Kaleidoscope probability theory

Matthias Löwe

April 4, 2003

These are the lectures notes for a series of four lectures given to first year students as their first encounter with probability (except for what has been done at school). In these lectures we will meet a couple of situations that can conveniently be described with the help of probabilistic models. We will refrain from giving an axiomatic description of probability theory – this is content of the courses in measure theory and integration and probability theory – and just define things loosely whenever we need them.

1 Poisson Approximation

In the first part of this course we will meet a situation where probability is naturally found in a setup of an experiment in physics. In a well known experiment the physicists Ernest Rutherford (a Noble Prize Winner) and E. Geiger (the one who invented the Geiger counter) investigated a radioactive substance. More precisely, they measured the emission of α - particles by this substance. The investigated 2608 times how many α–particles were emitted in a time period of 7.5 seconds. The following table shows their results. Here, for $i \in \mathbb{N}$, the number n_i indicates the number of time intervals, where precisely i α–particles were emitted. $\nu_i$denotes the relative frequency of these time intervals

i	n_i	ν_i
0	57	0.02186
1	203	0.0778
2	383	0.1469
3	525	0.2013
4	532	0.2040
5	408	0.1564
6	273	0.1047
7	139	0.0533
8	45	0.0173
9	27	0.0103
10	10	0.0038
11	4	0.0015
12	0	0
13	1	0.0004
14	1	0.0004

Probabilists who look at this table may have the idea that the number ν_i are close to a well known probability, the Poisson–distribution. The Poisson distribution is a probability that assigns probabilities $\pi_\lambda(k)$ to the natural numbers $k \in \mathbb{N}_0$. More precisely, the Poisson distribution is defined as follows.

Definition 1 *Let $\lambda > 0$. The Poisson distribution with parameter λ is defined as the probability*

$$\pi_\lambda(k) = e^{-\lambda}\lambda^k/k!$$

for $k \in \mathbb{N}_0$.

First we remark that indeed $\pi_\lambda(k)$ is a probability distribution on the natural numbers $\mathbb{N}_0$.This means that

$$\sum_{k=0}^{\infty} \pi_\lambda(k) = 1.$$

Indeed,

$$\sum_{k=0}^{\infty} \pi_\lambda(k) = \sum_{k=0}^{\infty} e^{-\lambda}\lambda^k/k! = e^{-\lambda}e^{\lambda} = 1.$$

Now we will first try to study this probability a bit more in detail. Assume we had a lottery where numbers $k \in \mathbf{N}_0$ are drawn according to the probabilities $\pi_\lambda(k)$. What would be the average of the numbers drawn? Since we draw $k \in \mathbf{N}_0$with probability $\pi_\lambda(k)$ we would expect this average to be equal to

$$\begin{aligned}\mathbb{E}_{\pi_\lambda} &= \sum_{k=0}^{\infty} k\pi_\lambda(k) = \sum_{k=0}^{\infty} ke^{-\lambda}\lambda^k/k! \\ &= e^{-\lambda}\lambda\sum_{k=1}^{\infty}\lambda^{k-1}/(k-1)! = e^{-\lambda}\lambda\sum_{k=1}^{\infty}\lambda^k/k! = \lambda e^{-\lambda}e^{\lambda} = \lambda.\end{aligned}$$

Hence we just proved the following

Lemma 2 *The expectation value (i.e. the value $\mathbb{E}_{\pi_\lambda}$) of the Poisson distribution π_λ with parameter λ is the parameter λ itself.*

The interpretation of the expectation value is the following: if we realize many numbers according to the π_λ–distribution we will expect that their average is close to λ (the expectation).
From the data in the table probabilists will conjecture that the number of emissions of α–particles is Poisson distributed. To check this we first have to guess the parameter λ. From what we learned above the parameter λ then needs to be close to the average number of emitted α–particles. In the experiment of Rutherford and Geiger this average number is equal to

$$a = \frac{10097}{2608} \sim 3.87.$$

In the following table we compare the numbers ν_k from the first table to those of a Poisson distribution with parameter $\lambda = 3,87$.

k	ν_k	$\pi_\lambda(k)$
0	0.0219	0.0208
1	0.0778	0.0807
2	0.1469	0.1561
3	0.2013	0.2015
4	0.2040	0.1949
5	0.1564	0.1509
6	0.1047	0.0973
7	0.0533	0.0538
8	0.0173	0.0260
9	0.0103	0.0112
10	0.0038	0.0043
11	0.0015	0.0015
12	0	0.0005
13	0.0004	0.0002
14	0.0004	4×10^{-5}

We see that the values of $\nu_k(n)$ and $\pi_\lambda(n)$ differ only by very little. We want to understand why this is the case. This insight is bases on the following ideas.

