Information Theory Notes

Alex Creiner

Contents

Part I
Shannon Information Theory

1 Entropy

1.1 Defining Entropy

Suppose we have an information source. This source produces **variety**; one out of many possible outcomes, with varying probabilities for each possibility. Each time the information source produces an outcome, we are receiving information. How could one measure the amount of information which the source produces per pulse of output? Clearly, the answer has to scale with the variety of the system, i.e. the number of possible outcomes. **The more possibilities there are, the more information can be produced per pulse**. This has a startling implication though - that to measure the rate of information transfer from a source is also to measure the amount of uncertainty that we have about what the outcome will be.

Simply counting the number of possible states is perfectly sufficient at measuring the uncertainty. That much is clear. But it should also be noted that any monotonic function of this number will also do. And there are a few big reasons why it makes more intuitive sense to use the logarithm of the number. Suppose we have two different information sources, which we are choosing to see as a single composite information source. Say the first source has m many possible states, and the second has variety n many. The number of possible states of this composite source is the product of the number of states, mn, and not the sum. However, if we take as our measure of uncertainty the *logarithm* of the total number of states, then by properties of logarithms we *can* simply add them together, since

$$\log(mn) = \log(m) + \log(m)$$

We will see other reasons which further justify this choice later on. We will typically use base 2 logarithms, so that the variety of a system is measured in bits. If there are 16 possible states of a system, then variety is 4 bits. Later on we will see that this is because conceivably I could ask that many yes or no questions to completely pin down what the outcome was. Equivalently, I could represent that state using a 4 bit string.

Things get trickier when some outcomes are likelier than others. For example, the letter e appears way more often than any other in the English language. This should *reduce* our uncertainty about the information produced, and therefore our measure should shrink. This establishes a principle: our measure of uncertainty should be maximal when all outcomes are equally likely.

Regardless of what we're talking about, the possible outcomes coming from the information source constitute what can be seen as the sample space of a probability experiment. For the sake of initial simplicity, we will assume a countable sample space. This sample space we will call an **alphabet**, and typically denote it $\mathcal{X}$. Random variables are typically seen as mappings from a sample space to a set of numbers. Since our sample spaces are countable, we can simply assume a bijection between the numbers and the outcomes themselves, and assume that a random variable ends up equaling the symbol of the alphabet directly.

With that said, let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = P(X = x)$, where $x \in \mathcal{X}$.

Definition 1.1. The **entropy** of the random variable X (as fixed directly above), denoted $H(X)$, is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x)) \tag{1}$$

Where the logarithm here is base 2. We are using the convention here that $0 \log(0) = 0$.

Immediately we see something quite odd about this definition, which is that the random variable X is being fed *back in* to it's own probability mass function. This means that entropy is only *indirectly* a function of the random variable X. Really, it is a function of the probability distribution itself $p(X)$.

The first thing we should note is that things greatly simplify as we said they should when all outcomes are equally likely. Suppose they are: that all probabilities p_i for the $x_i \in \mathcal{X}$ are the same, i.e. $\frac{1}{|\mathcal{X}|} = \frac{1}{n}$. Then

$$\begin{aligned} H(X) = -\sum_{i=1}^{n} \frac{1}{n} \log(\frac{1}{n}) &= -\frac{1}{n} \sum_{i=1}^{n} \log(1) - \log(n) \\ &= \frac{1}{n} \sum_{i=1}^{n} - \log(n) \\ &= \log(n) \end{aligned}$$

Thus at the very least this definition of entropy can be seen as an extension of our original measure of the amount of information per pulse of activity from an information source, i.e. the number of bits needed to encode the outcome. We will see later that the above definition is the only possible extension which maintains two other basic axioms about such a measure.

Note that the entropy X can be seen as an expected value, in particular that of $\log(\frac{1}{p(X)})$. I.e.

$$H(X) = E\left(\log\left(\frac{1}{p(X)}\right)\right) = -E(\log(p(X))) \tag{2}$$

Lemma 1.1. *$H(X) \geq 0$.*

Proof. $\log(\frac{1}{p(x)}) = -\log(p(x))$. Since probabilities are between 0 and 1, this logarithm is always non-positive, and thus the term itself is always non-negative. □

To define entropy using a logarithm base other than 2 changes the unit of measurement, but shouldn't actually change the measurement itself. For a number b, let $H_b(X)$ denote the entropy in which the logarithm of the definition is taken as base b.

Lemma 1.2. *For any possible bases a and b, $H_b(X) = (\log_b(a))H_a(X)$*

Proof. This follows from the basic formula for changing a logarithm base:

$$\log_a(x) = \frac{\log_b(x)}{\log_b(a)}$$

Multiplying both sides of this gives that $\log_b(x) = \log_a(x)\log_b(a)$. Substituting this for $\log_b(x)$ in the formula for H_b and pulling the term out of the sum gives the desired result. □

Suppose we are concerned with the outcome of a system consisting of two different information sources. For this, we must have a definition of the joint entropy of the overall system.

Definition 1.2. The **joint entropy** $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint probability mass function $p(x,y)$ is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x,y)) \tag{3}$$

Once again, if all outcomes from both sources are equally likely, and the number of possibilities for each are n and m respectively, then the total number of possibilities for the system is the product nm, and the joint entropy of the system becomes simply $\log(nm)$, as we would hope and expect.

Recall that for a pair of random variables, one can think about the conditional distribution of one variable Y given a fixed outcome of the other (say $X = x$):

$$p(y|x) = p(Y = y|X = x)$$

Fixing an x, we obtain a conditional random variable $Y|X = x$, which has it's own expected value different for each x. Accordingly, each has it's own entropy $H(Y|X = x)$.

Definition 1.3. The **conditional entropy of Y given X** is defined as

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \tag{4}$$

I.e. it is the average entropy of $Y|X=x$, weighted according to the probability of getting each particular x. Intuitively, it is our uncertainy

Note that

$$\begin{align} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \tag{5}\\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\log(p(y|x)) \tag{6}\\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p(y|x)\log(p(y|x)) \tag{7}\\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\log(p(y|x)) \tag{8}\\ &= -E(\log(p(Y|X)) \tag{9} \end{align}$$

where $\log(p(Y|X))$, the logarithm of the conditional probability of a particular Y given a particular X, is seen as a function of the joint random pair (X,Y), and the last step following from the law of the unconscious statistician.

Theorem 1.1 (Chain rule). *The joint entropy of a random pair (X,Y) is the entropy of 1 variable plus the conditional entropy of the second variable given the first. That is,*

$$H(X,Y) = H(X) + H(Y|X)$$

Proof.

$$\begin{align} H(X,Y) &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\log(p(x,y)) \tag{10}\\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)p(y|x)) \tag{11}\\ &= -\sum_{x\in\mathcal{X}} \sum_{\cancel{y\in\mathcal{Y}}} \overset{p(x)}{\cancel{p(x,y)}}\log(p(x)) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\log(p(y|x)) \tag{12}\\ &= -\sum_{x\in\mathcal{X}} p(x)\log(p(x)) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\log(p(y|x)) \tag{13}\\ &= H(X) + H(Y|X) \tag{14} \end{align}$$

□

The chain rule for entropy can be extended easily to any number of joint variables. Consider for instance a triplet (X,Y,Z). It can easily be shown in identical manner to the above that

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

So that

$$H(X,Y,Z) = H(X) + H(Y,Z|X) = H(X) + H(Y|X) + H(Z|X,Y) \tag{15}$$

And so on for however many variables.

1.2 Conditional Entropy and Mutual Information

Definition 1.4. The **relative entropy** or **Kullback-Leibler distance** between two probability mass functions $p(x)$ and $q(x)$ (defined over the same alphabet $\mathcal{X}$) is defined as

$$\begin{aligned} D(p||q) &= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right) \qquad (16) \\ &= E_p\left(\log\left(\frac{p(X)}{q(X)}\right)\right) \qquad (17) \end{aligned}$$

where E_p denotes that we are taking the expected value of a function of the random variable X under the assumption that the true distribution of X is p.

In other words, the relative entropy between the two distributions p and q is defined to be the expected error in assuming that the distribution is q given that it is actually p. It can in some sense be seen as a measure of the distance between two distributions, but it isn't actually a metric due to it's implicit bias towards one over the other (meaning that $D(p||q) \neq D(q||p)$) as well as the fact that it doesn't satisfy the triangle inequality. We will however see that it is non-negative and 0 iff $p = q$. Before that though, one more definition.

Definition 1.5. Let X and Y be random variables with joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The **mutual information between** X **and** Y, denoted $I(X; Y)$ is the relative entropy between the joint distribution $p(x, y)$ and the product of the marginal distributions $p(x)$ and $p(y)$. I.e.

$$\begin{aligned} I(X;Y) &= D(p(x,y)||p(x)p(y)) \qquad (18) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (19) \\ &= E_{p(X,Y)}\log\left(\frac{p(X,Y)}{p(X)p(Y)}\right) \qquad (20) \end{aligned}$$

The mutual information between X and Y is easily related to our notions of entropy:

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (21) \\ &= \sum_{x,y} p(x,y)\log\left(\frac{\cancel{p(y)}p(x|y)}{p(x)\cancel{p(y)}}\right) \qquad (22) \\ &= \sum_{x,y} p(x,y)\log(p(x|y)) - \sum_{x,y} p(x,y)\log(p(x)) \qquad (23) \\ &= -\sum_{x} p(x)\log(p(x)) - \left(-\sum_{x,y} p(x,y)\log(p(x|y))\right) \qquad (24) \\ &= H(X) - H(Y|X) \qquad (25) \end{aligned}$$

Note that we could have just as easily replaced $p(x, y)$ with $p(x)p(y|x)$ to obtain $H(Y) - H(X|Y)$. Thus

$$I(X;Y) = H(X) - H(Y|X) = H(Y) - H(X|Y) \qquad (26)$$

But remember by the chain rule that $H(X, Y) = H(Y) + H(X|Y) \implies H(X|Y) = H(X, Y) - H(Y)$, so that

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \qquad (27)$$

sparking immediate comparison with the classic addition rule for probabilities. Finally, note that $I(X; X) = H(X) - H(X|X)$. Since $H(X|X = x) = 0$ clearly, $H(X|X) = 0$, so that $I(X; X) = H(X)$. The mutual information between a source of information and itself is just the entropy, as one would hope. For this reason, entropy is sometimes referred to as **self-information**. We summarize these identites as a theorem:

Theorem 1.2 (Mutual information and entropy)**.**

We have the following:

(1) $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

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

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

(4) $I(X;X) = H(X)$

We already have a chain rule for entropy. Towards deriving similar chain rules for mutual information and relative entropy, we define conditional versions of these as we did for entropy.

Definition 1.6. The **conditional mutual information** of the random variables X and Y given Z is defined by

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{28}$$

$$= E_{p(x,y,z)} \log \left(\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \right) \tag{29}$$

Definition 1.7. For joint probability mass functions $p(x,y)$ and $q(x,y)$, the **conditional relative entropy** $D(p(y|x)||q(y|x))$ is the expected error assuming that the conditional distribution of Y given X is $q(x|y)$ when it is actually $p(x|y)$. More precisely:

$$D(p(y|x)||q(y|x)) = \sum_x p(x) \sum_y \log \left(\frac{p(y|x)}{q(y|x)} \right) \tag{30}$$

$$= E_{p(x,y)} \log \left(\frac{p(Y|X)}{q(Y|X)} \right) \tag{31}$$

Theorem 1.3 (Chain rules for entropy, mutual information, and relative entropy)**.** *For entropy, we have:*

$$H(X_1, X_2, X_3, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \ldots + H(X_n|H_{n-1}, \ldots, H_1) \tag{32}$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \tag{33}$$

For mutual information, we have

$$I(X_1, X_2, \ldots, X_n; Y) = I(X_1; Y) + I(X_2; Y|X_1) + \ldots + I(X_n; Y|X_{n-1}, \ldots, X_1) \tag{34}$$

$$= \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1) \tag{35}$$

And for relative entropy, only the base case is necessary. We have

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x), q(y|x)) \tag{36}$$

Proof. The entropy formula was already demonstrated earlier. As for mutual information, note that

$$I(X_1, X_2; Y) = H(X_1, X_2) - H(X_1, X_2|Y) \tag{37}$$

$$= H(X_1) + H(X_2|X_1) - (H(X_1|Y) + H(X_2|X_1, Y)) \tag{38}$$

$$= H(X_1) - H(X_1|Y) + H(X_2|X_1) - H(X_2|X_1, Y) \tag{39}$$

$$= I(X_1; Y) + I(X_2; Y|X_1) \tag{40}$$

The case for general n is identical using the general chain rule formula for entropy, just messier. Finally for conditional relative entropy,

$$D(p(x,y)||q(x,y)) = \sum_x \sum_y p(x,y) \log \left(\frac{p(x,y)}{q(x,y)} \right) \tag{41}$$

$$= \sum_x \sum_y p(x,y) \log \left(\frac{p(x)p(y|x)}{q(x)q(y|x)} \right) \tag{42}$$

$$= \sum_x \sum_y p(x,y) \log \left[\left(\frac{p(x)}{p(y)} \right) \left(\frac{p(y|x)}{q(y|x)} \right) \right] \tag{43}$$

$$= \sum_x \sum_y p(x,y) \log \left(\frac{p(x)}{q(x)} \right) + \sum_x \sum_y p(x,y) \log \left(\frac{p(y|x)}{q(y|x)} \right) \tag{44}$$

$$= D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \tag{45}$$

□

1.3 General Theory of Entropy

To derive some general relationships between these values, it will be helpful to recall some basic ideas regarding convexity. A function $f(x)$ is convex over an interval (a,b) if for every $x_1, x_2 \in (a,b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

and is strictly convex if equality holds only when $\lambda = 0$ or 1. Visually, convex functions lie underneath any chord (i.e. line connecting two points on the curve), and in terms of calculus, a differentiable function is convex over some interval if it has a non-negative second derivative over it (and strictly convex if it has a positive second derivative). A function f is concave if $-f$ is convex, and convex if $-f$ is concave.

