Ralph's channel

Ralph is interested in relating entropy to communication. In particular, what is the best data compression (shortest average code word or entropy) and what is the fastest transmission rate (the channel capacity C)?

Data compression. Suppose the code is binary $(0,1)$. Then, a bit (0 or 1) is $\log_2 2^1 = 1$, two bits $(00, 01, 10$, or $11)$ is $\log_2 2^2 = 2$, and so on. If we're dealing with 32 uniformly distributed codewords then the average codeword is $\log_2 2^5 = 5$ bits or $H_2 = -\sum_{i=1}^{32} \frac{1}{32} \log_2 \frac{1}{32} = \log_2 32 = 5$ or entropy (H_2 refers to entropy with base 2 logarithms). Alternatively, suppose we're dealing with eight codewords distributed as $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. If we make each codeword equal length this requires an average codeword of 3 bits $(2^3 = 8)$. However, if we assign shorter codewords to more likely words then we can gain better data compression.

$$H_2 = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64}$$
$$= \log_2 4 = 2 \text{ bits}$$

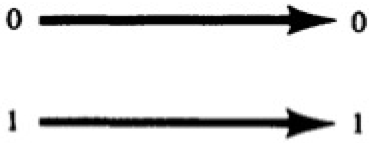
The codewords associated with this scheme might be (in order of probabilities): $0, 10, 110, 1110, 111100, 111101, 111110, 111111$. Again, the answer is entropy.

Channel capacity asks what is the maximum rate we can effectively transmit input X to output Y with arbitrarily small error. Cover and Thomas (p. 183) describe channel capacity as "the maximum number of distinguishable signals for n uses of a communication channel. This number grows exponentially with n, and the exponent is known as the channel capacity." Naturally, this relates to mutual information $I(X;Y)$ (see the introduction to questions 5 and 6 for more details) and is given by C bits per transmission or use of the channel where

$$C = \begin{array}{c} \max \\ p(x) \end{array} I(X;Y)$$

Consider three examples.

Noiseless binary channel. Suppose input $X = 0$ is received as output $Y = 0$ and $X = 1$ is received as $Y = 1$.



The maximum transmission rate is achieved when X is assigned a uniform distribution. As the channel is noiseless, this results in the following joint distribution.

	$Y=0$	$Y=1$
$X=0$	$\frac{1}{2}$	0
$X=1$	0	$\frac{1}{2}$

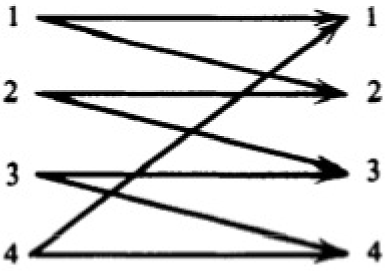
and a transmission rate equal to

$$I\left(X;Y\right)=H\left(X\right)+H\left(Y\right)-H\left(X,Y\right)=\log_2 2=1 \text{ bit per transmission}$$

-

Noisy four-symbol channel. Suppose we have an error prone channel with the following joint distribution.

	$Y=1$	$Y=2$	$Y=3$	$Y=4$
$X=1$	$\frac{1}{8}$	$\frac{1}{8}$	0	0
$X=2$	0	$\frac{1}{8}$	$\frac{1}{8}$	0
$X=3$	0	0	$\frac{1}{8}$	$\frac{1}{8}$
$X=4$	$\frac{1}{8}$	0	0	$\frac{1}{8}$



The transmission rate is

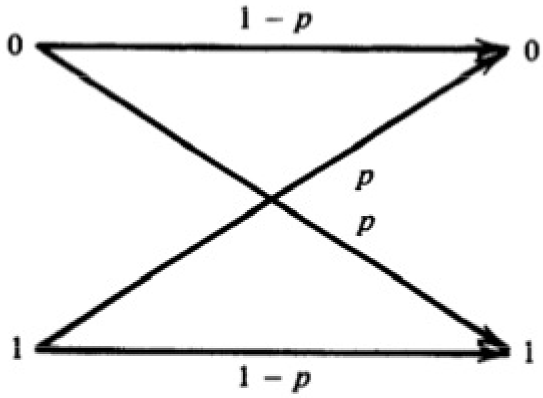
$$I\left(X;Y\right)=\log_2 4+\log_2 4-\log_2 8=2+2-3=1 \text{ bit per transmission}$$

but the error rate is likely unacceptably high. This can be remedied by restricting input to two codewords, say, $X=1,3$ (we've assigned zero probability to $X=2,4$). Now, there is no ambiguity about the input when Y is received. The joint distribution is

	$Y=1,2$	$Y=3,4$
$X=1$	$\frac{1}{2}$	0
$X=3$	0	$\frac{1}{2}$

and we've replicated an error-free transmission rate of one bit per transmission like the noiseless binary channel above. This conveys the intuition for Shannon's noisy channel theorem. The idea is that for sufficiently large block lengths, every channel has a subset of inputs that produce disjoint sequences as the output.

