Two Theorems by Claude Shannon

Kent Quanrud

December 14, 2020

1 Introduction

Today we will discuss two papers by Claude Shannon published in the Bell Systems Technical Journal at roughly the same time.

- Claude E. Shannon. "A mathematical theory of communication". In: *Bell Syst. Tech. J.* 27.3 (1948), pp. 379–423
- Claude E. Shannon. "The synthesis of two-terminal switching circuits". In: *Bell Syst. Tech. J.* 28.1 (1949), pp. 59–98

While we will not discuss it today, we point out that Shannon another important paper at the same time, this time laying foundations for cryptography.

- Claude E. Shannon. "Communication theory of secrecy systems". In: *Bell Syst. Tech. J.* 28.4 (1949), pp. 656–715

2 Circuit lower bounds

The first result of Shannon that we will discuss circuits, and in particular, lower bounds for circuits. This result was not Shannon's first famous encounter with circuits. His 1936 Master's thesis, born out of his experience as a research assistant working on Vannevar Bush's differential analyze, identified the logical equivalence between the switching circuits (that the differential analyzer was composed of) and boolean algebra.

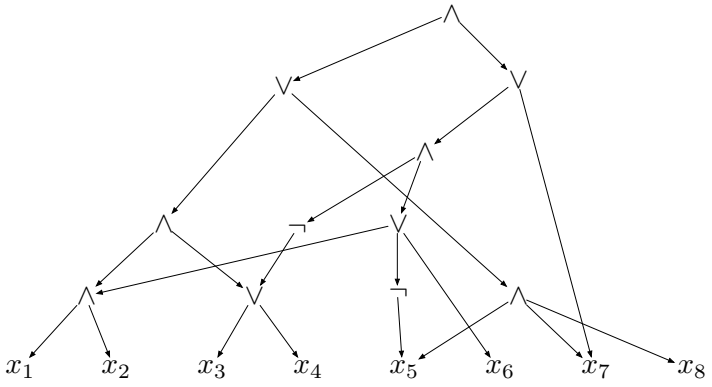


Figure 1: A circuit.

The current discussion on circuit fast forwards a little more than ten years for Shannon's masters thesis. It comes at an exciting time for Bell Labs: in late 1947, Bardeen, Brattain, and Shockley in the Solid States Physics group invented the first *transistor*. This sets off the digital revolution, as transistors rapidly improve and coallesce into the CPU's we know today. For the sake of this discussion, we can think of transistors as extremely fast and efficient physical circuits.

For us, a circuit is a family of logical *gates* arranged in a directed acyclic graph. We assume the reader has had some acquaintance with gates before (see, e.g., [7, 8]), and we briefly review. An example of a circuit is given in Figure 1. We have four types of gates. One is an *input gate*, representing a bit from the input. In the picture above, we have input gates for 8 bits, $x_1, \ldots, x_8$. We then have *logical gates*, that take as input a finite number of bits and output a bit based on applying either an "and", $\wedge$, an "or", $\vee$; or a negation, $\neg$. The $\neg$ takes only one input bit and the other two logical gates can take any finite number of bits. The input bits of a gate come from the outputs of some other gates. If we draw a directed edge from one gate to another based on the flow of information, this graph must form a DAG.

Theorem 1. *For all $n \in \mathbb{N}$, there exists a function $f : \{0,1\}^n \to \{0,1\}$ that cannot be computed by any circuit of size $< 2^n/cn$ for some universal constant $c > 0$.*

Proof. We first count the number of circuits of some size. We first observe that any circuit of size k can be encoded in $ck\log(k)$ bits, for some universal constant $c > 0$, by interprettign a circuit of size k as a graph with k vertices and edges. Each edge and each vertex requires $O\big(\log(k)\big)$ bits to encode. In turn, there are at most $2^{ck\log(k)}$ different circuits of size k. For $k = 2^n/cn$, this is less than 2^n.

Now, let $f : \{0,1\}^n \to \{0,1\}$ be a *uniformly random* boolean function. For each circuit C of size $\leq 2^n/cn$, the probability that $f(x) = C(x)$ for all $x \in \{0,1\}^n$ is $< 1/2^n$. Thus the expected number of circuits of size $\leq 2^n/cn$ that agree with f on all inputs is < 1. By the probabilistic method, there exists an f that is unequal to all circuits of size $\leq 2^n/cn$. ∎

2.1 A Circuit Hierarchy Theorem

An important application is the following Hierarchy theorem for circuits.

