Information Theory and Signal Processing

Muchang Bahng

Spring 2024

Contents

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].

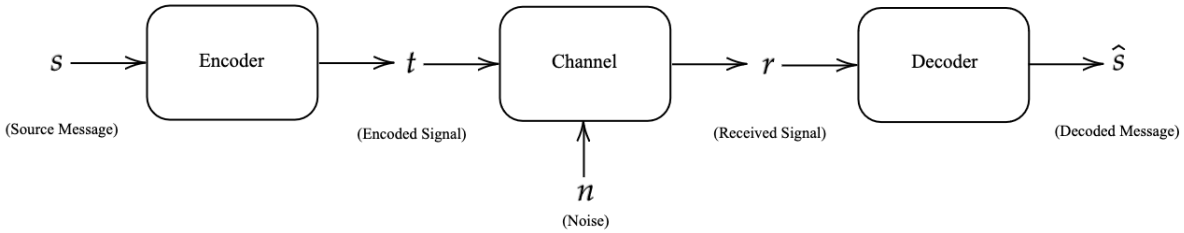


Figure 1: A channel diagram.

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.

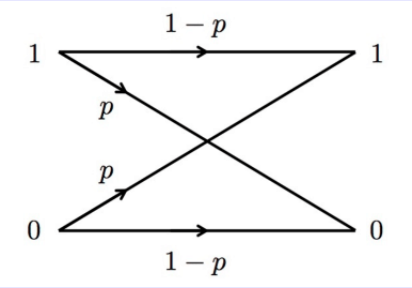


Figure 2: A simple example of noise.

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.

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

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.

