Information Theory & Statistical Physics

Lecture: Prof. Dr. Johannes Berg
Exercises: Stephan Kleinbölting **Sheet 1**

Due date: 04.05.17 12:00 *To be discussed on: 10.05.17*

5 Entropy I

30 pts.

(a) (10 pts) Show that the entropy $H(p_1, \ldots, p_n)$ assumes its maximum value for a uniform distribution.
Note: You need to account for the constraint of total probability $\sum p_n = 1$ via a Lagrange multiplier.

(b) (10 pts) Now consider a random variable A and maximize H under the additional constraint $\langle A \rangle = a$. What is the maximizing distribution? Does it look familiar?

In the lecture the entropy was introduced axiomatically. One of the axioms ensures that it can be decomposed into entropies of smaller ensembles.

$$H(p_1, p_2, ..., p_n) = H(p_1, 1-p_1) + (1-p_1)H(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, ..., \frac{p_n}{1-p_1}). \tag{1}$$

(c) (10 pts) A biased coin is tossed until a "tail" is thrown. The probability for the coin to come as "head" in a single toss is f. Calculate the entropy of the (random) number of necessary tosses in two ways: using the definition of entropy, and using the decomposition of entropy (1).

Solution

(a) Find the maximum of $H(\{p_j\}) - \lambda(\sum_j p_j - 1)$ wrt. p_j:

$$0 = \mathrm{d}\left(-\sum_j (p_j \log p_j + \lambda p_j)\right) = \sum_j (1 + \lambda + \log p_j)\mathrm{d}_{p_j} = 0$$
$$\Rightarrow \quad p_j = \exp(-1-\lambda) = \text{const}$$

Its a maximum due to concavity.

(b) Similarily $H(\{p_j\}) - \lambda(\sum_j p_j - 1) - \beta(\sum_j p_j A_j - a)$ has an extremum at

$$0 = \log p_j + \lambda + 1 + \beta A_j \quad \Leftrightarrow p_j = \exp(-1-\lambda)\exp(-\beta A_j)$$

which is just the Boltzmann distribution.

(c) The number of throws until "tail" comes up X is geometrically distributed

$$P(X = k) = f^{k-1}(1-f)$$

On the one hand one calculates directly

$$\begin{aligned}
H &= -\sum_{k=1}^{\infty} f^{k-1}(1-f)\left((k-1)\log f + \log(1-f)\right) \\
&= -(1-f)\log f \sum_k f^{k-1}(k-1) - (1-f)\log(1-f)\sum_k f^{k-1} \\
&= -(1-f)\log f \left\{\frac{1}{(1-f)^2} - \frac{1}{1-f}\right\} - (1-f)\log(1-f)\frac{1}{1-f} \\
&= -\log f \frac{f}{1-f} - \frac{(1-f)\log(1-f)}{1-f} = \frac{H_2(f)}{1-f}
\end{aligned}$$

More easily one recognizes that after the first throw is unsuccesful the next throw follows the same distribution again. Hence, by the decomposition property, we have a recursion

$$H(X) = H_2(f) + fH(X)$$

Or, more explicitely (draw a bifurcating tree)

$$H(X) = H_2(f) + fH_2(f) + f^2H_2(f) + \cdots$$

6 Sum of random variables

20 pts.

Let X, Y be two discrete, independent random variables that take values in $\mathcal{A}$.

(a) (2 pts) Show that independence implies $p_X(x|y) = p_X(x)$.

Define a new random variable as their sum $Z = X + Y$.

(b) (10 pts) Show that Z is distributed according to the *convolution* of p_X and p_Y

$$p_Z(z) = \sum_{x\in\mathcal{A}} p_X(x)p_Y(z-x) \tag{2}$$

Hint: Start by considering $p_Z(z) := P(Z = z) = P(X + Y = z)$ and rewriting this as an expectation value over a conditional distribution.

(c) (8 pts) Consider two cubic dice that you may label from the set $\{0, \dots, 6\}$ as you like. Is it possible to label them in such a way, that the sum of a roll of both dice has a uniform distribution over $\{1, \dots 12\}$? If yes, find such a configuration.

Solution

(a) $p(x|y) = p(x, y)/p(y) = p(x)p(y)/p(y)$ if X, Y are independent.

(b)

$$\begin{aligned}
p_Z(z) = P(X+Y=z) = P(X = z - Y) &= \sum_x P(Y = z - X|X = x)P(X = x) \\
= \langle P(Y = z - X|X)\rangle_X \overset{(*)}{=} \langle P(Y = z - X)\rangle_X &= \sum_x p_X(x)p_Y(z-x))
\end{aligned}$$

$(*)$ because X, Y are independent.

(c) The unique solution is to label one die $1, \dots, 6$ and the other $0, 0, 0, 6, 6, 6$.

Denote the number of j's on the first/second die by m_j and n_j.

There are a total of 36 outcomes, each is supposed to be equally probable. Hence each outcome must have 36/12 microstates realizing it. A 12 can only be rolled with two

6s. Since 3 is prime, the only way is for one die to have one 6 and the other three 6s, w.l.o.g. we set $m_j = 3, n_j = 1$.

An 11 is only possible as a combination of 5 and 6. We may either set $m_5 = 3, n_5 = 0$ or $m_5 = 0, n_5 = 1$. The former choice would exhaust dice number one and would fix 5 as the smallest possible outcome. We must therefor choose the latter option. The same argument holds for the next lower sums and one is forced to set $m_5 = m_4, m_3, m_2 = 0$ and $n_5 = n_4 = n_3 = n_2 = 1$.

This leaves one with distributing 1s and 0s. The only viable choice is $m_0 = 3, n_0 = 0, m_1 = 0, n_1 = 1$.

7 Weak law of large numbers

10 pts.

Let X_n be a sequence of iid. random variables with mean μ and variance σ^2. Let $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ denote their arithmatic mean.

Show that in the limit of large N, the arithmatic mean converges (in mean square) to μ

$$\lim_{N\to\infty} \mathrm{Var}(\bar{X} - \mu) = 0 \tag{3}$$

Solution

$$\mathrm{Var}(\frac{1}{n}\sum X_n - \mu) = \mathrm{Var}(\frac{1}{n}\sum X_n) \overset{(*)}{=} \frac{1}{n^2}\sum \mathrm{Var}(X_n) = \frac{\sigma^2}{n^2} \to 0$$

$(*)$ because $\{X_j\}$ are independent. The weak law of large numbers is usually stated wrt. the weaker notion of *convergence in probability*

$$\forall \epsilon > 0 : \mathrm{P}(|1/n \sum X_n - \mu| > \epsilon) \to 0$$

But by Markov's inequality convergence in the mean (square) – actually convergence in any $L^r, r \geq 1$ norm – implies convergence in probability.

Actually, the assumptions about X_j can be relaxed to either one of: They have finite variance, but need only be uncorrelated instead of independent. Or, one insists on iid. variables, but may drop the finite variance assumption.