Definition 2. *Let $f : \mathbb{N} \to \mathbb{N}$ be a function. A* **circuit family of size f** *is a sequence of boolean circuits $C_1, C_2, \ldots$, where $C_n : \{0,1\}^n \to \{0,1\}$ and $|C_n| \leq f(n)$.*

Definition 3. *A language $L \subset \{0,1\}^*$ has* **circuit complexity f**, *denoted $L \in \textsc{CircuitSize}(f)$. if there is a circuit family of size f, $\{C_i : \{0,1\}^i \to \{0,1\}\}$, such that for each $x \in \{0,1\}^n$, whether $x \in L$ can be decided by the circuit C_n.*

Some basic observations about circuits.

1. Every polynomial time language has polynomial circuit complexity.
2. Circuit Satisfiability and Boolean Satisfiablity are polynomial time equivalent.
3. The first two points combine to give a proof of NP-Completeness [4, 6].

Given the importance of circuits, it is impmortant to understand what circuits of a given size can and cannot do. Shannon's theorem above gives some limits to the power of circuits. The following Hierarchy theorem, the proof of which uses Shannon's theorem, shows that the size of a circuit matters at a fine grained level: there always exists boolean functions that a larger size can implement, but a smaller size cannot.

Theorem 4. *There exists a universal constant $c > 0$ for which the following holds. Let $f : \mathbb{N} \to \mathbb{N}$ with $n \leq f(n) \leq 2^n/cn$. Then*

$$\textsc{CircuitSize}(f) \subsetneq \textsc{CircuitSize}(cf).$$

Proof. Here we will prove a weaker bound with $cf \log f$ instead of f. Fix $n \in \mathbb{N}$ and let $m = f(n)$. We show that there is a boolean function $g : \{0,1\}^n \to \{0,1\}$, such that f has a $O(m \log m)$-size circuit but not an m-size circuit.

For any $k \in \mathbb{N}$, we know that there is a function $g : \{0,1\}^k \to \{0,1\}$ that requires at least $2^k/ck$ size, and at most $ck2^k$ size, for some constant $c > 0$.

We choose k large as possible such that $2^k/ck \leq m$. Then

$$ck2^k \leq (ck)^2(m+2) \leq O\big(m \log(m)\big).$$

■

3 Coding theory

Coding theory concerns the very practical problem of digitial copmmunication over imperfect lines of communication. Here we consider a one particular model adopted by Shannon. In this model, we imagine two locations, A and B. We want to transmit a bit string $x \in \{0,1\}^m$ from point A to point B. While points A and B are connected by some kind of connection, this connection is imperfect. When sending a bit string from point A to point B, every bit gets flipped independently with some probability p. The goal is to reliably communicate bit strings even in the presence of this faulty connection.

It is not impossible to communicate over a bad connection, as anyone who uses a phone knows. Suppose you are calling a friend, and the reception is not very good. You say something. Your friend 'replies, 'sorry, I couldn't hear you". So you say it agian. Your friend again suggests that they didn't understand you. So you say it again and again and again and eventually you start yelling. Ultimately, you are adding *redundancy* to try to communicate your point.

The goal of coding theory is to add enough redundancy to reliably communicate, but otherwise minimize the amount of redundancy. In particular, we have two functions, an **encoder** $C : \{0,1\}^m \to \{0,1\}^n$ and a **decoder** $\mathcal{D} : \{0,1\}^n \to \{0,1\}^m$. The encoder takes the input message from $x \in \{0,1\}^m$ and maps to to a longer message $\{0,1\}^n$, where $n \geq m$. The encoded message $C(x)$ is transmitted. On the other end,

Let $\mathcal{N} : \{0,1\}^n \to \{0,1\}^n$ be a random function that flips each bit independently with probability p.

$$\underset{(\mathbb{R}^m)}{x} \xrightarrow{\text{encode}} \underset{(\mathbb{R}^n)}{C(x)} \xrightarrow{\text{bits flip w/ prob. } p} \underset{(\mathbb{R}^n)}{\mathcal{N}(C(x))} \xrightarrow{\text{decode}} \underset{(\mathbb{R}^m)}{\mathcal{D}(\mathcal{N}(C(x)))}$$

The **rate of transmission** is the ratio

$$\text{rate of transmission} = \frac{m}{n} = \frac{\text{\# input bits}}{\text{\# output bits}}.$$

The **average error rate** is the probability

$$\mathbf{P}[\mathcal{D}(\mathcal{N}(C(x))) \neq x]$$

over the randomness in $\mathcal{N}$ and over $x \in \{0,1\}^m$ chosen uniformly at random.

