COMMUNICATION RATES AND SHANNON ENTROPY

PARTIAL NOTES AS OF JANUARY 29, 2019

1. Shannon entropy and communication rates

These notes are about the connection between Shannon entropy and the minimal amount of digital information which must be used on average to send a message.

Suppose we have N messages, numbered $1, \ldots, N$, we would like to communicate digitally. Message number i has probability p_i of being sent. To each i we would like to associated a finite string $b(i)$ of 0's and 1's which will be our digital encoding of i.

Given a stream of successive elements of $1, \ldots, N$, we would like to transmit the corresponding bit strings $b(1), \ldots, b(N)$. We'd like the receiver of the transmission to be able to accurately decode the original sequence of elements of $1, \ldots, N$ we had in mind. To do this, the receiver needs to recognize when a given sequence of digits corresponds to some $b(i)$. In particular, they need to know when a given word $b(i)$ finishes and the next word begins. For this reason we have the following constraint on the $b(i)$:

Hypothesis 1.1. If $1 \leq i \neq j \leq N$ then $b(i)$ is never an initial sequence of the digits of $b(j)$

If this hypothesis were violated, then on seeing the successive digits of $b(i)$ appear, the listener would not know whether or not to continue listening for $b(j)$ to appear. If the hypothesis holds, then the listener knows that once a string $b(i)$ appears, it represents a full word, and that the next digit received is the start of the next word.

Suppose $b(i)$ has length $\ell(i)$, so that it involves exactly $\ell(i)$ 0's and 1's. The average number of digits per message that will sent when we use B to encode messages is

$$T(B, p_1, \ldots, p_N) = \sum_{i=1}^{N} p_i \cdot \ell(i). \tag{1.1}$$

This is because message i will be sent p_i of the time, and in this case we send $\ell(i)$ digits.

Problem 3 of the second homework is to show that that the following definition is well defined:

Definition 1.2. Given N and $p_1, \ldots, p_N$, there is a B for which $T(B, p_1, \ldots, p_N)$ is minimal over all possible choices of B. Let $T^{min}(p_1, \ldots, p_N)$ be this minimal value. It represents the optimal economy in average bits per message one can achieve by encoding the messages $1, \ldots, N$ into bit strings $b(1), \ldots, b(N)$ as above.

Recall that the Shannon entropy is defined by

$$H(p_1, \ldots, p_N) = -\sum_{i=1}^{N} p_i \cdot \log_2(p_i) \tag{1.2}$$

A problem on the second homework is to show that if

$$H(1/N, \ldots, 1/N) = T^{min}(1/N, \ldots, 1/N) \tag{1.3}$$

then N is a power of 2.

The object of these notes is to sketch proofs of the following facts:

Theorem 1.3. *If $N = 2^m$ is a power of 2 and each p_i has the form $p_i = 1/2^{m(i)}$, then*

$$T^{min}(p_1, \ldots, p_N) = H(p_1, \ldots, p_N).$$

Notice that the conclusion of this Theorem definitely does not hold for all $(p_1, \ldots, p_N)$ when N is not a power of 2 by the second homework. Even when $N = 2$, it does not hold for all (p_1, p_2), e.g. when $p_1 = 3/4$ and $p_2 = 1/4$. In this case, one must transmit at least one bit to distinguish between event 1 and event 2, so $T^{min}(3/4, 1/4) = 1$, but $H(3/4, 1/4) < 1$.

One can achieve a transmission rate approaching the Shannon entropy only by sending out long blocks of digits at a time:

Theorem 1.4. *Suppose N is arbitrary and that $(p_1, \ldots, p_N)$ is any probability vector of length N with real components. For each integer $k \geq 1$, consider all possible ordered sequences $q = (i_1, \ldots, i_k)$of k elements of $\{1, \ldots, N\}$. The probability $p(q)$ of $q = (i_1, \ldots, i_k)$ occurring is $p(q) = p_{i_1} \cdot p_{i_2} \cdots p_{i_k}$ if we assume successive messages are independent of one another. There are N^k such q, and if we number them $q_1, \ldots, q_{N^k}$, one has the optimum transmission rate $T^{min}(q_1, \ldots, q_{N^k})$ for the expected number of bits one will need to send a sequence of k messages. One has*

$$\lim_{k \to \infty} \frac{1}{k} T^{min}(q_1, \ldots, q_{N^k}) = H(p_1, \ldots, p_N) \tag{1.4}$$

