Encoding Strings of Symbols

John Wyatt
3/17/04

Example 1 Source Coding (no probability)

Alphabet = A, B, C, D, E M=5 symbols
Need b = 3 bits, even though $\log_2$ M = 2.322 bits < 3 bits.

Let's instead encode strings of length N
For $n = 2$, the strings are AA, AB, AC, AD, AE, BA, BB, BC, etc.

N = # symbols in string	# strings = 5^p	b = # bits	# bits/symbol = b/N
1	5	3	3
2	25	5(2^5 = 32)	5/2 = 2.50
3	125	7(2^7 = 128)	7/3 = 2.333 $\geq \log_2$M=2.322
4	625	10(2^{10} = 1024)	10/4 = 2.50
5	3,125	12(2^{12} = 4096)	12/5 = 2.40
10	9,765,625	24	2.40
20	9.54 x 10^{13}	47	2.35
30	9.31 x 10^{20}	70	2.333
500	3.05 x 10^{349}	1161	2.322

Def. $\lceil y \rceil$ = smallest integer $\geq y$

Note that $y \leq \lceil y \rceil < y+1$

$$b = \lceil \log_2(\#\ \text{strings}) \rceil = \lceil \log_2 5^N \rceil = \lceil \text{N} \bullet \log_2 5 \rceil < (N \bullet \log_2 5)+1$$

$$\text{\#bits/symbol} = \frac{b}{N} \leq \log_2 5 + \frac{1}{N} \underset{N\to\infty}{\to} \log_2 5$$

Thus the number of bits per symbol required to encode a long string of symbols from a given alphabet converges to the logarithm (base 2) of the size of the alphabet as the length of the string grows without limit. Note that the convergence is not monotonic in the length of the string, however, since strings of length 4 in our example require more bits per symbol than strings of length 3.

Example 2 Source Coding (Huffman)

In this example the source alphabet consists of symbols $S_1, S_2, \cdots S_M$ occurring independently with probabilities $p(S_1), \cdots, p(S_M)$.

Entropy of this alphabet is

$$H = \sum_{k=1}^{M} p(S_k) \log_2 \frac{1}{p(S_k)}$$

Have shown in previous lecture that for any binary prefix code for this alphabet

$$L = \sum_{k=1}^{M} p(S_k)\ l_k \geq H$$

Though we have not shown it, it is also true that for the Huffman code, the average length L_H always satisfies

$$H \leq L_H \leq H + 1$$

Suppose we encode strings of N symbols instead. For example, for $N = 2$, the expanded alphabet consists of all ordered pairs of two symbols, $S_1S_1, S_1S_2, S_1S_M, S_2S_1, \cdots, S_2S_M, \cdots, S_{MM}$. Let

H_N = entropy of alphabet consisting of independent N-length strings of $S_1, \cdots S_M$.

Lemma

$H_2 = H + H$

$H_N = NH$

Proof $(N = 2)$

$$H_2 = \sum_j \sum_k p\left(S_j, S_k\right) \log_2 \frac{1}{p\left(S_j, S_k\right)} = \text{ (using independence)}$$

$$\sum_j \sum_k p\left(S_j\right) p\left(S_k\right) \log_2 \frac{1}{p\left(S_j\right) p\left(S_k\right)} =$$

$$\sum_j \sum_k p\left(S_j\right) p\left(S_k\right) \left(\log_2 \frac{1}{p\left(S_j\right)} + \log_2 \frac{1}{p\left(S_k\right)} \right) =$$

$$\sum_j \sum_k p\left(S_j\right) p\left(S_k\right) \log_2 \frac{1}{p\left(S_j\right)} + \sum_j \sum_k p\left(S_j\right) p\left(S_k\right) \log_2 \frac{1}{p\left(S_k\right)} =$$

$$\sum_j p\left(S_j\right) \log_2 \frac{1}{p\left(S_j\right)} + \sum_k p\left(S_k\right) \log_2 \frac{1}{p\left(S_k\right)} =$$