Suppose p_1 and p_2 are numbers between 0 and 1 which sum to 1, and that f is a convex function on the relevant interval. Then $p_1 = \lambda$, $p_2 = 1 - \lambda$, and so

$$f(p_1 x_1 + p_2 x_2) \leq p_1 f(x_1) + p_2 f(x_2)$$

Fair enough, but now suppose we have some k many p_i's $p_1, p_2, \ldots, p_k$ between 0 and 1 which sum to 1. Then if we define $p_i' = \frac{p_i}{1-p_k}$, we have a collection of $k-1$ many p_i''s which sum to 1. So by the inductive hypothesis and the definition of convexity with $\lambda = p_k$ we have

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(x_i) &= p_k f(k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i) \\
&\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i \right) \\
&\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p_i' x_i \right) \\
&= f\left(\sum_{i=1}^{k} p_i x_i \right)
\end{aligned}$$

This is a very handy conception of convexity when dealing with probability theory. It also can be used to easily demonstrate the following:

Theorem 1.4 (Jensen's Inequality). *If f is a convex function and X is a random variable, then*

$$E(f(X)) \geq f(E(X)) \tag{46}$$

Moreover, we have equality iff X is constant.

Our first use of Jensen's inequality is the following:

Theorem 1.5 (Relative entropy is a kinda-sorta metric)**.** *For any pair of probability mass functions p and q over the same alphabet $\mathcal{X}$, we have*

$$D(p||q) \geq 0 \tag{47}$$

with equality iff $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Proof. It is worth first observing that $f(x) = \log(x)$ is a concave function. So $-\log(x)$ is convex, and from the above this means if we have a set of non-zero probabilities p_i summing to 1 (as we do with $p(x)$ over it's support $A = \{x : p(x) > 0\}$), we have

$$-\log\left(\sum_{i=1}^{n} p_i x_i\right) \leq -\sum_{i=1}^{n} p_i \log(x_i)$$

Therefore

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

In the context of this proof, the x_i will be the ratios $\frac{q_i}{p_i}$. Observe:

$$\begin{aligned}
-D(p||q) &= -\sum_{x\in A} p(x) \log\left(\frac{p(x)}{q(x)}\right) && (48)\\
&= -\sum_{x\in A} p(x)(\log(p(x)) - \log(q(x))) && (49)\\
&= \sum_{x\in A} p(x)(\log(q(x)) - \log(p(x))) && (50)\\
&= \sum_{x\in A} p(x) \log\left(\frac{q(x)}{p(x)}\right) && (51)\\
&\leq \log\left(\sum_{i\in A} p(x)\frac{q(x)}{p(x)}\right) && (52)\\
&= \log\left(\sum_{i=1}^{n} q(x)\right) && (53)\\
&= \log(1) = 0 && (54)
\end{aligned}$$

Thus we have $-D(p||q) \leq 0$, and so $D(p||q) \geq 0$. Moreover, $\log(x)$ is actually *strictly* concave, and so we have equality iff $\frac{q(X)}{p(X)}$ is not actually a random variable but rather constant, i.e. there exists a c such that $\frac{q(x)}{p(x)} = c$ for all x. This means that $q(x) = cp(x)$, so that

$$c = c\sum_{x\in A} p(x) = \sum_{x\in A} cp(x) = \sum_{x\in A} q(x) = 1$$

Thus we have equality iff $p(x) = q(x)$ for all x. □

This expected result required some technical finesse to prove, but using it we can demonstrate many other expected results.

Corollary 1.1. *For any two random variables X, Y,*

$$I(X;Y) \geq 0 \tag{55}$$

with equality iff X and Y are independent.

Proof. $I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$, so that follows immediately. Also we have equality iff $p(x,y) = p(x)p(y)$ for all x, y, i.e. iff X and Y are independent. □

Identically we also have the following inequalities for conditional mutual information and conditional relative entropy by their definitions:

Corollary 1.2.

$$D(p(y|x)||q(y|x)) \geq 0 \tag{56}$$

with equality iff $p(y|x) = q(y|x)$ *for all* y *and* x *such that* $p(x) > 0$*, and*

$$I(X;Y|Z) \geq 0 \tag{57}$$

with equality iff X *and* Y *are conditionally independent given* Z*.*

Finally, we can show something extremely important: that entropy is maximal precisely when all outcomes are equally likely.

Theorem 1.6 (Non-uniformity only reduces uncertainty)**.** $H(X) \leq \log(|\mathcal{X}|)$*, with equality iff* X *has a uniform distribution over* $\mathcal{X}$*.*

Proof. Let $|\mathcal{X}| = n$, and let $u(x) = \frac{1}{n}$ denote the uniform distribution over $\mathcal{X}$. Let $p(x)$ denote the actual probability mass function for X. Then

$$D(p||u) = \sum_x p(x) \log\left(\frac{p(x)}{u(x)}\right) \tag{58}$$

$$= \sum_x p(x)(\log(p(x)) - \cancelto{0}{\log(1)} + \log(n)) \tag{59}$$

$$= \sum_x p(x) \log(n) + \sum_x p(x) \log(p(x)) \tag{60}$$

$$= \log(n) \cancelto{1}{\sum_x p(x)} - H(X) \tag{61}$$

By non-negativity of relative entropy,

$$0 \leq D(p||u) = \log(n) - H(X) \tag{62}$$

$$\implies H(X) \leq \log(n) \tag{63}$$

Moreover since $D(p||u) = 0$ iff $p = u$, it follows that entropy only meets this upper bound when p is the uniform distribution. □

This also leaves us with a very instructive new equation for entropy:

$$H(p) = \log(|\mathcal{X}|) - D\left(p||\frac{1}{|\mathcal{X}|}\right) \tag{64}$$

In other words, our definition of entropy is exactly the simpler logarithmic measure of variety that we started with modified by the relative difference between the uniform distribution and the actual probability distribution. When one substitutes that for entropy, the substitution is precisely as accurate as the correctness of the statement "all states are equally likely".

Theorem 1.7 (Conditioning reduces entropy)**.** *For any random variables* X *and* Y*,*

$$H(X|Y) \leq H(X) \tag{65}$$

with equality iff X *and* Y *are independent.*

Proof. $0 \leq I(X;Y) = H(X) - H(X|Y)$, meaning that $H(X|Y) \leq H(X)$. As noted, equality is only achieved when X and Y are independent. □

This further solidifies that we have the correct definition of entropy, because it shows that gaining information can only reduce it. The idea is that a reduction of uncertainty corresponds to an acquisition of information, and this is exactly what we just said.

Theorem 1.8. *Let $X_1, X_2, \ldots, X_n$ be identically distributed random variables. Then*

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i) \tag{66}$$

with equality iff all of the X_i are independent.

Proof. By the chain rule for entropy,

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) & (67) \\ &\leq \sum_{i=1}^{n} H(X_i) & (68) \end{aligned}$$

As noted above, equality is only achieved when the X_i's are all independent. □

This settles the relevant inequalities. To inspect the concavity of our various tools, we need one more inequality pertaining to logarithms which follows from convexity.

Lemma 1.3 (Log-sum Inequality)**.** *For nonnegative numbers $a_!, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, we have*

$$\sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{n} a_i\right) \log\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1^n} b_i}\right) \tag{69}$$

using the convention that $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ for $a > 0$ and $0 \log \frac{0}{0} = 0$. (These are what continuity would imply anyway). Moreover we have equality iff $\frac{a_i}{b_i}$ is constant for each i.

At first glance this is a very confusing inequality to stare at. The best way to understand what it actually gives you is to look at it from right to left, and note that we are going from three different summations to just one. That's what this inequality really says - that if you're willing to get something bigger, you can push all of the summing of terms to the end of a calculation.

Proof. If all of the a's and b's are 0 then both sides of this inequality are 0, and so the statement holds. If all of the b's are 0 and at least one of the a's isn't, then the lefthand side is ∞ and so is the rigthand side, so the statement holds once more. Thus without loss of generality assume that all of the a_i's and b_i's are nonzero.

Let $\alpha_i = \frac{b_i}{\sum_{j=1}^{n} b_j}$ and $t_i = \frac{a_i}{b_i}$ for each i. Note that the α_i are all positive and sum to 1. Consider the function $f(t) = t\log(t)$. The second derivative of this function is $\frac{1}{t}\log(e)$, which is positive for all positive reals, and therefore $f(t)$ is convex. We therefore have

$$f\left(\sum_i \alpha_i t_i\right) \geq \sum_i \alpha_i f(t_i)$$

Flipping this around and plugging in the a_i's and b_i's gives us what we want:

$$\sum_i \frac{\cancel{b_i}}{\sum_j b_j}\frac{a_i}{\cancel{b_i}} \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_i \frac{\cancel{b_i}}{\sum_j b_j}\frac{a_i}{b_i}\right) \log\left(\sum_i \frac{\cancel{b_i}}{\sum_j b_j}\frac{a_i}{\cancel{b_j}}\right) \tag{70}$$

$$\implies \left(\cancel{\frac{1}{\sum_j b_j}}\right)\sum_i a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\cancel{\frac{1}{\sum_j b_j}}\right)\left(\sum_i a_i\right)\log\left(\frac{\sum_i a_i}{\sum_j b_j}\right) \tag{71}$$

$$\implies \sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{n} a_i\right)\log\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1^n} b_i}\right) \tag{72}$$

□

Theorem 1.9 (Convexity of relative entropy). *$D(p||q)$ is a convex function of the pair (p, q); that is to say if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then*

$$D[\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2] \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2) \tag{73}$$

for all $\lambda \in (0, 1)$.

Proof. For each $x \in \mathcal{X}$, note that

$$(\lambda p_1(x) + (1-\lambda)p_2(x))\log\left(\frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)}\right) \tag{74}$$

$$\leq \lambda p_1(x)\log\left(\frac{\lambda p_1(x)}{\lambda q_1(x)}\right) + (1-\lambda)p_2(x)\log\left(\frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}\right) \tag{75}$$

But $D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2)$ is the summation of all of these terms for each x. Thus

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \tag{76}$$

$$= \sum_{x\in\mathcal{X}}(\lambda p_1(x) + (1-\lambda)p_2(x))\log\left(\frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)}\right) \tag{77}$$

$$\leq \sum_{x\in\mathcal{X}}\lambda p_1(x)\log\left(\frac{\cancel{\lambda} p_1(x)}{\cancel{\lambda} q_1(x)}\right) + \sum_{x\in\mathcal{X}}(1-\lambda)p_2(x)\log\left(\frac{\cancel{(1-\lambda)}p_2(x)}{\cancel{(1-\lambda)}q_2(x)}\right) \tag{78}$$

$$= \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2) \tag{79}$$

□

As always, we are working with relative entropy because we actually want to show results about the other stuff:

Theorem 1.10 (Concavity of entropy). *$H(p)$ is a concave function of the probability distribution p. That is to say, if $\lambda \in (0, 1)$ and p_1 and p_2 are two different distributions over the same alphabet $\mathcal{X}$, then*

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2) \tag{80}$$

Proof. Recall that

$$H(p) = \log(|\mathcal{X}|) - D\left(p||\frac{1}{|\mathcal{X}|}\right)$$

Now $\log(x)$ is a concave function, $-D(p||\frac{1}{|\mathcal{X}|})$ is as well. Thus $H(p)$ is the sum of two concave functions, itself concave. □

This isn't the only way to prove that entropy is a concave function of the distribution, and the second way is instructive. Let X_1 be a random variable with distribution p_1, taking on values in a set $\mathcal{X}$, and let X_2 be another random variable with distribution p_2 on the same set. Let

$$\theta = \begin{cases} 1 & \text{with probability } \lambda \\ 2 & \text{with probability } 1-\lambda \end{cases} \tag{81}$$

Let $Z = X_\theta$, i.e. we select one of the random variables at random and then sample from it. Note that the distribution of Z is $r(x) = \lambda p_1(x) + (1-\lambda)p_2(x)$. Now we know that conditioning only reduces entropy, i.e.

$$H(Z) \geq H(Z|\theta)$$

But then by definition of conditional entropy,

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2) \tag{82}$$

giving us concavity. One consequence of this fact is that mixing together two gases of equal entropy results in a gas with higher entropy, as we would expect.