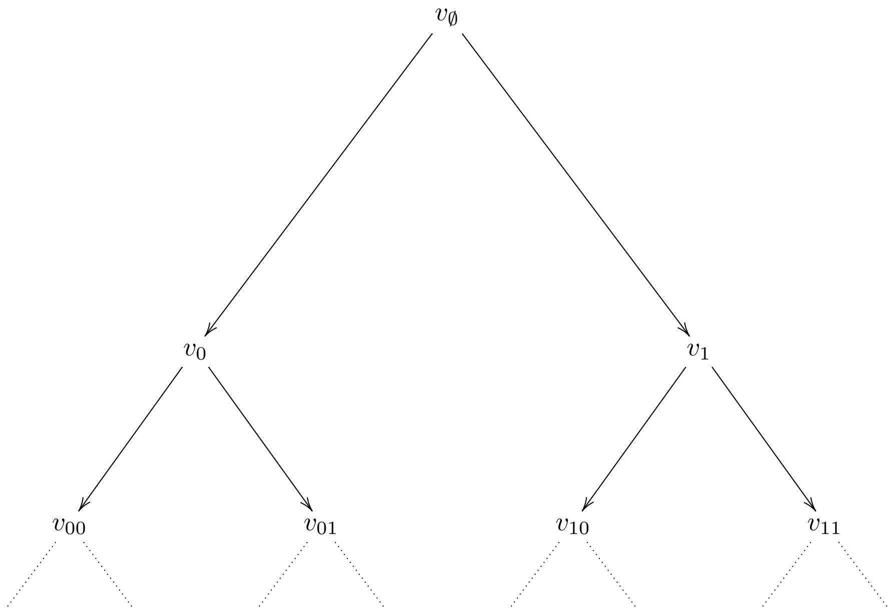
The $\frac{1}{k}$ factor on the left makes sense for the following reason. If one had an ideal digital encoding scheme, one would expect that the average amount of information needed to send out a sequence of k messages would be k times the amount needed to send out one message.

Exercise 1.5. Show that

$$\frac{1}{k} H(q_1, \ldots, q_{N^k}) = H(p_1, \ldots, p_N).$$

2. Graph theory

To understand the constraint represented by Hypothesis (1.1) it is useful to use the directed binary binary rooted tree in having vertices v_b in which b is a sequence of 0's and 1's describing how one walks down from the root vertex $v_\emptyset$ to reach v_b:



The following Lemma is clear:

Lemma 2.1. *The condition that $b(i)$ is not the initial string of digits appearing in some $b(j)$ is equivalent to the condition that $v_{b(j)}$ does lie along a path downward from $v_{b(j)}$ in the tree. Conversely, any collection C of N vertices in the tree which does not include $v_\emptyset$ and which has this property will define a set $B = \{b(1), \ldots, b(N)\}$ of binary digits which satisfies Hypothesis 1.1.*

Hence we can be use B to encode the messages $1,\ldots,N$. The length $\ell(i)$ of $b(i)$ is just the path distance from $v_\emptyset$ to $v_{b(i)}$.

If $N > 1$, note that $\emptyset$ can be seen as an initial part of every bit sequence, so in this case no B for which one of the $b(i)$ was the empty set could satisfy Hypothesis (1.1). Thus when $N > 1$, we could drop the condition that C not contain $v_\emptyset$ in Lemma 2.1. If $N = 1$, we still don't want to use $\emptyset$ as an element of B, since if one did then $B = \{\emptyset\}$ and we would never be able to know how many copies of the only possible message were being sent.

3. The case $N = 2^m$ and $p_i = 1/2^m$ for all $i = 1,\ldots,N$.

In this section we would like to prove:

Theorem 3.1. *If there is an integer $m \geq 1$ such that $N = 2^m$ and $p_i = 1/2^m$ for all $i = 1,\ldots,N$, then*

$$T^{min}(p_1,\ldots,p_N) = T^{min}(1/2^m,\ldots,1/2^m) = H(1/2^m,\ldots,1/2^m) = m \tag{3.5}$$

For the remainder of the section we suppose $N = 2^m$ and $p_i = 1/2^m$ for all $i = 1,\ldots,N$.

Let $B_0 = \{b(1),\ldots,b(N)\}$ be all digit strings of length m, so that $\ell(i) = m$ for all i. Then $C_0 = \{v_{b(1)},\ldots,v_{b(N)}\}$ consists of all vertices at distance m from the root $v_\emptyset$ in the tree, and none of these vertices lie above any other. So they satisfy the condition in Lemma 2.1, and B_0 satisfies Hypothesis 1.1. The transmission rate associated to B_0 is

$$T(B_0, 1/2^m,\ldots,1/2^m) = \sum_{i=1}^{2^m} \frac{1}{2^m}\cdot m = m \tag{3.6}$$

So we deduce that

$$T^{min}(1/2^m,\ldots,1/2^m) \leq m = H(1/2^m,\ldots,1/2^m) = -\sum_{i=1}^{2^m}\frac{1}{2^m}\cdot\log_2(\frac{1}{2^m}).$$

To show Theorem 3.1 we need to show

$$T(B, 1/2^m,\ldots,1/2^m) \geq m \tag{3.7}$$

for all possible choices of $B = \{b(1),\ldots,b(N)\}$.