Consider a game where we toss a coin n - times. Say p is the probability to toss heads and $1-p$ is the probability to toss tails. It is obvious that the coin has no memory, i.e. that knowing that we tossed head in the first toss, does not influence the probability to toss heads in the second trial. This phenomena is called independence. Two events A, B are called independent if the probability that both events occur at the same time, i.e. $P(A \cap B)$ is the product of the probabilities, i.e.

$$P(A \cap B) = P(A)P(B).$$

Hence for our n independent coin tosses with probability p for heads, the probability for each sequence with exactly k heads is $p^k(1-p)^{n-k}$. Since there are $\binom{n}{k}$ possibilities to locate the k heads the probability of seeing k heads in n tosses is $\binom{n}{k}p^k(1-p)^{n-k}$. A convenient way to write this is the following. For each $1 \leq i \leq n$, write $X_i = 1$, if the i'th toss shows heads and $X_i = 0$ if it is tails. Then $\sum_{i=1}^n X_i$ equals the number of heads in the n tosses. We have just seen

Lemma 3 *In the above situation it holds*

$$P\left(\sum_{i=1}^{N} X_i = k\right) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We will now consider a situation where we toss very many coins (n goes to infinity) which have an extremely small probability for heads (p will be of order $\frac{\lambda}{n}$). Then we prove the following

Theorem 4 *Consider the above situation of n coin tosses with success probability p. Assume that $p = \frac{\lambda}{n}$ (hence that p is n - dependent). Then $P\left(\sum_{i=1}^{n} X_i = k\right) \to \pi_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ as $n \to \infty$.*

Proof. We have seen that

$$P\left(\sum_{i=1}^{n} X_i = k\right) = \binom{n}{k} p^k (1-p)^{n-k}$$
$$= \frac{n \cdot (n-1) \cdot \cdot (n-k+1)}{k!} \frac{p^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k}$$
$$\to \frac{\lambda^k}{k!} e^{-\lambda}$$

as $n \to \infty$. ■

All of this may be very nice but what does it have to do with the experiment conducted by Rutherford and Geiger.

This question will be discussed next. To do so, we model the Rutherford–Geiger experiment as follows. Let $(0, \infty)$ be the time axes. In $(0, \infty)$ the times where there is a radioactive emission are marked by points. For $I = (t, t+s] \subseteq (0, \infty)$ we denote by N_I the number of points in I (number of particles emitted in I). Thus N_I is a random variable with values in $\mathbb{N}_0$. For $N_{(0,t]}$ we simply write N_t.

We assume the following

1. N_I depends on the length of I only (this is only true if our observation time is much shorter than the halftime of the substance).

2. For $I_1, ..., I_k$ pairwise disjoint the random variables $N_{I_1}, ...N_{I_k}$ are independent

3. For I of a finite lenght the average (expected) number $\mathbb{E}N_I$ is finite.

4. To avoid trivialities we assume that there is an interval I with $P(N_I > 0) > 0$.

   From (1) - (4) we can already conclude something. Denote

$$\lambda(t) = EN_t \geq 0.$$

Since we set $N_0 \equiv 0$ we have $\lambda_0 = 0$. Moreover, the number of points in disjoint time intervals is simply the sum of the points in the two intervals. In particular,

$$N_{t+s} = N_t + N_{(t,t+s]}$$

and thus

$$\begin{aligned} \lambda(t+s) &= \lambda(t) + EN_{(t,t+s]} \\ &= \lambda(t) + \lambda(s) \end{aligned}$$

the latter because of (1). From analysis it follows that $\lambda(\cdot)$ is a linear function, i.e. there is $\lambda \geq 0$ with $\lambda(s) = \lambda s$. The case $\lambda = 0$ can be excluded because of (4.). Otherwise we would have $EN_I = 0$ for all I and hence $P(N_I = 0) = 1$ for all I. For small intervals I the probability to find a point (have an emission) in I is very small. Indeed,

$$\begin{aligned} P(N_I \geq 1) &= \sum_{k=1}^{\infty} P(N_I = k) \leq \sum_{k=1}^{\infty} kP(N_I = k) \\ &= \mathbb{E}N_I. \end{aligned}$$

Therefore,

$$P\left(N_{(t,t+\varepsilon]} \geq 1\right) \leq \lambda\varepsilon$$

for all $t, \varepsilon \geq 0$. Our last assumptions basically says that there are no double emissions, hence that any two points can be separated. Mathematically speaking for $T > 0$ define

$$D_T := \inf_{t,s \leq T} \{|t-s| : |N_t - N_s| \geq 1\}.$$

Assumption (5) then is

5. $P(D_T \leq \alpha_n) \to_{n\to\infty} 0$ for each sequence $\alpha_n \to 0$ and all $T > 0$ finite. So, if we believe that the emissions in the Rutherford - Geiger experiment obey (1) - (5) we will show that they are approximately Poisson distributed with parameter λs.