Before introducing Shannon's theorem for codes, we have to introduce one more character: entropy.

Definition 5. *Let $X \in \mathcal{X}$ be a discrete random variable. The* **entropy** *of X, denoted $\mathbf{H}(X)$, is defined as*

$$\mathrm{H}(X) = \sum_{x \in \mathcal{X}} -\mathbf{P}[X = x] \log(\mathbf{P}[X = x]),$$

where the convention that $0/0 = 0$, and that log *denotes the logarithm base 2.*

For $p \in (0, 1)$, $\boldsymbol{H(p)}$ is defined as the entropy $H(X)$ of the binary variable $X \in \{0, 1\}$ with $\mathbf{P}[X = 1] = p$. That is,

$$H(p) = p \log\left(\frac{1}{p}\right) + (1 - p) \log\left(\frac{1}{1 - p}\right).$$

We will do a more thorough investigation of entropy later on in this article.

Theorem 6. *Consider transmission over a noisy channel where each bit is flipped independently with probability $p \in (0, 1/2)$.*

1. *For all $\delta > 0$, and n sufficiently large, there is a coding scheme $(C : \{0,1\}^m \to \{0,1\}^n, \mathcal{D} : \{0,1\}^n \to \{0,1\}^m)$ that has transmission rate $\geq 1 - H(p) - \epsilon$ and average error rate $\leq \epsilon$.*
2. *For all fixed $\delta > 0$, and n sufficiently large any coding scheme $(C : \{0,1\}^m \to \{0,1\}^n, \mathcal{D} : \{0,1\}^n \to \{0,1\}^m)$ with transmission rate $\geq 1 - H(p) + \delta$ has average error rate $> 1 - \delta$.*

3.1 A helpful inequality

Before proving Shannon's theorem, we mention a very helpful identity that relates entropy to sums of binomial coefficients. Entropy enters the analysis of Shannon's upper bound via this lemma.

Lemma 7. *Let $n \in \mathbb{N}$ and $p \in (0, 1)$. Then*

$$\sum_{i=0}^{pn} \binom{n}{i} \leq 2^{H(p)n}.$$

We will prove Lemma 7 below in Section 4 where we discuss entropy in greater detail.

3.2 Shannon's Upper Bound

In this section, we probe Shannon's upper bound - the first claim in Theorem 6. We first state the part that is relevant.

Theorem 8. *For all $\delta > 0$, there exists a coding scheme*

$$(C : \{0,1\}^m \to \{0,1\}^n, \mathcal{D} : \{0,1\}^n \to \{0,1\}^m)$$

that has average error $\leq \delta$ and transmission rate $\geq 1 - H(p) - \delta$.

Proof. Let $n \in \mathbb{N}$ be a large parameter TBD, and let $m = (1 - H(p) - \delta)n$. Let $C : \{0,1\}^m \to \{0,1\}^n$ be a *uniformly random* function. Define $\mathcal{D} : \{0,1\}^n \to \{0,1\}^m$ by setting $\mathcal{D}(y)$ to be the point x closest to $C(x)$, breaking ties arbitrarily.

Since $\mathrm{H}(p)$ is continuous, we can choose $\epsilon > 0$ sufficiently small such that

$$|\mathrm{H}((1 + \epsilon)p) - \mathrm{H}(p)| \leq \frac{\delta}{2}.$$

We claim the following for each fixed $x \in \{0,1\}^m$.

1. *For sufficiently large n, with probability of error* $\leq \delta/2$*, no other point* $x' \in \{0,1\}^m, x' \neq x$ *is within* $(1+\epsilon)pn$ *bits of* $C(\mathcal{N}(x))$.

2. *For sufficiently large n, with probability of error* $\leq \delta/2$*, the noisy transmission* $\mathcal{N}(C(x))$ *flips at most* $(1+\epsilon)pn$ *bits in* $C(x)$.

Suppose the above holds and let n be sufficiently large. Then with combined probability of error $\leq \delta$, $C(x)$ is the only point within $(1+\epsilon)pn$ bits of $C(x)$. That is, when we consider all of the randomness over the random of choice of $x \in \{0,1\}^m$, the random encoding function $C : \{0,1\}^m \to \{0,1\}^n$, and the noise $\mathcal{N}(C(x))$, we have

$$\mathop{\mathbf{P}}_{x,C,\mathcal{N}}[\mathcal{D}(\mathcal{N}(C(x))) \neq x] \leq \delta.$$

We can rewrite this as

$$\mathop{\mathbf{E}}_{C}[\text{average error of } C] = \mathop{\mathbf{E}}_{C}\left[\mathop{\mathbf{P}}_{x,\mathcal{N}}[\mathcal{D}(\mathcal{N}(C(x))) \neq x]\right] \leq \delta.$$

To dramatically complete the proof: *by the probabilistic method, there exists an encoding* $C : \{0,1\}^m \to \{0,1\}^n$ *such that the average error is* $\leq \delta$.