Lemma 3.2. *Suppose B is an arbitrary allowable choice of $\{b(1),\ldots,b(N)\}$. Then*

$$T(B, 1/2^m,\ldots,1/2^m) = \frac{1}{2^m}\sum_{i=1}^{2^m}\ell(i) \tag{3.8}$$

where $\ell(i)$ is the distance from $v_\emptyset$ to $v_{b(i)}$. If $\ell(i) \leq m$ for all $i = 1,\ldots,N$ then B must be the set B_0 of all digit strings of length m.

Proof. The formula (3.8) is just (1.1) since $p_i = 1/2^m$ for all i. Suppose now that $\ell(i) \leq m$ for all i, so that all the vertices $v_{b(i)}$ have distance $\leq m$ from the root $v_\emptyset$. For $i = 1,\ldots,2^m$, let $S(i)$ be the set of vertices at distance m from $v_\emptyset$ which can be reached by paths downward that begin at $v_{b(i)}$. Suppose there is a vertex v lying in the intersetion $S(i)\cap S(j)$ for some $1 \leq i \neq j \leq N$. There is a unique path from $v_\emptyset$ down to v, and both $v_{b(i)}$ and $v_{b(j)}$ must be on this path. But then one of $v_{b(i)}$ or $v_{b(j)}$ lies farther along the path than the other, and this violates the conditions of Lemma 2.1. So $S(i)$ and $S(j)$ are disjoint, and $S(1),\ldots,S(2^m)$ are disjoint non-empty subsets of the set of 2^m vertices at distance m from $v_\emptyset$. The only way such sets can exist is for each $S(i)$ to consist of just one vertex, and this vertix must be $v_{b(i)}$, so $B = B_0$. □

The following Lemma now completes the proof of Theorem 3.1.

Lemma 3.3. *Suppose B is a choice of $\{b(1),\ldots,b(N)\}$ for which*

$$T(B,1/2^m,\ldots,1/2^m) = T^{min}(1/2^m,\ldots,1/2^m) \tag{3.9}$$

and for which

$$q = \sum_{i=1}^{N} \max(0, m - \ell(i)) \tag{3.10}$$

is minimized. In other words, B achieves the minimum possible $T(B,1/2^m,\ldots,1/2^m)$, and among all such choices it minimizes the sum q of the distances from $v_\emptyset$ to those vertices $v_{b(i)}$ at distance less than m which arise from B. Then $q = 0$ and $B = B_0$ and (3.9) equals m.

Proof. Suppose $q > 0$, so that there is a $v_{b(i)}$ at distance $< m$ from $v_\emptyset$. Then $B \neq B_0$, so by Lemma 3.2 there must be a $j \neq i$ for which $b(j) > m$. The set $S(i)$ of vertices at distance m from $v_\emptyset$ which can be reached by paths down from $v_{b(i)}$ has $2^{m-\ell(i)} \geq 2$ elements. Let $v_{b(i)'}$ be one of the two vertices of the tree which are at distance one below $v_{b(i)}$ in the tree. Then exactly half of the elements of $S(i)$ are on paths downward from $v_{b(i)'}$, so there is a vertex v' in $S(i)$ which is not below $v_{b(i)'}$. We now let $B' = \{b(1)',\ldots,b(j)'\}$ be defined by letting $b(k)' = b(k)$ if $k \notin \{i,j\}$, by letting $b(i)'$ be as above, and by letting $b(j)'$ be the digit sequence with $v_{b(j)'} = v'$. One now checks that B' satisfies the condition in Lemma 3.2 because B does. Here $v_{b(i)'}$ is at distance from $v_\emptyset$ which is one greater than that of $v_{b(i)}$, while $v_{b(j)'} = v'$ is at distance at least one less from $v_\emptyset$ than $v_{b(j)}$ is. So (3.8) shows

$$T(B',1/2^m,\ldots,1/2^m) \leq T(B,1/2^m,\ldots,1/2^m) = T^{min}(1/2^m,\ldots,1/2^m)$$

so we in have to have $T(B',1/2^m,\ldots,1/2^m) = T^{min}(1/2^m,\ldots,1/2^m)$. However, if we replace B by B', then q in (3.10) does down by 1, since $\ell(i)$ increases by 1, $\ell(j) \geq m$ and $v_{b(j)'}$ has distance m from $v_\emptyset$. Hence if B had been chosen to minimize q, we get a contradiction if we suppose $q > 0$. This forces $q = 0$, so $\ell(i) \geq m$ for all $i = 1,\ldots,2^m$. But now (3.8) gives

$$T(B,1/2^m,\ldots,1/2^m) = \frac{1}{2^m}\sum_{i=1}^{2^m} \ell(i) \geq m.$$

The inequality must be an equality because of (3.6) and because B was chosen to minimize $T(B,1/2^m,\ldots,1/2^m)$. Hence all $\ell(i)$ must equal m, and this means $B = B_0$. □

