Notes on Information Theory

by Jeff Steif

1 Entropy, the Shannon-McMillan-Breiman Theorem and Data Compression

These notes will contain some aspects of information theory. We will consider some REAL problems that REAL people are interested in although we might make some mathematical simplifications so we can (mathematically) carry this out. Not only are these things inherently interesting (in my opinion, at least) because they have to do with real problems but these problems or applications allow certain mathematical theorems (most notably the Shannon-McMillan-Breiman Theorem) to "come alive". The results that we will derive will have much to do with the entropy of a process (a concept that will be explained shortly) and allows one to see the real importance of entropy which might not at all be obvious simply from its definition.

The first aspect we want to consider is so-called data compression which means we have data and want to compress it in order to save space and hence money. More mathematically, let $X_1, X_2, \ldots, X_n$ be a finite sequence from a stationary stochastic process $\{X_k\}_{k=-\infty}^{\infty}$. Stationarity will always mean strong stationarity which means that the joint distribution of $X_m, X_{m+1}, X_{m+2}, \ldots, X_{m+k}$ is independent of m, in other words, we have time-invariance. We also assume that the variables take on values from a finite set S with $|S| = s$. Anyway, we have our $X_1, X_2, \ldots X_n$ and want a rule which given a sequence assigns to it a new sequence (hopefully of shorter length) such that different sequences are mapped to different sequences (otherwise one could not recover the data and the whole thing would be useless, e.g., assigning every word to 0 is nice and short but of limited use). Assuming all words have positive probability, we clearly can't code all the words of length n into words of length $n-1$ since there are s^n of the former and s^{n-1} of the latter. However, one perhaps can hope that the expected length of the coded word is less than the length of the original word (and by some fixed fraction).

What we will eventually see is that this is always the case EXCEPT in the one case where the stationary process is i.i.d. AND uniform on S. It will turn out that one can code things so that the expected length of the output is $H/\ln(s)$ times the length of the input where H is the entropy of the process. (One should conclude from the above discussion that the only stationary process with entropy $\ln(s)$ is i.i.d. uniform.)

What is the entropy of a stationary process? We define this in 2 steps. First, consider a partition of a probability space into finitely many sets where (after some ordering), the sets have probabilities, $p_1, \ldots, p_k$ (which are of course positive and add to 1). The entropy of this partition is defined to be

$$-\sum_{i=1}^{k} p_i \ln p_i.$$

Sounds arbitrary but turns out to be very natural, has certain natural properties (after one interprets entropy as the amount of information gained by performing an experiment with the above probability distribution) and is in fact the only (up to a multiplicative constant) function which has these natural properties. I won't discuss these but there are many references (ask me if you want). (Here ln means natural ln, but sometimes in these notes, it will refer to a different base.)

Now, given a stationary process, look only at the random variables $X_1, X_2, \ldots X_n$. This breaks the probability space into s^n pieces since there are this many sequences of length n. Let H_n be the entropy (as defined above) of this partition.

Definition 1.1: *The* **entropy** *of a stationary process is* $\lim_{n\to\infty} \frac{H_n}{n}$ *where* H_n *is defined above (this limit can be shown to always exist using subadditivity).*

Hard Exercise: Define the conditional entropy $H(X|Y)$ of X given Y (where both r.v.'s take on only finitely many values) in the correct way and show that the entropy of a process $\{X_n\}$ is also $\lim_{n\to\infty} H(X_0|X_{-1}, \ldots X_{-n})$.

Exercise: Compute the entropy first of i.i.d. uniform and then a general i.i.d.. If you're then feeling confident, go on and look at Markov chains (they are not that hard).

Exercise: Show that entropy is uniquely maximized at i.i.d. uniform. (The previous exercise showed this to be the case if we restrict to i.i.d.'s but you need to consider all stationary processes.)

We now discuss the Shannon-McMillan-Breiman Theorem. It might seem dry and the point of it might not be so apparent but it will come very much alive soon.

Before doing this, we need to make an important (and not unnatural) assumption that our process is ergodic. What does this mean? To motivate it, let's construct a stationary process which does not satisfy the Strong Law of Large Numbers (SLLN) for trivial reasons. Let $\{X_n\}$ be i.i.d. with 0's and 1's, 0 with probability 3/4 and 1 with probability 1/4 and let $\{Y_n\}$ be i.i.d with 0's and 1's, 0 with probability 1/4 and 1 with probability 3/4. Now construct a stationary process by first flipping a fair coin and then if it's heads, take a realization from the $\{X_n\}$ process and if it's tails, take a realization from the $\{Y_n\}$ process. (By thinking of stationary processes as measures on sequence space, we are simply taking a convex (1/2,1/2) combination of the measures associated to the two processes $\{X_n\}$ and $\{Y_n\}$).

You should check that this resulting process is stationary with each marginal having mean 1/2. Letting Z_n denote this process, we clearly don't have that

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \to 1/2 \ \ a.s.$$

(which is what the SLLN would tell us if $\{Z_n\}$ were i.i.d.) but rather we clearly have that

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \to 3/4 \ (1/4) \text{ with probability } 1/2(1/2).$$

The point of ergodicity is to rule out such stupid examples. There are two equivalent definitions of ergodicity. The first is that if T is the transformation which shifts a bi-infinite sequence one step to the left, then any event which is invariant under T has probability 0 or 1. The other definition is that the measure (or distribution) on sequence space associated to the process is not a convex combination of 2 measures each also corresponding to some stationary process (i.e., invariant under T). It is not obvious that these are equivalent definitions but they are.

Exercise: Show that an i.i.d. process is ergodic by using the first definition.

Theorem 1.2 (Shannon-McMillan-Breiman Theorem): *Consider a stationary ergodic process with entropy H. Let $p(X_1, X_2, \ldots, X_n)$ be the probability that the process prints out $X_1, \ldots, X_n$. (You have to think about what this means BUT NOTE that it is a random variable). Then as $n \to \infty$,*

$$\frac{-\ln(p(X_1, X_2, \ldots, X_n))}{n} \to H a.s. \text{ and in } L^1.$$

NOTE: The fact that $E[\frac{-\ln(p(X_1,X_2,\ldots,X_n))}{n}] \to H$ is exactly (check this) the definition of entropy and therefore general real analysis tells us that the a.s. convergence implies the L^1–convergence.

We do not prove this now (we will in §2) but instead see how we can use it to do data compression. To do data compression, we will only need the above convergence in probability (which of course follows from either the a.s. or L^1 convergence). Thinking about what convergence in probability means, we have the following corollary.

Corollary 1.3: *Consider a stationary ergodic process with entropy H. Then for all $\epsilon > 0$, for large n, we have that the words of length n can be divided into two sets with all words C in the first set satisfying*

$$e^{-(H+\epsilon)n} < p(C) < e^{-(H-\epsilon)n}$$

and with the total measure of all words in the second set being $< \epsilon$.

One should think of the first set as the "good" set and the second set as the "bad" set. (Later it will be the good words which will be compressed).

Except in the i.i.d. uniform case, the bad set will have many more elements (in terms of cardinality) than the good set but its probability will be much lower. (Note that it is only in the i.i.d. uniform case that probabilities and cardinality are the same thing.) Here is a related result which will finally allow us to prove the data compression theorem.

Proposition 1.4: *Consider a stationary ergodic process with entropy H. Order the words of length n in decreasing order (in terms of their probabilities). Fix $\lambda \in (0,1)$. We select the words of length n in the above order one at a time until their probabilities sum up to at least λ and we let $N_n(\lambda)$ be the number of words we take to do this. Then*

$$\lim_{n\to\infty} \frac{\ln(N_n(\lambda))}{n} = H.$$

Exercise: Note that one obtains the same limit for all λ. Think about this. Convince yourself however that the above convergence cannot be uniform in λ. Is the convergence uniform in λ on compact intervals of $(0,1)$?

Proof: Fix $\lambda \in (0,1)$. Let $\epsilon > 0$ be arbitrary but $< 1-\lambda$ and λ. We will show both

$$\limsup_n \frac{\ln(N_n(\lambda))}{n} \leq H + \epsilon$$

and

$$\liminf_n \frac{\ln(N_n(\lambda))}{n} \geq H - \epsilon$$

from which the result follows.

By Corollary 1.3, we know that for large n, the words of length n can be broken up into 2 groups with all words C in the first set satisfying

$$e^{-(H+\epsilon)n} < p(C) < e^{-(H-\epsilon)n}$$

and with the total measure of all words in the second set being $< \epsilon$.

We now break the second group (the bad group) into 2 sets depending on whether $p(C) \geq e^{-(H-\epsilon)n}$ (the big bad words) or whether $p(C) \leq e^{-(H+\epsilon)n}$ (the small bad words). When we order the words in decreasing order, we clearly first go through the big bad words, then the good words and finally the small bad words.

We now prove the first inequality. For large n, the total measure of the good words is at least $1 - \epsilon$ which is bigger than λ and hence the total measure of the big bad words together with the good words is bigger than λ. Hence $N_n(\lambda)$ is at most the total number of big bad words plus the total number of good words. Since all these words have probability at least $e^{-(H+\epsilon)n}$, there cannot be more than $\lceil e^{(H+\epsilon)n} \rceil$ of them. Hence

$$\limsup_n \frac{\ln(N_n(\lambda))}{n} \leq \limsup_n \frac{\ln(e^{(H+\epsilon)n} + 1)}{n} = H + \epsilon.$$

For the other inequality, let $M_n(\lambda)$ be the number of good words among the $N_n(\lambda)$ words taken when we accumulated at least λ measure of the space. Since the total measure of the big bad words is at most ϵ, the total measure of these $M_n(\lambda)$ words is at least $\lambda - \epsilon$. Since each of these words has probability at most $e^{-(H-\epsilon)n}$, there must be at least $(\lambda - \epsilon)e^{(H-\epsilon)n}$ words in $M_n(\lambda)$ (if you don't see this, we have $|M_n(\lambda)|e^{-(H-\epsilon)n} \geq \lambda - \epsilon$). Hence

$$\limsup_n \frac{\ln(N_n(\lambda))}{n} \geq \limsup_n \frac{\ln(M_n(\lambda))}{n} \geq$$

$$\limsup_n \frac{\ln((\lambda - \epsilon)e^{(H-\epsilon)n})}{n} = H - \epsilon,$$

and we're done. □

Note that by taking λ close to 1, one can see that (as long as we are not i.i.d. uniform, i.e., as long as $H < \ln s$), we can cover almost all words (in the sense of probability) by using only a neglible percentage of the total number of words (e^{Hn} words instead of the total number of words $e^{(n \ln s)}$).