The book I'm currently reading out of has a corresponding result for mutual information but it looks annoying and not very useful so I'm not going to write it out until I see it used somewhere.

Definition 1.8. Random variables X, Y, Z are said to form a **Markov chain** in that order (denoted $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y, Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function can be written

$$p(x, y, z) = p(x)p(y|x)p(z|y) \tag{83}$$

Note that $X \to Y \to Z$ iff X and Z are conditionally independent given Y. This is because

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = \left(\frac{p(x, y)}{p(y)}\right) p(z|y) = p(x|y)p(z|y) \tag{84}$$

Also note that $X \to Y \to Z$ iff $Z \to Y \to X$. This is because $X \to Y \to Z$ implies

$$p(x, y, z) = p(z)p(y|z)p(x|z, y) = p(z)p(y|z)p(x|y) \tag{85}$$

Finally note that if Z is any function of Y, i.e. $Z = f(Y)$, then $X \to Y \to Z$ is clearly a Markov chain for any X.

Theorem 1.11 (Data-processing inequality)**.** *If $X \to Y \to Z$ then $I(X;Y) \geq I(X;Z)$. In other words, there is always more (at least as much) information shared between random variables nearer to each other in a Markov process.*

Proof. By the chain rule for mutual information, we can expand the mutual information of X with (Y, Z) in two different ways

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \quad (86)\\ &= I(X;Y) + I(X;Z|Y) \quad (87) \end{aligned}$$

Since X and Z are conditionally independent given Y, we have that $I(X;Z|Y) = 0$. Thus

$$I(X;Z) + I(X;Y|Z) = I(X;Y) \tag{88}$$

Since $I(X;Y|Z) \geq 0$, it follows that $I(X;Z) \geq I(X;Y)$. □

Corollary 1.3. *In particular if $Z = g(Y)$, we have $I(X;Y) \geq I(X;g(Y))$.*

Thus functions of the data (i.e. doing fancy tricks with it) cannot increase the information given about X. Quite the opposite, in fact; it can only destroy information one may have had.

Corollary 1.4. *If $X \to Y \to Z$ then $I(X;Y|Z) \leq I(X;Y)$.*

Thus, the dependence of X and Y is decreased by the observation of a 'downstream' random variable Z.

1.4 Fano's Inequality

To following is necessary for proving the noisy channel coding theorem later on. The idea is that we have a random variable X, which upon being fed through a communication channel produces a new random variable Y based on the imperfections of communication. The receiver of the information wants to recover the value of X and applies some function to it, in order to get a third variable $\hat{X} = g(Y)$, in order to estimate X. We therefore call $\hat{X}$ an **estimator** for X, and it is clear that $X \to Y \to \hat{X}$ forms a Markov chain. Define the **probability of error** to be $P_e = P(\hat{X} \neq X)$.

Theorem 1.12 (Fano's Inequality)**.** *For an estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, we have*

$$H(P_e) + P_e \log(|\mathcal{X}|) \geq H(X|\hat{X}) \geq H(X|Y) \tag{89}$$

where $H(P_e)$ denotes the entropy of a Bernoulli random variable with probability of success P_e.

Notably, this theorem's claim can be weakened. Since $H(P_e) \leq 1$, we also have

$$H(X|Y) \leq 1 + P_e \log(|\mathcal{X}|) \tag{90}$$

so that

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}|)} \tag{91}$$

Proof. Define the error random variable

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X \end{cases} \tag{92}$$

By the chain rule for entropy we have

$$H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) \tag{93}$$

as well as

$$H(E, X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}) \tag{94}$$

of these terms, we know that $H(E|X, \hat{X})$ is the uncertainty of an error given both the input and the decoded output, which is of course 0. By the fact that conditioning reduces uncertainty, $H(E|\hat{X}) \leq H(E) = H(P_e)$. Also

$$H(X|E, \hat{X}) = H(X|E = 0, \hat{X})P(E = 0) + H(X|E = 1, \hat{X})P(E = 1) \tag{95}$$

of these terms, $H(X|E = 0, \hat{X})$ is the uncertainty in X given we have our estimator and are sure it is correct, i.e. 0. $P(E = 1) = P_e$. As for $H(X|E = 1, \hat{X})$, we are discussing the uncertainty in X given we know there was an error and that our estimator is wrong. This rules out one of the $|\mathcal{X}|$ possibilities, leaving $|\mathcal{X}| - 1$ many. Thus $H(X|E = 1, \hat{X}) \leq \log(|\mathcal{X} - 1|)$. Thus we have

$$\begin{aligned} H(X|E, \hat{X}) &\leq H(P_e) + P_e \log(|\mathcal{X} - 1|) \qquad (96) \\ &\leq H(P_e) \log(\mathcal{X}|) \qquad (97) \end{aligned}$$

But $H(X|E, \hat{X}) = H(X|\hat{X})$ by our first application of the chain rule. Thus

$$H(X|\hat{X}) \leq H(P_e) + P_e \log(|\mathcal{X}|) \tag{98}$$

That $H(X|\hat{X}) \geq H(X|Y)$ follows directly from the data-processing inequality. □

2 The Asymptotic Equipartition Property

Recall the basics of convergence of sequences of random variables. A sequence of random variables $X_1, X_2, \ldots$ converges to the random variable X in probability if for every $\epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$. The sequence converges in mean square if $(E(X_n - X))^2 \to 0$. And the sequence converges with probability 1 (or almost surely) if $P(\lim_{n\to\infty} X_n = X) = 1$.

Theorem 2.1 (Asymptotic Equipartition Property (AEP))**.** *If $X_1, X_2, \ldots \overset{iid}{\sim} p(x)$ then*

$$-\frac{1}{n}\log(p(X_1, X_2, \ldots, X_n)) \xrightarrow{p} H(X) \tag{99}$$

where $p(x_1, \ldots, x_n)$ is the joint pmf.

Proof. This is really just the weak law of large numbers. The average converges in probability to the mean, as we know. Moreover single variable functions of independent random variables are still independent random variables. Thus

$$-\frac{1}{n}\log(p(X_1, X_2, \ldots, X_n)) = -\frac{1}{n}\log(p(X_1)p(X_2)\ldots p(X_n)) \tag{100}$$

$$= -\sum \frac{\log(p(X_i))}{n} \tag{101}$$

$$\xrightarrow{p} -E(\log(p(X_i)) = H(X) \tag{102}$$

□

Definition 2.1. A **typical set** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property that

$$\frac{1}{2^{n(H(X)+\epsilon)}} \leq p(x_1, x_2, \ldots, x_n) \leq \frac{1}{2^{n(H(X)-\epsilon)}} \tag{103}$$

Theorem 2.2. *(1) If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$ then*

$$H(X) - \epsilon \leq -\frac{1}{n}\log(p(x_1, x_2, \ldots, x_n) \leq H(X) - \epsilon$$

(2) $P(A_\epsilon^{(n)} > 1 - \epsilon$ for sufficiently large n

(3) $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$ (bars denote cardinality)

(4) $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ for sufficiently large n

Proof. For 1, simply invert all three sides of the inequality defining membership in the set $A_\epsilon^{(n)}$ (remembering to reverse the inequality), take the log of all three sides (remembering not to reverse the inequality), and then divide all three sides by n. For 2, note that for any $\epsilon > 0$

$$P(|-\frac{1}{n}\log(p(X_1, \ldots, X_n)) - H(X)| \leq \epsilon) = P(\frac{1}{2^{n(H(X)+\epsilon)}} \leq p(X_1, X_2, \ldots, X_n) \leq \frac{1}{2^{n(H(X)-\epsilon)}} \tag{104}$$

$$= P((X_1, \ldots, X_n) \in A_\epsilon^{(n)}) = P(A_\epsilon^{(n)}) \tag{105}$$

$$\tag{106}$$

By AEP, the probability of being in the complement of this set goes to 0 for sufficiently large n, and therefore there must be a sufficient large n that this probability exceeds $1 - \epsilon$. For 3, note by definition of $A_\epsilon^{(n)}$ that

$$1 = \sum_{\boldsymbol{x} \in \mathcal{X}^n} p(\boldsymbol{x}) \geq \sum_{\boldsymbol{x} \in A_\epsilon^{(n)}} p(\boldsymbol{x}n) \geq \sum_{\boldsymbol{x} \in A_\epsilon^{(n)}} \frac{1}{2^{n(H(X)+\epsilon)}} = \frac{|A_\epsilon^{(n)}|}{2^{n(H(X)+\epsilon)}} \tag{107}$$

so that $|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$. Finally for 4, as we know already that $P(A_\epsilon^{(n)}) > 1-\epsilon$ for sufficiently large n, and by definition of $A_\epsilon^{(n)}$ we have that

$$1-\epsilon < P(A_\epsilon^{(n)} \le \sum_{\boldsymbol{x} \in A_\epsilon^{(n)}} \frac{1}{2^{n(H(X)-\epsilon)}} = \frac{|A_\epsilon^{(n)}|}{2^{n(H(X)-\epsilon)}} \tag{108}$$

so that $|A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X)-\epsilon)}$. □

The AEP has big implications for data compression. Let $\boldsymbol{X}$ be a sequence of n iid random variables with probability mass function $p(x)$. We wish to find short descriptions for such sequences of random variables. Fixing an $\epsilon > 0$, we can divide $\mathcal{X}^n$ into two sets: the typical set $A_\epsilon^{(n)}$, and it's complement. Consider a naive attempt at coding sequences in $\mathcal{X}^n$. To use bit strings we would need $\log(|\mathcal{X}|$ many bits, and so $n\log(|\mathcal{X}|)$ for sequences of n of them. However we now have an upper bound on the number of elements of the typical set: $2^{n(H(X)+\epsilon)}$, meaning coding elements of this set would only require $n(H(X)+\epsilon)$ many bits, which is smaller than $n\log(|\mathcal{X}|)$ for sufficiently small epsilon (or at the very least, when all outcomes are equally likely, no larger).

Consider the following simple coding scheme: first, a single bit 0 or 1 to mark whether or not the sequence is in $A_\epsilon^{(n)}$ or out of it. Next, if it is in the typical set, use $n(H(X)+\epsilon)$ many bits to code it under some fixed ordering, and if it isn't, just use the extra however many bits are needed. Let $l(\boldsymbol{x})$ denote the length of the code for $\boldsymbol{x}$. Then $l(\boldsymbol{X})$ is a random variable with an expected value, i.e. an expected length of the code for the sequence received from the information source. Let's see what it is for a fixed ϵ and n:

$$\begin{aligned} E(l(\boldsymbol{X})) = \sum_{\boldsymbol{x}} p(\boldsymbol{x})l(\boldsymbol{x}) &= \sum_{\boldsymbol{x} \in A_\epsilon^{(n)}} p(\boldsymbol{x})l(\boldsymbol{x}) + \sum_{\boldsymbol{x} \notin A_\epsilon^{(n)}} p(\boldsymbol{x})l(\boldsymbol{x}) & (109)\\ &\le \sum_{\boldsymbol{x} \in A_\epsilon^{(n)}} p(\boldsymbol{x})(n(H(X)+\epsilon)+1) + \sum_{\boldsymbol{x} \notin A_\epsilon^{(n)}} p(\boldsymbol{x})(n\log(|\mathcal{X}|)+1) & (110)\\ &= P(\boldsymbol{x} \in A_\epsilon^{(n)})(n(H(X)+\epsilon)+1) + P(\boldsymbol{x} \notin A_\epsilon^{(n)})(n\log(|\mathcal{X}|)+1) & (111)\\ &\le n(H(X)+\epsilon)+1+\epsilon(n\log(|\mathcal{X}|)+1) \to n(H(X)+\epsilon)+1 & (112) \end{aligned}$$

where the final step is using the simple fact that $P(A_\epsilon^{(n)}) \le 1$ as well as the fact that $P(x \notin A_\epsilon^{(n)}) \le \epsilon$. Dividing both sides of this by n gives us

$$E\left(\frac{1}{n}l(\boldsymbol{X})\right) = H(X) + \epsilon + \frac{1}{n} \le H(X) + \epsilon \tag{113}$$

Thus it appears that this very simple coding scheme has brought down the average length of a code word from $\log(\mathcal{X})$ to something very near $H(X)$. We therefore have the following theorem:

Theorem 2.3. *Let $X_1, X_2, \dots, X_n$ be an iid collection of random variables with pmf $p(x)$, and $\epsilon > 0$. Then there exists a coding for elements over the shared alphabet $\mathcal{X}$ (i.e. a 1-1 mapping from $\mathcal{X}$ into bit-strings) such that*

$$E\left(\frac{1}{n}l(\boldsymbol{X})\right) \le H(X) + \epsilon \tag{114}$$

Thus we can represent sequences $\mathcal{X}^n$ using $nH(X)$ bits on average.

