuss some fundamental properties of entropy before applying them to a cryptosystem.

A real-valued function f on an interval I is said to be concave if, for any x and y in the interval I and for any $\alpha \in [0,1]$ the inequality $(1-\alpha)f(x) + \alpha f(y) \leq f((1-\alpha)x + \alpha y)$ holds.

.

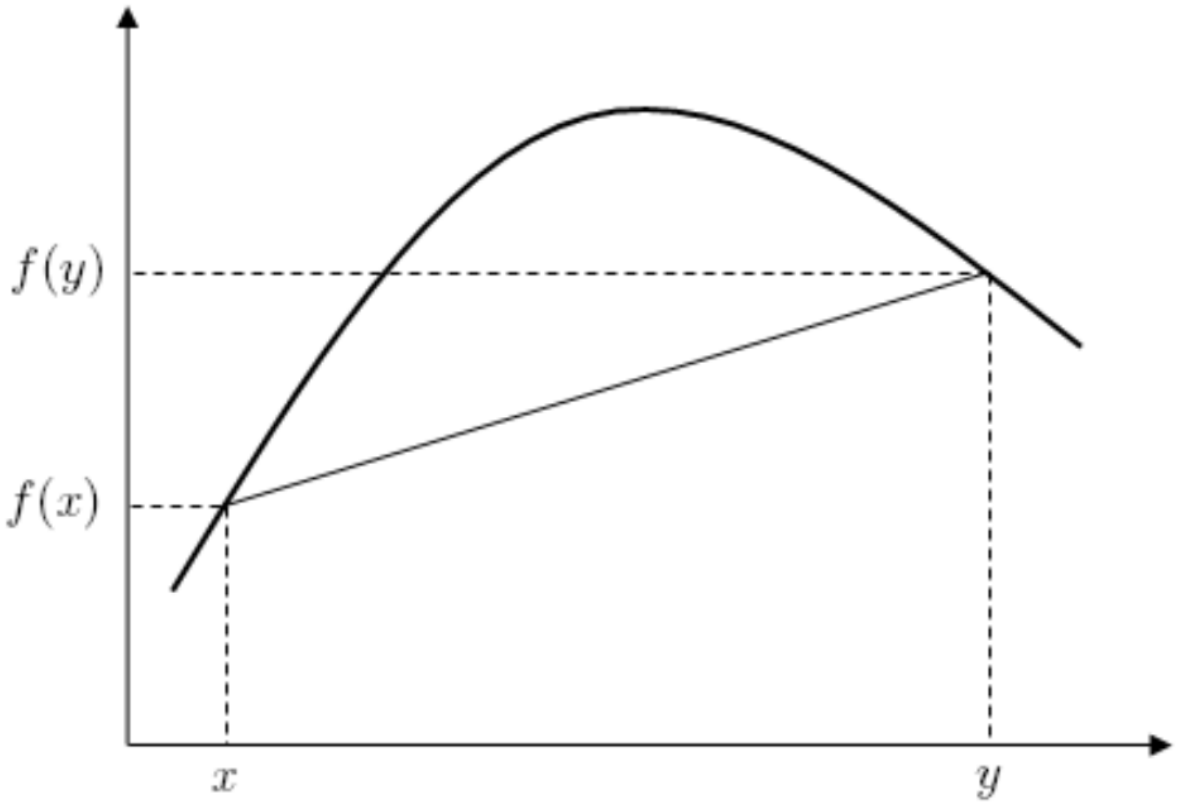
Remark: For continuous functions, we may restrict to $\alpha = 1/2$, that is,

$$\frac{f(x)+f(y)}{2} \leq f\left(\frac{x+y}{2}\right)$$

.

If the strict inequality holds for all $x \neq y$ in the above definition, then f is called a strictly concave function.

§3.4.1. Concave Function



For a function $f : \mathbb{R} \to \mathbb{R}$, this definition merely states that for every z between x and y, the point $(z, f(z))$ on the graph of f is above the straight line joining the points $(x, f(x))$ and $(y, f(y))$.

§3.4.1. Jensen's Inequality

There is a more general inequality, which we will state without the proof:

Jensen's Inequality: Let $\alpha_1, \alpha_2, \cdots, \alpha_n$ be an arbitrary list of positive numbers such that $\sum_{i=1}^{n} \alpha_i = 1$. Let $x_1, \ldots, x_n$ be an arbitrary list of numbers in the interval I. If f is a continuous strictly concave function on the interval I, then

$$\sum_{i=1}^{n} \alpha_i f(x_i) \leq f\left(\sum_{i=1}^{n} \alpha_i x_i\right)$$

Further, equality occurs if and only if $x_1 = \cdots = x_n$.

