

Information Theory and Signal Processing

Muchang Bahng

Spring 2024

Contents

1	Introduction	2
1.1	Channels	2
1.2	Coding Schemes	3
2	Entropy	5
2.1	Discrete Random Variables	6
2.2	Joint and Conditional Entropy	7
2.3	Source Coding Theorem	9
3	Symbol Codes	11
3.1	Huffman Coding	15
4	Differential Entropy	17
4.1	Differential Entropy	17
4.2	Kullback Leibler Divergence	17
4.3	Entropy of Probability Measures	18
	References	19

1 Introduction

1.1 Channels

In a *communication system*, we have a *transmitter* and *receiver*, with *signals* going through a *channel*. Let's briefly define what these terms are, which are pretty much taken verbatim from Shannon's famous paper [1].

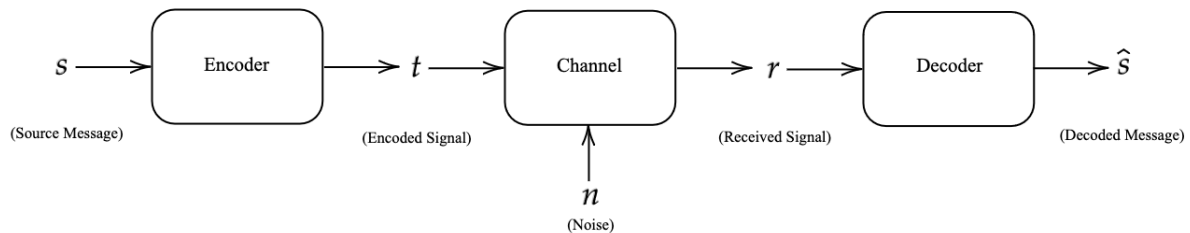


Figure 1: A channel diagram.

Definition 1.1 (Information Source)

An **information source** produces a message or sequence of messages to be communicated to the receiving terminal.

Definition 1.2 (Encoder)

A **transmitter**, or **encoder**, operator on the message in some way to produce a signal suitable for transmission over the channel.

Definition 1.3 (Channel)

The **channel** is the medium used to transmit the signal from the encoder to the decoder. Some examples of channels are:

1. A copper wire is a channel connecting one phone to another phone.
2. Air is a channel connecting your voice to another's ear.
3. Vacuum is a channel connecting an antenna on earth to the Mars rover.

Definition 1.4 (Decoder)

The **decoder**, or the **receiver** performs the inverse operation of that done by the transmitter, reconstructing the message from the signal.

Definition 1.5 (Destination)

The **destination** is the person (or thing) for whom the message is intended.

All the channels have the property that the received signal is maybe similar, but not identical, to the transmitted signal. This noise is not preferable, and we would ideally like to have perfect communications systems. To reduce this noise, we can improve physical systems (e.g. better insulation in copper wires) or we can improve our systems, such as our encoding/decoding schemes.

Example 1.1 (Binary Symmetric Channel)

Given a 1-bit input x , there is a certain probability p such that the input is flipped.^a This can be sometimes seen in practical applications, e.g. the salt-and-pepper noise in images.

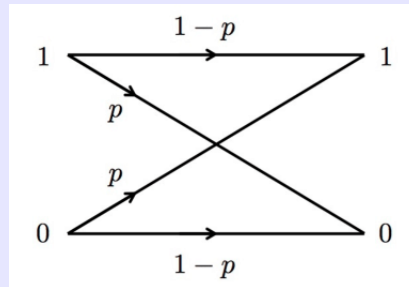


Figure 2: A simple example of noise.

1.2 Coding Schemes

To reduce the probability of $\hat{s} \neq s$, we can devise many schemes of the encoder and decoder. Depending on how much additional information we add, our channel throughput, or **rate**, becomes lower.

Definition 1.6 (Parity Encoding)

Given a string of bits, we can simply add a parity bit.

$$\text{encoder}(x_1, x_2, \dots, x_n) = x_1, \dots, x_n, (x_1 \oplus \dots \oplus x_n) \quad (1)$$

This has a rate of $n/(n+1)$.

Definition 1.7 (Repetition)

The encoder can just repeat each bit k times, which we will denote as R_k .

$$\text{encoder}(x_1, \dots, x_n) = x_1, x_1, x_1, x_2, \dots, x_n \quad (2)$$

For example, with $k = 3$ we have

1	$s = 01101$
2	$t = 000\ 111\ 111\ 000\ 111$
3	$n = 000\ 100\ 000\ 101\ 000$
4	$r = 000\ 011\ 111\ 101\ 111$

The decoder then can take the best of 3 to get 01111. Note that the second bit had a flip but was fixed, but the second to last bit was an error. We can then compute the probability of these errors with basic computations.^a This has a rate of $1/k$.

We can already predict that these encoding schemes can get quite sophisticated. Here's another one.

^aIn 2014 disk drives, the standard was that p should not be greater than 10^{-18} .

^aIt turns out that we need $k = 61$ to get a probability of error below 10^{-15} .