Claim 1. *For sufficiently large n, with probability of error* $\leq \delta/2$*, no other point* $x' \in \{0,1\}^m, x' \neq x$ *is within* $(1+\epsilon)pn$ *bits of* $\mathcal{N}(C(x))$.

We prove the claim conditional on $y = \mathcal{N}(C(x))$; the unconditional claim immediately follows. Fix $y = \mathcal{N}(C(x))$. Consider any other input point $x' \in \{0,1\}^m$. By construction, $C(x')$ is selected uniformly at random from $\{0,1\}^n$. Therefore

$$\mathbf{P}[\|C(x') - C(x)\|_0 \leq (1+\epsilon)pn] = 2^{-n} \sum_{i=0}^{(1+\epsilon)pn} \binom{n}{i} \leq 2^{(H((1+\epsilon)p)-1)n}.$$

By the union bound, we have

$$\mathbf{P}[\|C(x') - C(x)\|_0 \leq (1+\epsilon)pn \text{ for some } x' \neq x] \leq 2^{m+(H((1+\epsilon)p)-1)n}.$$

The RHS is $\leq \delta/2$ iff

$$m \leq (1 - H((1+\epsilon)p))n - 1 - \log(1/\delta),$$

which occurs iff

$$(\text{transmission rate}) = \frac{m}{n} \leq 1 - H((1+\epsilon)p) - O\left(\frac{\log(1/\epsilon)}{n}\right).$$

By choice of ϵ, we have

$$1 - H((1+\epsilon)p) - O\left(\frac{\log(1/\epsilon)}{n}\right) \geq 1 - H(p) - \delta/2 - O\left(\frac{\log(1/\epsilon)}{n}\right) \geq 1 - H(p) - \delta$$

for n sufficiently large. The claim now follows from the choice of m.

Claim 2. *Fix $y \in \{0,1\}^n$. For sufficiently large n, with probability of error $\leq \delta/2$, a noisy transmission $\mathcal{N}(y)$ flips at most $(1+\epsilon)pn$ bits in y.*

We have

$$\lim_{n\to\infty} \mathbf{P}[(\#\text{ bits flipped}) \geq (1+\epsilon)pn] \overset{(a)}{\leq} \lim_{n\to\infty} e^{-\epsilon^2 pn/2} = 0.$$

where (a) applies the Chernoff inequality. ∎

3.3 Lower bounds for Shannon's capacity

Theorem 9. *Let $\delta > 0$. For n sufficiently large, no coding scheme $(C : \{0,1\}^m \to \{0,1\}^n, \mathcal{D} : \{0,1\}^n \to \{0,1\}^m)$ can have transmission rate*

Proof. Let $(C : \{0,1\}^m \to \{0,1\}^n, \mathcal{D} : \{0,1\}^n \to \{0,1\}^m)$ be a coding scheme with transmission rate $> 1 - H(p) + \delta$. We claim that this code has error rate $> \delta$. In fact, will show that error rate can be made arbitrarily large for sufficiently large n.

Let $\epsilon > 0$ be a parameter TBD. Let E be the event that $\mathcal{N}(C(x))$ differs in at least $(1-\epsilon)pn$ bits from $C(x)$ and no more than $(1+\epsilon)pn$ bits from $C(x)$. Let $\bar{E}$ be the complementary event.

We have

$$(\text{average correctness rate}) \leq \mathbf{P}\left[\bar{E}\right] + \mathbf{P}[\mathcal{D}(\mathcal{N}(C(x))) = x \mid E].$$

For the first term, we have

$$\mathbf{P}\left[\bar{E}\right] = \mathbf{P}[\|\mathcal{N}(C(x)) - C(x)\|_0 < (1-\epsilon)pn] + \mathbf{P}[\|\mathcal{N}(C(x)) - C(x)\|_0 > (1+\epsilon)pn] \leq 2e^{-\epsilon^2 pn/3}$$

by Chernoff bounds. In particular, for fixed $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left[\bar{E}\right] = 0.$$

To bound the second term, for each x, let

$$Y_x = \left\{y \in \mathcal{D}^{-1}(x) : (1-\epsilon)pn \leq \|x - y\|_0 \leq (1+\epsilon)pn\right\}.$$