To conclude, it is clear to see that the typical sets are fairly small sets which contain most of the probability. But are they the smallest such sets? The answer is yes, and we proceed to show this.

For each $n = 1, 2, \dots, n$, let $B_\delta^{(n)} \subseteq \mathcal{X}^n$ be the smallest set such that

$$P(B_\delta^{(n)}) \ge 1 - \delta \tag{115}$$

Theorem 2.4. *Let $X_1, X_2, \dots, X_n$ be iid with pmf $p(x)$. For $\delta < \frac{1}{2}$ and any $\delta' > 0$, if $P(B_\delta^{(n)}) > 1 - \delta$ then for sufficiently large n we have*

$$\frac{1}{n}\log(|B_\delta^{(n)}|) > H - \delta' \tag{116}$$

Note this means that $B_\delta^{(n)}$ must have at least 2^{nH} elements. But $A_\epsilon^{(n)}$ has $2^{n(H\pm\epsilon)}$ elements. Thus it follows that $A_\epsilon^{(n)}$ has about the same number of elements as the smallest high-probability set.

Proof. need to do □

That statement, that one set has 'about the same number of elements' as another, deserves some elaboration. Define the notation $a_n \doteq b_n$ to mean

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\frac{a_n}{b_n}\right) = 0 \tag{117}$$

Need to finish this

3 Entropy Rates of Stochastic Processes

A **stochastic process** $\{X_i\}_{i\in I}$ is an indexed sequence of random variables. (Typically I is either the natural numbers or the positive reals, i.e. time elapsing. We will assume that I is the naturals in this section.) We will assume that all X_i in such a process share an alphabet $\mathcal{X}$. The process is characterized by the join probability mass functions $p(x_1, x_2, \ldots, x_n)$. A stochastic process is **stationary** if shifting the index on all variables in an initial segment simultaneously doesn't change anything. More formally, we mean that for all $n, l \in \omega$:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_{1+l}, X_2 = x_{2+l}, \ldots, X_n = x_{n+l}) \tag{118}$$

A special and simple case of a stochastic process is the **Markov chain** we discussed earlier. To extend it past three variables, a Markov chain (or a Markov process) is a stochastic process in which each random variable in the sequence is independent of all others except the one immediately preceding it. That is to say

$$P(X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1}|X_n = x_n) \tag{119}$$

A Markov process is **time invariant** if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n. That is, for $n = 1, 2, \ldots$ and $a, b \in \mathcal{X}$, we have

$$P(X_{n+1} = b|X_n = a) = P(X_2 = b|X_1 = a)$$

If $\{X_i\}$ is a Markov chain, then X_n is called the **state** at time n. A time invariant Markov chain is characterized by an initial state and a **probability transition matrix**

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & & & \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}$$

where p_{ij} is the probability of entering state j given the current state is i, and $1, 2, \ldots, m$ index the possible symbols of the alphabet $\mathcal{X}$. A Markov chain is said to be **irreducible** if it there is always a nonzero probability of going from any state to any other state in a finite number of steps. It is possible for certain states of a Markov chain to eventually return back to themselves after a certain number of steps. Say that a state i has **period** k if any such loop can only occur in a number of steps which is a multiple of k. Otherwise it is **aperiodic**. A Markov chain itself is referred to as periodic or aperiodic if all of it's states have that property.

By the law of total probability, we have the following distribution on the states at time n:

$$p(X_{n+1} = j) = \sum_{i=1}^{n} P(X_n = i)P(X_{n+1} = j|X_n = i) = \sum_{i=1}^{n} P(X_n = i)p_{ij}$$

Call a distribution on the states a **stationary distribution** if $p(X_{n+2} = j) = P(X_{n+1} = 1)$, and so on. Clearly not all Markov processes are stationary, and we need to be a bit careful. First of all, we should be clear that one transition matrix defines *multiple* Markov processes, one for every different initial state and indeed every initial probability distribution. One transition matrix for a Markov chain could easily have *multiple* stationary distributions. It will be shown that the distribution of X_n always *tends towards* a stationary distribution as $n \to \infty$ when there is one. It can be shown that when a finite-state Markov process is aperiodic and irreducible, then it has a unique stationary distribution. Finally, if the initial probability distribution of a system is this stationary distribution, then the distribution of states will always be the same from that point on. As a result, the initial state of a stationary Markov process is drawn according to a stationary distribution, then the entropy will be constant:

$$H(X_1) = H(X_2) = \ldots$$

An example will be helpful. Consider the two-state Markov chain given by the transition matrix

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

where $\alpha, \beta \in (0,1)$. Note that the rows of this matrix must sum to one; this is a general property of transition matrices. Suppose our initial state is $X_1 = 1$. It is a bit silly, but we can say that there is an initial probability distribution of $p(x_1)$ where $P(X_1 = 1) = 1$ and $P(X_1 = 2) = 0$, which we can represent with the vector

$$\vec{x}_1 = \begin{pmatrix} x_1^1 \\ x_1^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

What is the probability distribution $p(x_2)$? Clearly it is the vector

$$\vec{x}_2 = \begin{pmatrix} 1-\alpha \\ \alpha \end{pmatrix}$$

But note that the top entry is

$$\begin{aligned} x_2^1 = P(X_2 = 1) &= P(X_2 = 1|X_1 = 1)P(X_1 = 1) + P(X_2 = 1|X_1 = 2)P(X_1 = 2) && (120) \\ &= p_{11}x_1^1 + p_{12}x_1^2 && (121) \end{aligned}$$

which is the top entry of $P\vec{x}_1$. Thus in general we have

$$P\vec{x}_{n+1} = P\vec{x}_n$$

Suppose this matrix had a stationary distribution. Call it $\vec{\mu}$. By definition then, we have that $P\vec{\mu} = \vec{\mu}$, i.e. $\vec{\mu}$ is an eigenvector of the transition matrix with eigenvalue 1. Such a μ does exist, and can be easily solved for to find

$$\mu_1 = \frac{\beta}{\alpha+\beta} \qquad\qquad \mu_2 = \frac{\alpha}{\alpha+\beta}$$

Definition 3.1. The **entropy rate** of a stochastic process $\{X_i\}$ is defined by

$$\begin{aligned} H(\mathcal{X}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) && (122) \\ &= \lim_{n\to\infty} \frac{1}{n} \left[H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, X_2, \ldots, X_{n-1}) \right] && (123) \end{aligned}$$

Note that this limit is not guaranteed to exist. Entropy rate is only defined when the limit exists.

Some examples are in order. Consider a typewriter producing a sequence one characters over some alphabet $\mathcal{X}$ where $|\mathcal{X}| = m$, all equally likely. After producing n many characters we have a possibility space of m^n possible sequences, all still equally likely, and so the joint entropy $H(X_1, X_2, \ldots, X_n)$ is $\log(m^n) = n\log(m)$. We see then that the entropy rate is just $\log(m)$. What sense is there to make of this? The X_i's in this case are independent and identically distributed. As such, our uncertainty over what comes next shouldn't be increasing or decreasing at all regardless of how long the sequence gets. That is exactly what we see here. More generally, if the X_i of a stochastic process are independent and identically distributed, then they all individually have entropy $H(X_1)$. Then the entropy rate of the process is

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) = H(X_1)$$

which is simply the same intuitive result as the typewriter example but expressed more generally. On the other hand consider a sequence of independent random variables which are *not* identically distributed. One can hopefully see that the entropy rate of such a process is $\lim_{n\to\infty} \frac{1}{n} \sum H(X_i)$. We can glean a bit more intuition about entropy rates by considering how this limit might diverge. Intuition guides us here: we should be getting less and less certain as the process continues, and it needs to happen regardless of the particular sequence as it emerges due to our constraint of independence. Consider a sequence of Bernoulli random variables in which the probability of success p_i changes depending on where in the sequence we are. In particular, consider a sequence of success probabilities which are initially 0 (so that the corresponding X_i have entropy 0 independently of what else has happened so far), followed by being $\frac{1}{2}$ for some stretch of time (so that corresponding X_i have entropy 1), followed by being 0 for a number of steps which 2 to the power of the number of steps that it was previously $\frac{1}{2}$, followed by being $\frac{1}{2}$ for an exponentially longer number of steps, and so on. This limit will never converge because the sequence $\frac{1}{n}\sum H(X_i)$ seems to alternate between approaching 0 and approaching 1.

Another possible candidate to define the entropy rate of a process is the following:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_1, X_2, \ldots, X_{n-1}) \tag{124}$$

Thus the first definition of entropy rate $H(\mathcal{X})$ is the *per-character* entropy of the *total sequence* of random variables, while the second definition is the limiting conditional entropy of the next symbol given what the process has produced up to this point. A small point of subtlety about this definition: it might seem to be a nonincreasing function since by the intuition that conditioning reduces entropy, but in fact this requires the sequence to be stationary. For any stochastic process, we can drop terms that we are conditioning on to get bigger numbers:

$$H(X_{n+1}|X_1, X_2, \ldots, X_n) \leq H(X_{n+1}|X_2, X_3, \ldots, X_n)$$

But then we need that the process is stationary in order to subtract 1 from all of the variables at once and get what we actually want:

$$\begin{aligned} H(X_{n+1}|X_1, X_2, \ldots, X_n) &\leq H(X_{n+1}|X_2, X_3, \ldots, X_n) && (125)\\ &= H(X_n|X_1, X_2, \ldots, X_n) && (126) \end{aligned}$$

From which it follows that this sequence of conditional entropies is nonincreasing. This is important because it means that the limit $H'(\mathcal{X})$ is guaranteed to exist by the monotonic convergence theorem. Moreover it turns out that for stationary stochastic processes, these definitions coincide, providing a condition for the entropy rate to be well defined under our original definition. We need the following lemma:

Lemma 3.1 (Cesaro mean)**.** *If $a_n \to a$ and $b_n \to \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$.*

Theorem 3.1. *For a stationary stochastic process, $H(\mathcal{X}) = H'(\mathcal{X})$.*

Proof. By the chain rule for entropy we have that

$$\begin{aligned} H(\mathcal{X}) = \frac{H(X_1, X_2, \ldots, X_n)}{n} &= \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) && (127)\\ &\to \frac{1}{n}\sum_{i=1}^{n} H'(\mathcal{X}) = H'(\mathcal{X}) && (128) \end{aligned}$$

where the final equality follows from the lemma. □

Note that by the alternative definition, the entropy rate for a stationary Markov chain is

$$\begin{aligned} H(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_1, \ldots, X_{n-1}) &= \lim_{n\to\infty} H(X_n|X_{n-1}) && (129)\\ &= H(X_2|X_1) && (130) \end{aligned}$$

which is very simple.

3.1 The Second Law of Thermodynamics

In statistical thermodynamics, entropy is often defined as the log of the number of microstates in the system. When all states are equally likely this is exactly entropy for the physical system (seeing that system as an abstract source of information) as we've defined it. The second law is the statement that the entropy of a closed physical system can only increase. (Physical entropy is a *special case* of information entropy.