Definition 1.8 (7, 4 Hamming Code)

Given an input string of bits \mathbf{s} , we divide it up into sequences of 4.

$$\mathbf{s}_{i:i+4} = (s_i, s_{i+1}, s_{i+2}, s_{i+3}) \tag{3}$$

Then we can place them in a Venn diagram as shown below and fill out the rest of the three empty spots such that the parity within each circle is 0.

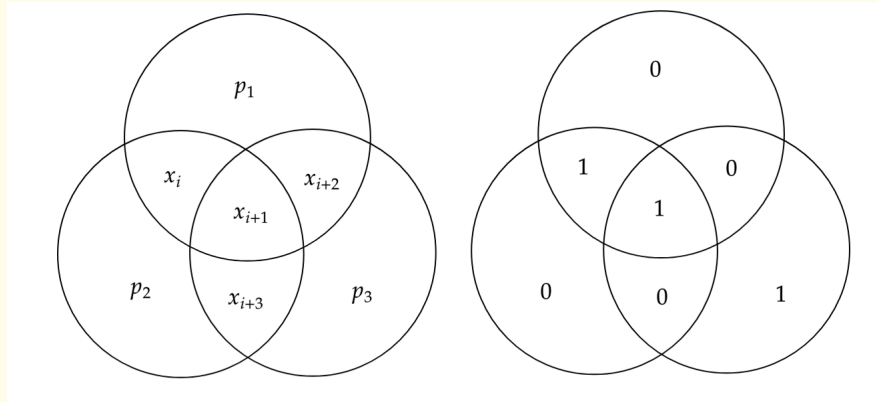


Figure 3: (7, 4) hamming code visual with example on the right.

This gives us the encoder.

$$\text{encoder}(x_1, x_2, x_3, x_4) = (x_1, x_2, x_3, x_4, p_1, p_2, p_3) \tag{4}$$

As for the decoder, we can fill up the Venn diagram with the received bits r_1, \dots, r_7 and then look at the minimum number of bits needed to flip to achieve the same rules we had to fill the inputs out in the Venn diagram. Given any combination of circles that have parity 1, we can then flip exactly one of the r_i to satisfy the rules again (i.e. find the bit that is outside all the valid circles and inside all the invalid circles). This has a rate of $4/7$.

Theorem 1.1 (Conditions for Detection and Correction)

The (7,4) Hamming code can correct an input if up to 1 bit is flipped in each sequence of 4 bits, but if there are more than 1 bit flip, the decoded sequence will be incorrect.

More specifically, the probability of a block error is $21p^2$ on the most significant order and a bit error is $9p^2$. If we look at these different algorithms and plot their rate vs probability of error, we can see some sort of dependency.

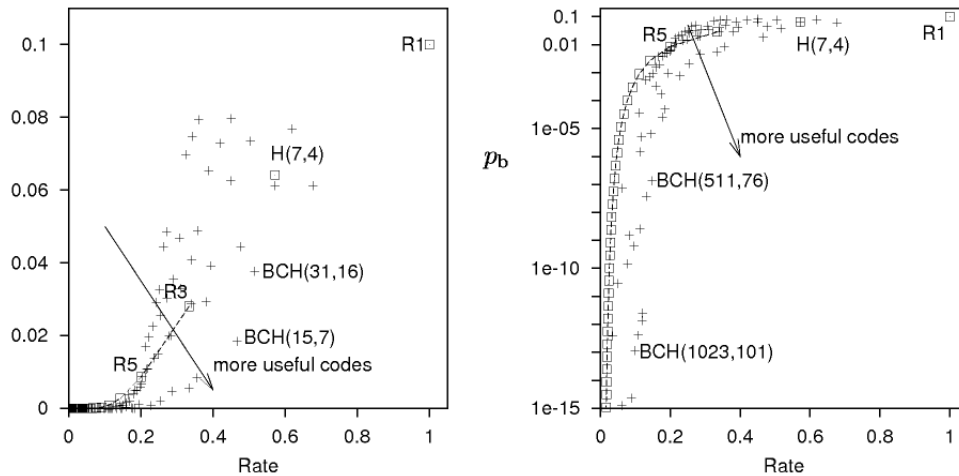


Figure 4: The rate of an encoding/decoding scheme vs probability of bit error.

It was reasonable to assume that we can make schemes that “hit” the upper-left portion of the left graph, i.e. we can make schemes that have a low rate (lots of repetition and such) yet still have a low probability of error. The question was how well we can reach the bottom-right corner containing the more useful codes. The general consensus assumed that as the probability of error goes to 0, the rate must also tend towards 0, and so we had a boundary that intersected through the origin that separated achievable and non-achievable schemes. However, Claude Shannon remarkably proved that this was not the case, through his *noisy-channel coding theorem*. Rather, we can achieve arbitrarily low probabilities without having to go below some non-zero rate, i.e. this boundary crosses the x-axis at some positive number C .

Definition 1.9 (Capacity)

C is the **capacity** of the channel.

