20 Aug 2024

Contents

Review of what we did last time
Shannon's channel coding theorem

Last time

A source or a probabilistic scheme consists of a set of symbols the source emits and their corresponding probabilities.

$X = [(a_i, p_i): i = 1, 2, ..., n]$

The instantaneous source coding problem

$T[X]$ = transmission cost associated with X
= expected number of bits Alice needs to send Bob to communicate the output of the source

Theorem I: $H[X] \le T[X] < H[X] + 1$

$H[X] = \sum_i p_i \log 1/p_i$

****
T[X] gives an operational motivation for studying H[X].
****

However, this motivation is not fully satisfactory because of the difference of 1 between the lower bound and upper bound. In particular, for biased coin, e.g.,

$X = \{(0,1/4), (1,3/4)\}$

$T[X] = 1$ but $H[X] = 0.8111$

What does H[X] mean in this case, and in general? To better understand the connection between entropy and compression, we consider encoding not just one symbol at a time but several of them together.

BLOCK CODING

Suppose the source X emits symbols according to distribution P.

A sequence x-bar is (P,eps)-typical if the number of emperical distribution is within eps of P.

Alice and Bob fix a small eps (something like $\exp(-k^{1/3})$) and decide that they would ignore x-bar that are not typical, and focus on the rest. Then then H[X] shows up in two things (ignoring lower order terms in k):

(i) The number of typical sequences grows as 2^{kH[X]}
(ii) The probabilitiy of any fixed typical sequence x-bar
grows as 2^{-kH[X]}

These considerations lead to the following.

Let s(k,delta) = min |S| where is a subset of A^k st P^k(S) > 1-delta.

Theorem: For all delta in (0,1),

lim_{k->infty} (1/k) log s(k,delta) = H[X]

So if Alice and Bob accept a probability eps of error, then they may decide to assign codewords to only the typical sequences and ignore the rest. Then they would encode blocks of k sybmols by long stings of about k H[X] bits, and thereby send about H[X] bits per symbol. If they try to spend less than H[X] per symbol, they will make erros with probability --> 1.

Note how allowing a negligible but positive probability of error brought down the cost from 1 bit per symbol to just 0.811 bit per symbol.

Alternatively, Alice after sampling X many times chooses to store the information in memory as bits. She could compress the data by ignore the non-typical strings, and store 1/H[X] symbols of the source for every bit of memory.

****
H[X] truly determines the best compression rate we can achieve while recording the data obtained by sampling the source repeatedly. Even with the slightly more, we get vanishing probability of error; with slightly less, error probability approaches 1.
****

Conditional entropy

Suppose (X,Y) are random variabels with some joint distribution. Say, Pr[X=a and Y=b] = p(a,b).

H[Y|X] = sum_a p(a) H[Y | X=a]
= sum_a p(a) [sum_b p(b|a) log 1/p(b|a)]

(Explain on the tree.)

What does H[Y|X] mean 'operationally'?

Suppose x-bar has type P, or emperical distribution P_X, the marginal distribution of X. We may ask how many sequences y-bar are such that (x-bar,y-bar) are jointly typical accoording to P (up to some tolerance eps, which we do not explicitly mention).

A small calculation gave us the answer.

    prod_a   2^{(k p(a)) H[Y|X=a]} = 2^{k H[Y|X]}

We have the following important equality H[(X,Y)] = H[X] + H[Y|X]. We will soon use this quantity in our understanding of Shannon's channel coding theorem.

CHANNEL CODING

We consider coding and decoding when the communication channel is distorts what is sent through it. The channel has an input alphabet A and an output alphabet B. We assume that the channel's behaviour can be modelled probabilistically. The rule (channel characteristics) are described by coditional probability of receiving the symbol b when the symbol a is sent into it. That is, the channel is specified by numbers {p(b|a): a in A, b in B}. We assume that the channel is memoryless, that is, its behaviour does not change over time. In particular, we may consider k uses of the channel and conclude that for x-bar in A^k and y-bar in B^k.