To understand this law from our own perspective we can model the isolated system as a Markov process whose transitions obey the physical laws governing the system. Why a Markov chain? The reason is that physicists presume that the future state of a system is entirely dependent on the present state of the system, and so whatever past states led the system to that present state are irrelevant towards predicting the future. That is exactly what a Markov process assumes. For the sake of simplicity (so as to not have to define differential entropy yet), we assume this process has finitely many possible states. Then it has the transition matrix

$$P = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ r_{21} & r_{22} & \dots & r_{2m} \\ \vdots & & & \\ r_{m1} & r_{m2} & \dots & r_{mm} \end{pmatrix}$$

where r_{ij} is the probability that the system enters state j given that it is in state i, and $|\mathcal{X}| = m$. Consider two different 'initial' distributions for the system, labelled $\vec{\mu}_0$ and $\vec{\sigma}_0$ respectively. (So the i^{th} entry of $\vec{\mu}_0$, μ_0^i, is the initial probability of entering state i, and so on.) According to the laws of physics, the distributions at the next moment will be $\vec{\mu}_1 = P\vec{\mu}_0$ and $\vec{\sigma}_1 = P\vec{\sigma}_0$. More importantly, we can think about the *joint* probability mass functions for the system at the two different moments (seeing it as an evolving process, i.e. a stream of information over time). Call the joint pmf's for the system under the two initial conditions $p(x_1, x_2)$ and $q(x_1, x_2)$ respectively. So

$$p(i,j) = P(X_1 = i, X_2 = j) = P(X_1 = i)P(X_2 = j|X_1 = i) = \mu_0^i r_{ij}$$

and similarly $q(i,j) = \sigma_0^i r_{ij}$. Consider the *relative entropy* between these two joint probability mass functions: $D(p(x_1,x_2)||q(x_1,x_2))$. Specifically, we are interested in comparing this with the relative entropy between the initial distributions: $D(p(x_1)||q(x_1))$. The chain rule for relative entropy allows us to easily make this comparison:

$$D(p(x_1,x_2)||q(x_1,x_2)) = D(p(x_1)||q(x_1)) + D(p(x_2|x_1)||q(x_2|x_1)) \tag{131}$$

Note however the second term, and recall that we are dealing with a Markov chain. These conditional probabilities are the same! Thus $D(p(x_2|x_2)||q(x_2|x_1)) = 0$. The chain rule can be used with privilege to x_2 instead of x_1 to produce a second identity:

$$D(p(x_1,x_2)||q(x_1,x_2)) = D(p(x_2)||q(x_2)) + D(p(x_1|x_2)||q(x_1|x_2)) \tag{132}$$

Unlike in the first identity, Markovity implies nothing about $D(p(x_1|x_2)||q(x_1|x_2))$. We do however know that it is non-negative. Combining the two right sides therefore gives

$$D(p(x_1)||q(x_1)) = D(p(x_2)||q(x_2)) + D(p(x_1|x_2)||q(x_1|x_2)) \geq D(p(x_2)||q(x_2)) \tag{133}$$

We are finally left with

$$D(p(x_1)||q(x_1)) \geq D(p(x_2)||q(x_2)) \tag{134}$$

Or to abuse the notation just a bit:

$$D(\vec{\mu}_0||\vec{\sigma}_0) \geq D(\vec{\mu}_1||\vec{\sigma}_1) \tag{135}$$

The same logic would apply to show that $D(p(x_2)||q(x_2)) \geq D(p(x_3)||q(x_3))$, and that $D(p(x_3)||q(x_3)) \geq D(p(x_4)||q(x_4))$. We have thus shown that the relative entropy between the distributions p and q is non-increasing. The error in assuming that the distribution is q when in fact it is really p can only get smaller. This feels to me very akin to the observation that Stafford Beer makes in chapter 2 of Brain of the Firm in which feedback tends to dominate and render the initial input to a control system irrelevant. Start two of the same system off in two different states. The distinction between them will generally tend to matter less and less over time.

Next, we bring in stationary states. Suppose that our Markov process is stationary, and let $\vec{\lambda}$ be a (the?) stationary distribution. If we substitute $\vec{\lambda}$ for $\vec{\sigma}_0$ in the previous case, then since it is stationary it follows that $\vec{\sigma}_1 = P\vec{\lambda} = \vec{\lambda}$. Thus we obtain the identity

$$D(\vec{\mu}_0||\vec{\lambda}) \geq D(\vec{\mu}_1||\vec{\lambda}) \tag{136}$$

In other words, any initial state will tend to approach indistinguishability with the stationary distribution over time as the Markov process plays out. In fact, it can be shown that this sequence of relative entropies approaches 0 in all cases.

Having relative entropy decreasing isn't quite the same as having actual entropy increase. What it means is that it really depends on what the stationary distribution of the physical system is. Suppose that the stationary distribution of the Markov process is non-uniform, and start with a uniform distribution as the initial condition. Then the above results show that we will approach a non-uniform distribution over time *from* a uniform distribution, implying an overall entropy which is *decreasing*! If however the stationary distribution is uniform $u(x)$, and we start from a non-uniform distribution $p_n(x)$, then it follows that

$$D(p_n(x)||u(x)) = \sum_x p(x) \log\left(\frac{p_n(x)}{u(x)}\right) = \log(\mathcal{X}) - H(X_n) \xrightarrow{n} 0 \tag{137}$$

$$\implies \lim_{n\to\infty} H(X_n) = \log(\mathcal{X}) \tag{138}$$

We know that $\log(\mathcal{X})$ is the maximum entropy of the system. Therefore the entropy must increase over time. One final note can be made of this theory and it's relation to thermodynamic entropy, by considering conditional entropies.

Definition 3.2. A probability transition matrix P is called **doubly stochastic** if not only all of it's rows sum to one, but also all of it's columns.

Theorem 3.2. *The uniform distribution is a stationary distribution of P iff the probability transition matrix is doubly stochastic.*

Proof. todo □

We know already that the conditional entropy $H(X_n|X_1, X_2, \ldots, X_{n-1})$ for a stationary Markov process is non-increasing. What, however, can be said about $H(X_n|X_1)$? Using the fact that additional conditioning reduces entropy, the definition of a Markov process, and the definition of a stationary process (in that order), we have that

$$H(X_n|X_1) \geq H(X_n|X_1, X_2) \tag{139}$$

$$= H(X_n|X_2) \tag{140}$$

$$= H(X_{n-1}|X_1) \tag{141}$$

Thus our conditional uncertainty of the future increases over time. The farther out in time we go, the less certain we are, if we only know the initial condition. The same result can be obtained through the data-processing inequality. Note that the $X_1 \to X_{n-1} \to X_n$ form a Markov chain in the sense described in section 1. By the data-processing inequality then we have

$$I(X_1; X_{n-1}) \geq I(X_1; X_n) \tag{142}$$

Expanding the mutual informations in terms of entropies we have then

$$H(X_{n-1}) - H(X_{n-1}|X_1) \geq H(X_n) - H(X_n|X_1) \tag{143}$$

Since we are dealing with a stationary Markov process, we are assuming that the initial probability distribution was the stationary distribution, and so $H(X_n)$ itself is constant. In particular $H(X_{n-1}) = H(X_n)$, and so we can cancel these out in the above inequality. Doing so and then multiplying both sides by -1 and flipping the inequality around, we are left with

$$H(X_{n-1}|X_1) \leq H(X_n|X_1) \tag{144}$$

Which is exactly what we just showed more directly above.

4 Data Compression

Definition 4.1. A **source code** C for a random variale X is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{D}^*$, the set of finite-length strings of symbols from a D-ary alphabet. Let $C(x)$ denote the codeword corresponding to x and let $l(x)$ denote the length of $C(x)$.

Definition 4.2. The expected length $L(C)$ of a source code $C(x)$ for a random variable x with pmf $p(x)$ is given by

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x) \tag{145}$$

where l is the length of the codeword associated with x. WLOG we will assume the D-ary alphabet is $D = \{0, 1, \dots, D-1\}$.

Definition 4.3. A code is said to be **non-singular** if every element of the range of $\mathcal{X}$ maps to different strings in $\mathcal{D}*$ (so just 1-1?)

This is obviously the place to start when thinking about coding and decoding, since it is the minimal condition for the ability to recover a symbol from a code. However, usually we are interested in coding sequences of symbols. One option is to distinguish between by adding a special symbol (usually a comma) which is only found in between codewords. But this is not very efficient. We can do better by introducing the idea of self-punctuating or instantaneous codes.

Definition 4.4. The **extension** C^* of a code C is the mapping from finite-length strings of $\mathcal{X}$ to finite-length strings of $\mathcal{D}$, defined by

$$C(x_1 x_2 \dots x_n) = C(x_1) C(x_2) \dots C(x_n) \tag{146}$$

A code is called **uniquely decodable** if it's extension is non-singular (again, why do we have this word non-singular it's just the dumb linear algebra stuff again).

As a minimal condition for unique codings of strings over an alphabet, this leaves us still short of any kind of qualifications for a *good* coding. The main issue still is that one may have to read over the entire coded string just to determine the first symbol of the original string. We therefore refine things with the following definition:

Definition 4.5. A code is called a **prefix code** or an **instantaneous code** if no codeword is a prefix of any other codeword.

If no codeword is a prefix of any other codeword then an interpreter can move from left to right across the string, and whenever it finds a full codeword it can be sure that it has found a full symbol of the original string. It can therefore fully determine the original message with one single scan. An instantaneous code is therefore self-punctuating; there is no need for a special comma symbol.

Theorem 4.1 (Kraft Inequality)**.** *For any prefix code over an alphabet of size D, the codeword lengths* $l_1, l_2, \dots, l_m$ *must satisfy*

$$\sum_i \frac{1}{D^{l_i}} \leq 1 \tag{147}$$

Conversely, given a set of codeword lengths that satisfy this inequality, there always exists a prefix code with these word lengths.

Note that this sum gets larger the smaller the code lengths. Thus the optimal coding would be something which satisfies this equality perfectly.

5 Data Transmission

5.1 Channel Capacity

Definition 5.1. A **discrete channel**, denoted by $(\mathcal{X}, p(y|x), \mathcal{Y})$, consists of two finite alphabets $\mathcal{X}$ and $\mathcal{Y}$ and a collection of pmf's $p(y|x)$, one for each $x \in \mathcal{X}$, with the interpretation that the random variable X is the input and Y is the output of the channel. A channel is **memoryless** if the distribution of the output depends only on the input at that time and is independent of previous channel outputs.

Definition 5.2. The (information) **channel capacity** of a discrete memoryless channel is

$$\max_{p(x)} I(X;Y) \tag{148}$$

i.e. it is the maximum possible mutual information between input and output ranging over all possible distributions of the input variable.

Some examples of this are in order.

- **The noiseless binary channel**: This is the simplest possible channel which does something. $\mathcal{X} = \mathcal{Y} = \{0,1\}$, and here information transfers perfectly with no possible loss. Thus $p(0,0) = P(Y = 0|X = 0) = 1$, $p(1,1) = 1$, and $p(1,0) = p(0,1) = 0$. Any reasonable definition of channel capacity should have C equaling 1 bit. We have

  $$I(X;Y) = I(X) - I(X|Y) \tag{149}$$

  Note that $I(X|Y) = 0$ since knowing Y removes all uncertainty about X. Thus maximizing the mutual information is equivalent here to maximizing the uncertainty of X, i.e. the uniform distribution $p(0) = p(1) = \frac{1}{2}$. In that case, $I(X;Y) = H(X) = 1$ bit, and so $C = 1$ bit as expected.

- **Noisy binary channel with non-overlapping outputs**: Let now $\mathcal{X} = \{0,1\}$ but $\mathcal{Y} = \{1,2,3,4\}$, and suppose that 0 inputs map to either 1 or 2 with equal probability, while 1 inputs map to either 3 or 4 with equal probability. The channel appears noisy but it really isn't, because whatever the output is there is no uncertainty about the input. Once again we can see that $H(X|Y) = 0$, and so mutual information is maximized at 1 bit when X has the uniform distribution. The channel capacity here remains 1.

- **Noisy typewriter**: Suppose $\mathcal{X}$ is the standard 26 letter alphabet, as is $\mathcal{Y}$, but nonetheless there any letter has an equal probability of being mapped to itself or the next letter in the alphabet. So if the input message is an a, there is a $\frac{1}{2}$ chance of the receiver getting either an a and a b. (Assume z has a $\frac{1}{2}$ chance of looping back around to a.) Note then first that the conditional entropy $H(Y|X = x)$ is that of a coin flip - the receiver knows it's either x or the thing after x. Thus $H(Y|X = x) = 1$ bit. Therefore

  $$H(Y|X) = \sum_x p(x)H(Y|X = x) = \sum_x p(x) = 1 \tag{150}$$

  Therefore the mutual information is expressible as $I(X;Y) = H(Y) - 1$. Clearly $H(Y) \leq \log(26)$ since at worst all letters are equally likely. Furthermore this is achieved when $p(x)$ is itself the uniform distribution. Thus the maximum mutual information is $\log(26) - 1 = \log(26) - \log(2) = \log(\frac{26}{2}) = \log(13)$. Why would it be $\log(13)$ bits? The reason is that despite the noise, X can still reliable communicate one of 13 different outcomes by simple skipping every other letter. If X only tries to send $a, c, e, g, \ldots$ then there will be no ambiguity. So we still haven't found a channel in which error is unavoidable.

- **Binary Symmetric Channel**: The simplest example of a communication channel with unavoidable error is the binary symmetric channel. Here 0 has a small chance of getting flipped to a 1 (call this probability p), and 1 has an equal chance of getting flipped to a 0. Let $H(p)$ denote the entropy of a Bernoulli random variable with parameter p. This is exactly $H(Y|X=x)$, regardless of X. Thus

$$H(Y|X) = \sum_x p(x)H(Y|X=x) = \left(\sum_x p(x)\right) H(p) = H(p)$$

Moreover $H(Y)$ is maximally 1 bit, which is achieved when X has the uniform distribution. This is because if $P(X=0) = P(X=1) = \frac{1}{2}$, then

$$P(Y=0) = P(X=0)P(Y=0|X=0) + P(X=1)P(Y=0|X=1) = \frac{1}{2}(1-p) + \frac{1}{2}p = \frac{1}{2}$$

which shows that Y is also uniformally distributed and therefore $H(Y)$ is maximally 1. It follows that the channel capacity is $C = 1 - H(p)$. Now $H(p)$ is a curve which starts and ends at 0, rising and falling and maximizing at $p = \frac{1}{2}$. Thus the channel capacity is 0 when there is an equal probability of flipping or not flipping, but the channel becomes equivalent to a normal binary channel at the two extremes. This is because if the bit has a 100 percent chance of flipping, then X can just send a 0 when they want to communicate a 1 and send a 1 when they want to communicate a 0.