4. The case in which $p_i = 1/2^{m(i)}$ for all $i = 1,\ldots,N$.

In this section we will generalize the results of previous section by showing:

Theorem 4.1. *Suppose there are integers $m(i) \geq 1$ such that $p_i = 1/2^{m(i)}$ for all $i = 1,\ldots,N$. Then*

$$T^{min}(p_1,\ldots,p_N) = H(p_1,\ldots,p_N) = -\sum_{i=1}^{N} p_i \log_2(p_i) = \sum_{i=1}^{m} \frac{1}{2^{m(i)}} \cdot m(i). \tag{4.11}$$

Notice that we don't have to assume N is a power of 2. If all the $m(i)$ were equal to some m, then one would have to have $N = 2^m$ since $p_1 + \ldots + p_N = 1$, but this is not so for other choices of the $m(i)$.

We suppose the hypotheses of the Theorem for the rest of the section. The proof breaks into two steps:

Step 1. Show that there is a choice of B for which $T(B,p_1,\ldots,p_N) = H(p_1,\ldots,p_N)$. This is is an explicit construction.

Step 2. Show that for all B, one has

$$T(B,p_1,\ldots,p_N) \geq H(p_1,\ldots,p_N)$$

This step is more difficult and uses the ideas which go into the proof of Shannon's theorem. As we discussed in class, it is not clear that the composition law for $H(p_1,\ldots,p_N)$ in

Shannon's theorem holds if $H(p_1, \ldots, p_N)$ is replaced by $T^{min}(p_1, \ldots, p_N)$. But if one weakens the composition law to an inequality, then one can show $T^{min}(p_1, \ldots, p_N)$ does satisfy this inequality. This is enough to get the required lower bound for $T^{min}(p_1, \ldots, p_N)$ once one can handle the case in Theorem 3.1.

4.1. **Step 1: A construction.** In view of the right hand formula in (4.11), it will be enough to show that we can pick $B = \{b(1), \ldots, b(N)\}$ satisfying Hypothesis 1.1 so that the vertex $v_{b(i)}$ is at distance $\ell(i) = m(i)$ from $v_\emptyset$, where $p_i = 1/2^{m(i)}$. Then we will have shown

$$T^{min}(p_1, \ldots, p_N) \le T(B, p_1, \ldots, p_N) = \sum_{i=1}^{m} \frac{1}{2^{m(i)}} \cdot m(i) = H(p_1, \ldots, p_N) \tag{4.12}$$

We can reorder $1, \ldots, N$ so that $p_i \ge p_j$ for $i \le j$. For $z \ge 1$, let $d(z)$ be the number of i for which $p_i = 1/2^z$. Then $f(z) = d(1) + \cdots + d(z)$ is the number of i with $p_i \ge 1/2^z$. We will show by induction on z that we can pick $b(1), \ldots, b(f(z))$ so the following is true:

i. For $1 \le i \le f(z)$, $v_{b(i)}$ has distance $m(i)$ from $v_\emptyset$,
ii. No $v_{b(j)}$ lies below $v_{b(i)}$ if $1 \le i \ne j \le f(z)$.

When z is sufficiently large we will have $f(z) = N$, so this will show (4.12).

To do the induction, suppose first that $z = 1$. If $d(1) = 0$, there are no $b(i)$ to be chosen. The next possibility is that $d(1) = 1$, so that $p_1 = 1/2 > p_j$ if $1 < j$. We then pick $b(1)$ to be the digit string 0 and we are done. Finally, the only other possibility is that $d(1) = 2$, $p_1 = p_2 = 1/2$ and $N = 2$. Then we pick $b(1) = 0$ and $b(2) = 1$ and we are done.

We now suppose by induction that $z > 1$ and that we have picked $b(1), \ldots, b(f(z-1))$ having the above properties when z is replaced by $z - 1$. We now need to pick an additional set of $d(z) = f(z) - f(z-1)$ vertices b in the tree which have distance z from $v_\emptyset$ in such a way that no v_b is below $v_{b(j)}$ if $1 \le j \le f(z-1)$. Since no two vertices at the same distance from $v_\emptyset$ can lie above one another or above a vertex nearer to $v_\emptyset$, this will complete the inductive step.