We finally arrive at data compression, now that we have things set up and have a little feeling about what entropy is. We now make some definitions.

Definition 1.5: *An n–**code** is an injective (i.e. 1-1) mapping σ which takes the set of words of length n with alphabet S to words of any finite length (of at least 1) with alphabet S.*

Definition 1.6: *Consider a stationary ergodic process $\{X_k\}_{k=-\infty}^{\infty}$. The **compressibility** of an n–code σ is $E[\frac{\ell(\sigma(X_1,\ldots,X_n))}{n}]$ where $\ell(W)$ denotes the length of the word W.*

This measures how well an n–code compresses **on average**.

Theorem 1.7: *Consider a stationary ergodic process $\{X_k\}_{k=-\infty}^{\infty}$ with entropy H. Let μ_n be the minimum compressibility over all n–codes. Then the limit $\mu \equiv \lim_{n\to\infty} \mu_n$ exists and equals $\frac{H}{\ln s}$.*

The quantity μ is called the **compression coefficient** of the process $\{X_k\}_{-\infty}^{\infty}$. Since H is always strictly less than $\ln s$ except in the i.i.d. uniform case, this says that data compression is always possible except in this case.

Proof: As usual, we have two directions to prove.

We first show that for arbitrary ϵ and $\delta > 0$, and for large n, $\mu_n \geq (1-\delta)\frac{H-2\epsilon}{\ln s}$.

Fix any n–code σ. Call an n–word C short (for σ) if $\sigma(C) \leq \frac{n(H-2\epsilon)}{\ln s}$. Since codes are 1-1, the number of short words is at most

$$s + s^2 + \ldots + s^{\lfloor \frac{n(H-2\epsilon)}{\ln s} \rfloor} \leq s^{\lfloor \frac{n(H-2\epsilon)}{\ln s} \rfloor} (\sum_{i=1}^{\infty} s^{-i}) \leq \frac{s}{s-1} e^{n(H-2\epsilon)}.$$

Next, since

$$\frac{\ln(N_n(\delta))}{n} \to H$$

by Proposition 1.4, for large n, we have that $N_n(\delta) > e^{n(H-\epsilon)}$. This says that for large n, if we take $< e^{n(H-\epsilon)}$ of the most probable words, we won't get δ total measure and so certainly if we take $< e^{n(H-\epsilon)}$ of any of the words, we won't get δ total measure. In particular, since the number of short words is at most

$$\frac{s}{s-1}e^{n(H-2\epsilon)}$$

which is for large n less than

$$e^{n(H-\epsilon)},$$

the short words can't cover δ total measure for large n, i.e., P(short word) $\leq \delta$ for large n. Now note that this argument was valid for any n–code. That is, for large n, any n-code satisfies P(short word) $\leq \delta$. Hence for any n-code σ

$$E[\frac{\sigma(C)}{n}] \geq (1-\delta)\left(\frac{(H-2\epsilon)}{\ln s}\right)$$

and so

$$\mu_n \geq (1-\delta)\frac{H-2\epsilon}{\ln s},$$

as desired.

For the other direction, we show that for any $\epsilon > 0$, we have that for large n, $\mu_n \leq 1/n\lceil\frac{n(H+\epsilon)}{\ln s}\rceil + \delta$. This will complete the argument.

The number of different sequences of length exactly $\lceil\frac{n(H+\epsilon)}{\ln s}\rceil$ is at least $s^{\frac{n(H+\epsilon)}{\ln s}}$ which is $e^{n(H+\epsilon)}$. By Proposition 1.4 the number $N_n(1-\delta)$ of most probable n–words needed to cover $1-\delta$ of the measure is $\leq e^{n(H+\epsilon)}$ for large n. Since there are at least this many words of length $\lceil\frac{n(H+\epsilon)}{\ln s}\rceil$, we can code these high probability words into words of length $\lceil\frac{n(H+\epsilon)}{\ln s}\rceil$. For the remaining words, we code them to themselves. For such a code σ, we have that $E[\sigma(X_1, \ldots, X_n)]$ is at most $\lceil \frac{n(H+\epsilon)}{\ln s} \rceil + \delta n$ and hence the compression of this code is at most $1/n \lceil \frac{n(H+\epsilon)}{\ln s} \rceil + \delta$. □

This is all nice and dandy but

EXERCISE: Show that for a real person, who really has data and really wants to compress it, the above is all useless.

We will see later on (§3) a method which is not useless.

2 Proof of the Shannon-McMillan-Breiman Theorem

We provide here the classical proof of this theorem where we will assume two major theorems, namely, the ergodic theorem and the Martingale Convergence Theorem. This section is much more technical than all of the coming sections and requires more mathematical background. Feel free if you wish to move on to §3. As this section is independent of all of the others, you won't miss anything.

We first show that the proof of this result is an easy consequence of the ergodic theorem in the special case of an i.i.d. process or more generally of a multistep Markov chain. We take the middle road and prove it for Markov chains (and the reader will easily see that it extends trivially to multi-step Markov chains).

Proof of SMB for Markov Chains: First note that

$$-\frac{\ln(p(X_1, X_2, \ldots, X_n))}{n} = -\frac{\sum_1^n \ln p(X_j | X_{j-1})}{n}$$

where the first term $\ln p(X_1|X_0)$ is taken to be $\ln p(X_1)$. If the first term were interpreted as $\ln p(X_1|X_0)$ instead, then the ergodic theorem would immediately tell us this limit is a.s. $E[-\ln p(X_1|X_0)]$ which (you have seen as an exercise) is the entropy of the process. Of course this difference in the first term makes an error of at most *const*/n and the result is proved. □

For the multistep markov chain, rather than replacing the first term by something else (as above), we need to do so for the first k terms where k is the look-back of the markov chain. Since this k is fixed, there is no problem.

To prove the general result, we isolate the main idea into a lemma which is due to Breiman and which is a generalization of the ergodic theorem which is also used in its proof. Actually, to understand the statement of the next result, one really needs to know some ergodic theory. Learn about it on your own, ask me for books, or I'll discuss what is needed or just forget this whole section.

Proposition 2.1: *Assume that $g_n(x) \to g(x)$ a.e. and that*

$$\int \sup_n |g_n(x)| < \infty.$$

Then if ϕ is an ergodic measure preserving transformation, we have that

$$\lim_{n\to\infty} \frac{1}{n} \sum_0^{n-1} g_j(\phi^j(x)) = \int g \text{ a.e. and in } L^1.$$

Note if $g_n = g \in L^1$ for all n, then this is exactly the ergodic theorem.

Proof: (as explained in the above remark), the ergodic theorem tells us

$$\lim_{n\to\infty} \frac{1}{n} \sum_0^{n-1} g(\phi^j(x)) = \int g \text{ a.e. and in } L^1.$$

Since

$$\frac{1}{n}\sum_{0}^{n-1} g_j(\phi^j(x)) = \frac{1}{n}\sum_{0}^{n-1} g(\phi^j(x)) + \frac{1}{n}\sum_{0}^{n-1}[g_j(\phi^j(x)) - g(\phi^j(x))],$$

we need to show that

$$(*)\frac{1}{n}\sum_{0}^{n-1} |g_j(\phi^j(x)) - g(\phi^j(x))| \to 0 \text{ a.e. and in } L^1.$$

Let $F_N(x) = \sup_{j\geq N} |g_j(x) - g(x)|$. By assumption, each F_N is integrable and goes to 0 monotonically a.e. from which dominated convergence gives $\int f_N \to 0$.

Now,

$$\frac{1}{n}\sum_{0}^{n-1} |g_j(\phi^j(x)) - g(\phi^j(x))| \leq \frac{1}{n}\sum_{0}^{N-1} |g_j(\phi^j(x)) - g(\phi^j(x))| +$$

$$\frac{n-N-1}{n}\frac{1}{n-N-1}\sum_{0}^{n-N-1} f_N(\phi^j\phi^N(x)).$$

For fixed N, letting $n \to \infty$, the first term goes to 0 a.e. while the second term goes to $\int f_N$ a.e. by the ergodic theorem. Since $\int f_N \to 0$, this proves the pointwise part of (*). For the L^1 part, again for fixed N, letting $n \to \infty$, the L^1 norm of the first term goes to the 0 while the L^1 norm of the second term is always at most $\int f_N$ (which goes to 0) and we're done. □

Proof of the SMB Theorem: To properly do this and to apply the previous theorem, we need to set things up in an ergodic theory setting. Our process X_n gives us a measure μ on $U = \{1, \ldots, S\}^{\mathbf{Z}}$ (thought of as its distribution function). We also have a transformation T on U, called the left shift, which simply maps an infinite sequence 1 unit to the left. The fact that the process is stationary means that T preserves the measure μ in that for all events A, $\mu(TA) = \mu(A)$.

Next, letting $g_k(w) = -\ln p(X_1|X_0, \ldots, X_{-k+1})$,
(with $g_0(w) = -\ln p(X_1)$), we get immediately that

$$-\frac{\ln(p(X_1(w), X_2(w), \ldots, X_n(w)))}{n} = \frac{1}{n}\sum_{0}^{n-1} g_j(T^j(w)).$$

(Note that w here is an infinite sequence in our space X.) We are now in the set-up of the previous proposition although of course a number of things need to be verified, the first being the convergence of the g_k's to something. Note that for any i in our alphabet $\{1, \ldots, S\}$, $g_k(w)$ on the cylinder set $X_1 = i$ is simply $-\ln P(X_1 = i|X_0, \ldots, X_{-k+1})$ which (here we use a second nontrivial result) converges by the Martingale convergence theorem (and the continuity of ln) to $-\ln P(X_1 = i|X_0, \ldots)$. We therefore get that

$$g_k(w) \to g(w)$$

where $g(w) = -\ln P(X_1 = i|X_0, \ldots)$ on the cylinder set $X_1 = i$, i.e.,
$g(w) = -\ln P(X_1|X_0, \ldots)$.