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

For example, with $k = 3$ we have

```
s = 01101
t = 000 111 111 000 111
n = 000 100 000 101 000
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.

[a] In 2014 disk drives, the standard was that p should not be greater than 10^{-18}.
[a] It 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.

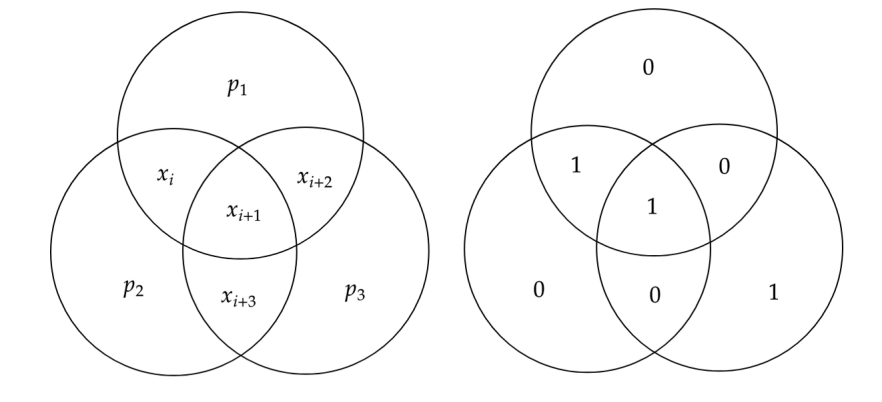


Figure 3: (7, 4) hamming code visual with example on the right.

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 $r_:$ 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.

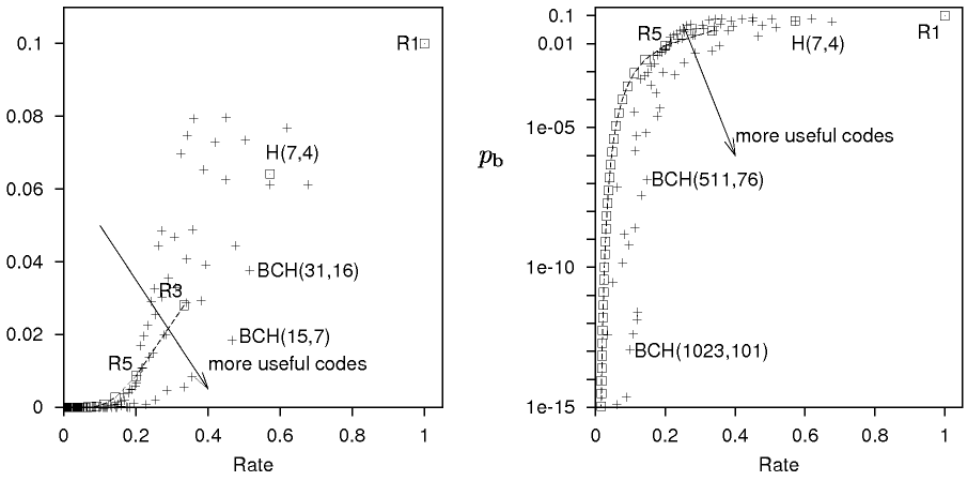


Figure 4: The rate of an encoding/decoding scheme vs probability of bit error.

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

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

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 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 2.1 (Surprisal)

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

$$\sigma_{\mathbb{P}}(E) \coloneqq -\log \mathbb{P}(E) \tag{6}$$

and the **expected surprisal** of E is

$$h_{\mathbb{P}}(E) = \mathbb{P}(E)\sigma_{\mathbb{P}}(E) \tag{7}$$

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

Definition 2.2 (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

$$H_{\mathbb{P}}(\mathcal{G}) \coloneqq \sum_{G \in \mathcal{G}} h_{\mathbb{P}}(G) \tag{8}$$

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

$$H_{\mathbb{P}}(\mathcal{F}) = \sup_{\mathcal{G} \subset \mathcal{F}} H_{\mathbb{P}}(\mathcal{G}) \tag{9}$$

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

$$H[X] \coloneqq H_{\mathbb{P}_X}(\mathcal{H}) = \sup_{G \subset \mathcal{H}} H_{\mathbb{P}_X}(\mathcal{G}) \tag{10}$$

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.

2.2 Discrete Random Variables

With the general version, we can prove the following subdefinition for discrete random variables.

Lemma 2.1 (Entropy of Discrete RV)

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

$$H[X] \coloneqq \mathbb{E}_X[-\ln p(X)] = -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \ln \mathbb{P}(X = x) \tag{11}$$

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

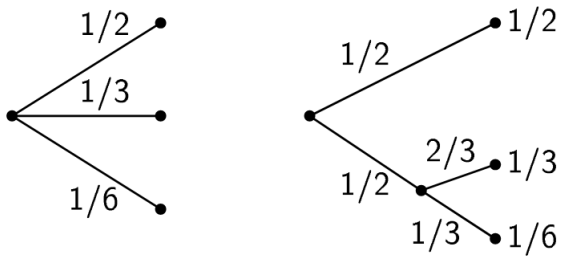
Proof.

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

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,

$$H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3}) \tag{12}$$

Which then leads to the definition of entropy above. 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.

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.

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

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.

2.3 Joint and Conditional Entropy

Definition 2.3 (Joint, Conditional Entropy)

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

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

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

Theorem 2.2 (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.

$$H(X,Y) \leq H(X) + H(Y) \tag{13}$$

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 2.3 (Conditional Entropy)

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

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y) \tag{14}$$

Theorem 2.4 (Conditioning Never Decreases Uncertainty)

Since

$$H(X) + H(Y) \geq H(X,Y) = H(X) + H(Y \mid X) \tag{15}$$

we have $H(Y) \geq H(Y \mid X)$. 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 2.4 (Mutual Information)

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

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \tag{16}$$

This can be conditioned on another random variable Z.

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) = H(Y \mid Z) - H(Y \mid X, Z) \tag{17}$$

The entropy also demonstrates the average length (if base is 2) number of bits required to transmit the state of a random variable.

2.4 Differential Entropy

Definition 2.5 (Differential Entropy)

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

$$H[\mathbf{X}] = -\int p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x} \tag{18}$$

2.5 Kullback Leibler Divergence

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

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

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

$$\int p(\mathbf{x})\ -\ln \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} \right) d\mathbf{x} \geq -\ln \int p(\mathbf{x})\, \frac{q(\mathbf{x})}{p(\mathbf{x})}\, d\mathbf{x} = -\ln \int q(\mathbf{x})\, d\mathbf{x} = -\ln(1) = 0 \tag{19}$$

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

$$\ell(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log P_\theta(y_i \mid x_i) \tag{20}$$

which asymptotically converges to

$$\mathbb{E}_{P^*}[-\log P_\theta(y_i \mid x_i)] = \mathrm{KL}(P^*||P) + H[P^*] \tag{21}$$

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

References

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