Here $v_{b(i)}$ for $i \le f(z-1)$ is at distance $m(i) = \ell(i)$ from $v_\emptyset$, so the set $S(i)$ of vertices at distance z which are below $v_{b(i)}$ has $2^{z-m(i)}$ elements. The sets $S(1), \ldots, S(f(z-1))$ must be disjoint subsets of the 2^z vertices which are at distance z from $v_\emptyset$. So if we can show

$$2^z - \sum_{j=1}^{f(z-1)} \#S(i) \ge d(z)$$

then we will have a sufficient number vertices to use at distance z to continue the induction.

We know $f(z-1) = d(1) + \cdot + d(z-1)$, so the inequality we want to show is

$$0 \le 2^z - \left(\sum_{j=1}^{f(z-1)} \#S(i) \right) - d(z) = 2^z - \left(\sum_{h=1}^{z-1} d(h) \cdot 2^{m-h} \right) - 2^z = 2^z - \sum_{h=1}^{z} d(h) \cdot 2^{z-h}$$

because $\#S(i) = 2^{m-h}$ if $v_{b(i)}$ has distance $h < z$ from $v_\emptyset$ and there are $d(h)$ such i. Here

$$2^z - \sum_{h=1}^{z} d(h) \cdot 2^{z-h} = 2^z \cdot (1 - \sum_{h=1}^{z} d(h) \cdot 2^{-h}) = 2^z \cdot (1 - \sum \{p_j : p_j \ge 2^{-z}\}) \tag{4.13}$$

because $d(h)$ is the number of j for which $p_j = 2^{-h}$. Since $(p_1, \ldots, p_N)$ is a probability vector we have $p_1 + \ldots + p_N = 1$. So the sum on the right side of (4.13) is always non-negative, and this completes the proof.

4.2. **Step 2: The partial composition law lower bound.** In this subsection we need to show that for all $B = \{b(1), \ldots, b(N)\}$ one has

$$T(B, p_1, \ldots, p_N) \ge H(p_1, \ldots, p_N) = -\sum_{i=1}^{N} p_i \log_2(p_i) = \sum_{i=1}^{m} \frac{1}{2^{m(i)}} \cdot m(i) \tag{4.14}$$

under our standing hypotheses that $p_i = 2^{-m(i)}$ for all i. This will prove $T^{min}(p_1, \ldots, p_N) \geq H(p_1, \ldots, p_N)$, so we will have proved Theorem 4.1 because we know (4.12).

The key to showing (4.14) is the following partial composition law bound:

Lemma 4.2. *Suppose $e = \max\{m(i) : 1 \leq i \leq N\}$. Then for all $B = \{b(1), \ldots, b(N)\}$,*

$$T^{min}(1/2^e, \ldots, 1/2^e) \leq T(B, p_1, \ldots, p_N) + \sum_{i=1}^{N} p_i \cdot T^{min}(1/2^{e-m(i)}, \ldots, 1/2^{e-m(i)}) \tag{4.15}$$

Here the terms involving T^{min} pertain to probability vectors having all components equal. If in (4.15) one replaced T^{min} and $T(B, \cdots)$ by the Shannon entropy H of the corresponding probability vector, the resulting inequality would in fact be an equality that is one of the axioms of Shannon entropy. The proof that (4.15) holds follows the same reasoning behind requiring this axiom for Shannon entropy.

To begin the proof, we suppose we are given an allowable $B = \{b(1), \ldots, b(N)\}$. The corresponding set of vertices $\{v_{b(1)}, \ldots, v_{b(N)}\}$ in the binary tree satisfies the constraints of Lemma 2.1. The bound in (4.15) says that we can come up with a transmission scheme associated to the probability vector $(1/2^e, \ldots, 1/2^e)$ such that the expected number of digits needed per message is bounded by the right side of (4.15). Recall that in the proof of Shannon's theorem, this was done by breaking equally likely events $1, 2, 3, \ldots, 2^e$ into clumps of sizes $r_1, \ldots, r_N$ with $p_i = r_i/2^e$. The idea was to specify which event occurred by first specifying the clump containing the event and then from within clump which was the specific event that occurred. We prove (4.15) by an analogous construction inside the binary tree.