Theorem 1.2 (Capacity of Binary Switch Channel)

The capacity of the BSC with flip probability f is

$$C_{BSC,f} = 1 - H(X), \quad X \sim \text{Bernoulli}(f) \quad (5)$$

This means that rather than needing 61 times our input to get past 10^{-15} error in the BSC with $f = 0.1$ (which we derive through repetition), we only need 2 disk drives, which is amazing.

2 Entropy

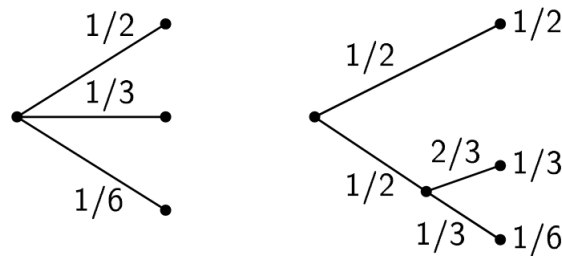
We have hinted at the fact through Shannon’s noisy encoding theorem that there is an optimal way to add redundancies to compress some input. Given a string of random variables X_1, \dots, X_n generated iid from a Bernoulli(p) distribution, we want to start to formalize this by introducing a metric to measure the information content of this stochastic process. We motivate the necessity of such a measure using general probability measures and then focus on the discrete case.

2.1 Discrete Random Variables

In Shannon's famous paper [1], he talks first on discrete channels, focusing on examples of transmitting languages through n-gram models as "higher order approximations" of language.¹ This is an ergodic Markov chain with some stationary distribution.

He then asks whether we can have some sort of measure on how much information is produced by the process, or at what rate the information is produced? Borrowing his terminology, if we have some measure $H(p_1, \dots, p_n)$, he states that it is reasonable to require the following properties, which are slightly different than ours. H should measure the uncertainty of the outcome.

1. H should be continuous in p_i .
2. If all p_i are equal, then H should be a monotonic increasing function of n . With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice is broken down into two successive choices, the original H should be a weighted sum of the individual values of H . For example, the uncertainty of both distributions should be the same.



That is,

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

Which then leads to the definition of entropy below.

Lemma 2.1 (Entropy of Discrete RV)

For a discrete random variable, the entropy reduces to the expectation of the information content.

$$H[X] := \mathbb{E}_X[-\ln p(X)] = - \sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \ln \mathbb{P}(X = x) \quad (7)$$

where we use $p(x)$ as the PMF.

Proof.

TBD. Since we are working with the power set, ...

Note the following properties.

Theorem 2.1 (Bounds on Entropy)

H is bounded by 0 and 1, attaining its minimum if and only if all the p_i but one are 0. It attains its maximum if p is uniform.

¹In fact, this is where n-gram models were first referenced.

Theorem 2.2 (Entropy of Independent Events)

If E_1 and E_2 are independent events, then $H(E_1 \cap E_2) = H(E_1) + H(E_2)$. The information from two independent events is the sum of their informations since information gain from one does not increase information from another independent variable.

Example 2.1 (Bits)

Given $X \sim \text{Bernoulli}(p)$, if we observe a value of 1, then we have received $\log_2\left(\frac{1}{p}\right)$ bits of information.

Now Shannon's claim is that this information content is the optimal encoding length that we should aim for. For example, given $p = 0.9$, then a 0 has 3.32 bits of information content and a 1 has 0.15 bits. This means that 0's, which occur infrequently, should be encoded with longer strings and 1 with shorter strings.

Exercise 2.1 (Weighing Problem)

You are given 12 balls, all equal in weight except for one that is either heavier or lighter. Design a strategy to determine which is the odd ball *and* whether it is heavier or lighter in as few uses of the balance as possible.

Proof.

We can tackle this by looking at the first action. We can choose to weigh n vs n balls for $n = 1, \dots, 6$. Shannon would advise you to choose such that we maximize our entropy, or expected information gain. Let's go through them one at a time. Our three outcomes for all scenarios are A (left is lighter), B (both equal), and C (right is lighter).

1. 6 v 6. The probability distribution is $(A, B, C) = (1/2, 0, 1/2)$ and so the entropy is $H = 1$ bit.
2. 5 v 5. The distribution is $(5/12, 1/6, 5/12)$ giving us $H = 1.48$ bits.
3. 4 v 4. The distribution is $(1/3, 1/3, 1/3)$ giving us $H = 1.58$ bits.
4. We go on.

We already know that entropy must be maximized in the uniform distribution, so it is best to choose 4 v 4. This is indeed the correct first step. As for the next step. Let's think about what to do in each of the events.

1. A . This means that there are four H 's (possibly heavies), four L 's, and four G 's (possibly good). We are left with 8 balls and we want to maximize the entropy. It turns out that if we measures HHL vs HHL, then the events turn out to have distribution $(3/8, 2/8/3/8)$, which is quite uniform.
2. B . We have eight G 's, so to maximize the entropy we can weigh one against another, which has a distribution of $(1/4, 1/2, 1/4)$.
3. C . By symmetry, we use the same method as A .

We just continue this process which is a stepwise optimization of entropy. It turns out that we just need 3 steps.