If we can show that $\int \sup_k |g_k(w)| < \infty$, the previous proposition will tell us that

$$-\frac{\ln(p(X_1, X_2, \ldots, X_n))}{n} \to \int g \text{ a.e. and in } L^1.$$

Since, by definition, $E[-\frac{\ln(p(X_1,X_2,\ldots,X_n))}{n}] \to h(\{X_i\})$ by definition of entropy, it follows that $\int g = h(\{X_i\})$ and we would be done.

EXERCISE: Show directly that $E[g(w)]$ is $h(\{X_i\})$.

We now proceed with verifying $\int \sup_k |g_k(w)| < \infty$. Fix $\lambda > 0$. We show that $P(\sup_k g_k > \lambda) \leq se^{-\lambda}$ (where s is the alphabet size) which implies the desired integrability.

$P(\sup_k g_k > \lambda) = \sum_k P(E_k)$ where E_k is the set where the first time g_j gets larger than λ is k, i.e., $E_k = \{g_j \leq \lambda, j = 0, \ldots k-1, g_k > \lambda\}$. These

are obviously disjoint for different k and make up the set above. Now, $P(E_k) = \sum_i P((X_1 = i) \cap E_k) = \sum_i P((X_1 = i) \cap F_k^i)$ where F_k^i is the set where the first time f_j^i gets larger than λ is k where
$f_j^i = -\ln p(X_1 = i|X_0, \ldots, X_{-j+1})$.

Since F_k^i is $X_{-k+1}, \ldots, X_0$–measurable, we have

$$P((X_1 = i) \cap F_k^i) = \int_{F_k^i} P(X_1 = i|X_{-k+1}, \ldots, X_0)$$
$$= \int_{F_k^i} e^{-f_k^i} \leq e^{-\lambda} P(F_k^i).$$

Since the F_k^i are disjoint for different k (but not for different i),
$\sum_k P(E_k) \leq \sum_i \sum_k e^{-\lambda} P(F_k^i) \leq s e^{-\lambda}$. □

We finally mention a couple of things in the higher–dimensional case, that is, where we have a stationary random field where entropy can be defined in a completely analogous way. It was unknown for some time if a (pointwise, that is, a.s.) SMB theorem could be obtained in this case, the main obstacle being that it was known that a natural candidate for a multidimensional martingale convergence theorem was in fact false and so one could not proceed as we did above in the 1-d case. (It was however known that this theorem held in "mean" (i.e., in L^1), see below.) However, recently, (in 1983) Ornstein and Weiss managed to get a pointwise SMB not only for the multidimensional lattice but in the more general setting of something called amenable groups. (They used a method which circumvents the Martingale Convergence Theorem).

The result of Ornstein and Weiss is difficult and so we don't present it. Instead we present the Shannon–McMillan Theorem (note, no McMillan here) which is the L^1 convergence in the SMB Theorem. Of course, this is a weaker result than we just proved but the point of presenting it is two-fold, first to see how much easier it is than the pointwise version and also to see a result which generalizes (relatively) easily to higher dimensions (this higher dimensional variant being left to the reader).

Proposition 2.2 (Mean SMB Theorem or MB Theorem): *The SMB theorem holds in L^1.*

Proof Let g_j be defined as above. The Martingale convergence theorem implies that $g_j \to g_\infty$ in L_1 as $n \to \infty$. We then have (with $\|\|$ denoting the L_1 norm),

$$\| - \frac{\ln(p(X_1, X_2, \ldots, X_n))}{n} - H\| = \|\frac{1}{n}\sum_0^{n-1} g_j(T^j(w)) - H\| \leq$$

$$\|\frac{1}{n}\sum_0^{n-1} |g_j(T^j(w)) - g_\infty(T^j(w))|\| + \|\frac{1}{n}\sum_0^{n-1} g_\infty(T^j(w)) - H\|.$$

The first term goes to 0 by the L_1 convergence in the Martingale Convergence Theorem (we don't of course always get L_1 convergence in the Martingale Convergence Theorem but we have so in this setting). The second term goes to 0 in L^1 by the ergodic theorem together with the fact that $\int g_\infty = H$, an easy computation left to the reader. □

3 Universal Entropy Estimation?

This section simply raises an interesting point which will be delved into further later on.

One can consider the problem of trying to estimate the entropy of a process $\{X_n\}_{n=-\infty}^{\infty}$ after we have only seen $X_1, X_2, \ldots, X_n$ in such a way that these estimates converge to the true entropy with prob 1. (As above, we know that to have any chance of doing this, one needs to assume ergodicity).

Obviously, one could simply take the estimate to be

$$\frac{-\ln(p(X_1, X_2, \ldots, X_n))}{n}$$

which will (by SMB) converge to the entropy. However, this is clearly undesirable in the sense that in order to use this estimate, we need to know what the distribution of the process is. It certainly would be preferable to find some estimation procedure which does not depend on the distribution since we might not know it. We therefore want a family of functions $h_n : S^n \to \mathbf{R}$ ($h_n(x_1, \ldots, x_n)$ would be our guess of the entropy of the process if we see the data $x_1, \ldots, x_n$) such that for any given ergodic process $\{X_n\}_{n=-\infty}^{\infty}$,

$$\lim_{n\to\infty} h_n(X_1, \ldots, X_n) = h(\{X_n\}_{n=-\infty}^{\infty}) \text{ a.s. } .$$

(Clearly the suggestion earlier is not of this form since h_n depends on the process). We call such a family of functions a "universal entropy estimator", universal since it holds for all ergodic processes.

There is no immediate obvious reason at all why such a thing exists and there are similar types of things which don't in fact exist. (If entropy were a continuous function defined on processes (the latter being give the usual weak topology), a general theorem (see §9) would tell us that such a thing exists BUT entropy is NOT continuous.) However, as it turns out, a universal entropy estimator does in fact exist. One such universal entropy estimator comes from an algorithm called the "Lempel-Ziv algorithm" which was later improved by Wyner-Ziv. We will see that this algorithm also turns out to be a universal data compressor, which we will describe later. (This method is used, as far as I understand it, in the real world).

4 20 questions and Relative Entropy

We have all heard problems of the following sort.

You have 8 coins one of which has a different weight than all the others which have the same weight. You are given a scale (which can determine which of two given quantities is heavier) and your job is to identify the "bad" coin. How many weighings do you need?

Later we will do something more general but it turns out the problem of how many questions one needs to ask to determine something has an extremely simple answer which is given by something called Kraft's Theorem. The answer will not at all be surprising and will be exactly what you would guess, namely if you ask questions which have D possible answers, then you will need $\ln_D(n)$ questions where n is the total number of possibilities. (If the latter is not an integer, you need to take the integer right above it).

Definitin 4.1: *A mapping C from a finite set S to finite words with alphabet $\{1, \ldots, D\}$ will be called a* **prefix-free D–code on S** *if for any $x, y \in S$, $C(x)$ is not a prefix of $C(y)$ (which implies of course that C is injective)*

Theorem 4.2 (Kraft's Inequality): *Let C be a prefix-free D–code defined on S. Let $\{\ell_1, \ldots \ell_{|S|}\}$ be the lengths of the code words of C. Then*

$$\sum_{i=1}^{|S|} D^{-\ell_i} \leq 1.$$

Conversely, if $\{\ell_1, \ldots \ell_n\}$ are integers satisfying

$$\sum_{i=1}^{n} D^{-\ell_i} \leq 1,$$

then there exists a prefix-free D–code on a set with n elements whose code words have length $\{\ell_1, \ldots \ell_n\}$.

This turns out to be quite easy to prove so try to figure it out on your own. I'll present it in class but won't write it here.

Corollary 4.3: *There is a prefix–free D–code on a set of n elements whose maximum code word has length $\lceil \ln_D(n) \rceil$ while there is no such code whose maximum code word is $< \lceil \ln_D(n) \rceil$.*

I claim that this corollary implies that the number of questions with D possible answers one needs to ask to determine something with n possibilities is $\ln_D(n)$ (where if this is not an integer, it is taken to be the integer right above it). To see this one just needs to think a little and realize that asking questions until we get the right answer is exactly a prefix–free D–code. (For example, if you have the code first, your first question would be "What is the first number in the codeword assigned to x?" which of course has D possible answers.) Actually, going back to the weighing question, we did not really answer that question at all. We understand the situation when we can ask any D–ary question (i.e., any partition of the outcome space into D pieces) but with questions like the weighing question, while any weighing breaks the outcome space into 3 pieces, we are physically limited (I assume, I have not thought about it) from constructing all possible partitions of the outcome space into 3 pieces. So this investigation, while interesting and useful, did not allow us to answer the original question at all but we end there nonetheless.

Now we want to move on and do something more general. Let's say that X is a random variable taking on finitely many values $\mathcal{X}$ with probabilities $p_1, \ldots, p_n$ (so $|\mathcal{X}| = n$). Now we want a successful guessing scheme (i.e., a prefix–free code) where the expected number of questions we need to ask (rather than the maximum number of questions) is not too large. The following theorem answers this question and more. For simplicity, we talk about prefix–free codes.

Theorem 4.4: *Let C be a prefix–free D–code on $\mathcal{X}$ and N the length of the codeword assigned to X (which is a r.v.). Then*

$$H_D(X) \leq E[N]$$

where $H_D(X)$ is the D–entropy of X given by $-\sum_{i=1}^{k} p_i \ln_D p_i$.

Conversely, there exists a prefix–free D–code satisfying

$$E[N] < H_D(X) + 1.$$

(We will prove this below but must first do a little more development.)

EXERCISE: Relate this result to Corollary 4.3.

An important concept in information theory which is useful for many purposes and which will allow us to prove the above is the concept of relative entropy. This concept comes up in many other places–below we will explain three other places where it arises which are respectively

(1) If I construct a prefix–free code satisfying $E[N] < H_D(X) + 1$ under the assumption that the true distribution of X is $q_1, \ldots, q_n$ but it turns out the true distribution is $p_1, \ldots, p_n$, how bad will $E[N]$ be?

(2) (level 2) large deviation theory for the Glivenko–Cantelli Theorem (which is called Sanov's Theorem)

(3) universal data compression for i.i.d. processes.

The first 1 of these will be done in this section, the 2nd in the §5 while the last will be done in §6.

Definition 4.5: *If $p = p_1, \ldots, p_n$ and $q = q_1, \ldots, q_n$ are two probability distributions on X, the* **relative entropy** *or* **Kullback Leiber distance** *between p and q is*

$$\sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}.$$

Note that if for some i, $p_i > 0$ and $q_i = 0$, then the relative entropy is ∞. One should think of this as a metric BUT it's not-it's neither symmetric nor satisfies the triangle inequality. One actually needs to specify the base of the logarithm in this definition.