$H + H$ ■

Now apply Huffman coding to the set of all symbol strings of length *N*. Let the average length of the Huffman code words for the strings of length *N* be called $L_H(N)$. Then by the Huffman coding theorem,

$$H_N \leq L_H(N) \leq H_N + 1$$

Using the previous lemma, we have

$$NH \leq L_H(N) \leq NH + 1$$

Average # bits per symbol $= \dfrac{L_H(N)}{N}$

$$H \leq \frac{L_H(N)}{N} \leq H + \frac{1}{N} \underset{N\to\infty}{\to} H$$

Note the similarity between the results in Example 1 and Example 2.

In both cases, for long strings of symbols from a given alphabet, the number of bits per symbol required to code the string converges to a quantity that is easily computed from knowledge of the alphabet alone ($\log_2 M$ in the first example and *H* in the second example). But it is greater than or equal to that quantity for any fixed length string, and in particular for coding single symbols $(N = 1)$. Longer strings reduce the number of bits per symbol toward the theoretical minimum.

Most Likely Head Count Sequences

Example 3

Consider a sequence of n independent tosses of a biased coin with $p(H)=3/4, p(T)=1/4$. Let h, $0\le h\le n$, be the total number of heads in the string. Then for large n, most sequences will have $h\approx\frac{3n}{4}$, and sequences with h quite different from $\frac{3n}{4}$ are rare enough that, for large n, we can neglect them.

Each sequence of n tosses with h heads has probability

$$(p(H))^{h}(p(T))^{n-h}$$

Now suppose n is large and, for convenience, chosen so that $h=\frac{3n}{4}$ is an integer. It is not hard to show that $h=\frac{3n}{4}$ is the most likely number of heads in such a sequence as well as the mean or expected number of heads. We call any sequence of n tosses with $h=3n/4$ heads a "most likely head count sequence." Note that while a "most likely head count sequence" has the most probable number of heads, the probability of that number of heads may be less than ½ and the probability of any specific most likely head count sequence may be much smaller yet. The point is that other numbers of heads are even less likely. Each "most likely head count sequence" has probability

$$p_{MLS}=\left(3/4\right)^{\frac{3n}{4}}\left(1/4\right)^{n/4}$$

Lets relate this to H, the entropy of a single toss:

$$H=(3/4)\log\left(\frac{1}{3/4}\right)+\left(1/4\right)\log\left(\frac{1}{1/4}\right)\approx 0.81$$

Note that

$$2^{-nH} = \left(2^{\log\left(\frac{3}{4}\right)}\right)^{\frac{3n}{4}} \left(2^{\log\left(\frac{1}{4}\right)}\right)^{\frac{n}{4}} = \left(\tfrac{3}{4}\right)^{\frac{3n}{4}} \left(\tfrac{1}{4}\right)^{\frac{n}{4}} = p_{MLS}.$$

Thus in this example every most likely head count sequence is equally likely and has probability

$$p_{MLS} = 2^{-nH} \geq 2^{-n}.$$

It is easy to see that this result holds in general. Let p represent the probability of heads in any sequence of n independent tosses of a (biased) coin. Again the most likely head count sequences have $h = np$ heads, which we assume to be an integer. Each sequence of n tosses with h heads has probability

$$p^h (1-p)^{n-h}$$

and each most likely head count sequence has probability

$$p_{MLH} = p^{np} (1-p)^{n(1-p)}.$$

Since

$$\log p_{MLH} = np \log p + n(1-p)\log(1-p) =$$
$$n\left(p \log p + (1-p)\log(1-p)\right) = -nH,$$

where H is the entropy of a single toss,

$$p_{MLH} = 2^{-nH} = \left(2^{-n}\right)^{H} \geq 2^{-n}$$

in general. The inequality holds because $H \leq 1$ for a coin toss with any probability p. Thus H quantifies exactly how much more likely the most likely head count sequences are than sequences with $p = 1/2$, all of which have probability 2^{-n}.