For each $b(i) \in B$, the vertex $v_{b(i)}$ has distance $\ell(i)$ from $v_\emptyset$ and no $v_{b(j)}$ lies below $v_{b(i)}$ if $1 \leq i \neq j \leq N$. We do not now that $\ell(i) = m(i)$ however. It could happen that some $\ell(i)$ have been chosen much smaller than $m(i)$ for example. This would amount to picking an unusually short digit string $b(i)$ to represent an event i for which $p_i = 2^{-m(i)}$ is small, e.g. for which i is rare. What we do know is that no vertex in the set $\tilde{S}(i)$ of all vertices below $v_{b(i)}$ in the tree is below a vertex $v_{b(j)}$ with $j \neq i$. In fact, no element of $\tilde{S}(i)$ is below any element of $\tilde{S}(j)$ if $j \neq i$.

We will prove (4.15) by using B to construct a set $B' = \{b'(1), \ldots, b'(2^e)\}$ of 2^e binary digit strings such that $T(B', 1/2^e, \ldots, 1/2^e)$ is bounded by the right hand side of (4.15). To do this, recall that we showed in Lemma 3.3 that for any integer $f \geq 1$, one has

$$T^{min}(1/2^f, \ldots, 1/2^f) = T(B_f, 1/2^f, \ldots, 1/2^f) = f = H(1/2^f, \ldots, 1/2^f) \tag{4.16}$$

when B_f is the set of 2^f digit strings of length f. The set $\{v_{b(i)} : b(i) \in B_f\}$ is the set of vertices at distance f from the root $v_\emptyset$. This suggests that to produce B' from B, we should replace each vertex $v_{b(i)}$ associated to $b(i) \in B$ by the set of vertices which are at distance $e - m(i)$ below $v_{b(i)}$ in the tree, in order to end up with the right side of (4.15). We now check that this works.

Define B' in the following way. For each $i = 1, \ldots, N$, let $B'(i)$ be the set of $2^{e-m(i)}$ vertices b such that v_b is at distance $e - m(i)$ below $b(i)$ along a path from $b(i)$. We assign to the $b \in B'(i)$ the probability $p_i \cdot 2^{m(i)-e} = 2^{-m(i)} \cdot 2^{m(i)-e} = 2^{-e}$. Thus $p_i = 2^{-m(i)}$ is the sum over the $2^{e-m(i)}$ elements b of $B'(i)$ of the probability 2^{-e} of b occuring. We let $B' = \cup_{i=1}^N B'(i)$, with each of the

$$\sum_{i=1}^{N} 2^{e-m(i)} = 2^e \sum_{i=1}^{N} 2^{-m(i)} = 2^e \sum_{i=1}^{N} p_i = 2^e$$

elements of B' having probability 2^{-e} of occurring.

Since no element of B lies above a different element of B in the tree, no element of B' lies above another element of B'. Thus B' is an allowable choice of 2^e binary digits with which to encode the events $1, \ldots, 2^e$, and we are giving each of these events equal probability 2^{-e}. We have

$$T(B', 1/2^e, \ldots, 1/2^e) = \sum_{b \in B'} 2^{-e} \cdot \text{length}(b) = \sum_{i=1}^{N} \sum_{b \in B'(i)} 2^{-e} \cdot \text{length}(b). \tag{4.17}$$

Here $B'(i)$ has $2^{e-m(i)}$ elements b, and for each such b we have

$$\text{length}(b) = \text{length}(b_i) + e - m(i).$$

Thus

$$\sum_{b \in B'(i)} 2^{-e}\text{length}(b) = 2^{e-m(i)}2^{-e} \cdot (\text{length}(b_i) + e - m(i))$$

Plugging this into (4.17) gives

$$\begin{aligned} T(B', 1/2^e, \ldots, 1/2^e) &= \sum_{i=1}^{N} 2^{-m(i)} \cdot (\text{length}(b_i) + e - m(i)) \\ &= \sum_{i=1}^{N} p_i \cdot \text{length}(b_i) + \sum_{i=1}^{N} p_i(e - m(i)) \\ &= T(B, p_1, \ldots, p_N) + \sum_{i=1}^{N} p_i T^{min}(1/2^{e-m(i)}, \ldots, 1/2^{e-m(i)}) \end{aligned} \tag{4.18}$$

when we use the equality in (4.16) with $f = e - m(i)$ to rewrite the second term of the last line. The equalities in (4.18) show (4.15) since $T^{min}(1/2^e, \ldots, 1/2^e)$ is the minimum of $T(B', 1/2^e, \ldots, 1/2^e)$ over all possible choices of B'.