The first thing we need is the following whose proof is an easy application of Jensen's inequality which is left to the reader.

Theorem 4.6: *Relative entropy (no matter which base we use) is nonnegative and 0 if and only if the two distributions are the same.*

Proof of Theorem 4.4: Let C be a prefix–free code with ℓ_i denoting the length of the ith codeword and N denoting the length of the codeword as a r.v.. Simple manipulation gives

$$E[N] - H_D(X) = \sum_i p_i \ln_D\left(\frac{p_i}{D^{-\ell_i}/\sum_i D^{-\ell_i}}\right) - \ln_D\left(\sum_i D^{-\ell_i}\right).$$

The first sum is nonnegative since it's a relative entropy and the the second term is nonnegative by Kraft's Inequality.

To show there is a prefix–free code satisfying $E[N] < H_D(X) + 1$, we simply construct it. Let $\ell_i = \lceil \ln_D(1/p_i) \rceil$. A trivial computation shows that the ℓ_i's satisfy the conditions of Kraft's Inequality and hence there is a prefix–free D code which assigns the ith account a codeword of length ℓ_i. A trivial calculation shows that $E[N] < H_D(X) + 1$. □

How did we know to choose $\ell_i = \lceil \ln_D(1/p_i) \rceil$? If the integers ℓ_i were allowed to be nonintegers, Lagrange Multipliers show that the ℓ_i's minimizing $E[N]$ subject to the contraint of Kraft's inequality are $\ell_i = \ln_D(1/p_i)$.

There is an easy corollary to Theorem 4.4 whose proof is left to the reader, which we now give.

Corollary 4.7: *The mimimum expected codeword length L_n^* per symbol satisfies*

$$\frac{H(X_1, \ldots, X_n)}{n} \leq L_n^* \leq \frac{H(X_1, \ldots, X_n)}{n} + \frac{1}{n}.$$

Moreover, if $X_1, \ldots$ is a stationary process with entropy H, then

$$L_n^* \to H.$$

This particular prefix–free code, called the Shannon Code is not necessarily optimal (although it's not bad). There is an optimal code called the Huffman code which I will not discuss. It turns out also that the Shannon Code is close to optimal in a certain sense (and in a better sense than that the mean code length is at most 1 from optimal which we know).

We now discuss our first application of relative entropy (besides our using it in the proof of Theorem 4.4). We mentioned the following question above. If I contruct the Shannon prefix–free code satisfying $E[N] < H_D(X) + 1$ under the assumption that the true distribution of X is $q_1, \ldots, q_n$ but it turns out the true distribution is $p_1, \ldots, p_n$, how bad will $E[N]$ be? The answer is given by the following theorem.

Theorem 4.8: *The expected length $E_p[N]$ under $p(x)$ of the Shannon code*

$\ell_i = \lceil \ln_D(1/q_i) \rceil$ *satisfies*

$$H_D(p) + D(p\|q) \le E_p[N] < H_D(p) + D(p\|q) + 1.$$

Proof: Trivial calculation left to the reader. □

5 Sanov's Theorem: another application of relative entropy

In this section, we give an important application (or interpretation) of relative entropy, namely (level 2) large deviation theory.

Before doing this, let me remind you (or tell you) what (level 1) large deviation theory is (this is the last section of the first chapter in Durrett's probability book).

(Level 1) large deviation theory simply says the following modulo the technical assumptions. Let X_i be i.i.d. with finite mean m and have a moment generating function defined in some neighborhood of the origin. First, the WLLN tells us

$$P(|\frac{\sum_{i=1}^n X_i}{n} - m| \ge \epsilon) \to 0$$

as $n \to \infty$ for any ϵ. (For this, we of course do not need the exponential moment). (Level 1) large deviation theory tells us how fast the above goes to 0 and the answer is, it goes to 0 exponentially fast and more precisely, it goes like $e^{-f(\epsilon)n}$ where $f(\epsilon) > 0$ is the so-called Frechel transform of the logarithm of the moment generating function given by

$$f(\epsilon) = \max_\theta(\theta\epsilon - \ln(E[e^{\theta X}])).$$

Level two large deviation theory deals with a similar question but at the level of the empirical distribution function. Let X_i be i.i.d. again and let F_n be the empirical distribution of $X_1, \ldots, X_n$. Glivenko–Cantelli tells us that $F_n \to F$ a.s. where F is the the common distribution of the X_i's. If C is a closed set of measures not containing F, it follows (we're now doing weak convergence of probability measures on the space of probability measures on $\mathbf{R}$ so you have to think carefully) that

$$P(F_n \in C) \to 0$$

and we want to know how fast. The answer is

$$\lim_{n\to\infty} -\frac{1}{n} \ln(P(F_n \in C)) = \min_{P\in C} D(P\|F).$$

The above is called Sanov's Theorem which we now prove.

There will be a fair development before getting to this which I feel will be useful and worth doing (and so this might not be the quickest route to Sanov's Theorem). All of this material is taken from Chapter 12 of Cover and Thomas's book.

If $x = x_1, \ldots x_n$ is a sample from an i.i.d. process with distribution Q, then P_x, which we call the type of x, is defined to be the empirical distribution (which I will assume you know) of x.

Definition 5.1: *We will let $\mathcal{P}_n$ denote the set of all types (i.e., empirical distributions) from a sample of size n and given $P \in \mathcal{P}_n$, we let $T(P)$ be all $x = x_1, \ldots x_n$ (i.e., all samples) which have type (empirical distribution) P.*

EXERCISE: If Q has its support on $\{1, \ldots, 7\}$, find $|T(P)|$ where P is the type of 11321.

Lemma 5.2: $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$ *where* $\mathcal{X}$ *is the set of possible values are process can take on.*

This lemma, while important, is trivial and left to the reader.

Lemma 5.3: *Let* $X_1, \ldots$ *be i.i.d. according to* Q *and let* P_x *denote the type of* x*. Then*

$$Q^n(x) = 2^{-n(H(P_x)+D(P_x||Q))}$$

and so the Q^n *probability of* x *depends only on its type.*

Proof: Trivial calculation (although a few lines) left to the reader. □

Lemma 5.4: *Let* $X_1, \ldots$ *be i.i.d. according to* Q*. If* x *is of type* Q*, then*

$$Q^n(x) = 2^{-nH(Q)}.$$

Our next result gives us an estimate of the size of a type class $T(P)$.

Lemma 5.5:

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}.$$

Proof: One method is to write down the exact cardinality of the set (using elementary combinatorics) and then apply Stirling's formula. We however proceed differently.

The upper bound is trivial. Since each $x \in T(P)$ has (by the previous corollary) under P, probability $2^{-nH(P)}$, there can be at most $2^{nH(P)}$ elements in $T(P)$.

For the lower bound, the main step is is to show that type class P has the largest probability of all type classes (note however, that an element from type class P does not necessarily have a higher probability than elements from other type classes as you should check), i.e.,

$$P^n(T(P)) \geq P^n(T(P'))$$

for all P'. I leave this to you (although it takes some work, see the book if you want).

Then we argue as follows. Since there are at most $(n+1)^{|\mathcal{X}|}$ type classes, and $T(P)$ has the largest probability, its P^n–probability must be at least $\frac{1}{(n+1)^{|\mathcal{X}|}}$. Since each element of this type class has P^n–probability $2^{-nH(P)}$, we obtain the lower bound. □

Combining Lemmas 5.3 and 5.5, we immediately obtain

Corollary 5.6:

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}.$$

Remark: The above tells us that if we flip n fair coins, the probability that we get exactly 50% heads, while going to 0 as $n \to \infty$, does not go to 0 exponentially fast.

The above discussion has set us up for an easy proof of the Glivenko–Cantelli Theorem (analogous to the fact that once one has level 1 large deviation theory set up, the Strong Law of Large Numbers follows trivially).

The key that makes the whole thing work is the fact (Corollary 5.6) that the probability under Q that a certain type class (different from Q) arises is exponentially small (the exponent being given by the relative entropy) and the fact that there are only polynomially many type classes.

Theorem 5.7: *Let $X_1, \ldots$ be i.i.d. according to P. Then*

$$P(D(P_{x_n}||P) > \epsilon) \leq 2^{-n(\epsilon - |\mathcal{X}|\frac{\ln(n+1)}{n})}$$

and hence (by Borel Cantelli)

$$D(P_{x_n}||P) \to 0 \text{ a.s. } .$$

Proof: Using

$$P^n(D(P_{x_n}||P) > \epsilon) \leq \sum_{Q:D(Q||P)\geq\epsilon} 2^{-nD(Q||P)} \leq$$

$$(n+1)^{|\mathcal{X}|}2^{-n\epsilon} = 2^{-n(\epsilon - |\mathcal{X}|\frac{\ln(n+1)}{n})}.$$

□

We are now ready to state Sanov's Theorem. We are doing everything under the assumption that the underlying distribution has finite support (i.e., there are only a finite number of values our r.v. can take on). One can get rid of this assumption but we stick to it here since it gets rid of some technicalities.

We can view the set of all prob. measures (call it $\mathcal{P}$) on our finite outcome space (of size $|\mathcal{X}|$) as a subspace of $\mathbf{R}^{|\mathcal{X}|}$ (or even of $\mathbf{R}^{|\mathcal{X}|-1}$). (In the former case, it would be simplex in the positive cone.) This immediately gives us a nice topology on these measures (which of course is nothing but the weak topology in a simple setting). Assuming that P gives every $x \in \mathcal{X}$ positive measure (if not, change $\mathcal{X}$), note that $D(Q||P)$ is a continuous function of Q on $\mathcal{P}$ and hence on any closed (and hence compact) set E in $\mathcal{P}$, this function assumes a minimum (which is not 0 if $P \notin E$).

Theorem 5.8 (Sanov's Theorem): *Let $X_1, \ldots$ be i.i.d. with distribution P and E be a closed set in $\mathcal{P}$. Then*

$$P^n(E)(= P^n(E \cap \mathcal{P}_n)) \leq (n+1)^{|\mathcal{X}|} 2^{-nf(E)}$$

where $f(E) = \min_{Q \in E} D(Q||P)$. ($P^n(E)$ means of course $P^n(\{x : \ P_x \in E\})$.)