Proof. Because of (1) it suffies to consider $N_s = N_{[0,s)}$. Fix s. For $k \in \mathbb{N}, 1 \leq j \leq k$ define

$$\begin{aligned} \chi_j^{(k)} &:= N_{(s(j-1)/k, sj/k]} \\ \bar{\chi}_j^{(k)} &:= \begin{cases} 1 & \text{if } \chi_j^{(k)} > 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Due to (2) the random variable $\chi_j^{(k)}$ are indpenedent and thus also $\bar{\chi}_j^{(k)}$ are independent. Let us collect some properties of these random variables. First $N_s = \sum_{j=1}^k \chi_j^{(k)}$. If $\bar{N}_s^{(k)} := \sum_{j=1}^k \chi_j^{(k)}$ then for all k

$$\bar{N}_S^{(k)} \leq N_s.$$

Hence

$$P\left(\bar{N}_s^{(k)} \geq m\right) \leq P\left(N_s \geq m\right).$$

Define

$$\begin{aligned} p_k &:= P\left(\chi_i^{(k)} = 1\right) = P\left(\chi_i^{(k)} \geq 1\right) \\ & \quad P\left(N_{\frac{1}{k}} \geq 1\right). \end{aligned}$$

Hence $\bar{N}_s^{(k)}$ is binomially distributed with parameters k and p_k. We will use (5) to show that for large k the random variables N_s and $\bar{N}_s^{(k)}$ differ only by very little. Indeed, if $N_s \neq \bar{N}_s^{(k)}$, then there is at least one time interval of lenght $\frac{1}{k}$ containly at least two points. Hence due to (5)

$$P\left(\bar{N}_s^{(k)} \neq N_s\right) \leq P\left(D_s \leq \frac{1}{k}\right) \to 0.$$

as $k \to \infty$. Hence

$$P\left(N_s = m\right) \text{ and } P\left(\bar{N}_s^{(k)} = m\right)$$

are approximately the same for large k, i.e.

$$P\left(N_s = m\right) = \lim_{k\to\infty} P\left(\bar{N}_s^{(k)} = m\right).$$

But $\bar{N}_s^{(k)}$ was binomially distributed with parameters k and p_k. We will now show that

$$\lim_{k\to\infty} kp_k = \lambda s$$

hence that $p_k \sim \frac{\lambda s}{k}$ such that the previous theorem can be applied. Indeed

$$kp_k = E\bar{N}_s^{(k)} = \sum_{j=1}^{\infty} jP\left(\bar{N}_s^{(k)} = j\right) = \sum_{m=1}^{\infty} P\left(\bar{N}_s^{(k)} \geq m\right).$$

Hence

$$\begin{aligned}\lim_{k\to\infty} k_{p_k} &= \lim_{k\to\infty} \sum_{m=1}^{\infty} P\left(\bar{N}_s^{(k)} \geq m\right) \\ &= \sum_{m=1}^{\infty} P\left(N_s \geq m\right) = EN_s = \lambda_s.\end{aligned}$$

This shows that

$$p_k \sim \frac{\lambda s}{k}.$$

Therefore we can apply the previous theorem and conclude that N_s is Poisson distributed with parameter λs. ■

2 Huffman coding

The following is part of a mathematical discipline called "information theory". Information theory says nothing about the content and value of an information. The main issue in this paragraph will be the following question: Given an alphabet $\Omega = \{\omega_1, ..., \omega_n\}$ of letters we want to encode them in such a way in $0-1$– sequences that the average length of the codeword is minimal. The notion "average length of the codewords" of course implies that there is a probability distribution $p_1, ...p_n$ (p_i the probability of ω_i) on Ω. It will turn out that the average codeword length cannot be made arbitrary small and that for the lower bound a quantity, called the entropy of $p\,(p = (p_1, ..., p_n))$ is of interest. The entropy could also (but isn't) be called the "mean surprise". For $p_i \in (0, 1)$, the number $-\log p_i$ could denote the "surprise" we feel when we see ω_i realized. The "mean surprise" would then be

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i.$$

This function is usually called the entropy of p. It is intrinsically related (but not quite the same) to the game where one person selects a letter $\omega_i \in \Omega$ according to p and another person is supposed to find ω_i with as few questions as possible [Here questions are only such questions that can be answered with "yes" or "no"; of course we are not allowed to ask "which of the ω_i is it?"].

Definition 5 *: The function $H_0(p)$, called the **true entropy** denotes the mean number of questions when we use an optional strategy.*

Example 6 *a) For $\Omega = \{\omega, \omega_2\}$ obviously for any p, it is optional to ask "Is it ω_1"? Therefore*

$$H_0(p) = 1 \text{ for all } p, \text{ if } |\Omega| = 2.$$

b. Also for $|\Omega| = 3$ and $p = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$ the best strategy is easily guessed. We first ask: "Is it ω_1?" In case the answer is no, we continue and ask: "Is it ω_2?". This strategy has average lenght:

$$H_0(p) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{3}{2}$$

First asking for ω_2, on the other hand, needs

$$\frac{1}{4} \cdot 1 + \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{7}{4}$$

questions in average.