- **Binary Erasure Channel**: Here, rather than the bits having a probability of flipping, they have a probability of getting lost entirely. The receiver will know if the bit was lost by receiving an error, which we'll denote e. So $\mathcal{X} = \{0,1\}$ while $\mathcal{Y} = \{0,1,e\}$. Denote the chance of getting this error occurring α, and assume it's the same for both possible inputs. Consider first $H(Y|X=x)$. We clearly have $p(1|0) = 0, p(0|0) = 1-\alpha$, and $p(e|0) = \alpha$. Thus

$$H(Y|X=0) = \alpha\log\left(\frac{1}{\alpha}\right) + (1-\alpha)\log\left(\frac{1}{1-\alpha}\right) + 0 = H(\alpha)$$

where again $H(\alpha)$ denotes the entropy of a Bernoulli random variable with parameter α. $H(Y|X=1) = H(\alpha)$ in an identical way. Thus

$$H(Y|X) = p(0)H(Y|X=0) + p(1)H(Y|X=1) = \cancelto{1}{(p(0)+p(1))}H(\alpha) = H(\alpha)$$

The mutual information between X and Y is then $I(X;Y) = H(Y) - H(\alpha)$. The question then becomes what $p(x)$ gives us a maximal $H(Y)$. Let $P(X=1) = \pi$. Then $P(Y=0) = P(X=0)P(Y=0|X=0) + P(X=1)P(Y=0|X=1) = (1-\pi)(1-\alpha)$. Similarly $P(Y=1) = \pi(1-\alpha)$, and $P(Y=e) = (1-\pi)\alpha + \pi\alpha = \alpha$. Therefore

$$H(Y) = -\alpha\log(\alpha) - (1-\pi)(1-\alpha)(\log((1-\pi)(1-\alpha)) - \pi(1-\alpha)(\log(\pi(1-\alpha)) \tag{151}$$

$$= -\alpha\log(\alpha) - (1-\pi)(1-\alpha)(\log((1-\pi)+\log(1-\alpha)) - \pi(1-\alpha)(\log(\pi)+\log(1-\alpha)) \tag{152}$$

$$= \alpha\log(\alpha) - [(1-\pi)(1-\alpha)\log(1-\alpha) + \pi(1-\alpha)\log(1-\alpha)] - (1-\pi)(1-\alpha)\log(1-\pi) - \pi(1-\alpha)\log(\pi) \tag{153}$$

$$= [\alpha\log(\alpha) - (1-\alpha)\log(1-\alpha)] + (1-\alpha)[-(1-\pi)\log(1-\pi) - \pi\log(\pi)] \tag{154}$$

$$= H(\alpha) + (1-\alpha)H(\pi) \tag{155}$$

We then have

$$I(X;Y) = H(Y) - H(Y|X) = H(\alpha) + (1-\alpha)H(\pi) - H(\alpha) = (1-\alpha)H(\pi)$$

This is clearly maximized when $H(\pi)$ is maximized, which happens when $\pi = \frac{1}{2}$, in which case $H(\pi) = 1$ and we finally find that the channel capacity is $C = (1-\alpha)$.

Consider specifically for a moment the binary symmetric channel. We can consider the 2×2 matrix consisting of the conditional probabilities $p(y|x)$:

$$\begin{pmatrix} p(0|0) & p(1|0) \\ p(0|1) & p(1|1) \end{pmatrix} = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} = \begin{pmatrix} \vec{r}_1 \\ \vec{r}_2 \end{pmatrix}$$

So x values are indexed by the rows and y values by the columns (somewhat backwards but it works out better this way). It is tempting to note that this matrix is symmetric, but the property of matrices defining symmetric *channels* is unfortunately more subtle than this. Consider the calculation for $I(X;Y) = H(Y) - H(Y|X)$. First considering the second term we see

$$H(Y|X) = \sum_x p(x)H(Y|X=x) = \sum_x \left[-\sum_y p(y|x)\log(p(y|x)) \right]$$

For our particular matrix, all rows are their own probability distributions (always the case) but in particular *every row is a permutation of every other row.* This means that this summation will be *the same* for all possible choices of x. Fix $\vec{r} = \vec{r}_1$ without loss of generality, and denote the entropy of any of these distributions $H(\vec{r})$. (Note these are different distributions, with different probabilities of each possible y value, but all nonetheless have the same entropy.) Then

$$H(Y|X) = \sum_x p(x)H(\vec{r}) = H(\vec{r})$$

since the $p(x)$ all sum to 1. We then have

$$I(X;Y) = H(Y) - H(\vec{r}) \leq \log(|\mathcal{Y}|) - H(\vec{r})$$

Now $\vec{r}$ is independent of any choice of $p(x)$. Thus maximizing the mutual information is purely a matter of maximizing $H(Y)$. The above upper bound can be achieved in all cases by having $p(x)$ be the uniform distribution, i.e. $p(x) = \frac{1}{|\mathcal{X}|}$ for all x:

$$p(y) = \sum_{x\in\mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} p(y|x)$$

Now notice a second property of our matrix for the binary symmetric channel: *all columns are also permutations of each other.* It follows that $\sum_{x\in\mathcal{X}} p(y|x)$ is the same for all y, and therefore $p(y)$ is the same for all y, i.e. $p(y)$ is the uniform distribution and therefore $H(Y) = \log(|\mathcal{Y}|)$. We therefore have two properties of the matrix which must hold in order to carry out what we just did. We summarize those in the following definitions:

Definition 5.3. A channel is said to be **symmetric** if the rows of the channel's transition matrix are all permutations of one another and the columns are as well. More importantly, a channel is said to be **weakly symmetric** if every row of the channel's transition matrix is a permutation of every other row, and the columns of the matrix all sum to the same number.

Theorem 5.1. *For a weakly symmetric channel, the channel capacity is given by*

$$C = \log(|\mathcal{Y}|) - H(\vec{r}) \tag{156}$$

where $\vec{r}$ is any row of the channel's transition matrix. Furthermore, this channel capacity is achieved when the input alphabet is given the uniform distribution.

Clearly, calculating channel capacity is rather difficult and there doesn't seem to be a clear-cut way to always accomplish the task. However, it does always exist, and can always be found.

Theorem 5.2 (Basic Properties of Channel Capacity)**.** *We have the following:*

(1) $C \geq 0$

(2) $C \leq \log(|\mathcal{X}|)$ *and* $C \leq \log(|\mathcal{Y}|)$

(3) C always exists and can be found using standard optimization techniques (although there is no closed form solution to always finding it).

Proof. 1 follows from the fact that $I(X;Y) \geq 0$. 2 follows from the fact that $C = \max I(X;Y) \leq \max H(X) = \log(\mathcal{X})$ and also $C \leq \max H(Y) \leq \log(|\mathcal{Y}|)$. To understand 3 we must recall some properties of mutual information that I skipped in chapter 1. Namely that $I(X;Y)$ is a continuous concave function of $p(x)$. Because it is a concave function over a closed and (trivially) convex set, local maximums are always global maximums. The boundedness of C from 2 therefore means that we can always use the word maximum instead of supremum. (This is lazy but I don't care right now.) □

Right now our conceptual setup has a channel in which just one symbol from the alphabet $\mathcal{X}$ is sent over to Y. We wish to consider the following more general and elaborate situation:

(1) Alice has a message that she wants to send, W, drawn from some finite index set of possible messages $\{1, 2, \ldots, M\}$. She first encodes the message via some uniquely decodable coding scheme, producing the string in the input alphabet: $W \to X^n(W) \in \mathcal{X}^n$.

(2) That message $X^n(W)$ is transmitted across the channel $p(y|x)$, one symbol at a time, producing a message in the output alphabet: $X^n(W) \to Y^n$.

(3) Bob, receiving this coded message, then guesses the index of the original message via an established decoding scheme, producing the final received message $\hat{W}$. $Y^n \to \hat{W}$.

We therefore need a few more definitions. First we need to define the extension of a channel to one which transmits multi-character messages.

Definition 5.4. The n^{th} **extension** of the discrete memoryless channel (DMC) $C = (\mathcal{X}, p(y|x), \mathcal{Y})$, is the channel $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$, where in order to maintain the principle of being memoryless we require that

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k) \tag{157}$$

for $k = 1, 2, \ldots, n$.

Related to all of this is the concept of feedback, which is a separate consideration from memorylessness. To have feedback would be to have inputs be effected by outputs of the communication channel. If the system does *not* have and feedback, then we can safely assume that

$$p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$$

I.e. that the next input character is independent of the output characters produced up to that point in the transmission. Note that in this case the channel transition function for the n^{th} extension of a DMC can be found in a simple way:

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i) \tag{158}$$

(I'm not actually seeing right now why the feedback condition is necessary for this.)

Definition 5.5. An (M, n) **code** for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following:

(1) An index set $\{1, 2, \ldots, M\}$.

(2) An encoding function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$. The set of codewords is called the **codebook**.

(3) A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$.

Note that we are now distinguishing between two numbers: the number of messages which someone might want to send, M (alternatively M is the variety of a system), and the number of successive times which the channel needs to be used in order to send that information, n. It stands to reason that the larger the channel capacity, the less times it needs to be used to send a 'full' message. For a fixed n, we are employing the channel for use in transmitting information - a **transmission** of an (M, n) code is a sequence of n successive uses of it. Consider the noiseless binary channel, which transmits just one bit at a time. If we have an information source which produces $M = 1024$ possible outcomes, then it would require $\log(1024) = 10$ uses of the channel in order to communicate what happened to the receiver on the other side. Consider the number

$$R = \frac{\log(M)}{n} \tag{159}$$

A few things to note about this number. Firstly, one can see that the smaller it is, the less efficiently we are using our channel. It's units are clearly bits per transmission. As we noted before, $M = 1024$ means we should only need to use a noiseless binary channel 10 times in order to send any message without any possibility of error. R here is therefore $\frac{10}{10} = 1$: 1 bit per one transmission, exactly as seems the maximal capability of the channel. If on the other hand we were using some overly complicated coding scheme for sending messages which required 20 uses of the binary channel, then $R = \frac{10}{20} = \frac{1}{2}$. This tells us we are only sending half a bit of information per use of the channel, an obvious drop in the **rate** of transmission. On the other hand, the bigger the rate R is, the more we are demanding of any channel. If $R = 10$, we are asking it to send 10 bits per one transmission. This is clearly impossible to achieve for our binary channel.

A point of subtlety: This number, *R has nothing to do directly with the channel itself.* It is purely a property of the code (M, n). Consideration of the channel is being considered in the choice of n, but that is itself a choice and not a property of the channel. One can define any code of any rate. The question we want to consider is whether or not a particular rate is **achievable** by a particular channel. The linkage between these notions and direct properties of a channel are given by the following definitions. Fix a channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ and a code $(M, n) = (\{1, 2, \ldots, M\}, X^n, g)$. First, define the **conditional probability of error** given the message i:

$$\lambda_i = P(g(Y^n) \neq i | X^n = x^n(i)) = \sum_{y^n \in \mathcal{Y}^n} p(y^n | x^n(i) I(g(y^n) \neq i) \tag{160}$$

This is the probability that the message is received without error given that the message was i. It could be different depending on the message. This allows us to define a 'worst-case' measure of performance:

$$\lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i \tag{161}$$

This we call what it is: the **maximum probability of error**. Also of interest is the **average probability of error**:

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i \tag{162}$$

What should it mean for a particular rate R to be achievable by a channel? It should mean that we can *always* send messages at that rate *regardless* of the variety of the messages. A binary channel with a rate of 1 bit per 1 transmission should care if $M = 1000$ or $M = 1000000$; it should be able to achieve it's rate independently of that. Thus our definition of achievability is going to have to involve a sequence of codes in which the probability of error is always small, or at the very least goes to 0.

Definition 5.6. A rate R is said to be **achievable** if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the maximal probability of error $\lambda^{(n)}$ tends to 0 as $n \to \infty$. For simplicity we will tend to write $(2^{nR}, n)$ in place of $(\lceil 2^{nR} \rceil, n)$.

So if the rate R is, say, 3, then the sequence of codes would be

$$(8, 1), (64, 2), (512, 3), \ldots$$

i.e. one would have to demonstrate the rate being achieved sending a variety $8 = (2^3)^1$ signal using one transmission of the channel, a variety 64 signal using $(2^3)^2$ signal using 2 transmissions of the channel, and so on. Finally, we can define the following

Definition 5.7. The **capacity** of a channel is the supremum of all achievable rates.

We are almost ready to state and prove the noisy channel coding theorem, the fundamental theorem of information theory. We only need one more theoretical concept: an extension of the concept of typicality to a joint set of random variables.