Question Using the probabilities in Exercise 3, find the probability of a most likely head count sequence for $n = 8$ and give its probability. In light of these numbers, explain why it is more probable that 8 tosses yields a most likely head count sequence than that it yields the most likely sequence.

A closely related approximate result holds for N_{MLH}, the number of most likely sequences:

$$N_{MLH} \approx 2^{nH}$$

This is a very coarse (but useful) approximation for large *n*, in contrast with the previous result for p_{MLH}, which was exact.

n	N_{MLH}	2^{nH}
4	4	9.5
8	28	90
20	15,504	76,627
100	2.42×10^{23}	2.64×10^{24}
252	2.02×10^{60}	3.49×10^{61}

Table: The exact number of most likely head count sequences and the value of 2^{nH} for sequences of *n* independent tosses of a biased coin with probability of heads = ¾ and a corresponding value of $H = 0.8113$.

This table (which extends up to the overflow level of my calculator) shows that for this example N_{MLH} and 2^{nH} grow similarly. But $N_{MLH} \approx 2^{nH}$ is a very coarse approximation, even for large *n*, since

$$\left(N_{MLH} - 2^{nH}\right)$$

grows without bound as *n* grows, and the even the ratio

$$\left(N_{MLH} / 2^{nH}\right)$$

does not approach 1 for large *n*. With a bit more work, however, one can show that the approximation satisfies the weaker condition

$$\frac{\log N_{MLH}}{nH} \xrightarrow[n\to\infty]{} 1 .$$

One feature of the approximation is that

$$N_{MLH} < 2^{nH}$$

for all $n \geq 1$. To see why this must be so, note that for each *n* the set of all sequences of heads and tails is a partition, and therefore

$$1 = \sum_{\text{all sequences } \alpha} p(\alpha) = \sum_{\text{all most likely head count sequences } \alpha_m} p(\alpha_m) + \sum_{\text{all less likely sequences } \alpha l} (\alpha_l) =$$

(since every most likely sequence has probability exactly 2^{-nH})

$$N_{MLH}\left(2^{-nH}\right) + p\left(\text{set of all less likely sequences}\right) = 1,$$

so

$$N_{MLH} = \left(1 - p\left(\text{set of all less likely sequences}\right)\right) 2^{nH} < 2^{nH}$$

But ignoring the coarseness of this approximation, the results

$$p_{MLH} = 2^{-nH} = \left(2^{-n}\right)^{H} \geq 2^{-n}$$

$$N_{MLH} \approx 2^{nH} = \left(2^{n}\right)^{H} \leq 2^{n}$$

taken together, indicate that the most likely sequences are more probable than a sequence with *p* = ½ (and therefore entropy $H = 1$ and probability for each sequence of 2^{-n}) and fewer in number than the set of all 2^n sequences by the same power laws, $\left(2^{-n}\right)^{H}$ and $\left(2^{n}\right)^{H}$, respectively. The latter approximation can be viewed heuristically as if there were only 2^H rather than 2 outcomes for each toss in the most likely sequences.

To see why this latter approximation might hold, let *p* be the probability of heads and *n* be a large number, with the number of heads *h* = *np* in a most likely sequence being an integer. Then the exact number of most likely sequences is

$$\binom{n}{h} = \frac{n!}{(n-h)!h!}.$$

Now we use the simplest form of Stirling's approximation for the factorial:

$$k! \approx \left(\frac{k}{e}\right)^{k}$$

for large integers k. (More accurate approximations are available, but we shall not need them here.)

