MDL exercises, third handout

(due March 3rd)

1. Consider the uniform distribution P on the set $\{a, b, c\}$.

   a) Show that there exists a prefix code C such that $L_C(x) = \lceil -\log P(x) \rceil$. Now consider strings from $\{a, b, c\}^{100}$. An easy way to code such strings is by using C a hundred times in a row. The result of this procedure is a new prefix code which we will call C'. Similarly, P can be extended to a distribution on 100 outcomes by defining $P'(x^n) = \prod_{i=1}^{100} P(x_i)$.

   b) Show that we no longer have for all strings z of length 100 that $L_{C'}(z) = \lceil -\log P'(z) \rceil$.

   c) Does this mean that there is no prefix code that corresponds to P' in the sense that its codelengths for each x are equal to to $\lceil -\log P'(x) \rceil$? If you think that there is one after all, then describe this code. And/or does this mean that there is no distribution that corresponds to C'? If you think that there is one after all, then describe this distribution.

2. A weird die has $P(1) = 1/24$, $P(2) = 1/12$, $P(3) = 1/8$, $P(4) = 1/6$, $P(5) = 1/4$, $P(6) = 1/3$.

   a) What is the entropy of this die?

   b) We know from the $1-1$ correspondence between code length functions and probability distributions that there exist prefix codes for which the expected codelength of an outcome is equal to the entropy, rounded up. Construct such a prefix code for the weird die above.

   c) Of all possible dice, pick the one that maximizes the entropy. What are the probabilities of each of the faces landing on top?

3. Let $f(n) = 2^{-n}$ and $g(n) = 1/(n(n+1))$ be two probability mass functions on the positive natural numbers. The corresponding (idealized) code-length functions are denoted $L_f(n) = -\log f(n)$ and $L_g(n) = -\log g(n)$.

   a) Show that both f and g are valid probability mass functions. In other words, show that for $n \geq 1$, we have $f(n)$ and $g(n)$ positive and $f(1) + f(2) + \ldots = g(1) + g(2) + \ldots = 1$.

   b) Draw $L_f(n)$ and $L_g(n)$ in a single graph.

   c) Let $\Delta(n) := L_f(n) - L_g(n)$. For which n do we have $\Delta(n) < 0$ and for which n do we have $\Delta(n) > 0$? For which n is Δ maximized and for which is it minimized? What are the corresponding values?

   d) Which would be more suitable as a code for the natural numbers from a data compression point of view? Why? **(see back side!)**

e) Let N be distributed according to the distribution with mass function g. What is $\mathbf{E}_{N \sim g}[N]$, i.e. what is the expected value of N? What is its entropy $\mathbf{E}_{N \sim g}[-\log g(N)]$? If you cannot calculate it exactly, give an upper bound using that $\sum_{i=1}^{\infty} i^{-3/2} < 3$ and that for all $i \geq 1$, $\log(i(i+1)) \leq 1.53\sqrt{i}$.

f) Now let N be distributed according to a distribution with mass function $p(n) = C/(n \cdot (\log(n+1))^2)$ for some $C < \infty$ (such a distribution exists). What is the expected value of N? What is its entropy?