Binary symmetric channel. Suppose we have a noisy binary symmetric channel as depicted below.



The channel capacity is

$$C = 1 + p\log_2 p + (1-p)\log_2 (1-p) \text{ bits per transmission}$$

This follows from assigning a uniform distribution to input X, by symmetry output Y also is uniformly distributed and the joint distribution is

	$Y=0$	$Y=1$
$X=0$	$\frac{1-p}{2}$	$\frac{p}{2}$
$X=1$	$\frac{p}{2}$	$\frac{1-p}{2}$

Mutual information is

$$\begin{aligned} I(X;Y) &= \log_2 2 + \log_2 2 + (1-p)\log_2 \frac{1-p}{2} + p\log_2 \frac{p}{2} \\ &= 1 + p\log_2 p + (1-p)\log_2 (1-p) \end{aligned}$$

For $p > 0$, the zero-error capacity for this channel is zero. However, repeated transmissions at rate $R < C$ can achieve an arbitrarily low average error rate by the law of large numbers.

Suggested:

1. Suppose we have a code with four words distributed as $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$. If we assign equal length to each codeword, what is the level of data compression

(how many bits are required, on average)? Can we improve data compression? If so, suggest a code and determine the average number of bits required.

2. Verify the transmission rate for the noiseless binary channel.

3. Verify the transmission rate for the noisy four-symbol channel.

4. Verify the transmission rate for the binary symmetric channel.

Shannon's noisy channel coding theorem employs typical sequences (of a large number of transmissions) to produce arbitrarily small error rates. If a sequence of transmissions is not a typical sequence then there is an error. Typical sequences are defined by sequences for which the negative average logarithmic frequency almost surely converges by the law of large numbers (of transmissions) to the negative expected logarithmic probability (or entropy). For the binary symmetric channel, the probability distribution associated with a transmission has a Bernoulli distribution with entropy

$$H_2(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

and the (negative) sample average frequency of a sequence of n transmissions with s errors is

$$T_2(p) = -\frac{1}{n} \log_2 p^s (1-p)^{n-s}$$

$$= -\frac{1}{n} \{s \log_2 p + (n-s) \log_2 (1-p)\}$$

-

$$\begin{matrix} \lim \\ n \to \infty \end{matrix} \; T_2(p) \to H_2(p)$$

Applying the above to X^n, Y^n and (X^n, Y^n) defines jointly typical sequences.

Near error-free decoding involves creating disjoint or distinct encodings of the n input sequence such that decoding produces the input message. Decoding treats any jointly typical sequence as the transmitted code word. For each input n-sequence, there are approximately $2^{nH(Y|X)}$ possible equally likely typical Y sequences. Ralph wishes to ensure no two X sequences produce the same Y output sequence. The number of typical Y sequences is about $2^{nH(Y)}$. This set is divided into typical sequences, that is, sets of size $2^{nH(Y|X)}$ corresponding to the different input X sequences. The total number of disjoint sets is less than or equal to $2^{nH(Y)-nH(Y|X)} = 2^{nI(X;Y)}$. Hence, Ralph knows the upper bound on distinguishable length-n legal codewords is approximately $2^{nI(X;Y)}$ and the probability of any other (than the encoded message) jointly typical sequence is no greater than $2^{-nI(X;Y)}$ which can be made arbitrarily small for large n.

5. For the binary symmetric channel with $p = 0.1$ and $n = 10$, how many typical sequences are there for equi-probable input X^n? (hint: $2^{nH(X)}$)

— number of typical output sequences Y^n? (hint: $2^{nH(Y)}$)

— number of typical sequences Y^n for each input $X^n = x^n$? (hint: $2^{nH(Y|X)}$)

— number of typical joint sequences (X^n, Y^n)? (hint: $2^{nH(X,Y)}$)

— and number of distinguishable codewords? (hint: $2^{nI(X;Y)}$)

(the inverse of the total is the probability associated with a particular sequence)

6. Repeat 5 for $p = 0$ and $n = 10$.
7. Repeat 5 for $p = 0.5$ and $n = 10$.