Solutions:

1:

The mutual information between X and Y is

$$I(X;Y) \equiv H(X) - H(X|Y),$$

and satisfies $I(X;Y) = I(Y;X)$, and $I(X;Y) \geq 0$. It measures the average reduction in uncertainty about x that results from learning the value of y, **or vice versa**. Equivalently, it measures the amount of information that y conveys about x. [1]

The Capacity of a channel Q is the maximum mutual information between its input and its output, maximizing over the input distribution $\mathcal{P}_X$:

$$C(Q) = \max_{\mathcal{P}_X} I(X;Y).$$

The **optimal input distribution** is the distribution that achieves this optimum.

The first input distribution achieves $I = 1$, trivially.
The second distribution creates a binary erasure channel, whose mutual information is $1 - \frac{1}{2} = 0.5$. (This result may be quoted without proof.)
The probabilities for y are $a + b/2$, b, $a + b/2$. By decomposability of the entropy, imagining we learn "is it `M`?" first,

$$H(Y) = H_2(b) + (2a + b)H_2(0.5) = H_2(b) + 2a + b.$$

Because only inputs b and c create uncertain outputs,

$$H(Y|X) = 2b$$

So

$$I(X;Y) = H_2(b) + 2a + b - 2b = H_2(b) + 2a - b = H_2(b) + 1 - 3b.$$

The derivative is

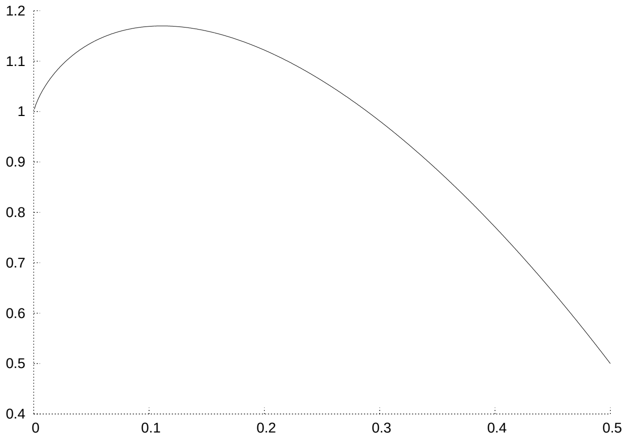
$$dI/db = \log_2\left(\frac{1-b}{b}\right) - 3,$$

which is zero when

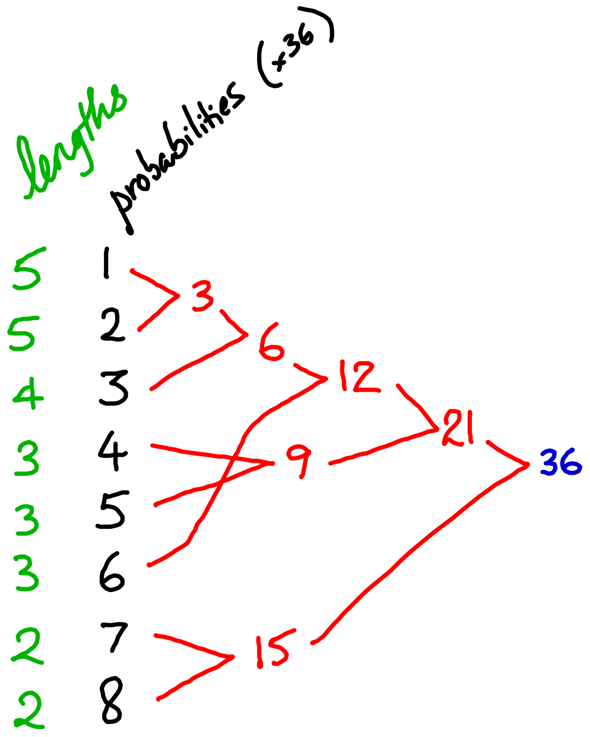
$$b = \frac{1}{1+2^3} = \frac{1}{9}$$

The sketch should feature correct values at the left and right sides, and a maximum in the right place. Technically the slope at zero is infinite, but this need not be noted.

(28 April 2007)



2: a) The implicit probability is $2^{-l(x)}$; the implicit probability is $A^{-l(x)}$.
b) The Huffman algorithm should be used to construct the optimal symbol code.