To link the game to our original question about coding and to make the questions more precise, we introduce the notion of a code. Instead of "yes" or "no" we write 0 and 1. A word is a finite sequence of 0's and 1's. For a word μ_1, $|\mu|$ denotes its length. A word μ_1 is called a prefix of a word μ_2, if $|\mu_1| < |\mu_2|$ and the first $|\mu_1|$ letters of μ_2 agree with μ_1. The relation between codes and the above game is given by the strategy in which we ask questions. Once we know this strategy the code of every $\omega_i \in \Omega$ is unique.

Definition 7 *A prefix code κ for (Ω, p) is an injective mapping*

$$\kappa : \Omega \to \{0,1\}^{\mathbb{N}}$$

such that $|\kappa(\omega_i)| < \infty$ for all i and no $\kappa(\omega_i)$ is the prefix of another word $\kappa(\omega_j)$.

The injectiveness of κ is needed to ensure that no two letters have the same codeword. If the code were not a prefix code, the code would not be uniquely decipherable. E.g: $\Omega = \{a, b, c\}, \kappa(a) = 0, \kappa(b) = 01, \kappa(c) = 1$. For this code the word 001 could correspond to "ab" or also to "aac".

Now let (Ω, p) be such a probability space and κ be a code. The expected codeword lenght is defined as

$$E(|\kappa|) = \sum_{i=1}^{n} p_i |\kappa(\omega_i)|.$$

with this definition we can also make our definition of $H_0(.)$ more precise.

Definition 8 *For (Ω, p) the true entropy $H_0(p)$ is defined by*

$$H_0(p) = \min\{E(|\kappa|) : k \text{ is a code for } (\Omega, p)\}.$$

Sometimes it is useful to visualize codes as binary trees. Here the vertices of the tree is the set of all codewords and all its prefixes. This set is called $V(\kappa)$. We draw an edge between a word μ and μa, $a \in \{0, 1\}$, if $\mu, \mu_a \in V(\kappa)$. These connections we collect in $\mathcal{E}(\kappa)$. $V(\kappa)), (\mathcal{E}(\kappa))$ then is a connected graph without circles, hence a tree. We order the elements of $V(\kappa)$ increasing in their lenght. So, on the lowest level of the tree we put the empty word at the root of the tree. Then, in increasing order the words of lenght one, two, etc. Here we will connect μ to μ_1 by drawing an edge to the up - right end μ_0 will be an up - left connection.

Example 9 $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$.

$$\kappa(\omega_1) = 00, \quad \kappa(\omega_2) = 010, \quad \kappa(\omega_3) = 10, \kappa(\omega_4) = 110, \kappa(\omega_5) = 1111$$

Then

$$K(\kappa) = \{\emptyset, 0, 1, 00, 01, 10, 11, 010, 110, 111, 1111\}.$$

The corresponding tree is found in Fugure 1 below.

From the tree of a code the corresponding questions can be directly derived. In the above example one would first ask: Is is it $\omega_3, \omega_4, \omega_5$?" If the answer were "yes" we are in state "1" otherwise in "0" and we go on.

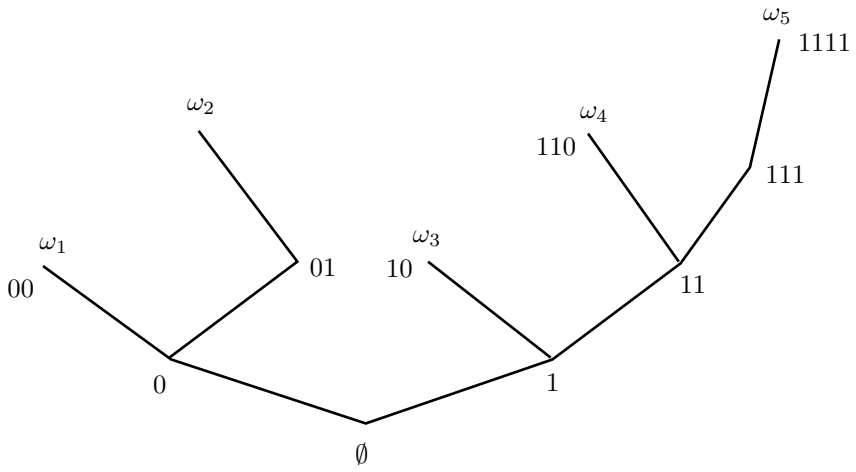

Figure 1: The tree

We will call such a binary tree complete, if for each word $\mu \in V(\kappa)$, which is not a leaf, i.e. which is none of the codewords, both μ_0 and μ_1 belong to $V(\kappa)$. It is evident that when searching for a code with minimum lenght, we only need to concentrate on such codes with complete trees. Other trees contain superfluous questions. We call a code complete if the corresponding code is complete. Uncomplete trees can be shortened and completed by deleting superfluous vertices.