If, in addition, E is the closure of its interior, then the above exponential upper bound also gives a lower bound in that

$$\lim_{n \to \infty} \frac{1}{n} \ln P^n(E) = -f(E).$$

Proof: The derivation of the upper bound follows easily from Theorem 5.7 which we leave to the reader. The lower bound is not much harder but a little care is needed (as should be clear from the topological assumption on E). We need to be able to find guys in $E \cap \mathcal{P}_n$. Since the $\mathcal{P}_n$'s become "more and more" dense as n gets large, it is easy to see (why?) that $E \cap \mathcal{P}_n \neq \emptyset$ for large n (this just uses the fact that E has nonempty interior) and that there exist $Q_n \in E \cap \mathcal{P}_n$ with $Q_n \to Q^*$ where Q^* is some distribution which minimizes $D(Q||P)$ over $Q \in E$. This immediately give that $D(Q_n||P) \to D(Q^*||P)(= f(E))$.

We now have

$$P^n(E) \geq P^n(T(Q_n)) \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(Q_n||P)}$$

from which one immediately concludes that

$$\liminf_n \frac{1}{n} \ln P^n(E) \geq \liminf_n -D(Q_n||P) = -f(E).$$

The first part gives this as an upper bound and so we're done. □

Note that the proof goes through if one of the Q^*'s which minimize $D(Q||P)$ over $Q \in E$ (of which there must be at least one due to compactness) is in the closure of the interior of E. It is in fact also the case (although not att all obvious and related to other important things) that if E is also convex, then there is a unique minimizing Q^* and so there is a "Hilbert space like" picture here. (If E is not convex, it is easy to see that there is not necessarily a unique minimizing guy). The fact that there is a unique minimizing guy (in the convex case) is suggested by (by not implied by) the convexity of relative entropy in its two arguments.

Hard Exercise: Recover the level 1 large deviation theory from Sanov's Theorem using Lagrange multipliers.

6 Universal Entropy Estimation and Data Compression

We first use relative entropy to prove the possibility of universal data compression for i.i.d. processes. Later on, we will do the much more difficult universal data compression for general ergodic soures, the so–called Ziv-Lempel algorithm which is used in the real world (as far as I understand it). Of course, from a practical point of view, the results of Chapter 1 are useless since typically one does NOT know the process that one is observing and so one wants a "universal" way to compress data. We will now do this for i.i.d. processes where it is quite easy using the things we derived in §5.

We have seen previously in §4 that if we know the distribution of a r.v. x, we can encode x by assigning it a word which has length $-\ln(p(x))$. We have also seen if the distribution is really q, we get a penalty of $D(p||q)$. We now find a code "of rate R" which suffices for all i.i.d. processes of entropy less than R.

Definition 6.1: *A* **fixed rate block code of rate** R *for a process* $X_1, \ldots$ *with known state space* $\mathcal{X}$ *consists of two mappings which are the encoder,*

$$f_n : \mathcal{X}^n \to \{1, \ldots, 2^{nR}\}$$

and a decoder

$$\phi_n : \{1, \ldots, 2^{nR}\} \to \mathcal{X}^n.$$

We let

$$P_e^{(n,Q)} = Q^n(\phi_n(f_n(X_1, \ldots, X_n)) \neq (X_1, \ldots, X_n))$$

be the probability of an error in this coding scheme under the assumption that Q is the true distribution of the x_i's.

Theorem 6.2: *There exists a fixed rate block code of rate R such that for all Q with $H(Q) < R$,*

$$P_e^{(n,Q)} \to 0 \text{ as } n \to \infty.$$

Proof: Let $R_n = R - |\mathcal{X}|\frac{\ln(n+1)}{n}$ and let $A_n = \{x \in \mathcal{X}^n : H(P_x) \leq R_n\}$. Then

$$|A_n| = \sum_{P \in \mathcal{P}_n, H(P) \leq R_n} |T(P)| \leq$$

$$(n+1)^{|\mathcal{X}|} 2^{nR_n} = 2^{nR}.$$

Now we index the elements of A_n in an arbitrary way and we let $f_n(x) =$ index of x in A_n if $x \in A_n$ and 0 otherwise. The decoding map ϕ_n takes of course the index into corresponding word. Clearly, we obtain an error (at the nth stage) if and only if $x \notin A_n$. We then get

$$P_e^{(n,Q)} = 1 - Q^n(A_n) = \sum_{P: H(P) > R_n} Q^n(T(P)) \leq$$

$$\sum_{P: H(P) > R_n} 2^{-nD(P||Q)} \leq (n+1)^{|\mathcal{X}|} 2^{-n \min_{P: H(P) > R_n} D(P||Q)}.$$

Now $R_n \to R$ and $H(Q) < R$ and so $R_n > H(Q)$ for large n and so the exponent $\min_{P: H(P) > R_n} D(P||Q)$ is bounded away from 0 and so $P_e^{(n,Q)}$ goes exponentially to 0 for fixed Q. Actually, this last argument was a little sloppy and one should be a little more careful with the details. One first should note that compactness together with the fact that $D(P||Q)$ is 0 only when $P = Q$ implies that if we stay away from any neighborhood of Q, $D(P||Q)$ is bounded away from 0. Then note that if $R_n > H(Q)$, the continuity of the entropy function tells us that $\{P : H(P) > R_n\}$ misses some neighborhood of Q and hence $\min_{P: H(P) > R_n} D(P||Q) > 0$ and clearly this is increasing in n (look at the derivative of $\frac{\ln x}{x}$). □

Remarks: (1) Technically, 0 is not an allowable image. Fix this (it's trivial).
(2) For fixed Q, since the above probability goes to exponentially fast, Borel–Cantelli tells us the coding scheme will work eventually forever (i.e., we have an a.s. statement rather than just an "in probability" statement).
(3) There is no uniformity in Q but there is a uniformity in Q for $H(Q) \leq R - \epsilon$. (Show this).
(4) For Q with $H(Q) > R$, the above guessing scheme works with probability going to 0. Could there be a different guessing scheme which works?

The above theorem produces for us a “universal data compressor” for i.i.d. sequences.

Exercise: Construct an entropy estimator which is universal for i.i.d. sequences.

We now construct a “universal entropy estimator” which is universal for all ergodic processes. We will now construct something which is not techniquely a “universal entropy estimator” according to the definition given in §2 but seems analogous and can easily be modified to be one. We will later discuss how this procedure is related to a universal data compressor.

Theoreom 6.3: *Let* $(x_1, x_2, \ldots)$ *be an infinite word and let*

$$R_n(x) = \min\{j \geq n : x_1, x_2, \ldots, x_n = x_{j+1}x_{j+2}\ldots x_{j+n}\}$$

(which is the first time after n *when the first* n*–block repeats itself). Then*

$$\lim_{n\to\infty} \frac{\log R_n(X_1, X_2, \ldots)}{n} = H a.s.$$

where H *is the entropy of the process.*

Rather then writing out the proof of that, we will read the paper “Entropy and Data Compression Schemes” by Ornstein and Weiss. This paper will be much more demanding than these notes have been up to now.

7 Shannon’s Theorems on Capacity

In this section, we discuss and prove the fundamental theorem of channel capacity due to Shannon in his groundbreaking work in the field. Before going into the mathematics, we should first give a small discussion.

Assume that we have a channel to which one can input a 0 or a 1 and which outputs a 0 or 1 from which we want to recover the input. Let’s assume that the channel transmits the signal correctly with probability $1 - p$ and corrupts it (and therefore sends the other value) with probability p. Assume that $p < 1/2$ (if $p = 1/2$, the channel clearly would be completely useless). Based on the output we receive, our best guess at the input would of course be the output in which case the probability of making an error would be p. If we want to decrease the probability of making an error, we could do the following. We could simply send our input through the channel 3 times and then based on the 3 outputs, we would guess the input to be that output which occurred the most times. This coding/decoding scheme will work as long as the channel does not make 2 or more errors and the probability of this is (for small p) much much smaller ($\approx p^2$) than if we were to send just 1 bit through. Great! The probability is much smaller now BUT there is a price to pay for this. Since we have to send 3 bits through the channel to have one input bit guess, the "rate of transmission" has suddenly dropped to 1/3. Anyway, if we are still dissatisfied with the probability of this scheme making a error, we could send the bit through 5 times (using an analogous majority rule decoding scheme) thereby decreasing the probability of a decoding mistake much more ($\approx p^3$) but then the transmission rate would go even lower to 1/5. It seems that in order to achieve the probability of error going to 0, we need to drop the rate of transmission closer and closer to 0. The amazing discovery of Shannon is that this very reasonable statement is simply false. It turns out that if we are simply willing to use a rate of transmission which is less than "channel capacity" (which is simply some number depending only on the channel (in our example, this number is strictly positive as long as $p \neq \frac{1}{2}$ which is reasonable)), then as long as we send long strings of input, we can transmit them with arbitrarily low probability of error. In particular, the rate need not go to 0 to achieve arbitrary reliability.

We now enter the mathematics which explains more precisely what the above is all about and proves it. The formal definition of capacity of a channel is based on a concept called mutual information of two r.v.'s which is denoted by $I(X;Y)$.

Definition 7.1: *$I(X;Y)$, called the* **mutual information** *of X and Y, is the relative entropy of the joint distribution of X and Y and the product distribution of X and Y.*

Note that $I(X;Y)$ and $I(Y;X)$ are the same and that they are not infinite (although recall in general relative entropy can be infinite).

Definition 7.2: *The* **conditional entropy** *of Y given X, denoted $H(Y|X)$ is defined to be $\sum_x P(X=x)H(Y|X=x)$.*

Proposition 7.3:
(1)$H(X,Y) = H(X) + H(Y|X)$
(2) $I(X;Y) = H(X) - H(X|Y)$.

The proof of these are left to the reader. There is again no idea involved, simply computation.

Definition 7.4: *A* **discrete memoryless channel** *(DMC) is a system which takes as input elements of an input alphabet $\mathcal{X}$ and prints out elements from an output alphabet $\mathcal{Y}$ randomly according to some $p(y|x)$ (i.e., if x is sent in, then y comes out with probability $p(y|x)$). Of course the numbers $p(y|x)$ are nonnegative and for fixed x give 1 if we sum over y. (Such a thing is sometimes called a* **kernel** *in probability theory and could be viewed as* *a markov transition matrix if the sets $\mathcal{X}$ and $\mathcal{Y}$ were the same (which they need not be)).*

Definition 7.5: *The* **information channel capacity** *of a discrete memoryless channel is*

$$C = \max_{p(x)} I(X;Y).$$

Exercise: Compute this for different examples like $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and the channel being $p(x|x) = p$ for all x.

