Problems of the Week – 4 and 5

4. Huffman encoding and entropy

Entropy is a mathematical formulation of the uncertainty and/or the amount of information in a data set. Consider a data set D consisting of n characters, each character independently chosen from a set C according to a specified probability distribution p. That is, for $c \in C$ and $0 \leq i < n$, the probability that the ith character of D is c is $p(c)$. Note that $\sum_{c \in C} p(c) = 1$. The entropy of data set D is then defined to be

$$n \sum_{c \in C} p(c) \log_2(1/p(c)).$$

Intuitively, the entropy measures the *information-theoretic minimum* number of bits needed to represent the data set.

Prove that if all the probabilities are powers of 2 (i.e., for every c there exists an $i \geq 0$ such that $p(c) = 1/2^i$), then the expected number of bits used in the Huffman encoding of D equals its entropy.

5. Splitting an art heist

Two art thieves share a collection S of n paintings. They have agreed on a value for each painting, say $x_1, x_2, \ldots, x_n$. They would like to split the paintings into two sets as evenly as possible. That is, split S into two sets S_1 and S_2 that minimize $|\text{value of } S_1 - \text{value of } S_2|$ subject to the constraint that $S_1 \cup S_2 = S$ and $S_1 \cap S_2 = \emptyset$. (The value of a set of paintings is simply the sum of the values of the paintings in the set.)

Design an algorithm that determines S_1 and S_2 that achieve the above. Your algorithm should run in $O(nT)$ time, where T is the sum of the x_i's.