We have

$$\mathbf{P}[\mathcal{D}(\mathcal{N}(C(x))) = x \mid E] = \frac{1}{2^m} \sum_{x\in\{0,1\}^m} \sum_{y\in Y_x} \mathbf{P}[\mathcal{N}(C(x)) = y].$$

For each x, and for each $y \in Y_x$, we have

$$\mathbf{P}[\mathcal{N}(C(x)) = y] \overset{(a)}{\leq} p^{(1-\epsilon)pn}(1-p)^{(1-(1-\epsilon)p)n} = 2^{-\mathrm{H}(p)n}\left(\frac{1-p}{p}\right)^{\epsilon pn}$$

Here (a) is because because y differs in at least $(1-\epsilon)pn$ bits. We now have

$$\begin{aligned}
\mathbf{P}[\mathcal{D}(\mathcal{N}(C(x))) = x \mid E] &\leq \frac{1}{2^m} \sum_{x\in\{0,1\}^m} e^{-\mathrm{H}(p)n}\left(\frac{1-p}{p}\right)^{\epsilon pn} |Y_x| \\
&\overset{(b)}{\leq} 2^{(1-\mathrm{H}(p))n-m}\left(\frac{1-p}{p}\right)^{\epsilon pn} \\
&\overset{(c)}{\leq} 2^{-\delta n}\left(\frac{1-p}{p}\right)^{\epsilon pn} = 2^{\left(\epsilon \log\left(\frac{1-p}{p}\right)-\delta\right)n}.
\end{aligned}$$

Here (b) is because the sets Y_x partition $\{0,1\}^n$, so their cardinalities sum to at most 2^n. (c) is by assumption on the transmission rate. For $\epsilon > 0$ sufficiently small, the RHS tends to 0 as $n \to \infty$. ∎

4 Entropy

In this section, we explore the some of the many interesting properties of entropy. We restate the definition for the reader's convenience.

Definition 5. *Let $X \in \mathcal{X}$ be a discrete random variable. The* **entropy** *of X, denoted $\mathbf{H}(X)$, is defined as*

$$\mathrm{H}(X) = \sum_{x \in \mathcal{X}} -\mathbf{P}[X = x] \log(\mathbf{P}[X = x]),$$

where the convention that $0/0 = 0$, and that log *denotes the logarithm base 2.*

For $p \in (0, 1)$, $\boldsymbol{H(p)}$ is defined as the entropy $H(X)$ of the binary variable $X \in \{0, 1\}$ with $\mathbf{P}[X = 1] = p$. That is,

$$H(p) = p \log\left(\frac{1}{p}\right) + (1 - p) \log\left(\frac{1}{1 - p}\right).$$

For an alternative definition of the entropy of $X \in \mathcal{X}$, let X' be an independent and identically distributed copy of X. Then

$$\mathrm{H}(X) = \mathop{\mathbf{E}}_{X}\left[\log\left(\frac{1}{\mathbf{P}[X' = X]}\right)\right].$$

Put another way, given a discrete random variable $X \in \mathcal{X}$, let us define[1] the **shock** of X, $S_X > 0$, as the random variable that, if $X = x$, takes the value

$$S_X = \frac{1}{\mathbf{P}[X = x]}.$$

Here $\mathbf{P}[X = x]$ refers to the *a priori* probability of X equaling x. Then the entropy of X is

$$\mathrm{H}(X) = \mathbf{E}\big[\log(S_X)\big].$$

4.1 Concavity and the maximality principle.

Recall that a function $f : [a, b] \to \mathbb{R}$ is **concave** if for all $x, y \in [a, b]$ and $p \in [0, 1]$, we have

$$pf(x) + (1 - p)f(y) \leq f(px + (1 - p)y).$$

By induction, we can extend this to finite convex combinations of points. Let $x_1, \ldots, x_n \in [a, b]$ and $p_1, \ldots, p_n \geq 0$ with $p_1 + \cdots + p_n = 1$. Then we have

$$p_1 f(x_1) + \cdots + p_2 f(x_n) \leq f(p_1 x_1 + \cdots + p_n x_n). \tag{1}$$

Let $f : [a, b] \to \mathbb{R}$ be a function and let $X \in [a, b]$ be a discrete random variable taking on a finite number of valuesq. Say X takes on n values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$ respectively. Then (1) is the same as saying that

$$\mathbf{E}[f(X)] \leq f(\mathbf{E}[X]).$$

We can extend this to continuous distributions of X by approximation by finite distributions. Thus we have *Jensen's inequality*, which is basically rewriting the definition of concavity.