Thus

$$\binom{n}{h} \approx \frac{\left(\frac{n}{e}\right)^{n}}{\left(\frac{n-h}{e}\right)^{n-h}\left(\frac{h}{e}\right)^{h}} = \frac{n^{n}}{(n-h)^{n-h}\, h^{n}} =$$

(letting $h = np$ for the most likely head count sequences)

$$\frac{n^{n}}{\left(n(1-p)\right)^{n(1-p)}(np)^{np}} = \frac{1}{(1-p)^{n(1-p)}\, p^{np}} =$$

$$\left[\left(\frac{1}{p}\right)^{p}\left(\frac{1}{1-p}\right)^{1-p}\right]^{n}$$

Note that this expression agrees exactly with our approximation for N_{MLH}

$$2^{nH} = \left(2^{H}\right)^{n} = \left[2^{p\log(1/p)+(1-p)\log\left(\frac{1}{1-p}\right)}\right]^{n} = \left[\left(\frac{1}{p}\right)^{p}\left(\frac{1}{1-p}\right)^{1-p}\right]^{h}.$$

Asymptotic Equipartition Theorem

The situation would be vastly simpler if the most likely head count sequences were overwhelmingly probable, so that all the others could safely be ignored in our calculations. This is sadly not the case, as we can see from the previous table for sequences of length, say, 100 where the

p (set of most likely head count sequences) =

$$N_{MLH} \bullet p_{MLH} = N_{MLH} \bullet 2^{-NH} = 0.091$$

and the probability grows even smaller for $n = 252$.

We can save the day by also including sequences with head counts slightly different from np. Sequences with $h \approx np$ are called *typical sequences,* and they are truly typical in that the probability a sequence will have $h \approx np$ is almost 1. It turns out that including these extra sequences does not much alter our earlier conclusions, although

p(sequence will be typical) > p(sequence will have the most likely head count) = 2^{-nH}

The exact statement and proof of the general result are given in the Appendix, but here is a useful paraphrase.

Asymptotic Equipartition Theorem (Paraphrase)

Consider a sequence on n independent tosses of a possibly biased coin with probability p of turning up heads. Let *TS* be the *set of all typical sequences of outcomes*, i.e., those in which the $h \approx np$, where h is the number of heads. For large n,

p(sequence will be typical) ≈ 1

p(sequence will not be typical) ≈ 0

p(each typical sequence) $\approx 2^{-nH}$

the number of typical sequences $\approx 2^{nH} \leq 2^{n}$.

In plain English, the theorem says that for large n the outcome is quite likely to be a typical sequence, all typical sequences are about equally likely with probability of about 2^{-nH}, and therefore the number of typical sequences is about 2^{nH}, which is exponentially less than 2^{n} unless $p = 0$ or 1.

This simple outcome will be very helpful in understanding the capacity of a channel.

Exercise

Consider 100 independent tosses of a fair coin.

a) Find the average number of heads.
b) Give an estimate for the probability the number of heads will be near the average.
c) Give an estimate for the probability of any single sequence of tosses that has near the average number of heads.

d) Give an estimate for the total number of sequences of tosses that have near the average number of heads.

Channel Capacity

Suppose the raw data we wish to transmit is binary. Instead of transmitting single bits, we transmit "symbol strings" consisting of binary input strings of length p. For error correction purposes the channel encoder assigns to each binary symbol string a codeword of length $n \geq p$. We want to argue that, despite noise in the channel, it is possible (in the limit that p and n both become large) to transmit with negligible probability of error (after error correction at the receiver) if $\left(p/n\right) < C$, the channel capacity, and that it is impossible if $\left(p/n\right) > C$.

Let I_{in} be the entropy of a single bit of input to the channel and I_{out} be the entropy of a single bit of output.

The asymptotic equipartition theorem tells us that, for large n, there are about $2^{nI_{out}}$ typical output sequences of length n.

For a given input bit the output entropy is N (the channel noise), so the number of typical output codewords for each input string (due to noise) is about 2^{nN}.