We assume now that if we send a sequence through the channel, the outputs are all independently chosen from $p(y|x)$ (i.e., the outputs occur independently, an assumption you might object to).

Definition 7.6: *An (M,n)* **code** *C consists of the following:*
1. An index set $\{1,\dots,M\}$.
2. An encoding function $X^n : \{1,\dots,M\} \to \mathcal{X}^n$ which yields the codewords $X^n(1),\dots,X^n(M)$ which is called the codebook.
3. A decoding function $g : \mathcal{Y}^n \to \{1,\dots,M\}$.

For $i \in \{1,\dots,n\}$, we let λ_i^n be the probability of an error if we encode and send i over the channel, that is $P(g(Y^n) \neq i)$ when i is encoded and sent over the channel. We also let $\lambda^n = \max_i \lambda_i^n$ be the **maximal error** and $P_e^n = \frac{1}{M}\sum_i \lambda_i^n$ be the **average** probability of error. We attach a superscript C to all of these if the code is not clear from context.

Exercise: Given any DMC, find an (M,n) code such that $\lambda_1^n = 0$.

Definition 7.7: *The* **rate** *of an (M,n) code is $\frac{\ln M}{n}$ (where we use base 2).*

We now arrive at one of the most important theorems in the field (as far as I understand it), which says in words that all rates below channel capacity can be achieved with arbitrarily small probability of error.

Theorem 7.8 (Shannon's Theorem on Capacity): *All rates below capacity C are achievable in that given $R < C$, there exists a sequence of $(2^{nR}, n)$–codes with maximum probability $\lambda^n \to 0$. Conversely if there exists a sequence of $(2^{nR}, n)$–codes with maximum probability $\lambda^n \to 0$, then $R \leq C$.*

Before presenting the proof of this result, we make a few remarks which should be of help to the reader. In the definition of an (M, n)–code, we have this set $\{1, \ldots, M\}$ and an encoding function X^n mapping this set into $\mathcal{X}^n$. This way of defining a code is not so essential. The only thing that really matters is what the image of X^n is and so one could really have defined an (M, n)–code to be a subset of $\mathcal{X}^n$ consisting of M elements together with the decoding function $g : \mathcal{Y}^n \to \mathcal{X}^n$. Formally, the only reason these definitions are not EXACTLY the same is that X^n might not be injective which is certainly a property you want your code to have anyway. The point is one should also think of a code in this way.

Proof: The proof of this is quite long. While the mathematics is all very simple, conceptually it is a little more difficult. The method is a randomization method where we choose a code "at random" (according to some distribution) and show that with high probability it will be a good code. The fact that we are choosing the code according to a special distribution will make the computation of this probability tractable.

The key concept in the proof is the notion of "joint typicality". The

SMB (for i.i.d.) tells us that a typical sequence $x_1, \ldots, x_n$ from an ergodic stationary process has $-\frac{1}{n}\ln(p(x_1, \ldots, x_n))$ being close to the entropy $H(X)$. If we have instead independent samples from a joint distribution $p(x, y)$, we want to know what $x_1, y_1, \ldots, x_n, y_n$ typically looks like and this gives us the concept of **joint typicality**. An important point is that $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ can be individually typical without being jointly typical. We now give a completely heuristic proof of the theorem.

Given the output, we will choose as our guess for the input something (which encodes to something) which is jointly typical with the output. Since the true input and the output will be jointly typical with high probability, there will with high probability be some input which is jointly typical with the output, namely the true input. The problem is maybe there is more than 1 possible input sequence which is jointly typical with the output. (If there is only one, there is no problem.) The probability that a possible input sequence which was not input to the output is in fact jointly typical with the output sequence should be more or less the probability that a possible input and output sequence chosen independently are jointly typical which is $2^{-nI(X;Y)}$ by a trivial computation. Hence if we use 2^{nR} with $R < I(X;Y)$ different possible input sequences, then, with high probability, none of the possible input sequences (other than the true one) will be jointly typical with the output and we will decode correctly.

We need the following lemma whose proof is left to the reader. It is easy to prove using the methods in §1 when we looked at consequences of the SMB Theorem. An arbitrary element of $\mathcal{X}^n$ $x_1, \ldots, x_n$ will be denoted by x^n below.

Lemma 7.9: *Let $p(x, y)$ be some joint distribution. Let A_ϵ^n be the set of*

ϵ–jointly typical sequences (with respect to $p(x,y)$) of length n, namely,

$$A_\epsilon^n = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : |-\frac{1}{n}p(x^n) - H(X)| < \epsilon,$$

$$|-\frac{1}{n}p(y^n) - H(Y)| < \epsilon, \text{ and } |-\frac{1}{n}p(x^n, y^n) - H(X,Y)| < \epsilon\}$$

where $p(x^n, y^n) = \prod_i p(x_i, y_i)$. Then if (X^n, Y^n) is drawn i.i.d. according to $p(x,y)$, 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 $(\tilde{X}^n, \tilde{Y}^n)$ is chosen i.i.d. according to $p(x)p(y)$, then

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

and for sufficiently large n

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

We now proceed with the proof.

Let $R < C$. We now construct a sequence of $(2^{nR}, n)$–codes (2^{nR} means $\lfloor 2^{nR} \rfloor$ here). whose maximum probability of error $\lambda^{(n)}$ goes to 0. We will first prove the existence of a sequence of codes which have a small average error (in the limit) and then modify it to get something with small maximal error (in the limit). By the definition of capacity, we can choose $p(x)$ so that $R < I(X;Y)$ and then we can choose ϵ so that $R + 3\epsilon < I(X;Y)$.

Let $\mathcal{C}_n$ be the set of all encoding functions $X^n : \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X}^n$. Our encoding function will be chosen randomly according to $p(x)$ in that each $i \in \{1, 2, \ldots, 2^{nR}\}$ is independently sent to x^n where x^n is chosen according to $\prod_{i=1}^n p(x)$. Once we have chosen some encoding function X^n, the decoding procedure is as follows. If we receive y^n, our decoder guesses that the word $W \in \{1, 2, \ldots, 2^{nR}\}$ has been sent if (1) $X^n(W)$ and y^n are ϵ–jointly typical and (2) there is no other $\hat{W} \in \{1, 2, \ldots, 2^{nR}\}$ with $X^n(\hat{W})$ and y^n being ϵ–jointly typical. We also let $\mathcal{C}_n$ denote the set of codes obtained by considering all encoding functions and their resulting codes.

Let $\mathcal{E}^n$ denote the event that a mistake is made when $W \in \{1, 2, \ldots, 2^{nR}\}$ is chosen uniformly. We first show that $P(\mathcal{E}^n) \to 0$ as $n \to \infty$.

$$P(\mathcal{E}^n) = \sum_{C \in \mathcal{C}_n} P(C) P_e^{(n),C} =$$

$$\sum_{C \in \mathcal{C}_n} P(C) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{C \in \mathcal{C}_n} P(C) \lambda_w(C).$$

By symmetry $\sum_{C \in \mathcal{C}_n} P(C)\lambda_w(C)$ does not depend on w and hence the above is $\sum_{C \in \mathcal{C}_n} P(C)\lambda_1(C) = P(\mathcal{E}^n | W = 1)$. One should easily see directly anyway that symmetry gives $P(\mathcal{E}^n) = P(\mathcal{E}^n | W = 1)$. Because of the independence in which the encoding function and the word W was chosen, we can instead consider the procedure where we choose the code at random as above but always transmit the first word 1. In the new resulting probability space, we still let $\mathcal{E}^n$ denote the event of a decoding mistake. We also for each $i \in \{1, 2, \ldots, 2^{nR}\}$ have the event $E_i^n = \{(X^n(i), Y^n) \in A_\epsilon^n\}$ where Y^n is the output of the channel. We clearly have $\mathcal{E}^n \subseteq ((E_1^n)^c \cup \cup_{i=2}^{2^{nR}} E_i^n)$. Now $P(E_1^n) \to 1$ as $n \to \infty$ by the joint typicality lemma. Next the independence of $X^n(i)$ and $X^n(1)$ (for $i \neq 1$) implies the independence of $X^n(i)$ and $Y^n(1)$ (for $i \neq 1$) which gives (by the joint typicality lemma) $P(E_i^n) \leq 2^{-n(I(X;Y)-3\epsilon)}$ and so $P(\cup_{i=2}^{2^{nR}} E_i^n) \leq 2^{nR} 2^{-n(I(X;Y)-3\epsilon)} = 2^{-n(I(X;Y)-3\epsilon-R)}$ which goes to 0 as $n \to \infty$.

Since we saw that $P(\mathcal{E}^n)$ goes to 0 as $n \to \infty$ and

$P(\mathcal{E}^n) = \sum_{C \in \mathcal{C}_n} P(C) P_e^{(n),C}$, it follows that for all large n there is a code $C \in \mathcal{C}_n$ so that $P_e^{(n),C}$ is small.

Now that we know such a code exists, one can find it by exhaustive search. There is much interest in finding such codes. We finally obtain a code with maximum probability of error being small. Since the average error is very small (say δ), it follows that if we take half of the codewords with smallest error, the maximum error of these is also very small (at most 2δ, why?). We therefore obtain a new code by throwing away the larger half of the codewords (in terms of their errors). Since we now have 2^{nR-1} codewords, the rate has been cut to $R - \frac{1}{n}$ which is negligible.

(Actually, by throwing away this set of codewords, we obtain a new code, and since it is a new code, we should check that the maximal error is still small. This is of course obvious but the point is that this non–problem should come to mind).

CONVERSE:

We are now ready to prove the converse which is somewhat easier. We first need two easy lemmas.

Lemma 7.10 (Fano's inequality): *Consider a DMC with code C with input message uniformly chosen. Letting $P_e^{(n)} = P(W \neq g(Y^n))$, we have*

$$H(X^n|Y^n) \leq 1 + P_e^{(n)} nR.$$

Hint of Proof: Let E be the event of error and expand $H(E, W|Y^n)$ in two possible ways. Then think a little. □

Lemma 7.11: *Let Y^n be the output when inputting X^n through a DMC.*

Then for any distribution $p(x^n)$ on X^n (the n bits need not be independent),

$$I(X^n; Y^n) \leq nC$$

where C is channel capacity.

