Information Theory & Statistical Physics

Lecture: Prof. Dr. Johannes Berg
Exercises: Stephan Kleinbölting **Sheet 2**

Due date: 18.05.17 12:00 *To be discussed on: 24.05.17*

8 Entropy II 20 pts.

Let X, Ξ be independent discrete random variables and consider their sum $Z = X + \Xi$. For the sake of intuition we may interpret X as a signal and Ξ as noise. Intuitively it is expected that noise increases uncertainty and hence entropy.

(a) *10pt* - First show that the entropy conditioned on Ξ is unaffected.

$$H(Z|\Xi) = H(X|\Xi) = H(X) \tag{1}$$

(b) *5pt* - Show then that

$$H(Z) \geq H(X) \quad \text{and} \quad H(Z) \geq H(\Xi) \tag{2}$$

and hence

$$\max\{H(X), H(\Xi)\} \leq H(X + \Xi)$$

(c) *5pt* - Demonstrate that the independence of signal and noise is crucial for (2) to hold by constructing a counterexample!

9 Entropy of the normal distribution 20 pts.

Let X be a continuous random variable on $\mathbb{R}$ with PDF $f(x)$.

(a) *2pt* - Show that the *differential entropy*

$$h(X) = -\int f(x) \ln f(x) \mathrm{d}x$$

is invariant under translations

$$h(X + c) = h(X), \quad c \in \mathbb{R} \tag{3}$$

(b) *8pt* - Let $Y \sim \mathcal{N}(0, \sigma)$ be normally distributed. Show that

$$h(Y) = \ln(\sigma\sqrt{2\pi e}) \tag{4}$$

(c) *10pt* - Assume that X also has variance σ^2. Show that $h(Y) \geq h(X)$, that is from all the distributions with a given (finite) variance, the Gaussian possesses the largest entropy. Due to (a) we may restrict ourself to a centered random variable ($\langle X \rangle = 0$).
Hint: Consider the Kullback-Leibler divergence $D(X||Y)$ and exploit its positivity.

10 Kraft inequality and optimal codes 20+5 pts.

A *source code* C over an ensemble $(X, \mathcal{X}, p_X)$ is a mapping from $\mathcal{X}$, the range of X, to a set of finite-length strings composed from an alphabet $\mathcal{D}$. For example a binary code over the (lower-case) Latin alphabet could be

$$C(a) = 00000, C(b) = 00001, C(c) = 00010, \ldots, C(z) = 11010$$

The *extended code* C^+ maps strings of source symbols onto strings of code symbols

$$C^+(x_1 x_2 \cdots x_n) = C(x_1)C(x_2)\cdots C(x_n)$$

A code is called *uniquely decodable* if

$$C^+(x) = C^+(y) \Rightarrow x = y$$

i.e no two source strings have the same encoding. It is called a *prefix-, or instantaneous, or self-punctuating code*, if no codeword is the prefix of another. Prefix codes are necessarily uniquely decodable.

Let $l(x)$ denote the length of a codeword.

We want to derive an important inequality due to Kraft which all prefix codes obey:

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1 \tag{5}$$

where $D = |\mathcal{D}|$ is the length of the code alphabet.

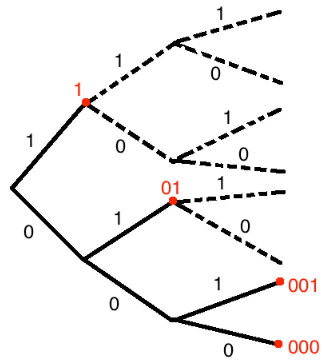

Figure 1: **Tree presentation of a binary prefix code.** A codeword of a prefix code terminates a branch of the tree at some node, which then becomes a leaf of the restricted tree (i.e excluding the dashed branches). A codeword is the series of letters in the branches from the root to the leaf.

(a) *0pt* - Convince yourself – not your tutor! – that a prefix code can be represented as a tree of depth $l_{max} = \max_x\{l(x)\}$, whereby the prefix property implies that each codeword terminates the respective branch of the tree. As an example consider the figure representing the binary code $\{1, 01, 000, 001\}$.

(b) *10pt* - To prove the Kraft inequality, consider the number of descendants a given codeword of length $l(x)$ would have on level l_{max} of the unrestricted tree. What is the total number of descendants at level l_{max}? Bound this by the total number of leafs to prove the inequality.

The expected length of a code is given by

$$L(C, X) = \sum_{x \in \mathcal{X}} p_X(x) l(x) \tag{6}$$

An optimal prefix code is one that minimizes $L(C, X)$.

(c) *10pt* - Minimize (6) constraint by the Kraft inequality (5). You may treat $l(x)$ as a real number instead of an integer and assume that the Kraft inequality is saturated. Show that the optimal code lengths are given by

$$l(x) = \log_D \frac{1}{p_X(x)} \tag{7}$$

and hence the entropy $H(X)$ (in base D) is a lower bound on L.

(d) **Bonus 5pt* - Assume we had erroneously assigned lengths according to a distribution $q(x)$ instead of the true one $p(x)$

$$l^*(x) = \left\lceil \log_D \frac{1}{q(x)} \right\rceil$$

Show that this choice incurs a penalty on the expected length of the code

$$L^* = \langle l^* \rangle_p \geq H(X) + D(p||q) \tag{8}$$

The reverse statement holds true as well. Given a set of codeword lengths $\{l(x), x \in \mathcal{X}\}$ obeying (5), there exists a prefix code with these lengths.

N.B.: One might think that dropping the prefix property could yield even better codes. Maybe somewhat surprisingly it turns out that the Kraft inequality must be obeyed by *all* uniquely decodable codes. Hence no improvement is gained by dropping the prefix property.