Completion of the proof of Theorem 4.1

Because of (4.16), we can rewrite (4.15) as

$$H(1/2^e, \ldots, 1/2^e) - \sum_{i=1}^{N} p_i H(1/r_i, \ldots, 1/r_i) \leq T(B, p_1, \ldots, p_N) \tag{4.19}$$

where $r_i = 2^{e-m(i)}$ is the number of elements in the subset $B'(i)$ of $1, \ldots, 2^e$ and $p_i \cdot \frac{1}{r_i} = 2^{-e}$. However, the left hand side is exactly $H(p_1, \ldots, p_N)$ by the composition law for the Shannon entropy. Since (4.19) holds for all B, we conclude that

$$H(p_1, \ldots, p_N) \leq T^{min}(p_1, \ldots, p_N).$$

Now (4.12) shows

$$H(p_1, \ldots, p_N) = T^{min}(p_1, \ldots, p_N)$$

which completes the proof of Theorem 4.1.

5. Shannon entropy as the maximum efficiency of sending long digit strings

In this section we will show Theorem 1.4, whose notations we now assume. The first step is to observe that Shannon entropy does satisfy the expected relationship

$$\frac{1}{k} \cdot H(q_1, \ldots, q_{N^k}) = H(p_1, \ldots, p_N) \tag{5.20}$$

between the expected amount of information $H(p_1, \ldots, p_N)$ needed to send a single message and the expected amount of information $H(q_1, \ldots, q_{N^k})$ needed to send a string of k messages. One checks this by expanding the formula for $H(q_1, \ldots, q_{N^k})$ using the fact that if q_j is the probaiblilty of a sequence $i_1, \ldots, i_k$ being sent, then

$$q_j = p_{i_1} \cdots p_{i_k}.$$

So

$$\log_2(q_j) = \sum_{\ell=1}^{k} \log_2(p_{i_\ell})$$

and (5.20) follows on regrouping and using $p_1 + \cdots + p_N = 1$.

The central idea in deducing Theorem 1.4 from Theorem 4.1 is now this. We can assume that no p_i equals 1, since this case is trivial. Then every q_j is bounded by $(\max_{i=1}^N p_N)^k$, and this bound goes to 0 as $k \to \infty$. Thus for large enough k, each q_j will satisfy an inequality

$$2^{-n(j)} \leq q_j \leq 2^{-n(j)+1} = 2^{-n(j)} \cdot 2 \tag{5.21}$$

for some integers $n(j)$, and there is a lower bound on all the $n(j)$ which goes to $+\infty$ as $k \to \infty$. It follows that

$$\frac{\log_2(2^{-n(j)})}{k} \leq \frac{\log_2(q_j)}{k} \leq \frac{\log_2(2^{-n(j)})}{k} + \frac{\log_2(2)}{k} \quad \text{where} \quad \frac{\log_2(2)}{k} \to 0 \quad \text{as} \quad k \to \infty \tag{5.22}$$

We will also have

$$\sum_j 2^{-n(j)} \leq \sum_j q_j = 1 \tag{5.23}$$

The essential idea now is to use $2^{-n(j)}$ as a reasonable approximation to q_j in order to use the arguments involved in showing Theorem 4.1 to show Theorem 1.4. We are in effect trying to replace $T^{min}(p_1, \ldots, p_N)$ in Theorem 4.1 by

$$\lim_{k \to \infty} \frac{T^{min}(q_1, \ldots, q_{N^k})}{k}$$

The main issue is to show that the same arguments and constructions go through, with appropriate modifications, because (5.22) holds.

We now sketch some details. Suppose first that one would like to prove an analog involving $\frac{1}{k} T^{min}(q_1, \ldots, q_{N^k})$ for Theorem 3.1. The hypothesis of Theorem 3.1 is that $p_i = 1/2^m$ for all $i = 1, \ldots, N$ with $N = 2^m$. The analog should concern $\frac{1}{k} T^{min}(q_1, \ldots, q_{N^k})$ under the additional hypotheses that the q_j are "almost equal" to one another, in the sense that there is an $\tilde{m}$ with

$$2^{-\tilde{m}} \leq q_j \leq 2^{-\tilde{m}+1} \tag{5.24}$$

for all j. We no longer assume that N is a power of 2. The first part of the proof of Theorem 3.1 is simply to observe that with the hypotheses of this Theorem, one can use the set B of all digit sequences of length m to have

$$T(B, 1/2^m, \ldots, 1/2^m) = H(1/2^m, \ldots, 1/2^m) = m.$$

In the analog, we similarly use a set $\tilde{B}$ of N^k digit sequences of length $\tilde{m}$. Here (5.24) shows

$$1 = \sum_j q_j \geq N^k 2^{-\tilde{m}} \quad \text{so} \quad N^k \leq 2^{\tilde{m}}$$

and there are enough vertices at distance $\tilde{m}$ to form $\tilde{B}$. Now

$$T(B', q_1, \ldots, q_{N^k}) = \sum_j q_j \cdot \tilde{m} \leq \sum_j q_j \cdot \log_2(q_j) \leq \sum_j q_j \cdot (\tilde{m} + 1) \tag{5.25}$$