Example 10 *We consider the code consisting of the code words* 01*,* 1101*,* 1110*,* 1111*. DarauThis gives the tree in Figure 2 below*

This tree can be schortened to the tree given in Figure 3 below in order to obtain the better code with the codewords 0*,* 10*,* 110*,* 111*.*

The mathematician Huffman gave a way to construct an optimal code. This code is called the Huffman code. The code is constructed recursively in n, where $n = |\Omega|$.
It suffies to assume that $p_i > 0$ for all i (otherwise we omit those ω_i with $p_i = 0$) and that $\Omega = \{1, ..., n\}$.
For $n = 2$, obviously, $\kappa(1) = 0$ and $\kappa(2) = 1$ is optimal for each $p = (p_1, p_2)$.
Now take $n > 2$ and assume we already constructed all Huffman codes of lenght $n - 1$. First of all note that the order of the letters $\omega_1, ..., \omega_n$ does

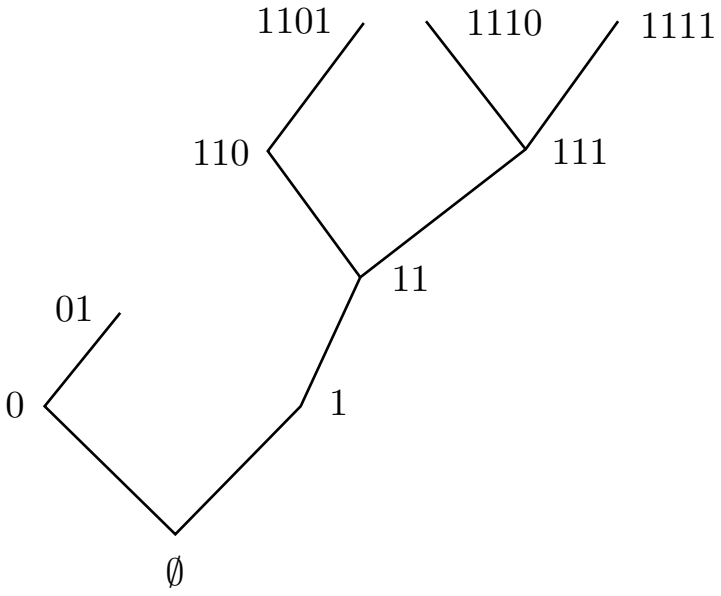

Figure 2: The tree of the longer code

not play any role. Hence we may assume that $p_1 \geq p_2 ... \geq p_n$. Now we will consider the two letters with the smallest probabilities as one and obtain the new vector $(p_1, ..., p_{n-2}, p_{n-1} + p_n)$ with $n-1$ components. By induction hypothesis this vector has a Huffman code $\kappa(1), ..., \kappa(n-1)$. The Huffman code for $p_1, ..., p_n$ is now defined as $\kappa(1), ..., \kappa(n-2), \kappa(n-1)\,0, k(n-1)\,1$. Obviously it is complete, but, of course, this does by no means imply that it is optimal.
Before proving optimally we consider an example:

Example 11 *In the table below the vector $p = (p_1, ..., p_8)$ to be coded is the first column. The following columns are obtained from the previous column by adding the two smallest probabilities and putting this sum at the correct*

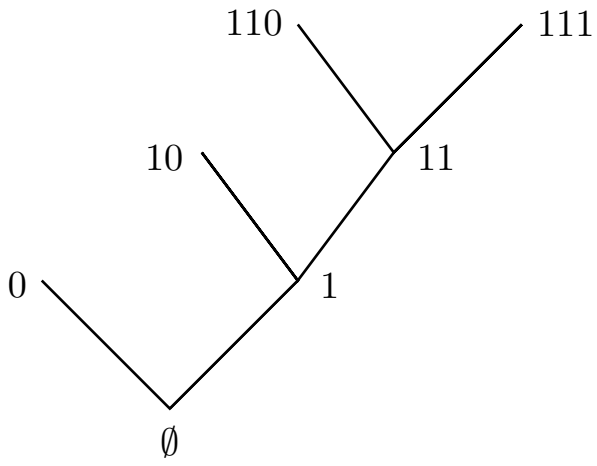

Figure 3: The tree of the shortened code

position. This sum is underlined in each of the columns.