2.2 Joint and Conditional Entropy**Definition 2.1 (Joint, Conditional Entropy)**

We can define the joint entropy and conditional entropy between two discrete random variables X, Y

as

$$H(X, Y) = \mathbb{E}_{X \times Y}[-\log p(x, y)] = \sum_{x, y \in \mathcal{X}, \mathcal{Y}} p(x, y) \cdot -\log p(x, y)$$

$$H(X | Y) = \mathbb{E}_{X \times Y}[-\log p(x | y)] = \sum_{x, y \in \mathcal{X}, \mathcal{Y}} p(x, y) \cdot -\log p(x | y)$$

Theorem 2.3 (Joint Entropy)

The uncertainty of a joint event is less than or equal to the sum of the individual uncertainties, with equality achieved only if the events are independent.

$$H(X, Y) \leq H(X) + H(Y) \quad (8)$$

Another property is that any change towards “equalization” of the probabilities p_i increases H . Since we don’t have a method of measuring how close to the uniform distribution, we will return back to this after defining the KL divergence.

Theorem 2.4 (Conditional Entropy)

The joint entropy is the entropy of X plus the conditional entropy of Y given X .

$$H(X, Y) = H(X) + H(Y | X) = H(Y) + H(X | Y) \quad (9)$$

Theorem 2.5 (Conditioning Never Decreases Uncertainty)

Since

$$H(X) + H(Y) \geq H(X, Y) = H(X) + H(Y | X) \quad (10)$$

we have $H(Y) \geq H(Y | X)$. That is, the uncertainty of Y is never increased by the knowledge of X .

In fact, the amount of uncertainty that decreases when conditioning has a well known name.

Definition 2.2 (Mutual Information)

The **mutual information** between random variables X, Y is the decrease in entropy when we condition X by Y .

$$I(X; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X) \quad (11)$$

This can be conditioned on another random variable Z .

$$I(X; Y | Z) = H(X | Z) - H(X | Y, Z) = H(Y | Z) - H(Y | X, Z) \quad (12)$$

Therefore, we can interpret $I(X; Y)$ as the partial information you learn about X from knowing Y . The entropy also demonstrates the average length (if base is 2) number of bits required to transmit the state of a random variable.

Theorem 2.6 ()

From simple substitution, we can derive

$$H(X, Y) = H(X | Y) + H(Y | X) + I(X; Y) \quad (13)$$

Unlike entropy, conditioning the mutual information on a third variable can have either a hiding or revealing

effect. It can *both* be the case that²

1. $I(X; Y | Z) > I(X; Y)$ happens when X and Y both are causes of some common effect Z , i.e. if you know Z has happened, then X and Y are more dependent than before. For example, a car's engine fails to start (event Z), it may be because of either blocked fuel pump (X) or that the battery is dead (Y). Normally, X and Y are independent, so $I(X; Y) = 0$, but if the engine doesn't start, they suddenly become very dependent, since now you can look at the battery (Y) and from that conclude the status of the pump (X) with much more confidence, making $I(X; Y | Z) > 0$.³
2. $I(X; Y) > I(X; Y | Z)$ happens when Z is the cause of both X and Y . For example, if clouds (Z) always cause rain (X) and blocks the sun, (Y), then we know that $I(X; Y | Z) = 0$ since Z already tells us everything about X and Y , so Y does not tell us anything more about X . But if we only observe whether the sun is blocked, this only tells us partially about whether it is rainy (may or may not be due to clouds or some other factor), making $I(X; Y) > 0$ due to some correlation revealed.

2.3 Source Coding Theorem

Let's do a few more puzzle examples to give some motivation for the source coding theorem.

Exercise 2.2 (63 Puzzle)

I am thinking of an integer $0 \leq x \leq 63$. You must identify this x by asking if it is at least a number y . How do you get it in the minimum number of questions?

Proof.

The answer is clearly binary search, which gets it done in $\log_2 64 = 6$ questions. More specifically, we can come up with the following predetermined questions which each gives the i th binary digit of the solution.

1. $C_1 : x \bmod 64 \geq 32?$
2. $C_2 : x \bmod 32 \geq 16?$
3. $C_3 : x \bmod 16 \geq 8?$
4. $C_4 : x \bmod 8 \geq 4?$
5. $C_5 : x \bmod 4 \geq 2?$
6. $C_6 : x \bmod 2 \geq 1?$

Note that if we are assuming a uniform distribution, this is the strategy that maximize stepwise entropy, since the outcome of each question has an equal probability of being 0 or 1, leading to 1 bit of information (and not less since all these random variables C_i are independent) This is indeed exactly how much information about the binary expansion of x we get. After 6 questions, the total information content was 6 bits.

From this, we can claim something.

Lemma 2.2 ()

All outcomes of a random variable X from a set of size S can be communicated in $\lceil \log_2 |S| \rceil$ bits.

Let's think more about what information means with another game.

²Read the Wikipedia article on *Interaction Information*.