[1] This is not a standard terminology.

Lemma 10. *Let $X \in [a,b]$ be a random variable, and let $f : [a,b] \to \mathbb{R}$ by concave. Then*

$$\mathbf{E}[f(X)] \leq f(\mathbf{E}[X]).$$

Proof. Suppose X takes on only two values, a and b, with probability p and $(1-p)$ respectively. Then

$$\mathbf{E}[f(X)] = pf(a) + (1-p)f(b) = f(pa + (1-p)b) = f(\mathbf{E}[X]).$$

We can extend the argument to any finite number of values by induction. ∎

Lemma 11. *Over all discrete distributions over n values, entropy is maximized by the uniform distribution, which has entropy is* $\log(n)$.

Proof. Suppose X takes on at most n different values x. Then

$$\mathrm{H}(X) = \mathbf{E}\big[\log(S_X)\big] \overset{\text{(a)}}{\leq} \log(\mathbf{E}[S_X]) \overset{\text{(b)}}{=} \log(|\mathcal{X}|)$$

(a) is by Jensen's inequality. (b) is because

$$\mathbf{E}[S_X] = \sum_x \frac{\mathbf{P}[X=x]}{\mathbf{P}[X=x]} = n.$$

On the other hand, if X is the uniform distribution over n values $x_1, \ldots, x_n$, then

$$H(x) = \sum_i \mathbf{P}[X=x_i] \log\left(\frac{1}{X=x_i}\right) = \sum_i \frac{1}{n}\log(n) = \log(n).$$

∎

4.2 Conditional entropy

Let (X,Y) be jointly distributed random variables. Conditional on X, Y is a random variable with a well defined entropy $H(Y)$ (given X).

Definition 12. *The conditional entropy of Y on X is defined as*

$$\mathrm{H}(Y \mid X) = \mathop{\mathbf{E}}_X[\mathrm{H}(Y) \mid X]$$

In terms of "shocks", we let $S_{Y|X}$ be the shock value of the conditional variable Y given X. To be precise, conditional on $X = x$ and $Y = y$, $S_{Y|X}$ takes the value

$$S_{Y|X} = \frac{1}{\mathbf{P}[Y=y \mid X=x]},$$

where $\mathbf{P}[Y=y \mid X=x]$ is the *a priori* probability given only $X=x$. Then

$$\mathrm{H}(Y \mid X) = \mathop{\mathbf{E}}_X\left[\mathop{\mathbf{E}}_Y\big[\log\big(S_{X|Y}\big)\big]\right] = \mathop{\mathbf{E}}_{X,Y}\big[\log\big(S_{X|Y}\big)\big]$$

Lemma 13.

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

Proof. Observe that conditional on $X=x$ and $Y=y$, we have

$$S_{X,Y} = \frac{1}{\mathbf{P}[X=x, Y=y]} = \frac{1}{\mathbf{P}[X=x]\,\mathbf{P}[Y=y \mid X=x]} = S_X S_{Y|X}.$$

Thus

$$\mathrm{H}(X,Y) = \mathop{\mathbf{E}}_{X,Y}\big[\log\big(S_{X,Y}\big)\big] = \mathop{\mathbf{E}}_{X,Y}\big[\log\big(S_{Y|X}\big) + \log\big(S_X\big)\big] = \mathrm{H}(Y \mid X) + \mathrm{H}(X),$$

as desired. ∎

5 Principle of Independence

Lemma 14. *Let (X, Y) be jointly distributed. Then* $\mathrm{H}(Y \mid X) \leq \mathrm{H}(Y)$

Proof. We have

$$\mathrm{H}(Y \mid X) = \mathop{\mathbf{E}}_{X,Y}\left[\log\left(S_{Y|X}\right)\right] = \mathop{\mathbf{E}}_{Y}\left[\mathop{\mathbf{E}}_{X}\left[\log\left(S_{Y|X}\right) \,\middle|\, Y\right]\right]$$
$$\overset{(a)}{\leq} \mathop{\mathbf{E}}_{Y}\left[\log\left(\mathop{\mathbf{E}}_{X}\left[S_{Y|X} \,\middle|\, Y\right]\right)\right] \overset{(b)}{\leq} \mathrm{H}(Y).$$

(a) is by Jensen's inequality. (b) is because, conditional on $Y = y$, we have

$$\mathop{\mathbf{E}}_{X}\left[S_{Y|X} \,\middle|\, Y\right] = \sum_{x} \frac{\mathbf{P}[X = x \mid Y = y]}{\mathbf{P}[Y = y \mid X = x]} \overset{(c)}{=} \sum_{x} \frac{p(x)}{p(y)} = \frac{1}{p(y)}.$$