$p_1 = 0,36$	*0,36*	*0,36*	*0,36*	*0,36*	*0,37*	*0,63*	*1*
$p_2 = 0,21$	*0,21*	*0,21*	*0,21*	*0,27*	*0,36*	*0,37*	
$p_3 = 0,15$	*0,15*	*0,15*	*0,16*	*0,21*	*0,27*		
$p_4 = 0,12$	*0,12*	*0,12*	*0,15*	*0,16*			
$p_5 = 0,07$	*0,07*	*0,09*	*0,12*				
$p_6 = 0,06$	*0,06*	*0,07*					
$p_7 = 0,02$	*0,03*						
$p_8 = 0,01$							

The Huffman code is obtained backwards. For the probability vector of length 2 *the corresponding codewords are* 0 *and* 1*. Then, in every step, the codeword belonging to the underlined probability in the first table is split by appending* 0 *and* 1 *in order to obtain new code words for the two last probabilities. In*

the following table the split codewords are underlined.

$\kappa(1) = 00$	*00*	*00*	*00*	*00*	*1*	*0*
$\kappa(2) = 10$	*10*	*10*	*10*	*01*	*00*	*1*
$\kappa(3) = 010$	*010*	*010*	*11*	*10*	*01*	
$\kappa(4) = 011$	*011*	*011*	*010*	*11*		
$\kappa(5) = 111$	*111*	*110*	*011*			
$\kappa(6) = 1100$	*1100*	*111*				
$\kappa(7) = 11010$	*1101*					
$\kappa(8) = 11011$						

We can copute that for this particular choice of p the Huffman code has expected codeword lenght of $2,55$. The corresponding tree is given in Figure 4 below.

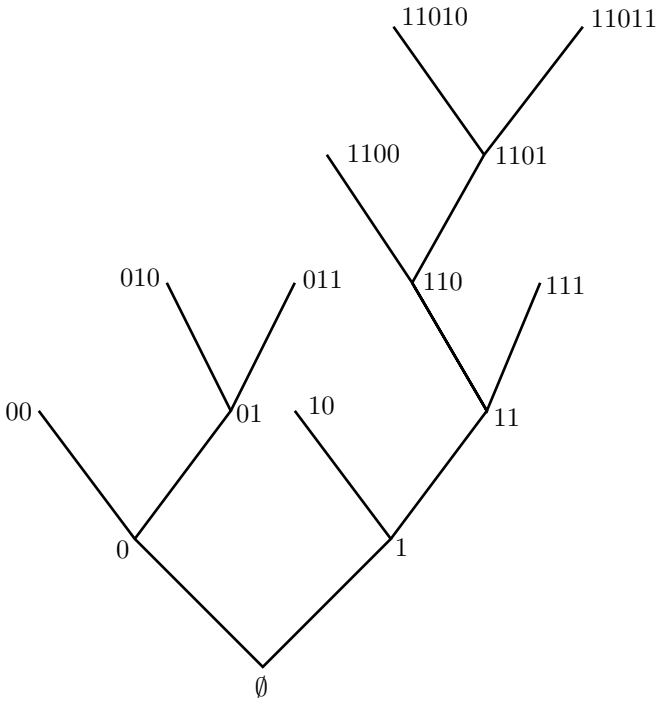

Figure 4: The tree of the Huffman code

Obviously, the Huffman code need not be unique, since it may happen that the sum of the two smallest probabilities equals a given p_i. However, for the

question of optimality of the Huffman code this is irrelevant, since any two Huffman codes have the same average lenght.

Theorem 12 *Every Huffman code is optimal.*

Proof. We prove the assertion inductively. For $n = 2$ this is obvious.
For the induction step assume we have proved the assertion for $n - 1 \geq 2$. Take an arbitrary probability vector $(p_1, ..., p_n)$ with $p_i > 0$ for all i and without lass of generality $p_1 \geq p_2 \geq p_n$. Let κ_H^n be the Huffman code for this vector. Let κ be an arbitrary other code for $(p_1, ..., p_n)$ with codewords $\mu_1, ..., \mu_n$. We will show that

$$E(|\kappa|) \geq E\left(\left|\kappa_H^{(n)}\right|\right).$$

First we order the codewords of κ such that their length is increasing. This ordered code we will call $\kappa' = (\mu_1', ..., \mu_n')$. For the codewords it holds $|\mu_1'| \leq |\mu_2'| \leq ... \leq |\mu_n'|$. Since the set of codewords of κ and κ' is the same and since the probabilities were in decreasing order we see that

$$E(|\kappa|) \geq E(|\kappa'|).$$

If now $|\mu_n'| > \left|\mu_{n-1}'\right|$ we cut the word μ_n' by omitting the last $(|\mu_n'|) - \left|\mu_{n-1}'\right|$ bits. This word we call $\mu_n^{''}$. Since we were only talking about prefix codes $\mu_n'' \notin \{\mu_1', ..., \mu_{n-1}\}$‘. Moreover μ_n'' also is not prefix of any of $\mu_1', ..., \mu_{n-1}'$, since its length is at least as long as those words. Hence $\kappa'' = \left(\mu_1^{'}, ..., \mu_{n-1}', \mu_n^{''}\right)$ is a prefix code. If $\left|\mu_{n-1}'\right| = |\mu_n'|$ we set $\kappa'' = \kappa'$ and at any rate