Definition 5.8. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be two collections of iid random variables. Recall the sample entropy of this random sample for, say X, is the statistic $h(\vec{x}) = -\frac{1}{n}\log(p(\vec{x})$. The set $A_\epsilon^{(n)}$ of **jointly typical sequences** (x^n, y^n) with respect to the distribution $p(x, y)$ is the set

$$A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : |h(\vec{x}) - H(X)| < \epsilon \wedge |h(\vec{y}) - H(Y)| < \epsilon \wedge |h(\vec{y}, \vec{z}) - H(X, Y)| < \epsilon\} \quad (163)$$

Thus it is the set of sequences (x^n, y^n) in which x^n would be typical with respect to $p(x)$, y^n would be typical with respect to $p(y)$, and the pair (x^n, y^n) would be typical with respect to $p(x^n, y^n)$, where we are assuming based on the context of our discussion of channel capacity that

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$$

We now unfortunately need to restate and reprove the asymptotic equipartition property for this new extended definition:

Theorem 5.3 (Joint AEP)**.** *Let (X^n, Y^n) be sequences of length n drawn iid according to $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$. (What we mean by drawing iid here is drawing (X, Y) pairs one at a time n times according to their joint probability distribution $p(x, y)$.) Then*

(1) $P((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ *as* $n \to \infty$

(2) $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon}$

(3) If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n, y^n)$ (i.e. $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with the same marginals as $p(x^n, y^n)$), then

$$P\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \leq \frac{1}{2^{n(I(X;Y)-3\epsilon)}}$$

and for sufficiently large n we also have

$$P\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \geq (1-\epsilon)\frac{1}{2^{n(I(X;Y)+3\epsilon}}$$

Proof. By the weak law of large numbers we have that the sample entropies converge in probability to their respective actual entropies. Thus we can pick an n_1 such that for all $n > n_1$,

$$P(|h(\vec{X}) - H(X)| \geq \epsilon) < \frac{\epsilon}{3}$$

an n_2 such that for all $n > n_2$,

$$P(|h(\vec{Y}) - H(Y)|) \geq \epsilon) < \frac{\epsilon}{3}$$

and an n_3 such that for all $n > n_3$,

$$P(|h(\vec{X}, \vec{Y}) - H(X, Y)| \geq \epsilon) < \frac{\epsilon}{3}$$

Picking $N = \max\{n_1, n_2, n_3\}$, we see that for all $n > N$, the probability of being in at least one of these sets is less than ϵ. However, these probabilities are precisely the probabilities of not being in the respective typical sets. Therefore the probability of being in all three of them at once has to be at least $1-\epsilon$, proving (1). Part (2) is identical to the arguments made in section 2: Simply note that

$$1 = \sum_{(\vec{x},\vec{y})\in\mathcal{X}^n\times\mathcal{Y}^n} p(\vec{x},\vec{y}) \tag{164}$$

$$\geq \sum_{(\vec{x},\vec{y})\in A_\epsilon^{(n)}} p(\vec{x},\vec{y}) \tag{165}$$

$$\geq |A_\epsilon^{(n)}| \frac{1}{2^{n(H(X,Y)+\epsilon)}} \tag{166}$$

Thus multiplying both sides we see $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$. For (3), if $\tilde{X}^n$ and $\tilde{Y}^n$ are independent but have the same marginals as X^n and Y^n, then

$$P((\tilde{X}^n,\tilde{Y}^n)\in A_\epsilon^{(n)}) = \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n) \tag{167}$$

$$\leq \left(2^{n(H(X,Y)+\epsilon)}\right)\left(\frac{1}{2^{n(H(X)-\epsilon)}}\right)\left(\frac{1}{2^{n(H(Y)-\epsilon)}}\right) \tag{168}$$

$$= \frac{1}{2^{n(I(X;Y)-3\epsilon)}} \tag{169}$$

Where the final equality follows from the identity that $I(X;Y) = H(X)+H(Y)-H(X,Y)$. For the second part of (3), we have already that $P(A_\epsilon^{(n)}) \geq 1-\epsilon$, and therefore

$$1-\epsilon \leq \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n,y^n) \tag{170}$$

$$\leq |A_\epsilon^{(n)}| \frac{1}{2^{n(H(X,Y)-\epsilon)}} \tag{171}$$

by definition of $A_\epsilon^{(n)}$. Thus by multiplying both sides we have

$$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}$$

But then it follows that for sufficiently large n, we have

$$P((\tilde{X}^n,\tilde{Y}^n)\in A_\epsilon^{(n)}) = \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n) \tag{172}$$

$$\geq \left((1-\epsilon)2^{n(H(X,Y)-\epsilon)}\right)\left(\frac{1}{2^{n(H(X)+\epsilon)}}\right)\left(\frac{1}{2^{n(H(X)+\epsilon)}}\right) \tag{173}$$

$$= (1-\epsilon)\frac{1}{2^{n(I(X;Y)+3\epsilon)}} \tag{174}$$

completing the proof. □

The really crucial thing is in the above theorem is fact (3), in conjunction with fact (1). What these two facts mean taken together is that if we are drawing pairs of sequences (X^n, Y^n), then we can be nearly certain of drawing jointly typical pairs according to the channel, and nearly certain of *not* drawing typical pairs if we are drawing the two sequences independently, and moreover this certainty increases with n. The exceptions to this rule concern the mutual information $I(X;Y)$ in relation to a channel. If the channel is extremely noisy, then X and Y will be nearly independent, and $I(X;Y)$ will be close to 0. In this case the (3) and (1) are not at all contradictory since the bounds converge to 1. On the other hand, if the channel is airtight, then $I(X;Y)$ will something decently sized and greater than 0 and this theorem confirms that there is a gulf of difference between drawing the pair (X,Y) according to the rules of the channel and drawing them independently.

From all of these results combined we have that there are about $2^{nH(X)}$ typical X sequences, about $2^{nH(Y)}$ typical Y sequences, and about $2^{nH(X,Y)}$ jointly typical sequences. But we know that $H(X,Y) \leq H(X) + H(Y)$. The number of typical X and typical Y pairs is therefore bigger than the number of jointly typical sequences; not all pairs of typical X^n and typical y^n sequences are jointly typical. We also have that the probability of a randomly chosen pair being jointly typical is about $\frac{1}{2^{n(I(X;Y)}}$. Consider the experiment of drawing pairs of sequences (x^n, y^n) uniformly at random until finding a jointly typical pair. Then the number of non-jointly typical pairs drawn before finding one would be a geometric random variable with $p = \frac{1}{2}^{nI(X;Y)}$, and with expected value $2^{nI(X;Y)} - 1$, i.e. we would expect to draw about $2^{nI(X;Y)}$ pairs before finding a jointly typical one.

We are now finally ready to state and prove the fundamental theorem of information theory.

Theorem 5.4 (Channel coding theorem)**.** *For a discrete memoryless channel, all rates below capacity C are achievable. Specifically, for any rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.*

Proof. Let $(\mathcal{X}, p(y|x), \mathcal{Y})$ be a channel with capacity C. Fix any probability distribution for $p(x)$ (later we will fix p to be the distribution which maximizes $I(X;Y)$, i.e. achieves channel capacity). Fix an R, and an n. We generate our codewords by simply drawing strings in $\mathcal{X}^*$ at random according to the distribution $p(x)$. Thus the probability of some string being the i^{th} codeword is

$$p(x) = \prod_{i=1}^{n} p(x_i)$$

The *ith* string we draw is the *ith* codeword, and we need 2^{nR} many. We can index the codewords vertically from left to right in order to construct a matrix of symbols in $\mathcal{X}$:

$$\mathcal{C} = \begin{pmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ x_1(2) & x_2(2) & \dots & x_n(2) \\ \vdots & & & \\ x_1(2^{nR}) & x_2(2^{nR}) & \dots & x_n(2^{nR}) \end{pmatrix} \tag{175}$$

Since the codes are generated independently according to $p(x)$, the probability that we get a particular code matrix $\mathcal{C}$ is

$$P(\mathcal{C}) = \prod_{j=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(j)) \tag{176}$$

Now we consider two people trying to communicate over this channel and using this codebook. From the perspective of the receiver, the message W chosen might as well be chosen uniformly at random from the collection of $M = 2^{nR}$ many possible ones. (Since the message W is being seen as a random variable we use lower case w as the fixed outcome of this.) Both the sender and the receiver have a copy of the codebook and understand the conditional probability transition matrix $p(y|x)$ for the channel. Thus if the sequence received is y^n, and we are assuming a memoryless channel with no feedback (which we are) then the receiver knows the conditional probability of this given any particular code $x^n(w)$ is

$$p(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w)) \tag{177}$$

Using knowledge of these probabilities the receiver makes a guess at which message was actually sent. Their strategy for making this choice is **jointly typical decoding**: they search through the codebook looking for a *unique* index $\hat{W}(y^n)$ such that $(X^n(\hat{W}), Y^n)$ is jointly typical. If the receiver can't find any such index, or finds more than one of them, then they declare that there was an error in the communication. Even if the receiver does find a unique index, there could still be an error in the decoding, since this is all probabilistic.

In this case they output the dummy sequence 0. Let $\mathcal{E}$ be the event that there is a decoding error, i.e. $\mathcal{E} = \{\hat{W}(Y^n) \neq W\}$.

$$\begin{aligned} P(\mathcal{E}) = P(\hat{W}(Y^n) \neq W) &= \sum_{\mathcal{C}} P(\mathcal{C}) P(\hat{W}(Y^n) \neq W | \mathcal{C}) && (178) \\ &= \sum_{\mathcal{C}} P(\mathcal{C}) \left(P(\bigvee_{w=1}^{2^{nR}} \hat{W}(Y^n) \neq W \wedge W = w | \mathcal{C} \right) && (179) \\ &= \sum_{\mathcal{C}} P(\mathcal{C}) \sum_{w=1}^{2^{nR}} P(W = w) P(\hat{W}(Y^n) \neq W | W = i \wedge \mathcal{C}) && (180) \\ &= \sum_{\mathcal{C}} P(\mathcal{C}) \sum_{w=1}^{2^{nR}} \frac{1}{2^{nR}} \lambda_w(\mathcal{C}) = \sum_{\mathcal{C}} P(\mathcal{C}) P_e^{(n)}(\mathcal{C}) && (181) \end{aligned}$$

where $\lambda_w(\mathcal{C})$ and $P_e^{(n)}(\mathcal{C})$ denote the conditional probability of error and average probability of error given the code $\mathcal{C}$. Rearranging a bit we have

$$P(\mathcal{E}) = \sum_{\mathcal{C}} P(\mathcal{C}) \sum_{w=1}^{2^{nR}} \frac{1}{2^{nR}} \lambda_w(\mathcal{C}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \left[\sum_{\mathcal{C}} P(\mathcal{C}) \lambda_w(\mathcal{C}) \right] \tag{182}$$

Note that because the term in the brackets. Because the symmetric nature of the code construction (what is meant by this???) the probability of error doesn't actually depend on the particular index which was sent. Thus this term in the brackets is the same for all w. Without loss of generality we set $w = 1$, so that this becomes

$$\begin{aligned} P(\mathcal{E}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}) = \sum_{\mathcal{C}} P(\mathcal{C}) \frac{1}{2^{nR}} 2^{nR} \lambda_1(\mathcal{C}) &= \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}) && (183) \\ &= P(\mathcal{E} | W = 1) && (184) \end{aligned}$$

Now for each $i = 1, \ldots, 2^{nR}$, let E_i be the event that the pair $(X^n(i), Y^n)$ is jointly typical. Suppose that the noise of the channel produces a Y^n such that a unique jointly typical sequence wouldn't be $X^n(1)$. Then the decoding scheme will produce an error, and this is precisely the event E_1^c (c denoting the complement). Note that this event also includes the cases in which the receiver finds multiple jointly typical sequences or none, since in that case they decode a string which isn't jointly typical with anything. Errors can also be produced when there is a unique jointly typical $X^n(i)$ which isn't correct. Thus for instance $E_2 \subseteq \mathcal{E}$, as well as $E_3, E_4, \ldots, E_{2^{nR}}$. All in all, E_1^c along with $E_2, \ldots, E_{2^{nR}}$ constitute a cover of $\mathcal{E}$ and also are contained within it. Therefore

$$(\mathcal{E} | W = 1) = E_1^c \cup E_2 \cup \ldots \cup E_{2^{nR}} \tag{185}$$

Thus we have

$$\begin{aligned} P(\mathcal{E}) = P(\mathcal{E} | W = 1) &= P(E_1^c \cup E_2 \cup \ldots \cup E_{2^{nR}} | W = 1) && (186) \\ &\leq P(E_1^c | W = 1) + \sum_{i=2}^{2^{nR}} P(E_i | W = 1) && (187) \end{aligned}$$