³Another example is given independent Bernoullis X, Y , with $Z = X \oplus Y \pmod{2}$, we can clearly see that $I(X; Y) = H(X) - H(X | Y) = 0$ since they are independent and so Y does not give any information about X . However, if we condition on Z further, this gives complete information on X . This is quite intuitive since you would know more about X from knowing both Y and Z rather than just knowing Y .

Example 2.2 (Submarines)

You are playing battleship on a 8×8 grid, but there is one 1×1 submarine that you are trying to hit. Say you choose some square and hit it.

1. You don't hit it, which happens with probability $63/64$. You then get $\log_2(64/63) \approx 0.0227$ bits of information.
2. You fire again and don't hit, which happens with probability $62/63$. You get $\log_2(63/62) \approx 0.0230$ bits of information. The total information gained is 0.04560 .
3. You keep firing off at squares. You obviously don't want to fire at an already hit square since the probability that you don't hit is 1, so no information is gained.
4. If you keep firing and don't hit after 32 tries, then you have gained a total information of

$$\log_2 \frac{64}{63} + \dots + \log_2 \frac{33}{32} = 1 \quad (14)$$

bit. This is similar to getting 1 bit from the first step of binary search, which is consistent with our intuition.

5. Let's keep firing and say on the 35th hit, we actually hit the submarine. Then our total information content is

$$\log_2 \frac{64}{63} + \dots + \log_2 \frac{33}{32} + \dots + \log_2 \frac{21}{20} + \log_2 \frac{20}{1} = 6 \quad (15)$$

We have then acquired 6 bits of information (around 4.3 bits for the hit) and gotten all possible information we can get from the grid.

Lemma 2.3 (Approximation of Binomial Distribution)

We can approximate

$$\binom{n}{k} = \log \frac{n!}{k!(n-k)!} \approx nH_2\left(\frac{p}{n}\right) \quad (16)$$

Exercise 2.3 (Bent Coin Lottery)

A coin with $p = 0.1$ is tossed 1000 times to get a random vector $\mathbf{x} \in \{0, 1\}^{1000}$. You can buy any of the 2^N possible tickets for \$1 each, before the coin tossing. If you own the correct ticket, you get a lot of money.

1. If you are forced to buy one ticket, which ticket would you buy?
2. To have a 99% chance of winning at a lowest possible cost, which tickets should you buy?
3. And how many tickets is that? Express it in the form of 2^n .

Proof.

Let's go through them.

1. Even though the expected number of 1's is 2, the all 0 ticket would be the most likely outcome.
2. From the previous problem, we can intuit that we should buy all the tickets with zero 1s, then one 1s, then two 1s, and so on until some threshold r where the probability is 99%. By CLT, we can approximate this to be normal with mean 100 and standard deviation $\sqrt{1000 \cdot 0.1 \cdot 0.9} \approx 2$, and therefore a z-score of about 2.3 will give us 123 tosses with a 99% chance of winning.
3. To find out how many tickets this is, we compute

$$1 + \binom{1000}{1} + \binom{1000}{2} + \dots + \binom{1000}{123} \quad (17)$$

the rightmost is the dominant term, and we use the approximation to get it approximately 2^{530}

tickets.

Therefore, we have essentially “compressed” the set of all 2^{1000} tickets up to 99% probability of hitting, into a set of approximately 2^{530} tickets, called the **typical set**, which can be encoded in 530 bits. Therefore,

1. the compressor takes the typical set of tickets and creates a bijection into a second set of 530 bit long strings.
2. The decompressor just undoes this bijection from the typical set back into the 1000 bit long strings.

In a general case lottery with n -length strings and a probability of p , we can compute that you will need approximately

$$\left(fn + 2.3\sqrt{Nf(1-f)} \right) \approx 2^{NH_2(p)+\epsilon} \quad (18)$$

where ϵ is a small term that scales with \sqrt{n} . We see a certain pattern that coincides with the source coding theorem on how well we can compress a certain set that scales with some probability. Note that this depends on precisely defining the typical set.

Definition 2.3 (Typical Set)

When a source X produces N independent outcomes

$$\mathbf{x} = x_1, x_2, \dots, x_n \quad (19)$$

This string is very likely to be in a **typical set** consisting of $\sim 2^{nH(X)}$ outcomes all of which have a probability of $\sim 2^{-nH(X)}$.

Theorem 2.7 (Source Coding Theorem)

N outcomes from a source X can be compressed into roughly $NH(X)$ bits.

3 Symbol Codes

Definition 3.1 (Symbol Code)

Let X be a discrete random variable over some finite **alphabet** \mathcal{S} with probability measure \mathbb{P} . A **symbol code** is a map $C : \mathcal{S} \rightarrow \{0,1\}^*$ of this **ensemble**. It maps the string

$$x_1, x_2, \dots, x_N \mapsto C(x_1), C(x_2), \dots, C(x_N) \quad (20)$$

It should satisfy the properties:

1. Every encoded string should be uniquely decodable.
2. It should be easy to decode in some sense.
3. The expected length

$$\mathbb{E}_{\mathbb{P}}[\ell(C)] = \sum_{s \in \mathcal{S}} \ell(C(s))\mathbb{P}(s) \quad (21)$$

of the encoded symbol should be small.