$$E(|\kappa|) \geq E(|\kappa''|).$$

For κ'' it holds that at least two of its words have length $m := |\mu_n''|$. Define α to be the word consisting of the first $m-1$ letters of μ_n''. Then either $\mu_n'' = \alpha 0$ or $\mu_n'' = \alpha 1$. Without loss of generality the latter is the case. Consider two cases:

(i) One of the words of κ'' of length m is $\alpha 0$. If $\alpha 0$ is not already μ_{n-1}' we interchange this word and μ_{n-1}' and call this code κ'''.

(ii) Non of the words of $\kappa''$is $\alpha 0$. In this case we replace $\mu_{n-1}^{'}$ by $\alpha 0$. This gives an admissible prefix code κ''', since $\alpha 1$ already was a codeword.

Obviously,

$$E\left(|\kappa''|\right) = E\left(|\kappa'''|\right)$$

since the codeword length are the same. Let us write

$$\kappa''' = (\nu_1, ..., \nu_n)$$

with $\nu_{n-1} = \alpha 0$ and $\nu_n = \alpha 1$. Then

$$(\nu_1, ..., \nu_{n-2}, \alpha)$$

is a code for

$$(p_1, ..., p_{n-2}, p_{n-1}, p_n).$$

To see that this is true we simply need to check the prefix condition. But α cannot be the prefix of $\nu_1, ..., \nu_{n-2}$, since the lengths of these codewords are at most $|\alpha| + 1$, and $\alpha 0, \alpha 1$ were different from $\nu_1, ..., \nu_{n-2}$, since the lengths of these codewords are at most $|\alpha|+1$, and $\alpha 0, \alpha 1$ were different from $\nu_1, ..., \nu_{n-2}$. By induction hypothesis is the mean length of $(\nu_1, ..., \nu_{n-2}, \alpha)$ at least as long as the length of the corresponding Huffman-code, hence

$$\sum_{i=1}^{n-2} p_i |\nu_i| + (p_{n-1} + p_n)|\alpha| \geq E\left(\left|\kappa_H^{(n-1)}\right|\right)$$

where $\kappa_H^{(n-1)}$ is the Huffman-code for $(p_1, ..., p_{n-2}, p_{n-1} + p_n)$. From the recursive construction of Huffman codes we see that

$$E\left(\left|\kappa_H^{(n)}\right|\right) = E\left(\left|\kappa_H^{(n-1)}\right|\right) + p_{n-1} + p_n.$$

Thus

$$\begin{aligned} E\left(|\kappa'''|\right) &= \sum_{i=1}^{n} p_i |\nu_i| = \sum_{i=1}^{n-2} p_i |\nu_i| + (p_{n-1} + p_n) |\alpha| + (p_{n-1} + p_n) \\ &\geq E\left(\left|\kappa_H^{(n-1)}\right|\right) + p_{n-1} + p_n = E\left(\left|\kappa_H^{(n)}\right|\right). \end{aligned}$$

This proves the theorem. ■

Finally we will state (but not prove) a theorem that compares the true entropy $H_0(p)$ to the entropy $H(p)$.

Theorem 13 *For all $p = (p_1, ..., p_n)$ and all n it holds*

$$H(p) \leq H_0(p) < H(p) + 1.$$

3 Branching processes

The story that was told when branching processes were mathematically studied for the first time goes back to the times when the surname of a family could just be inheristed by the male line of descendents: Suppose a man with a unique surname considers how many generations in the future his name might last. Since in his culture surnames pass down the male line only, the quantity of interest is, how many sons he has and the number of their sons, and the number of their sons' sons, etc.
As a mathematical model we can let time be indicated by $n \in \mathbb{N}_0$ and denote the number of generations. The random variable X_n is the size of the n'th generation, i.e. the number of times the name is represented n generations hence. We start with just one person, i.e. with $X_0 = 1$. Of course the above model can also represent the survival of a species or other questions.