The error correction decoding scheme must assign all 2^{nN} outputs to a single input string, so there can be at most

$$\frac{2^{nI_{out}}}{2^{nN}} = 2^{n(I_{out}-N)} = 2^{nC}$$

input strings, i.e.,

$$p < nC.$$

Appendix – Careful Statement and Proof of the Asymptotic Equipartition Theorem and Application to Source Coding

The theorem is stated here for sequences of n independent tosses of a possibly biased coin with a probability p of heads. Let h represent the number of heads in the sequence. The expected number of heads is $h = np$, and one can also show that this is the most likely number of heads. In the lecture notes we defined a *most likely head count sequence* of tosses as any sequence with $h = np$. In the more general theorem we want to allow for the fact that sequences with h very near to np are almost as likely as a most likely head count sequence and include them as well. Thus we define *typical sequences* of tosses as those in which

$$|h - np| < d\sqrt{n},$$

and all other sequences as *atypical*. We no longer require that np be an integer. The constant $d > 0$ characterizes the maximum deviation from the number of heads $h = np$ in a most likely head count sequence that we will allow in calling a sequence typical. The constant d need not be an integer, does not vary with n, and can be chosen as we wish. If $d\sqrt{n} \leq 1$ and np is an integer, the only typical sequences of tosses of length n are the most likely head count sequences.

This definition allows odd behavior for small values of n: if np is not an integer and $d\sqrt{n}$ is small, there may be *no* typical sequences, while if $d\sqrt{n}$ is large, *all* sequences may be typical. Nonetheless, for large n

$$\frac{(\text{\# possible values of } h \text{ in typical sequences})}{(\text{total \# of possible values of } n)} \lesssim \frac{2d\sqrt{n}}{n} = \frac{2d}{\sqrt{n}} \xrightarrow[n\to\infty]{} 0,$$

and thus for large n the typical sequences include only a vanishingly small fraction of the possible values of h, centered about the most likely and average value $h = np$.

The following theorem shows that for large d the atypical sequences become sufficiently improbable they can be neglected. Also for all $d > 0$ and for large n all typical sequences are about equally likely, and it gives the approximate number of typical sequences, which is exponentially less than 2^n for large n if $H < 1$. Understanding the proof requires some elementary background in probability, specifically Chebyshev's inequality.

Asymptotic Equipartition Theorem (AEP)

Given any choice of $d > \sqrt{p(1-p)}$, for all $n \geq 1$

A) the set of all typical sequences has a probability $p_{\text{all } TS}$ that is bounded below by

$$p_{\text{all } TS} \geq 1 - \frac{p(1-p)}{d^2} \underset{d \to \infty}{\to} 1$$

B) each individual typical sequence α has a probability $p(\alpha)$ that is bounded above and below by

$$2^{-n\left(H + \frac{b}{\sqrt{n}}\right)} < p(\alpha) < 2^{-n\left(H - \frac{b}{\sqrt{n}}\right)}$$

where

$$b = d \log\left(\frac{1}{p(1-p)}\right)$$

C) the total number of typical sequences N_{TS} is bounded above and below by

$$2^{n(H-g(n))} \leq N_{TS} \leq 2^{n(H+g(n))}$$

where the positive function $g(n)$ vanishes as $n \to \infty$ as $\frac{b}{\sqrt{n}}$.

Part A gives a lower bound on $P_{\text{all } TS}$, the probability of the set of *all* typical sequences, while part B bounds the probability $p(\alpha)$ of *each individual* typical sequence α.

For part A, we can choose *d* sufficiently large in the definition of typical sequences that the probability of the set of all atypical sequences (which is at most $\frac{1}{4d^2}$) becomes small enough to be neglected. This bound on the probability of atypical sequences depends only on *d* and *p* and holds for all *n*. This result is surprising because, for $p \neq 1/2$, one can show that the *fraction* of all sequences that are atypical approaches 1 as $n \to \infty$. Nonetheless their *total probability* becomes negligibly small for large *d*.

Part B shows that, to a certain approximation, every typical sequence has a probability of about 2^{-nH} for large *n*. This is reflected in the name of the theorem, where "Asymptotic" refers to the limit as $n \to \infty$ and "Equipartition" refers to the fact that all typical sequences are, to a coarse approximation, equally likely. One consequence of part B is that, given any error margin $\varepsilon > 0$, no matter how small, there is a value of N_o such that every typical sequence α has probability in the range

$$2^{-n(H+\varepsilon)} \leq p(\alpha) \leq 2^{-n(H-\varepsilon)}$$

for all $n \geq N_o$. (In fact, $N_o = (b/\varepsilon)^2$.)