(c) substitutes Bayes' law:

$$\mathbf{P}[X = x \mid Y = y]\,\mathbf{P}[Y = y] = \mathbf{P}[X = x, Y = y] = \mathbf{P}[Y = y \mid X = x]\,\mathbf{P}[X = x].$$

■

Maximality of independence

Lemma 15. *Fix marginal probabilities over two finite sets $\mathcal{X}$ and $\mathcal{Y}$. Over all joint distributions of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, the entropy of (X, Y) is maximized by taking X and Y to be independent.*

Subadditivity

Lemma 16. *Let (X, Y) be a joint distribution of discrete random variables. Then*

$$\mathrm{H}(X, Y) \leq \mathrm{H}(X) + \mathrm{H}(Y).$$

5.1 Bounding sums of binomial coefficients

Finally, let us prove Lemma 7 which we recall was the key to the proof of Shannon's upper bound.

Lemma 7. *Let $n \in \mathbb{N}$ and $p \in (0, 1)$. Then*

$$\sum_{i=0}^{pn} \binom{n}{i} \leq 2^{H(p)n}.$$

Proof. Let M denote the sum on the LHS. We interpret M as the number of subsets of $[n]$ with $\leq \alpha n$ elements. Define a discrete random variable X as a uniformly random set with $\leq \alpha n$ elements. X has entropy

$$H(X) = \log(M)$$

by the maximality principle. We can identify X with the joint distribution $(Y_1, \ldots, Y_n)$, where Y_i indicates whether element i appears in X. Each Y_i (taken alone) is a Bernoulli variable with probability p. Then

$$H(X) \leq \sum_{i=1}^{n} H(X_i) = nH(p).$$

Thus $\log(M) = nH(p)$, as desired. ■

5.2 What is entropy, really?

We close with a quote from Shannon [5].

> My greatest concern was what to call it. I thought of calling it "information", but the word was overly used, so I decided to call it "uncertainty". When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage".

6 Exercises

Exercise 1. In Section 3, we discussed coding schemes for achieving low *average error rates*. Recall that an average error rate of δ means that the error is averaged over all $x \in \{0,1\}^m$:

$$\text{average error} = \mathop{\mathbf{E}}_{x}\left[\mathop{\mathbf{P}}_{\mathcal{N}}[\mathcal{D}(\mathcal{N}(C(x)))] \neq x\right] = \mathop{\mathbf{P}}_{x,\mathcal{N}}[\mathcal{D}(\mathcal{N}(C(x))) \neq x].$$

By contrast, a **uniform error** of δ means that for *every* $x \in \{0,1\}^m$, the probability of transmitting and failing to decode x is at most δ.

$$\text{uniform error} = \max_{x} \mathop{\mathbf{P}}_{\mathcal{N}}[\mathcal{D}(\mathcal{N}(C(x))) \neq x].$$

Prove Shannon's theorem (Theorem 6) for uniform error instead of average error.

Exercise 2. In Section 3, we develop *redundant* codes that are extremely efficient w/r/t their transmission rate. Another problem, moving in sort of the opposite direction, is *compression*.

Here we consider compression in the following model. Let Σ be a finite alphabet of n letters. Our goal is to efficiently assign bit strings (codes) to each letter in Σ so that messages, composed of sequences of letters in Σ, are as efficient as possible. More specifically, we are only allowed to use **prefix-free codes**, which are mappings

$$C : \Sigma \to \{0,1\}^*$$

assigning bit strings (of varying length) to letters such that no code $C(x)$ ($x \in \Sigma$) is a prefix to another code $C(y)$ ($y \in \Sigma$). Prefix codes are particularly easy to decode. As we scan the bits of an encoded message, as soon as we see a string that matches the code of a letter, we immediately decode that the scanned bits to the letter. We then continue to scan the rest of the bits as the beginning of the code of a new letter.

For example, the most straight forward prefix code would be to assign each letter in Σ a different bit string with $\lceil \log n \rceil$ bits.

Prefix codes can be identified with binary trees where each branch represents a 0 or 1 bits, and the leaves correspond to a letter where the root to leaf path gives the encoding of a letter.

We assume that not all letters in Σ are distributed frequently. (This is where we have some opportunity for compression) Let $p \in \Delta^{\Sigma}$ be a fixed distribution over Σ. For each letter in $x \in \Sigma$, p_x represents the average frequency of the letter x in these messages.

