ΔI can act as an upper bound on information loss

Consider the following game: person A chooses, at random, one object out of a set of objects, and person B has to guess which one was chosen by asking yes/no questions. The i^{th} object is chosen with probability p_i. Person B, however, thinks the i^{th} object is chosen with probability q_i, and will base her question-asking strategy on that wrong distribution. Cover and Thomas (mainly Chap. 5) tells us that the average number of yes/no questions person B has to ask to guess the object, denoted $N(p,q)$, is

$$N(p,q) = -\sum_i p_i \log q_i = H(p) + \Delta I \tag{1}$$

where $H(p) \equiv -\sum_i p_i \log p_i$ is the entropy of the true distribution and $\Delta I \equiv \sum_i p_i \log p_i/q_i = D(p||q)$ is the Kullback-Leibler distance between p and q.

The quantity ΔI that appears in Eq. (1) has a natural interpretation: it is the penalty, in yes/no questions, that person B pays for using the wrong distribution to design her question-asking strategy. We can also interpret ΔI as an information loss, in the following sense: We can make up for the wrong distribution by supplying person B with ΔI bits. In other words, if we give ΔI bits to person B, on average she will do as well guessing what object is present as a person who knows the true distribution.

In fact, what we show below is stronger than that: if we give ΔI bits to person B, she will do *no worse* at guessing the object than a person who knows the true distribution, and she may do better. Alternatively, if we want person B to guess the object in $H(p)$ yes/no questions (the same number as a person who knows the true distribution), we could do that by supplying her with *at most* ΔI bits, and sometimes less than that. The actual number of bits she needs depends on the distributions p and q.

When we say "give bits to person B", we have in mind the following: Person A chooses an object, and then sends a string of symbols (0s and 1s, say) to person B through a noisy channel. Those strings provide information about which object was chosen using a pretty standard coding scheme: the objects are divided into groups, and each string tells which group the object is in. For example, a coding scheme for 6 objects might be: objects 1 and 2 are labeled with the string 1, objects 3 and 4 are labeled with the string 01, and objects 5 and 6 are labeled with the string 00. If, say, object 5 is chosen, then the string 00 would be sent. Since the channel is noisy, a string different than 00 might be received.

So here's the situation. Person A chooses object i, determines that it is labeled with string k, and then sends string k to person B. Because the channel through which the string is sent is noisy, person B receives string l. She then revises her estimate of the probability that object i is chosen. Letting $P(k|l)$ be the probability that string k was sent given that string l was received, her new estimate of the distribution of objects, which we'll call $q(i|l)$, is

$$q(i|l) = \frac{q_i}{Q_k} P(k|l) \tag{2}$$

where $Q_k \equiv \sum_{i \in k} q_i$ and the notation $i \in k$ means sum over only those i such that i is in group k. The number of yes/no questions she will have to ask to guess the object is now

$$N_I(p, q) = -\sum_l P(l) \sum_i p(i|l) \log q(i|l) \tag{3}$$

where the subscript I means that I bits were sent (I will be computed shortly), $P(l)$ is the probability that person B received string l, and $p(i|l)$ is the true probability that object i was chosen given that string l was received. Analogous to Eq. (2), this last quantity is given by

$$p(i|l) = \frac{p_i}{P_k} P(k|l) \tag{4}$$

where $P_k \equiv \sum_{i \in k} p_i$.

Inserting Eqs. (2) and (4) into (3), rearranging terms slightly, and using $P_k = \sum_l P(k|l)P(l)$, we find that

$$N_I(p, q) = N(p, q) - \sum_l P(l) \sum_k P(k|l) \log \frac{P(k|l)}{P_k} - \sum_k P_k \log \frac{P_k}{Q_k} \, .$$

The second term is $I(k; l)$, the amount of information transmitted through the noisy channel (also the average string length), and the third term is $D(P||Q)$, the Kullback-Leibler distance between P and Q. We thus have

$$N_I(p, q) = N(p, q) - I(k; l) - D(P||Q) \, .$$

Since $D(P||Q)$ is non-negative, giving I bits to person B reduces the number of yes/no questions she has to ask by *at least* I. The reduction could be larger, though. To find out how much larger, we minimize $N_I(p, q)$ with respect to q, subject to the constraint that the q_i sum to 1. When we do this, we find that the minimum value of $N_I(p, q)$, which occurs when $q_i = p_i Q_k / P_k$, is $H(p) - I(k;l)$. The maximum *reduction* in yes/no questions, $N(p,q) - [H(p) - I(k;l)]$, is thus $I(k;l) + \Delta I$ (see Eq. (1)). Consequently, giving I bits to person B reduces the number of yes/no questions she has to ask by an amount somewhere between I and $I + \Delta I$, inclusive. Alternatively,

> Giving ΔI bits **or less** to person B allows her to guess the object in exactly the same number of yes/no questions as someone who knows the true distribution.

This is our main result, and it's why we interpret ΔI as an upper bound on information loss.

This looks sort of odd: we can, in principle, give person B an arbitrarily small amount of information and produce a potentially large reduction in the number of yes/no questions. Can this really happen? The answer is yes, as the following example shows.

Let $p_i = 1/M$, $i = 1, ..., M$, $q_1 = 1/M - \alpha/M$, and $q_{i>1} = 1/M + \alpha/M(M-1)$. For these distributions

$$\Delta I = -\frac{1}{M}\log(1-\alpha) - \frac{M-1}{M}\log\left[1 + \frac{\alpha}{M-1}\right] .$$

For fixed M, we can make ΔI arbitrarily large by letting α approach 1.

Now let's inject a little information by telling person B whether or not element 1 was chosen. We'll use a lossless channel, so this results in a transmission of

$$I = -\frac{1}{M}\log\frac{1}{M} - \frac{M-1}{M}\log\frac{M-1}{M}$$

bits, which, for large M, approaches $\log M/M$. Thus, we can provide person B with an arbitrarily small amount of information (by letting M go to infinity), while reducing the number of yes/no questions she would have to ask by an arbitrarily large amount (by letting α go to 1).