Each typical sequence becomes exponentially unlikely as $n \to \infty$ (provided $p \neq 0$ or 1). The probability bounds in part B are useful but quite coarse n that they only guarantee

$$2^{-b\sqrt{n}} \leq \frac{p_{TS}}{2^{-nH}} \leq 2^{b\sqrt{n}},$$

where the lower bound goes to zero and the upper bound to infinity as $n \to \infty$. They do guarantee, however, that

$$\frac{\log p_{TS}}{-nH} \xrightarrow[n\to\infty]{} 1$$

Part C gives a similarly coarse bound on the number of typical sequences. It implies that, given any error margin $\delta > 0$, no matter how small, there is a value of N_1 such that the number of typical sequences N_{TS} satisfies

$$2^{n(H-\varepsilon)} \leq N_{TS} \leq 2^{n(H+\varepsilon)}$$

whenever $n \geq N_1$. They also guarantee that

$$\frac{\log N_{TS}}{nH} \xrightarrow[n\to\infty]{} 1$$

Proof

A) p(sequence is atypical) $= p\left(|h-np| \geq d\sqrt{n}\right) \leq$

(using the Chebyshev inequality and the fact that the variance of h is $np(1-p)$)

$$\frac{np(1-p)}{d^2 n} = \frac{p(1-p)}{d^2}.$$

B) Each sequence with h heads has probability

$$p(h) = p^h (1-p)^{(n-h)}.$$
$$\log p(h) = h \log p + (n-h)\log(1-p)$$

For each typical sequence,

$$np - d\sqrt{n} < h < np + d\sqrt{n}.$$
$$n(1-p) - d\sqrt{n} < n-h < n(1-p) + d\sqrt{n}.$$

Since $\log p \leq 0$ and $\log(1-p) \leq 0$, the probability $p(\alpha)$ for any typical sequence α satisfies

$$\left(np + d\sqrt{n}\right)\log p + \left(n(1-p) + d\sqrt{n}\right)\log(1-p) < \log p(\alpha) <$$

$$\left(np - d\sqrt{n}\right)\log p + \left(n(1-p) - d\sqrt{n}\right)\log(1-p)$$

i.e.,

$$n\left(p \log p + (1-p)\log(1-p)\right) + d\sqrt{n}\log\left(p(1-p)\right) < \log p(\alpha) <$$

$$n\left(p \log p + (1-p)\log(1-p)\right) - d\sqrt{n}\log\left(p(1-p)\right),$$

where $\log p(1-p) < 0$, i.e.,

$$-nH + d\sqrt{n}\log\left(p(1-p)\right) < \log p(\alpha) < -nH - d\sqrt{n}\log\left(p(1-p)\right),$$

i.e.,

$$2^{-n\left(H+\frac{b}{\sqrt{n}}\right)} < p(\alpha) < 2^{-n\left(H-\frac{b}{\sqrt{n}}\right)},$$

where

$$b = -d\log\left(p(1-p)\right) = d\log\left(\frac{1}{p(1-p)}\right)$$

C) From part A)

$$1-\frac{p(1-p)}{d^2} \le p(\text{set of typical sequences}) = \sum_{\text{all typical sequences } \alpha_k} p(\alpha_k) \le 1.$$