Proof: Left to reader-very straightforward computation with no thought needed. □

We now complete the converse. Assume that we have a sequence of $(2^{nR}, n)$ codes with maximal error (or even average error) going to 0.

Let the word W be chosen uniformly over $\{1, \ldots, 2^{nR}\}$ and consider $P_e^n = P(\hat{W} \neq W)$ where $\hat{W}$ is our guess at the input W. We have

$$nR = H(W) = H(W|Y^n) + I(W; Y^n) \leq$$

$$H(W|Y^n) + I(X^n(W); Y^n) \leq$$

$$1 + P_e^{(n)} nR + nC.$$

Dividing by n and letting $n \to \infty$ gives $R \leq C$. □

Remarks.

(1) The last step gives a lower probability on the average error of $P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}$ which of course is only interesting when the transmission rate R is larger than the capacity C.

(2) Note that (in the first half of the proof) we proved that if $p(x)$ is any distribution on X (not necessarily one which maximized mutual information), then choosing a code at random using the distribution $p(x)$ instead, will with high probability result in a code whose probability of error goes to 0 with n providing we use a transmission rate R which is smaller than the mutual information between the input and output when X has distribution

$p(x)$.
(3) One can prove a stronger converse which says that for rates above capacity, the average error of probability must go exponentially to 1 (What does this exactly mean?) which makes capacity an even stronger dividing line.

Exercise: Consider a channel with 1,2,3 and 4 as both input and output and where 3 and 4 are always sent to themselves while both 1 and 2 are sent to 1 and 2 with probability 1/2. What is the capacity of this DMC? Before computing, take a guess first and guess how it would compare to a channel with only 2 possible inputs but with perfect transmission. How would it compare to the same channel where 2 is not allowed as input?

Exercise: Given a DMC, construct a new one where we add a new symbol $*$ to both the input and output alphabets and where if we input $x \neq *$, the output has the same distibution as before (and so $*$ can't come out) and if $*$ is inputted, the output is uniform on all possible outputs (including $*$). What has happened to the capacity of the DMC?

We have two important theorems, namely Corollary 4.7 and Theorem 7.8. These two theorems seem to be somewhat disjoint from each other but the following theorem brings them together in a very nice way. We leave this theorem as a difficult exercise for the reader who using the methods developed in these notes should be able to carry out this proof.

Theorem 7.12 (Source–channel coding theorem): *Let $\mathcal{V} = V_1, \ldots$ be a stochastic process satisfying the AEP property (stationary ergodic suffices for example). Then if $H(\mathcal{V}) < C$, there exists a source channel code with $P_e^{(n)} \to 0$.*

Conversely, for any stationary stochastic process, if $H(\mathcal{V}) > C$, then the probability of error is bounded away from 0 (bounded away over all n and all source codes).

One should think why this theorem says that there is no better way to send a stationary process over a DMC resulting in small probability of error than to do it in 2 steps, first compressing (as we saw we can do) to get strings of length the order of 2^{nH} and then sending that over the channel using Theorem 7.8.

We end this section with a result which is perhaps somewhat out of place but we place it here since mutual information was introduced in this section. This result will be needed in the next section.

Theorem 7.13: *The mutual information $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.*

Proof: We give an outline of the proof which contains most of the details.

The first thing we need is the so called log sum inequality which says that if $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are nonnegative then

$$\sum_i a_i \log(\frac{a_i}{b_i}) \geq (\sum_i a_i) \log(\frac{\sum_i a_i}{\sum_i b_i})$$

with equality if and only if $\frac{a_i}{b_i}$ is constant.

This is proved by Jensens inequality as is the nonnegativity of relative entropy but in fact this also follows from the latter easily anyway as follows. A trivial computation shows that if the above holds for $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, then it holds if we multiple these vectors by any (possibly different) constants. Now just normalize them to be probability vectors and use the nonnegativity of relative entropy.

The log sum inequality can then be used to demonstrate the convexity of relative entropy in its two arguments, i.e.,

$$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2).$$

as follows.

The log sum inequality with $n = 2$ gives for fixed x

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

$$\lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{\lambda p_2(x)}{\lambda q_2(x)}.$$

Now summing over x gives the desired convexity.

Now we prove the theorem.

Fixing $p(y|x)$, we want to first show that $I(X;Y)$ is concave in $p(x)$. We have

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X = x).$$

If $p(y|x)$ is fixed, then $p(y)$ is a linear function of $p(x)$. Since $H(Y)$ is a concave function of $p(y)$ (a fact which follows easily from the convexity of relative entropy in its two arguments or can be proven more directly), it follows that $H(Y)$ is a concave function of $p(x)$ (since linear composed with concave is concave). The second term on the other hand is obviously linear in $p(x)$ and hence $I(X;Y)$ is concave in $p(x)$ (concave minus linear is concave).

For the other side, fix $p(x)$ and we want to show $I(X;Y)$ is convex in $p(y|x)$. A simple computation shows that the joint distribution obtained by using a convex combination of $p_1(y|x)$ and $p_2(y|x)$ is simply the same convex combination of the measures obtained by using these two conditional things.

Similarly, the product measure obtained when using a convex combination of $p_1(y|x)$ and $p_2(y|x)$ is simply the same convex combination of the product measures obtained by using these two conditional things. Using the definition of information, these two observations together with the convexity of relative entropy in its two arguments allows an easy computation (left to the reader) which shows that $I(X;Y)$ is convex in $p(y|x)$ as desired. □

EXERCISE: It seems that the second argument above could perhaps also be used to show that for fixed $p(y|x)$, $I(X;Y)$ is convex in $p(x)$. Where does the argument break down?

8 Rate Distortion Theory

The area of rate distortion theory is very important in information theory and entire books have been written about it. Our presentation will be brief but we want to get to one of the main theorems, whose development parallels the theory in the previous section.

The simplest context in which rate distortion theory arises is in quantization which is the following problem. We have a random variable X which we want to quantize, that is, we want to represent the value of X by say 1 of 2 possible values. You get to choose the values and assign them to X as you wish, that is, you get to choose a function $f : R \to R$ whose image has only two points such that $f(X)$ (thought of as the quantized version of X) represents X well. What does "represent X well" mean? Perhaps we would want $E[(f(X) - X)^2]$ to be small.

Recall that in the previous section, we needed to send alot of data (i.e., take n large) in order to have small probability of error for a fixed rate which is below the channel capacity. Similarly, in this section, we will be considering "large n" where n is the amount of data to obtain another elegant result.

We let $\mathcal{X}$ denote the state space for our random variable and $\hat{\mathcal{X}}$ denote the possible quantized versions (or representations) of our random variable. We will now consider a generalization of the quantization problem which we now formulate.

Definition 8.1: *A* **distortion function** *is a mapping*

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to R^+.$$

$d(x, \hat{x})$ should be thought of as some type of cost of representing the outcome x by $\hat{x}$. Given the above distortion function d, we take as the distortion between to finite sequences x_1^n and $\hat{x}_1^n$ to be

$$d(x_1^n, \hat{x}_1^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$$

which might be something you object to but which is required for our theory to go through as we will do it.

Definition 8.2: *A* $(2^{Rn}, n)$–**rate distortion code** *consists of an encoding function*

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$$

and a decoding (reproduction) function

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n.$$

If $X_1^n \equiv X_1, \ldots, X_n$ are random variables taking values in $\mathcal{X}$ and d is a distortion function, then the distortion associated with the above code is

$$D = E[d(X_1^n, g_n(f_n(X_1^n)))].$$

The way to think of f_n is that it breaks up the set of all possible outcomes $\mathcal{X}^n$ into 2^{nR} sets (this is the quantization part) and then each quantized piece is sent to some element of $\hat{\mathcal{X}}^n$ which is supposed to "represent" the original element of $\mathcal{X}^n$. Obviously, the larger we take R, the smaller we can make D (e.g. if $|\mathcal{X}| = 2$ and $R = 1$, we can get $D = 0$). In the previous section, we wanted to take R large, while now we want to take it small.

We now assume that we have a fixed rate distortion function d and a fixed distribution on $\mathcal{X}$, $p(x)$. We will assume that $X_1^n \equiv X_1, \ldots X_n$ are i.i.d. with distribution $p(x)$. As we did in the last chapter, for simplicity, we will assume that everything (i.e., $\mathcal{X}$ and $\hat{\mathcal{X}}$) are finite.

Definition 8.3: *The pair (R, D) is* **achievable** *if there exists a sequence of $(2^{Rn}, n)$–rate distortion codes, (f_n, g_n) with*

$$\limsup_n E[d(X_1^n, g_n(f_n(X_1^n)))] \leq D.$$

Definition 8.4: *The* **rate distortion function** *$R(D)$, (where we still have a fixed distribution $p(x)$ on X and a fixed distortion function d) is defined by*

$$R(D) = \inf\{R : (R, D) \text{ is achievable } \}.$$

One should view $R(D)$ as the coarsest possible quantization subject to keeping the distortion at most D. This is clearly a quantity that one should be interested in although it seems (analogous to (non-information) channel capacity) that this would be impossible to compute. We will now define something called the "information rate distortion" function which will play the role in our rate distortion theory that the information channel capacity did in the channel capacity theory. This quantity is something much easier to compute (as was the information channel capacity) and the main theorem will be that these two (the rate distortion function and the information rate distortion function) are the same.

Definition 8.5: *The* **information rate distortion function** $R^I(D)$, *(where we still have a fixed distribution $p(x)$ on X and a fixed distortion function d) is defined as follows. Let $\mathcal{P}^D$ be the set of all distributions $p(x,\hat{x})$ on $\mathcal{X} \times \hat{\mathcal{X}}$ whose first marginal is $p(x)$ and which satisfy $E_{p(x,\hat{x})}d(x,\hat{x}) \leq D$. Then we define*

$$R^I(D) = \min_{p(x,\hat{x}) \in \mathcal{P}^D} I(X;\hat{X}).$$

Long Exercise: Let $\mathcal{X} = \hat{\mathcal{X}} = \{0,1\}$ and consider the rate distortion function which is Hamming distance (which is 1 if the two guys are not the same and 0 otherwise). Assume that X has distribution $p\delta_1 + (1-p)\delta_0$. Compute $R^I(D)$ in this case.