Example 3.1 (Simple Code)

Let's try to create symbol codes for $\mathcal{S} = \{a, b, c, d\}$ with probabilities $\{1/2, 1/4, 1/8, 1/8\}$.

1. The most obvious one is

$$C(s) = \begin{cases} 1000 & s = a \\ 0100 & s = b \\ 0010 & s = c \\ 0001 & s = d \end{cases} \quad (22)$$

with an expected length of 4.

2. We can perhaps shorten this by realizing that the trailing zeros are not needed.

$$C(s) = \begin{cases} 1 & s = a \\ 01 & s = b \\ 001 & s = c \\ 0001 & s = d \end{cases} \quad (23)$$

It does have an expected length of $1\frac{7}{8}$.

3. This one is not a valid scheme since 10010 can be decoded into dc or abd .

$$C(s) = \begin{cases} 1 & s = a \\ 00 & s = b \\ 010 & s = c \\ 10 & s = d \end{cases} \quad (24)$$

It does have an expected length of $1\frac{5}{8}$.

4. Since we have 4 characters, we can just encode into a constant 2-bit string.

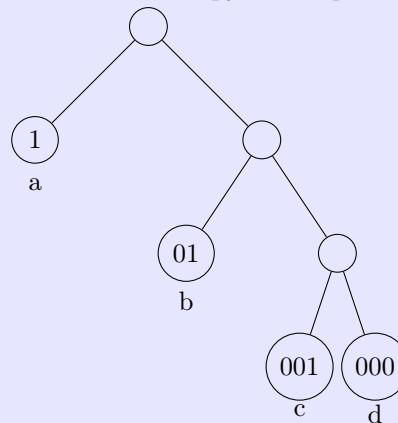
$$C(s) = \begin{cases} 00 & s = a \\ 01 & s = b \\ 10 & s = c \\ 11 & s = d \end{cases} \quad (25)$$

It does have an expected length of 2.

5. We can also see that if we have three 0s, then the next character must be a 1, so this is repetitive.

$$C(s) = \begin{cases} 1 & s = a \\ 01 & s = b \\ 001 & s = c \\ 000 & s = d \end{cases} \quad (26)$$

The expected length is $1\frac{3}{4}$, which turns out to be entropy of this probability distribution. We



can visualize this using a binary tree.

Definition 3.2 (Prefix Code)

Note that all of these encodings except for the nonvalid scheme has the property that no encoding of a character is a prefix of another character. A scheme with this property is called a **prefix code**.

Example 3.2 ()

By modifying the best scheme so far by swapping the 0s and 1s, we have

$$C(s) = \begin{cases} 1 & s = a \\ 10 & s = b \\ 100 & s = c \\ 000 & s = d \end{cases} \quad (27)$$

But this is not a prefix code, though it is a valid code (uniquely decodable since its symmetric counterpart is uniquely decodable). For example, we can sequentially decode the string

$$1000000\dots \quad (28)$$

since we don't know where the 0s end. While we may be able to decode this if we knew the length, it isn't really easy to decode.

The right intuition at this point is to give characters with large probabilities a short codeword and low ones longer ones. However, this is also not true.

Example 3.3 ()

Let's have the same alphabet but now slightly perturb the probabilities

$$\mathbb{P}(a) = \frac{1}{4} + \epsilon, \mathbb{P}(b) = \frac{1}{4} + \frac{\epsilon}{2}, \mathbb{P}(c) = \frac{1}{4} - \frac{\epsilon}{2}, \mathbb{P}(d) = \frac{1}{4} - \epsilon, \quad (29)$$

Then our prefix coding would be

$$C(s) = \begin{cases} 1 & s = a \\ 01 & s = b \\ 001 & s = c \\ 000 & s = d \end{cases} \quad (30)$$

which still has an expected length of 2.25, which is not enough to beat just the regular 2-bit encoding of each word.

Here is a better system of thinking about this. If all codewords have length l , then the number of codewords that we can make is 2^l . Then we can think of each codeword of length l having a "cost" of 2^{-l} in our *codeword supermarket*.

0	00	000	0000
			0001
		001	0010
	01		0011
		010	0100
			0101
	011	0110	
		0111	
1	10	100	1000
			1001
		101	1010
	11		1011
		110	1100
			1101
	111	1110	
		1111	

Figure 5: Our symbol code supermarket where we can buy code words of length l for a price of 2^{-l} . Our budget is 1.

With this visual, there are two constraints that we can reintroduce. First is Kraft’s inequality.

Theorem 3.1 (Kraft Inequality)

Every viable symbol code must have a budget ≤ 1 .

$$\sum_i 2^{-l_i} \leq 1 \tag{31}$$

If a symbol code achieves equality, then this is called a **complete symbol code**.

Second, a prefix code must have all codewords in different “rows” of the supermarket.

0	00	000	0000
		001	0001
			0010
	01	010	0011
			0100
			0101
	011	0110	
		0111	
1	10	100	1000
			1001
		101	1010
			1011
	11	110	1100
			1101
	111	1110	
		1111	

Figure 6: A symbol code that is a prefix code.