Using the bounds on $p(\alpha_k)$ from part B,

$$\frac{\left(1-\frac{p(1-p)}{d^2}\right)}{2^{-n\left(H-\frac{b}{\sqrt{n}}\right)}} \le N_{TS} \le \frac{1}{2^{-n\left(H+\frac{b}{\sqrt{n}}\right)}},$$

i.e.,

$$\left(1-\frac{p(1-p)}{d^2}\right)2^{n\left(H-\frac{b}{\sqrt{n}}\right)} = \left(\frac{d^2-p(1-p)}{d^2}\right)2^{n\left(H-\frac{b}{\sqrt{n}}\right)} \le N_{TS} \le 2^{n\left(H+\frac{b}{\sqrt{n}}\right)},$$

i.e.,

$$2^{n\left(H-\frac{b}{\sqrt{n}}-\frac{1}{n}\log\frac{d^2}{d^2-p(1-p)}\right)} \le N_{TS} \le 2^{n\left(H+\frac{b}{\sqrt{n}}\right)} \le 2^{n\left(H+\frac{b}{\sqrt{n}}+\frac{1}{n}\log\frac{d^2}{d^2-p(1-p)}\right)}.$$

Choosing

$$g(n) = \frac{b}{\sqrt{n}} + \frac{1}{n}\log\frac{d^2}{d^2-p(1-p)} > 0$$

we see that $g(n)$ vanishes as $b/\sqrt{n}$ for large n and

$$2^{n(H-g(n))} \le N_{TS} \le 2^{n(H+g(n))}$$

$\blacksquare$

Exercise

Consider a sequence of 10,000 tosses of a fair coin. Please answer each part by giving relevant formulas and then give numerical values.

a) What is the expected number of heads?

b) Give a lower bound on the probability the number of heads differs from the value you found above by less than 1% of the number of tosses.

c) Give an estimate and give upper and lower bounds for the probability of any single sequence of 10,000 tosses where the number of heads differs from the value you found in part a) by less than 1% of the number of tosses.

d) Give an estimate and give upper and lower bounds on the total number of sequences of heads and tails in which the total number of heads differs from the value you found in part a) by less than 1% of the number of tosses.

Exercise

In this more theoretical exercise you will show that the assumptions in the Asymptotic Equipartition Theorem can be weakened and that its conclusions can be made somewhat stronger at various points.

a) Show that part B of the theorem actually holds for any d >0, not just for $d > \sqrt{p(1-p)}$ as stated.

b) Show that part B of the theorem actually gives a tighter bound with b replaced by

$$b' = d\left|\log\frac{p}{(1-p)}\right| \le b$$

c) If you are familiar with the central limit theorem, use it to show that for large n the lower bound in part A of the theorem can be replaced by the following bound, which is much tighter for large d,

$$\lim_{n\to\infty} p_{\text{all } TS} \geq 1-\sqrt{\frac{2}{\pi}}e^{-\frac{d^2}{2p(1-p)}}$$

d) Show that the changes in parts b) and c) above imply that in part C of the theorem, $g(n)$ can be replaced by $\hat{g}(n)$ which vanishes as $n\to\infty$ as $b'/\sqrt{n}$.

In many applications the details of the Asymptotic Equipartition Theorem are not really important. The most useful features are captured in the following simplified corollary.

Corollary to Asymptotic Equipartition Theorem

Given any arbitrarily small choice of margin of tolerance $\varepsilon > 0$,

A) there is a value $d_{\min}$ such that for any $d \geq d_{\min}$,

$$p_{\text{all } TS} \geq 1-\varepsilon$$

B) for any choice of $d > \sqrt{p(1-p)}$ there is a value of $n_{\min}(d,\varepsilon)$ such that

$$2^{-n(H+\varepsilon)} < p(\alpha) < 2^{-n(H-\varepsilon)}$$

for all $n \geq n_{\min}(d,\varepsilon)$

C) for any choice of $d \geq d_{\min}$ there is a value of $n_{\min}$ such that

$$2^{n(H-\varepsilon)} < N_{TS} < 2^{n(H+\varepsilon)}$$

for all $n \geq n_{\min}(d,\varepsilon)$.