We are now set up to use the joint AEP theorem. Firstly, we know by it that $P(E_1^c | W = 1) \to 0$. Thus for n sufficiently large we can assume that $P(E_1^c | W = 1) \leq \epsilon$. Next, by definition of $\mathcal{C}$ we know that $X^n(1)$ and $X^n(i)$ are independent for $i \neq 1$. Thus $Y^n(X^n(1))$ and $X^n(i)$ are also independent. Nonetheless, they have the same marginals as $p(x^n, y^n)$. Therefore by the joint AEP again, we have that $P(E_i | W = 1) \leq$

$2^{-n(I(X;Y-3\epsilon)}$. Thus

$$\begin{aligned}
P(\mathcal{E}) &\leq P(E_1^c|W=1) + \sum_{i=2}^{2^{nR}} P(E_i|W=1) \qquad (188)\\
&\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \qquad (189)\\
&= \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \qquad (190)\\
&\leq \epsilon + 2^{nR}2^{-nI(X;Y)-3n\epsilon} \qquad (191)\\
&= \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)} \qquad (192)
\end{aligned}$$

Now we know by hypothesis that $R < I(X;Y)$. This means that $I(X;Y) - R - 3\epsilon > -3\epsilon$, so that $-n(I(X;Y) - R - 3\epsilon) < -3n\epsilon$, so that $2^{-n(I(X;Y)-R-3\epsilon)} < 2^{-3n\epsilon} = (2^{3\epsilon})^{-n}$. Choosing n large enough to make this number less than or equal to ϵ means that

$$\begin{aligned}
P(\mathcal{E}) &< \epsilon + 2^{3n\epsilon}2^{-n(I(X;Y)-R)} \qquad (193)\\
&< \epsilon + 2^{-3n\epsilon} \qquad (194)\\
&\leq 2\epsilon \qquad (195)
\end{aligned}$$

which can be made arbitrarily small by choice of ϵ. By choosing $p(x)$ to be the distribution which maximizes $I(X;Y)$ we see that we can choose any R below the channel capacity and have this method work. We still aren't done yet, however. This probability is conditional on the codebook being chosen. We are supposed to be fixing one. Moreover $\mathcal{E}$ is really the average probability of error conditioned on the codebook $\mathcal{C}$ when our definition of achievability requires that the *maximal* probability of error goes to 0. In fact, recall what $P(\mathcal{E})$ really is:

$$P(\mathcal{E}) = \sum_{\mathcal{C}} P(\mathcal{C}) P_\epsilon^{(n)}(\mathcal{C}) = E_{\mathcal{C}}(P_\epsilon^{(n)}) \leq 2\epsilon$$

for sufficiently large n. Such an expectation cannot possibly be achieved unless there is a specific codebook $\mathcal{C}^*$ such that the average probability of error is actually below 2ϵ for sufficiently large n. Pick that one, specifically. We need to now ensure that the *maximal* probability of error is getting arbitrarily small.

The trick is to simply throw out half of the codewords $\mathcal{C}^*$; specifically the half with the highest probability of error. Suppose that more than half of the codewords have probability of error greater than 4ϵ. Then even if all other codewords have probability of error equal to 0,

$$P_\epsilon^{(n)} > \frac{1}{2^{nR}}(2^{nR-1}4\epsilon) > 2\epsilon \overset{n}{\to} 0$$

a contradiction. We therefore have to have that at most half of the codewords have to have a probability of error less than 4ϵ. Throwing the worst half of the codewords out therefore leaves us with a set of codewords such that the maximal probability of error is still *less* than 4ϵ. What are we left with? Before throwing out half the codewords for each n, we had a sequence of codes

$$(2^R, 1), (2^{2R}, 2), (2^{3R}, 3), \ldots$$

Now we have a sequence of codes

$$2^{R-1}, 1), (2^{2R}-1, 2), (2^{3R}-1, 3), \ldots$$

i.e.

$$2^{(R-1)}, 1), (2^{2(R-\frac{1}{2})}), 2), (2^{3(R-\frac{1}{3})}, 3), \ldots$$

i.e. we have constructed a sequence of codes of rates $R' = R - \frac{1}{n}$. However, the distinction becomes negligible over time by definition of achievability, since achieving a rate R involves taking the ceiling of 2^{nR}. Thus the sequence of codes also works to achieve the rate R. Thus we have proven the first half of the theorem - that all rates $R < C$ are achievable.

It remains to show that all achievable rates are $\leq C$. Three ingredients are required for this. First, we need to recall Fano's inequality. We have the Markov chain $W \to X^n(W) \to Y^n \to \hat{W}$, where $\hat{W}$ is an estimator of the original W. Assume that the message chosen W is uniformly distributed (as we have been). Fano's inequality tells us that

$$H(W|\hat{W}) \leq 1 + P_e^{(n)} \log(2^{nR}) = 1 + P_e^{(n)} nR \tag{196}$$

where the probability of error in the context of the statement of Fano's inequality is exactly the average probability of error we've been dealing with $P_e^{(n)}$ because

$$P_e = P(W \neq \hat{W}) = \sum_i P(\hat{W} \neq W | W = i) P(W = i) = \frac{1}{2^{nR}} \sum_i \lambda_i = P_e^{(n)}$$

The next ingredient is a small lemma. We claim that if Y^n is the result of passing X^n through a discrete memoryless channel of capacity C (one letter at a time), then $I(X^n; Y^n) \leq nC$ for all probability distributions $p(x^n)$. To see this note

$$\begin{aligned} I(X^n; Y^n) &= H(Y^n) - H(Y^n | X^n) & (197) \\ &= H(Y^n) - \sum_{i=1}^n H(Y_i | Y_{i-1}, \dots, Y_1, X^n) & (198) \\ &= H(Y^n) - \sum_{i=1}^n H(Y_i | X_i) & (199) \\ &\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i | X_i) & (200) \\ &= \sum_{i=1}^n I(X_i; Y_i) \leq nC & (201) \end{aligned}$$

where the second line applies the chain rule for entropy, the third uses the memoryless property of the channel, and the fourth line uses the fact that joint entropy is smaller than the sum of the individual entropies.

Lastly, we need to be a bit fancy with the data processing inequality. It is clear that $W \to X^n \to Y^n \to \hat{W}$ is a Markov chain. Applying the data processing inequality directly gives us $I(W; \hat{W}) \leq I(W; Y^n)$ But we need a bit more than this. Note that we showed earlier that $\hat{W} \to Y^n \to X^n \to W$ is also a Markov chain. Using this as well as the fact that $I(A; B) = I(B; A)$ gives us that

$$I(W; Y^n) = I(Y^n; W) \leq I(Y^n, X^n = I(X^n, Y^n)$$

Thus we have that $I(W; \hat{W}) \leq I(X^n; Y^n)$. This is the final ingredient. With our ingredients prepared, observe the following:

$$\begin{aligned} nR = \log(2^{nR}) &= H(W) & (202) \\ &= H(W|\hat{W}) + I(W; \hat{W}) & (203) \\ &\leq 1 + P_e^{(n)} nR + I(W; \hat{W}) & (204) \\ &\leq 1 + P_e^{(n)} nR + I(X^n; Y^n) & (205) \\ &\leq 1 + P_e^{(n)} nR + nC & (206) \end{aligned}$$

where the first line is using our assumption of the uniform distribution, the second line uses the identity $I(A; B) = H(A) - H(A|B) \implies H(A) = H(A|B) + I(A; B)$, the third line applies our first ingredient (Fano's inequality), the fourth line uses our third ingredient (the data processing inequality), and the final line using our second ingredient (our little lemma). Dividing both sides by n gives

$$R \leq P_e^{(n)} R + \frac{1}{n} + C \xrightarrow{n} C \tag{207}$$

since the probability of error goes to 0 by definition of achieving a rate of transmission. This completes the proof. □

This converse to the channel coding theorem is sometimes referred to as the **weak** converse. The strong converse is an addendum. It adds that for all rates above capacity, the probability of error goes to 1 exponentially fast. Hence, channel capacity is a very clear dividing point. Rates of error go to 0 exponentially fast when we are below capacity, and go to 1 exponentially fast when we are above it.

5.2 Regulation and Ashby's Law of Requisite Variety

Our setup is the following. First, we have a system of some sort. What we are calling a system is whatever our regulator is seeking to **stabilize**, and the relevant states of the system are determined by the aspects of this system which we care about. We denote the state of this system E, and view it as a discrete random variable. It is important to see this system concretely, as the thing itself, but it is equally important to see it as an information source, or more specifically a *variety generator.* Our second information source generating variety is the external environment itself. This information source produces disturbances which will alter or disrupt the stability of the system. We thus denote this source of information D. Lastly, we have the regulator itself. We will refer to the regulator as a machine, even if it isn't. This regulator is an information source just like the other two; the information created is the action taken moment to moment. We denote this information source R.

The goal of regulation is in some sense the opposite of the goal of a channel. With a channel, we are seeking to maximize the mutual information between the information generated by a source and the information generated at the opposite side of the channel. With a regulator on the other hand, we are seeking to minimize the mutual information. If a regulator does it's job properly, it completely destroys any information about the disturbance D. A few assumptions about the control systems we are interested in are necessary. First, our regulator is to be seen as a machine. This is to say that it's action is a deterministic function of the disturbance. Because of this, the conditional uncertainty of the regulator given the disturbance is zero. That is to say, we will assume

$$H(R|D) = 0 \tag{208}$$

Next we consider the conditional uncertainty of the regulator given the disturbance. We definitely don't, in general, have that $H(D|R) = 0$ even if the machine is deterministic. This is because the regulator might want to take the same action for multiple different kinds of disturbances. In other words, if $R = f(D)$, then the function f need not be injective. Nonetheless, we are looking to derive conditions on the regulator which are necessary to achieve stability. We need to be careful not to hide anything from ourselves, and this requires putting a bound on the sophistication of the actions that our regulator can take. In particular, we need to assume that if we fix a particular action of the regulator, then any new disturbance will require the regulator to change it's action. If this weren't the case for some action of the machine, then that would mean that this action of the machine is safeguarding the system from multiple possible disturbances at once, and this would mean we are smuggling extra variety into our regulator without counting it. In order to properly keep track of the variety therefore, we add the condition that if we fix the action of the regulator, then the number of possible effects should therefore be at least as large as the number of possible disturbances. (Put another way, if we stop allowing the regulator to regulate, it should be as if the regulator isn't there in the first place.) Expressed in terms of entropy we have the assumption:

$$H(E|R) \geq H(D|R) \tag{209}$$

Now, using the chain rule for joint entropy we can express $H(D, R)$ in two different ways:

$$H(D, R) = H(D) + H(R|D) = H(R) + H(D|R) \tag{210}$$

Now $H(R|D) = 0$ as we said. And $H(D|R) \leq H(E|R)$. Therefore

$$\begin{aligned} H(D) &= H(R) + H(D|R) && (211) \\ &\leq H(R) + H(E|R) && (212) \\ &= H(E, R) && (213) \\ &\leq H(E) + H(R) && (214) \end{aligned}$$

So then we have

$$H(D) \leq H(E) + H(R) \tag{215}$$

Rearranging the terms then gives

$$H(E) \geq H(D) - H(R) \tag{216}$$

Recall that our goal here is *stability*. This is *all* that we want, and therefore we should have as little uncertainty about E as possible, regardless of the disturbance D. If the goal is therefore to minimize E and we have no control over D, then there is only one way to do it: our regulator must have sufficient variety generating potential to absorb the variety created by D. This is Ashby's law:

Theorem 5.5 (Ashby's Law of Requisite Variety)**.** *A regulator seeking to fully stabilize a system it is overseeing must have at least as much variety as that of the possible disturbances which could disrupt that stability. Moreover achieving partial stability amounts to approaching coming close to the variety of D.*

When transmitting data across a channel, we are seeking to maximize the mutual information between the two sides of the channel. There is a third party in this process which seeks to disrupt this transmission - a noise generator, seeking to destroy information, which we are working against when designing channels. In thinking about regulation, we are playing the role of the noise generator, seeking to destroy information before it gets anywhere. In other words, the variety seeks to disrupt this information flow, and *minimize* the mutual information between the two sides of the channel. Ashby's Law gives us a minimal condition for accomplishing this task. It tells us that the variety of the noise generator must match the variety of the source if it is to have any hope of success.

Bringing in the second law of thermodynamics completes the story. We are all machines, seeking to maximize our own local stability, i.e. reduce our local entropy. However, we are all part of a larger overall system, and the second law of thermodynamics tells us that entropy is non-decreasing. If I bring my own local entropy down from H to $H - \epsilon$, then the entropy of the surrounding system which excludes me must increase by that much. In the pursuit of reducing entropy local, I serve the cosmic purpose of increasing entropy, globally. In other words, all machines are variety attenuators relative to themselves, but are also variety generators relative to the larger system they exist in. In our pursuit of our own interest, we are all agents of the second law.

Part II
Algorithmic Information Theory