3.1 Huffman Coding

Now how well can we do with symbol codes? It turns out that the expected symbol code length cannot beat the entropy, and we will describe how to construct such a symbol code.

Theorem 3.2 (Ideal Code Lengths)

Given the expected length of a symbol code C on an ensemble X , the expected length cannot be less than the entropy.

$$H(X) \leq \mathbb{E}[\ell(C(X))] = \sum_{s \in \mathcal{S}} \mathbb{P}(s)l_i \tag{32}$$

Proof.

Such a code must exist. We first define the *ideal length* of the character s_i with probability p_i to be its surprisal

$$l_i^* = \log_2 \frac{1}{p_i} = \sigma_i \tag{33}$$

If you rearrange this and imagine someone that picked length l_i . We can pick an implicit probability q_i satisfying the ideal length. It is $q_i = 2^{-l_i}$, but the person may not have chosen a complete code, so we must normalize it.

$$q_i = \frac{2^{-l_i}}{Z} \tag{34}$$

where $Z = 1$ if we have a complete code and $Z < 1$ if not. Therefore, we have

$$l_i = \log_2 \frac{1}{q_i} - \log_2 Z \tag{35}$$

We can then plug this into the expected length formula.

$$\begin{aligned}\mathbb{E}[\ell(C(X))] &= \sum_i p_i \left[\log_2 \frac{1}{q_i} - \log_2 Z \right] \\ &= \sum_i p_i \log_2 \frac{1}{p_i} + \sum_i p_i \log_2 \frac{p_i}{q_i} - \log_2 Z \\ &= H(X) + D_{\text{KL}}(p \parallel q) - \log Z \\ &\geq H(X)\end{aligned}$$

since the KL divergence (as we will show later) is greater than 0, and since $Z \leq 1$, we are subtracting a negative number.

To get equality, the proof shows you that you must make your implicit probabilities equal to your true probabilities, so the length of each character should be equal to its surprisal or information content. But these lengths aren't integers, so we must modify this in practice, which will get close, but not exactly to the true minimum. It turns out that we can get the expected length L such that it is within 1 bit of the true minimum.

$$H(X) \leq L \leq H(X) + 1 \quad (36)$$

This is called the *Huffman algorithm*.

Theorem 3.3 (Huffman Algorithm)

The **Huffman algorithm** constructs an optimal prefix code for a given set of symbols and their probabilities. The algorithm proceeds as follows:

1. Create a leaf node for each symbol, and add it to a priority queue.
2. While there is more than one node in the queue:
 - (a) Remove the two nodes with the lowest probability from the queue.
 - (b) Create a new internal node with these two nodes as children, with a probability equal to the sum of their probabilities.
 - (c) Add the new node back into the queue.
3. The remaining node is the root of the Huffman tree.
4. Traverse the tree, assigning 0 to each left branch and 1 to each right branch.
5. The Huffman code for each symbol is the sequence of 0s and 1s on the path from the root to that symbol's leaf node.

The Huffman algorithm produces an optimal prefix code, meaning that for any given set of symbols and probabilities, no other prefix code produces a smaller expected codeword length.

1. The expected length L of the Huffman code satisfies:

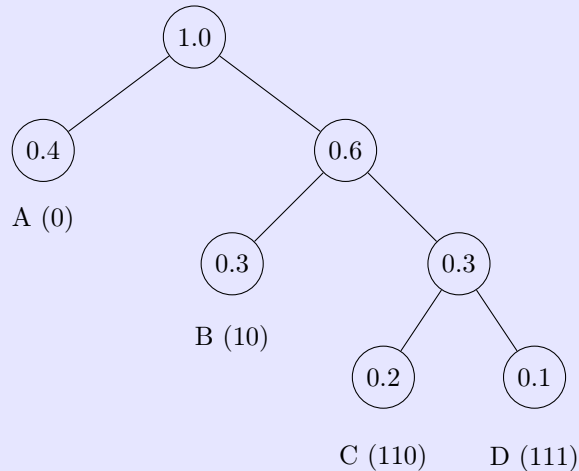
$$H(X) \leq L < H(X) + 1 \quad (37)$$

2. The Huffman code is optimal among all prefix codes.
3. Symbols with higher probabilities get shorter codewords.
4. The algorithm has a time complexity of $O(n \log n)$ where n is the number of symbols.

Example 3.4 (Huffman Coding)

Consider the alphabet $\mathcal{S} = \{A, B, C, D\}$ with probabilities $\{0.4, 0.3, 0.2, 0.1\}$.