The middle term here is just $H(q_1, \ldots, q_j) = kH(p_1, \ldots, p_N)$, and the right term is

$$\sum_j q_j \cdot (\tilde{m} + 1) = T(B', q_1, \ldots, q_{N^k}) + \sum_j q_j = T(B', q_1, \ldots, q_{N^k}) + 1$$

So we get

$$\frac{T(B', q_1, \ldots, q_{N^k})}{k} \leq H(p_1, \ldots, p_N) \leq \frac{T(B', q_1, \ldots, q_{N^k})}{k} + \frac{1}{k} \tag{5.26}$$

under the hypothesis (5.24) that the q_j are "almost equal". This hypotheses does depend on k, since $(q_1, \ldots, q_{N^k})$ depends on k. We will now show that (5.26) is enough to carry over the arguments used to prove Theorem 4.1 to show Theorem 1.4.

The first step in showing Theorem 4.1 was constructive: one needed to produce a set of digit sequences B for which

$$T(B, p_1, \ldots, p_N) \leq H(p_1, \ldots, p_N).$$

Similarly, the first step in showing Theorem 1.4 is to construct a digit sequence $\tilde{B}_k$ for each k for which

$$\frac{T(\tilde{B}_k, q_1, \ldots, q_k)}{k} \leq H(p_1, \ldots, p_N) + o(k) \tag{5.27}$$

where $o(k) \to 0$ as $k \to \infty$. To do this, we suppose that $n(j)$ is defined so (5.23) holds for each $j = 1, \ldots, N^k$, and we use a digit sequence of length $n(j)$ encode event j. The fact that $2^{-n(j)} \leq q_j$ will suffice to show as in the proof of Theorem 4.1 that there are enough vertices in the tree at distance $n(j)$ from the root to construct $\tilde{B}_k$. This requires the same calculations as in Theorem 4.1, so we will omit them. The reason the calculations work is that $n(j)$ is larger than $-\log_2(q_j)$, so we are looking for vertices farther away from the root than $-\log_2(q_j)$, and as one moves farther from the root there are more and more vertices to choose from. Now the bound (5.22) suffices to show (5.27).

The more difficult step in showing Theorem 4.1 is to show a lower bound

$$\frac{T(B_k, q_1, \ldots, q_k)}{k} \geq H(p_1, \ldots, p_N) + o(k) \tag{5.28}$$

for all possible choices B_k of digit strings used to encode $1, \ldots, N^k$ when i has probability q_j. In the proof of Theorem 4.1 this involved first showing the composition law inequality in Lemma 4.2. This composition law came about from mimicking in the binary tree the reasoning behind the composition law axiom for Shannon entropy. Namely, break up the specification of a particular event into first specifying which of a collection of disjoint subsets the event lies, and then specify the event to be chosen within that subset when all events in a subset have the same likelihood. In the tree, this amount to replacing vertices by the sets of all vertices which are at a certain distance below them; this distance is chosen to lead to equal probabilities for all of the final events, so one can apply Theorem 3.1.

To carry this out to show (5.28), we use the same ideas, but we need to use the "almost equal" version of Theorem 3.1. More specifically, we suppose the $n(j)$ are as in (5.22). Suppose $B_k = \{b(1), \ldots, b(N^k)\}$ encodes $\{1, \ldots, N^k\}$. Pick n so $n \geq n(j)$ for all $j = 1, \ldots, N^k$. We form an encoding of $\{1, \ldots, N(k)\}$ for an appropriate integer $N(k)$ by replacing the vertex $v_{b(j)}$ for $j = 1, \ldots, N^k$ by the set of $2^{n-n(j)}$ vertices which are at distance $n - n(j)$ below $v_{b(j)}$ in the tree. We assign these vertices probability

$$q_j \cdot 2^{n(j)-n}$$

Here (5.22) gives

$$2^{-n} = 2^{-n(j)} \cdot 2^{n(j)-n} \leq q_j \cdot 2^{n(j)-n} \leq 2^{-n+1}$$

so the resulting probabilites of each of the events $1, \ldots, N(k)$ are "almost equal". This is sufficient by our "almost equal" version of Theorem 3.1 to show the required composition law inequality up to an error $o(k)$ which goes to 0 as $k \to \infty$. Then the "almost equal" generalization of Theorem 3.1 is enough to show an inequality of the form (5.28). This is enough to prove Theorem 1.4 because we have already shown (5.27).