§3.4.1. Properties of Entropy

- We are interested in the above inequalities, because the function $\log_2 x$ is strictly concave on the interval $(0, \infty)$.
- This follows easily from elementary calculus since the second derivative of the logarithm function is negative on the interval $(0, \infty)$.
- We make use of this fact to derive several properties of entropy.

Property 1: Suppose X is a random variable having a probability distribution which takes on the values $p_1, p_2, \ldots, p_n$, where each $p_i > 0$. Then

$$H(X) \leq \log_2 n,$$

with equality if and only if $p_i = 1/n$ for each $i \in \{1, 2, \ldots, n\}$.

§3.4.1. Properties of Entropy

Proof:

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} p_i \log_2 p_i \\ &= \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \\ &\leq log_2 \sum_{i=1}^{n} \left(p_i \times \frac{1}{p_i} \right) \quad \text{(Jensen's inequality)} \\ &= \log_2 n \end{aligned}$$

Further more, equality occurs if and only if $p_i = 1/n$ for each $i \in \{1, 2, \ldots, n\}$.

§3.4.1. Properties of Entropy

Definition: Let X be a random variable on S and Y be a random variable on T, then the joint entropy $H(X, Y)$ is defined as

$$H(X, Y) = -\sum_{x \in S} \sum_{y \in T} p[x, y] \log_2 p[x, y]$$

Property 2: $H(X, Y) \leq H(X) + H(Y)$. The equality holds if and only if X and Y are independent random variables.

Proof: Suppose that $S = \{x_1, \cdots, x_m\}$ with $p[x_i] = p_i$.

Suppose also that $T = \{y_1, \cdots, y_n\}$ with $p[y_j] = q_j$.

Let $p[x_i, y_j] = r_{ij}$. Then observe that

$$p_i = \sum_{j=1}^{n} r_{ij} \quad \text{and} \quad q_j = \sum_{i=1}^{m} r_{ij}$$

§3.4.1. Properties of Entropy

Then

$$H(X,Y) = -\sum_{x\in S}\sum_{y\in T} p[x,y]\log_2 p[x,y]$$

$$= -\sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij}\log_2 r_{ij} \text{ and}$$

$$H(X)+H(Y) = -\sum_{i=1}^{m} p_i\log_2 p_i - \sum_{j=1}^{n} q_j\log_2 q_j$$

$$= -\sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij}\log_2 p_i - \sum_{j=1}^{n}\sum_{i=1}^{m} r_{ij}\log_2 q_j$$

$$= -\sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij}\log_2 p_i q_j$$

§3.4.1. Properties of Entropy

By combining, we get:

$$
\begin{aligned}
H(X+Y) - H(X) - H(Y) &= -\sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij} \log_2 r_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij} \log_2 p_i q_j \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij} \log_2 \frac{1}{r_{ij}} + \sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij} \log_2 p_i q_j \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij} \log_2 \frac{p_i q_j}{r_{ij}} \\
&\leq \log_2 \sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij} \frac{p_i q_j}{r_{ij}} = \log_2 \sum_{i=1}^{m}\sum_{j=1}^{n} p_i q_j \\
&= log_2 1 = 0
\end{aligned}
$$

§3.4.1. Properties of Entropy

We will leave the proof of equality part as exercise. We will also list two more properties without proof.

Definition: Suppose X and Y are two random variables on S and T respectively. Then the conditional entropy $H(X|Y)$ is defined by using the conditional probabilities as:

$$H(X|Y) = -\sum_{y}\sum_{x} p[y]p[x|y]\log_2 p[x|y]$$

This quantity measures the average amount of information about X that is not revealed by Y.

§3.4.1. Properties of Entropy

Remark: For any fixed value y of T, we get the associated (conditional) random variable $X|y$ with

$$H(X|y) = -\sum_x p[x|y] \log_2 p[x|y]$$

$$\text{So} \quad H(X|Y) = \sum_y p[y] H(X|y)$$

Property 3: (Chain rule) $H(X, Y) = H(Y) + H(X|Y)$. That is, the uncertainty of a joint experiment (X,Y) is equal to the uncertainty of the experiment Y + the uncertainty of the experiment X given that Y has happened.

Property 4: $H(X|Y) \leq H(X)$. The equality holds if and only if X and Y are independent (follows from property 2 and 3).