The expected length is

$$\frac{3 \times 5 + 3 \times 4 + 15 \times 3 + 2 \times 15}{36} = 102/36 = 2.83.$$

How much better is the optimal symbol code expected to compress N outcomes from this source than a simple code that assigns all outcomes a codeword of length 3 bits?

(28 April 2007)

$3 \times 36 = 108$, so (comparing with 102) it's better in terms of expected length per character by 6/36, or 1/6. So the expected length with the optimal symbol code is better by $N/6$ bits.

c) The number of plus codes of length l is 10^l. The number of distinct possibilities we want to encode is

(31) × (Number of start times) × (Number of durations) × (Number of channels). Assuming 5-minute precision for start times, and that durations could be specified to 5 minute precision up to 2 hours, then 10 minutes up to 6 hours,

$= 31 \times (24 \times 12) \times (24 + 24) \times 128 = 31 \times 288 \times 48 \times 128 = 55\,\text{million} < 10^8$. That requires 8 digits (followed by the 'end of word' code).

More discussion of "short lengths for probable channels, probable start times, probable durations". The ideal approach (assuming no computational constraints) is to sort all possible events by their assumed probabilities, and give the shortest codes to the most probable.

3: The likelihood ratio associated with one datum x_n is

$$\frac{P(x_n \,|\, \mathcal{H}_2)}{P(x_n \,|\, \mathcal{H}_1)} = 2x.$$

The likelihood ratio is thus

$$\prod_{n=1}^{N} (2x_n)$$

An answer saying something about the posterior or the log-likelihood ratio is expected here.

The example data set gives a ratio of 0.3 (thus the posterior probability of $\mathcal{H}_2$ is $0.3/1.3 = 0.23$).

The expected growth of the log likelihood ratio given $\mathcal{H}_1$ is linear growth with slope equal to the (negative) relative entropy between the two distributions,

$$\int_0^1 (1) \log \frac{2x}{1} dx = -0.3$$

The expected growth of the log likelihood ratio given $\mathcal{H}_2$ is linear growth with slope equal to the relative entropy between the two distributions,

$$\int_0^1 (2x) \log \frac{2x}{1} dx = 0.2$$

Thus the average log likelihood ratio after N data is either $-0.3N$ or $0.2N$ and to have a ballpark figure of 100:1, we need $N \simeq 4.6/0.2 \simeq$ **23**.

[Big bonus marks for anyone who does a fluctuation analysis. The standard deviation of the the log likelihood ratio is 1/2.]

Inference of α. The log likelihood is

$$\ln P(\{x\} \,|\, \alpha) = N \ln \alpha + (\alpha - 1) \sum \ln x.$$

Defining $G = \sum \ln x / N$,

$$\ln P(\{x\} \mid \alpha) = N \ln \alpha + (\alpha - 1) N G.$$

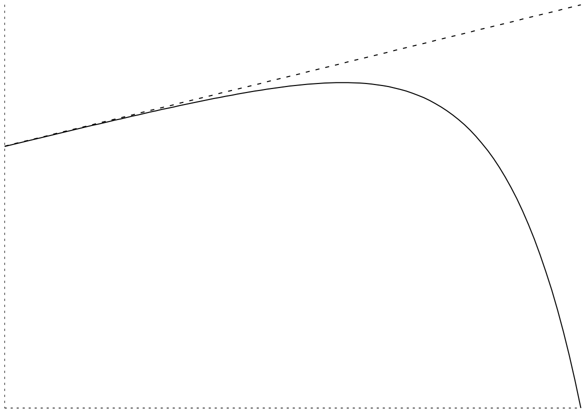
The derivative wrt $\ln \alpha$ is

$$(d/d \ln \alpha) \ln P(\{x\} \mid \alpha) = N + \alpha N G,$$

which is zero when

$$\alpha = -1/G.$$

The sketch should show a concave function of $\ln \alpha$. For large N, the exponential of this will be close to Gaussian.



The second derivative at the maximum is

$$-N$$

so the variance of $\ln \alpha$ (using a Gaussian approximation) is

$$\sigma^2_{\ln \alpha} = \frac{1}{N}$$

so finally the estimate and interval for α are

$$\ln \alpha \simeq \ln \left[\frac{N}{- \sum \ln x} \right] \pm \frac{1}{\sqrt{N}}.$$

(28 April 2007)