Proof of Corollary

A) We require

$$1-\frac{p(1-p)}{d^2}\geq 1-\varepsilon$$

$$\varepsilon \leq \frac{p(1-p)}{d^2}$$

$$d_{\min} \geq \sqrt{\frac{p(1-p)}{\varepsilon}}$$

B) We require

$$\frac{b}{\sqrt{n}}\leq \varepsilon$$

$$n \geq \frac{b^2}{\varepsilon^2}=\frac{d^2}{\varepsilon^2}\log^2\left(\frac{1}{p(1-p)}\right)$$

$$n_{\min} \geq \left[\frac{d}{\varepsilon}\log\left(\frac{1}{p(1-p)}\right)\right]^2$$

C) Since part C uses results from parts A and B, we require both $d \geq d_{\min}$, $n \geq n_{\min}$.

Source Coding Revisited

The Asymptotic Equipartition Theorem gives us another view of the source coding problem that does not rely on Huffman coding, the Kraft inequality or the Gibbs inequality. We will state it here only for alphabets with two symbols, 0 and 1, though it holds for sequences of symbols chosen from any alphabet. The essential idea is that there are only about 2^{nH} typical sequences, which only require about nH bits to code.

Consider sequences of n symbols, each chosen independently from an alphabet S_0, S_1 with probabilities

$$p(S_1) = p$$

$$p(S_0) = 1 - p.$$

The alphabet has entropy

$$H = p\log\frac{1}{p} + (1-p)\log\frac{1}{(1-p)} \leq 1 \text{ bit}$$

There are 2^n possible sequences, and it would take n bits to code them all (say by replacing S_0 by 0 and s_1 by 1). Nonetheless, for any small error margin $\delta > 0$, the AEP shows that if n is sufficiently large we can encode these strings with a code having an average number of bits

$$l \leq n(H + \delta)$$

Recall that $H < 1$ if $p \neq \frac{1}{2}$. Since δ can be arbitrarily small, we can find a code with average length

$$l \leq n(H + \delta) < n$$

for any $p \neq \frac{1}{2}$.

Using the corollary to the AEP, there are at most $2^{n(H+\varepsilon)}$ typical sequences. We can code them by, for example, considering them as binary sequences and coding them in sequential order. This takes

$$\lceil n(H + \varepsilon) \rceil \leq n(H + \varepsilon) + 1$$

bits, since $n(H + \varepsilon)$ may not be an integer. We can set an additional bit equal to 1 as prefix for each codeword to let the receiver know that it is a typical sequence and thereby know the remaining codeword length is $\lceil n(H+\varepsilon) \rceil$. Thus the length of the entire codeword for each typical sequence is

$$l_{typ} \leq n(H+\varepsilon)+2$$

We encode atypical sequences in a very lazy way, since they are so improbable: replace S_0 by 0 and S_1 by 1 and add a zero at the beginning to let the receiver know the sequence is atypical and thereby that the codeword length is $l_{atyp} = n+1$.

Note that this is not a prefix code, but it is uniquely decipherable provided there are no transmission or framing errors.

The average length of this code is

$$l = \sum_{\text{typical sequences } \alpha_m} l(\alpha_m)p(\alpha_m) + \sum_{\text{atypical sequences } \alpha_l} l(\alpha_l)p(\alpha_l) \leq$$

$$\left[n(H+\varepsilon)+2\right]p(\text{all } TS)+(n+1)(1-p(\text{all } TS)) \leq$$

$$\left[n(H+\varepsilon)+2\right](1)+(n+1)\varepsilon =$$

$$n(H+2\varepsilon)+2+\varepsilon = n\left(H+2\varepsilon+\frac{(2+\varepsilon)}{n}\right)$$

Since we can choose ε as small as we wish, we can choose it sufficiently small that

$$2\varepsilon+\frac{(2+\varepsilon)}{n}<\delta$$

for n sufficiently large, so the average codeword length l is bounded by

$$l \leq n(H+\delta)$$

This extremely elementary coding scheme works only because the number of typical sequences is about $2^{nH} < 2^n$ for $p \neq \frac{1}{2}$, (and thus $H < 1$) and yet typical sequences are (for sufficiently large d) overwhelmingly the most probable ones.