Pr[output = y-bar | input = x-bar] = prod_i p(y_i | x_i)

The idea of communication using such a channel is the following. We imaging that Alice has a large number of potential messages to send: say M_1, M_2, ..., M_N. She must map these words into codewords w_1, w_2, ..., w_N in A^k. When a message M_j is to be sent, its codeword w_j is fed into the channel. The transformation of M_j to the codeword is called encoding (the encoder
needs to be efficient, a concern we will ignore). Out comes the received word y-bar, which the decoder maps back to one of the messages (hopefully, M_j itself, but we allow some small probability of error).

A code is C for blocklength k is a subset of A^k.

Rate(C) = (1/k) log |C|

This represents the number of bits Alice is able to send per use of the channel.

We view a decoder for the code is a function from B^k to C, that is, it takes a received word and determines what codeword Alice had meant to transmit. We say that the decoder makes error at most delta (wrt code C) if

    for all words w in C, Pr[D(y-bar) = w] >= 1 - delta,

where y_b is distributed according to Y^k|X^k=w.

Fix k large. Suppose Alice an Bob claim to have a code C in A^k of rate R and a delta-error decoder D for C. The codewords in C might have various types. There are at most (k+1)^n types. So there must be

    2^{kR} / (k+1)^n =   2^{k (R - n log(k+1)/k) }

codewords of a common (most popular codeword type) P. Note that the rate of the code restricted to this type is essentially the same because n log(k+1)/k is neglible for large k.

Consider a codeword of type P, say x-bar. How many y-bar are jointly typical wit x-bar (wrt the given channel characteristic P_Y|X.

Answer: about 2^{k H[Y|X]}

When x-bar is fed into the channel, the received word is distributed 'essentially' uniformly in a set of size about 2^{k H[Y|X]}. If the decoder is to decode x-bar correctly, then most of these received words must be mapped back to x-bar.

Note also that when (x-bar, y-bar) is joinly typical, then y-bar is typical wrt to the distribution Q.

    q(b) = sum_a p(a) p(b|a)

Let X be drawn according to P, and let Y then be draw according to the
P_{Y|X}, so that

Pr[(X,Y) = (a,b)] = p(a) p(b|a).

So we have the following picture. For each codeword x-bar in C of type P, the decoder D maps about

2^{k H[Y|X]} (1-delta)

of the jointly typical sequences back to x-bar. But there are only about 2^{k H[Y]} typical sequences in B^k in all. So

    2^{k (R - n log(k+1)/k)} 2^{k H[Y|X]} (1-delta) <= 2^{k H[Y]}

Taking logs, dividing by k, etc.

    R <= H[Y] - H[Y|X] = H[X] + H[Y] - H[XY]

We call RHS I[X:Y], the mutual information of X and Y. So, Alice can do no better than picking the type P for X so that I[X:Y] is maximized. It turns out that the converse is also true.

Cap_delta(Channel) = Cap_delta(P_Y|X) = max_C Rate(C), where the maximum is taken over all codes for which there is a decoder with error at most delta.

Let C = max_X I[X:Y], C stands for capacity

Theorem (Shannon't Channel coding theorem):

(a) For all delta > 0 (however, small) and all R < C, for all large enough k, there is code C subset in A^k and a delta-error deoder D for
C such that Rate(C) > R.

(b) For all delta > 0 and all R > C, for all large k, for every code C
in A^k of rate R(C) > C, and decoder D, there is a w in C such that

    Pr[error(w)] >= 1 - delta.

(You cannot decode well if you operate above capacity.)

How does the proof go.

For (b), it is essentially what we did above. We show that if the code has rate more than the capacity, then for some codeword w far fewer than 2^{k H[Y|X]} typical received words it generates can map back to it. From our discussion last time, we conclude that when w is sought to be transmitted the deoder will succeed with miniscule probability.

For (a), we turn things around. We fix the distribution X obtained in the maximization implicit in the definition of C. Now, we pick 2^{nR} codewords at random according to the distribution of X, and argue that most received words will have only one codeword that is jointly typical with it. Some care is needed to ensure that we will transmit EVERY codeword with high probability.

(End of Lecture 2)