We now state out main theorem and go directly to the proof. Recall that both the rate distortion function and the information rate distortion function are defined relative to the distribution $p(x)$ on X and the distortion function d.

Theorem 8.6: *$R^I(D) = R(D)$, i.e., the rate distortion function and the information rate distortion function are the same.*

Lemma 8.7: *$R^I(D)$ is a nonincreasing convex function of D.*

Proof: The nonincreasing part is obvious. For the convexity, consider D_1 and D_2 and $\lambda \in (0,1)$. Choose (using compactness) $p_1(x,\hat{x})$ and $p_2(x,\hat{x})$ which minimize $I(X;\hat{X})$ over the sets $\mathcal{P}^{D_1}$ and $\mathcal{P}^{D_2}$. Then (by linearity of expectation)

$$p_\lambda \equiv \lambda p_1(x,\hat{x}) + (1-\lambda)p_2(x,\hat{x}) \in \mathcal{P}^{\lambda D_1 + (1-\lambda)D_2}.$$

By the second part of Theorem 7.13, we have

$$R^I(\lambda D_1 + (1-\lambda)D_2) \leq I_{p_\lambda}(X;\hat{X}) \leq$$

$$\lambda I_{p_1}(X;\hat{X}) + (1-\lambda)I_{p_2}(X;\hat{X}) =$$

$$\lambda R^I(D_1) + (1-\lambda)R^I(D_2).$$

□

Proof of $R(D) \geq R^I(D)$: We need to show that if we have a sequence of $(2^{Rn}, n)$–rate distortion codes, (f_n, g_n) with

$$\limsup_n E[d(X_1^n, g_n(f_n(X_1^n)))] \leq D,$$

then $R \geq R^I(D)$. The encoding and decoding functions clearly give us a joint distribution of $\mathcal{X}^n \times \hat{\mathcal{X}}^n$ (where the marginal on $\mathcal{X}^n$ is $\prod_{i=1}^n p(x)$) and we then have

$$nR \geq H(\hat{X}_1^n) \geq H(X_1^n) - H(X_1^n|\hat{X}_1^n) \geq \sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i|\hat{X}_i) =$$

$$\sum_{i=1}^n I(X_i;\hat{X}_i) \geq \sum_{i=1}^n R^I(E[d(X_i,\hat{X}_i)]) = n\sum_{i=1}^n \frac{1}{n} R^I(E[d(X_i,\hat{X}_i)]) \geq$$

$$nR^I(\sum_{i=1}^n \frac{1}{n} E[d(X_i,\hat{X}_i)])(\text{ by the convexity of } R^I\ ) = nR^I(E[d(X_1^n,\hat{X}_1^n)]).$$

This gives $R \geq R^I(E[d(X_1^n, \hat{X}_1^n)])$ for all n. Since $\limsup_n E[d(X_1^n, \hat{X}_1^n)] \leq D$ and $R^I(D)$ is nondecreasing, we should be able to conclude that $R \geq R^I(D)$ but note that to make this conclusion, we need the continuity of R^I or if you really think about it, you need only right–continuity. Of course if for some n, we had that $E[d(X_1^n, \hat{X}_1^n)] \leq D$, then we wouldn't have to worry about this technicality.

Exercise: Prove the needed continuity property. □

Remark: Note that we actually proved that even if $\liminf_n E[d(X_1^n, \hat{X}_1^n)] \leq D$, then $R \geq R^I(D)$.

We now proceed with the more difficult direction which uses (as in the channel capacity theorem) a randomization procedure. We introduced the definition of ϵ–jointly typical sequences in the previous section. We now need to return to this again but with one other condition.

Definition 8.8: *Let $p(x, \hat{x})$ be a joint distribution of $\mathcal{X} \times \hat{\mathcal{X}}$ and d a distortion function on the same set. We say (x^n, y^n) is ϵ–**jointly typical** (relative to $p(x, \hat{x})$ and d) if*

$$|-\frac{1}{n}p(x^n) - H(X)| < \epsilon, |-\frac{1}{n}p(y^n) - H(\hat{X})| < \epsilon,$$

$$|-\frac{1}{n}p(x^n, y^n) - H(X, \hat{X})| < \epsilon, \text{ and } |d(x^n, y^n) - E[d(X, \hat{X})]| < \epsilon$$

and we denote the set of all such pairs by $A_{d,\epsilon}^n$.

The following three lemmas are left to the reader. The first is immediate, the second takes a little more work and the third is pure analysis.

Lemma 8.9: *Let $(X_i, \hat{X}_i)$ be drawn i.i.d. according to $p(x, \hat{x})$. Then $P(A_{d,\epsilon}^n) \to 1$ as $n \to \infty$.*

Lemma 8.10: *For all* $(x^n, \hat{x}^n) \in A^n_{d,\epsilon}$,

$$p(\hat{x}^n) \geq p(\hat{x}^n|x^n)2^{-n(I(X;\hat{X})+3\epsilon)}.$$

Lemma 8.11: *For* $0 \leq x, y \leq 1$ *and* $n > 0$

$$(1 - xy)^n \leq 1 - x + e^{-yn}.$$

Proof of $R(D) \leq R^I(D)$**:** Recall that $p(x)$ (the distribution of X) and d (the distortion function) are fixed and $X_1, \ldots, X_n$ are i.i.d. with distribution $p(x)$. We need to show that for any D and any $R > R^I(D)$, (R, D) is achievable.

EXERCISE: Show that it suffices to show that $(R, D + \epsilon)$ is achievable for all ϵ and then of course it suffices to show this for ϵ with $R > R^I(D) + 3\epsilon$.

For this fixed D, choose a joint distribution $p(x, \hat{x})$ which minimizes $I(X; \hat{X})$ in the definition of $R^I(D)$ and let $p(\hat{x})$ be the marginal on $\hat{x}$. In particular, we then have that $R > I(X; \hat{X}) + 3\epsilon$.

Choose 2^{nR} sequences from $\hat{X}^n$ independently, each with distribution $\prod_{i=1}^n p(\hat{x})$ which then gives us a (random) decoding (or reproduction) function $g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n$. Once we have the decoding function g_n, we define the encoding function f_n as follows. Send X^n to $w \in \{1, 2, \ldots, 2^{nR}\}$ if $(X^n, g_n(w)) \in A^n_{d,\epsilon}$ (if there is more than one such w send it to any of them) while if there is no such w, send X^n to 1.

Our probability space consists of first choosing a $(2^{Rn}, n)$–rate distortion code in the above way and then choosing X^n independently from $\prod_{i=1}^n p(x)$. Consider the event E_n that there does not exist $w \in \{1, 2, \ldots, 2^{nR}\}$ with

$(X^n, g_n(w)) \in A^n_{d,\epsilon}$ (an event which depends on both the random code chosen and X^n). The key step is the following which we prove afterwards.

Lemma 8.12: *$P(E_n) \to 0$ as $n \to \infty$.*

Calculating things in this probability space, we get ($\hat{X}^n$ is of course $g_n(f_n(X^n))$ here)

$$E[d(X^n, \hat{X}^n)] \leq D + \epsilon + d_{max}P(E_n)$$

where d_{max} is of course the maximum value d can take on. The reason why this holds is that if E_n^c occurs, then $(X^n, \hat{X}^n)$ is in $A^n_{d,\epsilon}$ which implies by definition that $|d(X^n, \hat{X}^n) - E[d(X, \hat{X})]| < \epsilon$ which gives the result since $E[d(X, \hat{X})]| \leq D$. Applying the lemma, we obtain $\limsup_n E[d(X^n, \hat{X}^n)] \leq D + \epsilon$.

It follows that there must be a sequence of $(2^{Rn}, n)$–rate distortion codes (f_n, g_n) with $\limsup_n E[d(X_1^n, g_n(f_n(X_1^n)))] \leq D + \epsilon$, as desired. □

Proof of Lemma 8.12: This is simply a long computation. Let $\mathcal{C}_n$ denote the set of $(2^{Rn}, n)$–rate distortion codes. For $C \in \mathcal{C}_n$, let $J(C)$ be the set of sequences x^n for which there is some codeword $\hat{x}^n$ with $(x^n, \hat{x}^n) \in A^n_{d,\epsilon}$. We have

$$P(E_n) = \sum_{C \in \mathcal{C}_n} P(C) \sum_{x^n : x^n \notin J(C)} p(x^n) =$$

$$\sum_{x^n} p(x^n) \sum_{C \in \mathcal{C}_n : x^n \notin J(C)} P(C).$$

Letting $K(x^n, \hat{x}^n)$ be the indicator function of $(x^n, \hat{x}^n) \in A^n_{d,\epsilon}$ and fixing some x^n, we have

$$P((x^n, \hat{X}^n) \notin A^n_{d,\epsilon}) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n)$$

giving

$$\sum_{x^n} p(x^n) \sum_{C\in\mathcal{C}_n : x^n \notin J(C)} P(C) = \sum_{x^n} p(x^n)[1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n)]^{2^{nR}}.$$

Next, Lemmas 8.10 and 8.11 (in that order) and some algebra give

$$P(E_n) \leq 1 - \sum_{x^n, \hat{x}^n} p(x^n, \hat{x}^n) K(x^n, \hat{x}^n) + e^{-2^{-n(I(X;\hat{X})+3\epsilon)} 2^{nR}}.$$

Since $R > I(X;\hat{X}) + 3\epsilon$, the last term goes (super)–exponentially to 0. The first two terms are simply $P((A^n_{d,\epsilon})^c)$ which goes to 0 by Lemma 8.9. This completes the proof. □

9 Other universal and non–universal estimators

The following question I find to be interesting and a number of papers have been written about it.

Given a property of ergodic stationary processes, when does there exist a consistent estimator of it (consistent in the class of all ergodic stationary processes)?

Exercise: Sticking to processes taking on only the values $\{0, \ldots, 9\}$, show that if a property (i.e., a function) is continuous on the set of all ergodic processes (where the latter is given the usual weak topology), then there exists a consistent estimator.

(Olle Nerman pointed this out to me when I mentioned the more general question above).

I won't write anything in this section, but rather we will go through some recent research papers. I will attach zeroxed copies of some papers here.

The theme of these papers will be, among other things, to learn how to construct stationary processes with certain desired behavior.