Given an encoding $C : \Sigma \to \{0,1\}^*$, the average number of bits per letter is

$$\sum_{x \in \Sigma} p_x |C(x)|,$$

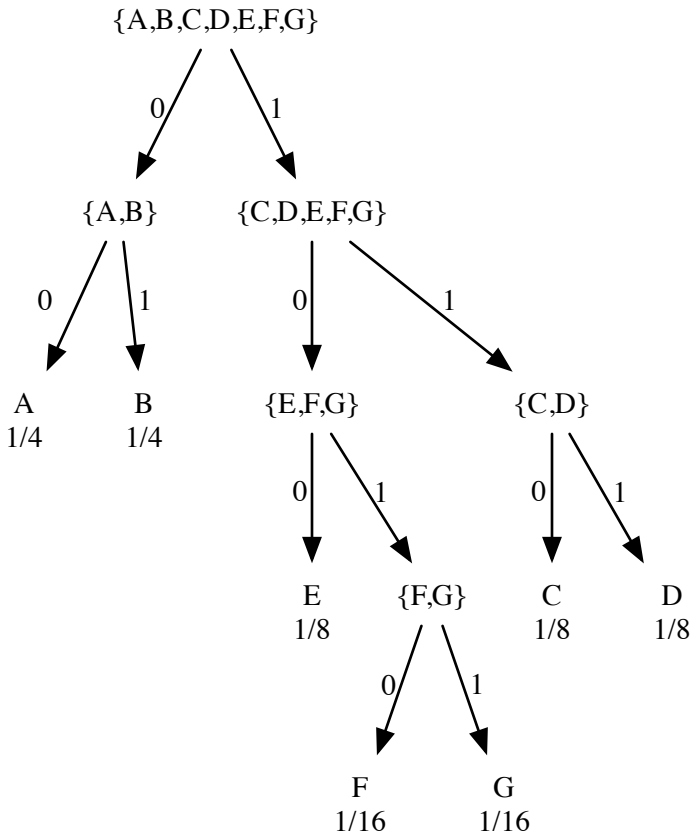
where $|C(x)|$ denotes the length of the bit string $C(x)$.

For this problem, consider a special case where every probability p_x is a power of 2, of the form $1/2^{i_x}$ for some integer $i_x \in \mathbb{N}$. Show that there exists a prefix code $C : \Sigma \to \{0,1\}^*$ where the average length is *exactly* $H(p)$, where $H(p)$ is the entropy of the random letter drawn from Σ in proportion to p:

$$H(p) = \sum_x p(x) \log\left(\frac{1}{p(x)}\right).$$

Here's a harder follow up question: can one do better than the entropy $H(p)$? Either prove that $H(p)$ is optimal, or give a counter example where the probabilities are powers of 2 and one can achieve better than $H(p)$ average bits per letter.

As an example, the following tree defines a prefix code over a distribution of 7 letters $\{A, B, C, D, E, F, G\}$ with probabilities $\{1/4, 1/4, 1/8, 1/8, 1/8, 1/16, 1/16\}$, respectively. One can see that the average length matches the entropy of the distribution.



References

[1] Claude E. Shannon. "A mathematical theory of communication". In: *Bell Syst. Tech. J.* 27.3 (1948), pp. 379–423.

[2] Claude E. Shannon. "Communication theory of secrecy systems". In: *Bell Syst. Tech. J.* 28.4 (1949), pp. 656–715.

[3] Claude E. Shannon. "The synthesis of two-terminal switching circuits". In: *Bell Syst. Tech. J.* 28.1 (1949), pp. 59–98.

[4] Stephen A. Cook. “The Complexity of Theorem-Proving Procedures”. In: *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing, May 3-5, 1971, Shaker Heights, Ohio, USA*. Ed. by Michael A. Harrison, Ranan B. Banerji, and Jeffrey D. Ullman. ACM, 1971, pp. 151–158.

[5] Myron Tribus and Edward C. McIrvine. “Energy and Information”. In: *Scientific American* 225 (1971), pp. 179–188.

[6] Leonid A. Levin. “Universal Sequential Search Problems”. In: *Problems of Information Transmission* 9.3 (1973).

[7] Michael Sipser. *Introduction to the theory of computation*. PWS Publishing Company, 1997.

[8] Sanjeev Arora and Boaz Barak. *Computational Complexity - A Modern Approach*. Cambridge University Press, 2009. URL: http://www.cambridge.org/catalogue/catalogue.asp?isbn=9780521424264.