# Diskrete Stochastik und Informationstheorie – 11 Jun 2014

**Exercise 36** (Optimal codes with one bit above entropy)**.** The source coding theorem shows that the optimal code for a random variable $X$ has an expected length less than $H(X)+1$. Give an example of a random variable such that the expected length of an optimal code is very close to $H(X)+1$ to show that this bound is tight. That is, construct a random variable $X_\varepsilon$ for every $\varepsilon > 0$ such that its optimal code $C_\varepsilon$ fulfills $\mathbb{E}(\ell(C_\varepsilon(X_\varepsilon))) > H(X_\varepsilon) + 1 - \varepsilon$.

**Exercise 37** (Coding for geometric distribution)**.** Let $X_n : \Omega \to \mathbb{N}$ be a geometric random variable with

$$\mathbb{P}[X_n = i] = (1-q)^{i-1} \cdot q,$$

where $0 \le q \le \frac{1}{2}$. The possible values of $X_n$ are encoded as follows:

$$C(1) = 0, \qquad C(2) = 10, \qquad C(3) = 110, \ldots,$$

that is, value $i$ is encoded to an $i$-bit string in a prefix-free way.

(a) Find $q^*$ such that for all values $q \ge q^*$, the expected length for this encoding is at most one bit above entropy (but not for any $q < q^*$).

*Hint*: Simplify the expressions for $\mathbb{E}(\ell(C(X)))$ and $H(X)$ and then find a solution for the remaining equation numerically.

(b) For $q < q^*$, Shannon coding is more efficient. There, the codeword lengths are given by

$$\ell(C(i)) = \left\lceil -\log_2 (1-q)^{i-1} \cdot q \right\rceil.$$

Show that for this encoding, the expected code length is at most one bit above entropy.

(c) For Shannon coding and $q = 0.01$, find $i_0$ and $i_1$ such that

$$i_0 = \max\{i \in \mathbb{N} : \ell(C(i)) = \ell(C(1))\}, \qquad i_1 = \max\{i \in \mathbb{N} : \ell(C(i)) = \ell(C(i_0+1))\},$$

that is, the values 1 to $i_0$ are all encoded to the shortest codeword length, while $i_0+1$ to $i_1$ are encoded to the second shortest. Argue that a prefix-free coding for the values $1, \ldots, i_1$ with the above lengths is possible.

(d) What is the difference between the codings of (a) and (b)? Which would you prefer?

**Exercise 38.** Let $X : \Omega \to \{a, b, c, d, e\}$ be a random variable with the distribution

$$\mathbb{P}[X = a] = \frac{1}{3}, \qquad \mathbb{P}[X = b] = \mathbb{P}[X = c] = \frac{1}{5}, \qquad \mathbb{P}[X = d] = \mathbb{P}[X = e] = \frac{2}{15}.$$

(a) Find a binary Huffman code for this random variable and compare $H(X)$ to $\mathbb{E}(\ell(C(X)))$.

(b) Argue that this code is also optimal for a uniformly distributed random variable with the same values, and compare $H(X)$ and $\mathbb{E}(\ell(C(X)))$ for this case.

**Exercise 39.** Let $X : \Omega \to \{1, \ldots, 10\}$ be a random variable with uniform distribution.

(a) Give three different binary prefix-free codings: one for each $k \in \{1, 3, 8\}$ such that $\max_{i,j \in \{1,\ldots,10\}} |\ell(C(i)) - \ell(C(j))| = k$. Give $H(X)$ and $\mathbb{E}(\ell(C(X)))$ for each code.

(b) Construct a binary Huffman code and compare it with your three codes. Is any of the three codes already a Huffman code?

**Exercise 40.** Which of the following Codes can certainly not be a Huffman code (with respect to any probability distribution): (i) $\{01, 10\}$, (ii) $\{0, 10, 11\}$, (iii) $\{00, 01, 10, 110\}$?