1. Initial nodes: A(0.4), B(0.3), C(0.2), D(0.1)
2. Combine D and C: (D,C)(0.3), A(0.4), B(0.3)
3. Combine (D,C) and B: ((D,C),B)(0.6), A(0.4)
4. Final tree: (A,((D,C),B))(1.0)

Figure 7: Huffman tree for the alphabet $\{A, B, C, D\}$

Resulting codes:

- A: 0
- B: 10
- C: 110
- D: 111

The expected length is $L = 0.4(1) + 0.3(2) + 0.2(3) + 0.1(3) = 1.9$ bits compared to the entropy of $H(X) = -\sum_i p_i \log_2 p_i \approx 1.85$ bits. This demonstrates that the Huffman code achieves a length very close to the entropy, and always within 1 bit of it:

$$H(X) \approx 1.85 \leq L = 1.9 < H(X) + 1 \approx 2.85 \quad (38)$$

4 Differential Entropy

4.1 Differential Entropy

Definition 4.1 (Differential Entropy)

For a continuous random vector, the **differential entropy** is defined

$$H[\mathbf{X}] = - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \quad (39)$$

4.2 Kullback Leibler Divergence

The **relative entropy**, or **Kullback-Leibler divergence**, of distributions $p(x)$ and $q(x)$ is defined

$$\begin{aligned} \text{KL}(p||q) &:= - \int p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} - \left(- \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \right) \\ &= - \int p(\mathbf{x}) \ln \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} \right) d\mathbf{x} \end{aligned}$$

We can show that this quantity is always greater than or equal 0 by Jensen's inequality using the fact that $-\ln(x)$ is concave

$$\int p(\mathbf{x}) - \ln\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} \geq -\ln \int p(\mathbf{x}) \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} = -\ln \int q(\mathbf{x}) d\mathbf{x} = -\ln(1) = 0 \quad (40)$$

and it is precisely 0 if $p = q$, so it behaves similarly to a metric. However, it isn't exactly since it is not symmetric.

Let's demonstrate how entropy and the KL divergence applies to maximum likelihood estimation. Suppose that iid samples $\mathcal{D} = \{(x^{(n)}, y^{(n)})\}$ are given in a regression problem. Let $P^* = (X, Y)$ be the true data generating function. Then, we want to compute an approximation of P^* with P_θ , where P_θ is some parameterized distribution. The negative log likelihood of the y 's being generated is

$$\ell(\theta) = \frac{1}{N} \sum_{n=1}^N \log P_\theta(y_i | x_i) \quad (41)$$

which asymptotically converges to

$$\mathbb{E}_{P^*}[-\log P_\theta(y_i | x_i)] = \text{KL}(P^* || P) + H[P^*] \quad (42)$$

and since the entropy is constant, this is equivalent to minimizing the KL divergence between P and P^* .

We assume that the $y^{(n)}$'s come from a conditional distribution P_{θ, x_i} , where the parameters of the distribution is θ and x_i

4.3 Entropy of Probability Measures

First, we want to quantitatively measure the "surprise" of an event E happening in a probability space by assigning it a value $H(E)$. We want it to satisfy the following:

1. $H(E) \geq 0$. The surprisal of any event is nonnegative.
2. $H(E) = 0$ iff $\mathbb{P}(E) = 1$. No surprisal is gained from events with probability 1.
3. If E_1 and E_2 are independent events, then $H(E_1 \cap E_2) = H(E_1) + H(E_2)$. The information from two independent events should be the sum of their informations.
4. H should be continuous, i.e. slight changes in probability correspond to slight changes in surprisal.

Definition 4.2 (Surprisal)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the **surprisal**, or **self-information**, of an event $E \in \mathcal{F}$ is

$$\sigma_{\mathbb{P}}(E) := -\log \mathbb{P}(E) \quad (43)$$

and the **expected surprisal** of E is

$$h_{\mathbb{P}}(E) = \mathbb{P}(E) \sigma_{\mathbb{P}}(E) \quad (44)$$

Now we can define entropy as the expected surprisal of a random variable, which seems now more motivated and intuitive.

Definition 4.3 (Entropy)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a \mathbb{P} -almost partition is a set family $\mathcal{G} \subset \mathcal{F}$ such that $\mu(\cup_{G \in \mathcal{G}} G) = 1$ and $\mathbb{P}(A \cap B) = 0$ for all distinct $A, B \in \mathcal{G}$ (this is a relaxation of the usual conditions for a partition).

The **entropy** of the subfamily \mathcal{G} is

$$H_{\mathbb{P}}(\mathcal{G}) := \sum_{G \in \mathcal{G}} h_{\mathbb{P}}(G) \quad (45)$$

The **entropy** of the σ -algebra \mathcal{F} is defined

$$H_{\mathbb{P}}(\mathcal{F}) = \sup_{\mathcal{G} \subset \mathcal{F}} H_{\mathbb{P}}(\mathcal{G}) \quad (46)$$

Now the entropy of a random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathcal{X}, \mathcal{H})$ will induce a measure \mathbb{P}_X on \mathcal{X} . Then the entropy of X is defined over this induced measure.

$$H[X] := H_{\mathbb{P}_X}(\mathcal{H}) = \sup_{\mathcal{G} \subset \mathcal{H}} H_{\mathbb{P}_X}(\mathcal{G}) \quad (47)$$

Intuitively, this represents the element of surprise of a certain data point, and distributions that have relatively sharp peaks will have lower entropy (since we expect most of the samples to come from the peaks) while uniform distributions have higher entropy.

References

- [1] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.