Homework problem set 2

Your homework must be handed in **electronically via Moodle before Friday November 18th, 20:00h**. This deadline is strict and late submissions are graded with a 0. At the end of the course, the lowest of your 6 weekly homework grades will be dropped. You are strongly encouraged to work together on the exercises, including the homework. However, after this discussion phase, you have to write down and submit your own individual solution. Numbers alone are never sufficient, always motivate your answers.

Problem 1: Deriving the weak law of large numbers (7pt)

(a) **(3pt)** (Markov's inequality) For any real non-negative random variable X and any $t > 0$, show that

$$P_X(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Exhibit a random variable (which can depend on t) that achieves this inequality with equality.

(b) **(2pt)** (Chebyshev's inequality.) Let Y be a random variable with mean μ and variance σ^2. By letting $X = (Y - \mu)^2$, show that for any $\varepsilon > 0$,

$$P(|Y - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

(c) **(2pt)** (The weak law of large numbers.) Let $Z_1, Z_2, ..., Z_n$ real i.i.d. random variables with mean $\mu = \mathbb{E}[Z_i]$ and variance $\sigma^2 = \mathbb{E}[(X_i - \mu)^2] < \infty$. Define the random variables $S_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$. Show that

$$P(|S_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Thus, $P(|S_n - \mu| \geq \varepsilon) \to 0$ as $n \to \infty$. This is known as the weak law of large numbers (Theorem 2.6.1 in the lecture notes).

Problem 2: Huffman Coding (7pt)

(a) **(4pt)** For a binary source P_X with $P_X(0) = \frac{1}{8}$ and $P_X(1) = \frac{7}{8}$, design a Huffman code for blocks of $N = 1, 2$ and 3 bits. For each of the three codes, compute the average codeword length and divide it by N, in order to compare it to the optimal length, i.e. the entropy of the source. What do you observe?

(b) **(1pt)** If you were asked in (a) to design a Huffman code for a block of $N = 100$ bits, what problem would you run into?

(c) **(2pt)** Consider the random variable Z with

z	1	2	3	4	5	6
$P_Z(z)$	1/10	3/10	2/10	2/10	1/10	1/10

Find an optimal *ternary* Huffman encoding for Z (i.e. using an alphabet with three symbols).

Problem 3: Inefficiency when using the wrong code (6pt)

(a) **(2pt)** Given are two distributions P_X and Q_X for the same set $\mathcal{X} = \{a, b, c, d\}$:

x	a	b	c	d
$P_X(x)$	1/4	1/4	1/4	1/4
$Q_X(x)$	1/2	1/4	1/8	1/8

Design an optimal code for $Q_X(x)$. What is its expected codeword length? What is the expected codeword length if you use this code to encode the source P_X?

(b) **(4pt)** Prove that in general, when using optimal code for any source source Q_X (with lengths $\ell(x) = \lceil -\log Q_X(x) \rceil$ for all $x \in \mathcal{X}$) to encode another source P_X, this incurs a penalty of $D(P_X || Q_X)$ in the average description length. More formally, prove that

$$H(P_X) + D(P_X || Q_X) \leq \mathbb{E}_{P_X}[\ell(X)] \leq H(P_X) + D(P_X || Q_X) + 1$$

for all P_X and Q_X.