Exercise: KL divergence and the number game

Consider a simplified version of the number game with only 4 possible hypotheses:

$$h_1 = \text{all numbers in 1..100} \tag{1}$$
$$h_2 = \text{even numbers } = \{2, 4, .., 98, 100\} \tag{2}$$
$$h_3 = \text{odd numbers } = \{1, 3, ..., 99\} \tag{3}$$
$$h_4 = \text{powers of 2 } = \{2, 4, 8, 16, 32, 64\} \tag{4}$$

Let $\mathcal{H} = \{1, 2, 3, 4\}$ represent the hypothesis space, and define the likelihood functions as

$$p_h(x_{1:n}) = p(x_{1:n}|h) = \left(\frac{1}{|\text{size}(h)|}\right)^n \prod_{i=1}^{n} I(x_i \in h) \tag{5}$$

This is the standard strong sampling model, which samples numbers uniformly at random from the extension (suppport) of the hypothesis, and gives 0 probability to data outside this set. Suppose the true distribution / hypothesis is

$$h^* = \text{powers of 4 } = \{4, 16, 64\} \tag{6}$$

Note that this is not in our hypothesis space, $h^* \notin \mathcal{H}$.

1. Compute the KL divergence from the true distribution to each possible model:

$$KL(h) = KL\left(p_{h^*}(\cdot)||p_h(\cdot)\right) = \sum_{x=1}^{100} p(x|h^*) \log \frac{p(x|h^*)}{p(x|h)} \tag{7}$$

   Use log base 2. (Hint: $\log_2 x = \frac{\log_{10} x}{\log_{10} 2}$.) Note that some answers may be infinite.

2. Which $h \in \mathcal{H}$ is the best approximation to the true distribution? Why?