The most convenient assumption is now that the number of sons is an independent quantity for all the men in the line and that this variable obeys the same probability distribution for each of the men. This assumption, of course, may be little realistic, but it allows for a clean mathematical treatment. It will turn out that a key quantity to study is the probability generating function (pgf)

$$g(z) = \sum_{k=0}^{\infty} p_k z^k.$$

One annoying trivial case is when $g(z) = z$, i.e. $p_k = 0$ when $k \neq 1$ and $p_1 = 1$. The resulting proces is dull: $X_n = 1$ for all $n \geq 0$ and there is really nothing to add. As this case would constitute an exception to some main results we will exclude it and assume that $p_1 \neq 1$. Let us define $g_n(z)$ to be the pgf of the process X_n. This means

$$g_n(z) = \sum_{k=0}^{\infty} p_k^{(n)} z^k$$

where $p_k^{(n)}$ is the probability to have k men in the n'th generation. Since $X_0 \equiv 1$ we know that $g_0(z) = z$ and that $g_1(z) = g(z)$. Note that $g_1(z) = g\left(g_0\left(z\right)\right)$. We will soon see that this is a general principle:

Theorem 14 *: For all $n \geq 1$,$g_{n+1}(z) = g_n\left(g\left(z\right)\right) = g\left(g_n\left(z\right)\right)$. Hence, using either expression,*

$$g_n(z) = g\left(g\left(...\left(z\right)...\right)\right),$$

the n'th interation of g.

Proof. First note that if Y has the probability distribution $p = (p_0, p_1, ...)$, then

$$g(z) = \mathbb{E}z^Y$$

by definition of the expectation. Using that for independent random variables the expectation factorizes and that with $Y_1, ..., Y_n$ also $z^{Y_1}, ..., z^{Y_n}$ are independent we find that for $S_n = Y_1 + ... + Y_n$ (where each of the X_i has pgf g and the Y_i are independent) the pgf looks like

$$\begin{aligned} \mathbb{E}z^{S_n} &= \mathbb{E}z^{Y_1+...+Y_n} = \mathbb{E}\prod_{i=1}^{n} z^{Y_i} = \prod_{i=1}^{n} \mathbb{E}z^{Y_i} \\ &= \prod_{i=1}^{n} g(z) = g(z)^n. \end{aligned}$$

Now we can think of X_{n+1} as the offspring of generation n, so as the sum of X_n independent random variables $Y_i, i = 1, ..., X_n$ each of which has pgf g. Then the pgf of X_{n+1} is

$$\begin{aligned} g_{n+1}(z) &= \mathbb{E}z^{Y_1+...Y_{x_n}} = \sum_{m=0}^{\infty} \mathbb{E}\left(z^{Y_1+...Y_{X_n}} \mid X_n = m\right) P\left(X_n = m\right) \\ &= \sum_{m=0}^{\infty} \mathbb{E}\left(z^{Y_1+...+Y_m}\right) P\left(X_n = m\right) \\ &\quad \sum_{m=0}^{\infty} P\left(X_n = m\right) g^m\left(z\right), \end{aligned}$$

where the last step follows from the considerations above. Now

$$\sum_{m=0}^{\infty} P\left(X_n = m\right) g^m(z) = \sum_{k=0}^{\infty} q_k^{(n)} \left(g\left(z\right)\right)^k = g_n\left(g\left(z\right)\right)$$

by definition. The other identity $g_{n+1}\left(z\right) = g\left(g_n\left(z\right)\right)$ follows by the same considerations if now we think of X_{n+1} as the offspring of generation one over n generations. ■

Example 15 *Suppose there are either no offspring or two offspring, with respective probabilities q and p, so that* $g\left(z\right)=q + pz^2$. *This would corresond*

to a model in which a cell dies with probability q, or splits into two identical daughter cells with probability p. then

$$g_2(z) = g(g(z)) = q + p\left(q + pz^2\right)^2 = q + pq^2 + 2p^2qz^2 + p^3z^4.$$

The exact expression for $g_n(z)$ can be quite complicated, but with the help of the above formulae one could find them e.g. with a computer algebra package.

The importance of the pgf lies, among others, in a way to compute expectation value and variance of a given random variable.

Theorem 16 *Let Y be a random varianle with pgf*

$$g(z) = \sum_{n=0}^{\infty} p_n z^n.$$

Then

$$\mathbb{E}Y = g'(1)$$

and

$$V(Y) = \mathbb{E}\left[(Y - \mathbb{E}Y)^2\right] = g''(1) + g'(1) - (g'(1))^2.$$

Remark 17 *: $V(Y)$ is called the variance of Y and gives a measure of how much the expectation value tells about Y.*

Proof.

$$g'(1) = \sum_{n=0}^{\infty} np_n z^{n-1} \mid_{z=1} = \sum_{n=0}^{\infty} np_n = \mathbb{E}Y.$$

Moreover

$$\begin{aligned} g''(1) &= \sum_{n=0}^{\infty} n(n-1)p_n z^{n-2} \mid_{z=1} = \sum_{n=0}^{\infty} n^2 p_n - \sum_{n=0}^{\infty} np_n \\ &= EY^2 - g'(1). \end{aligned}$$

On the other hand

$$\begin{aligned} V(Y^2) &= \mathbb{E}(Y - \mathbb{E}Y)^2 = \mathbb{E}\left(Y^2\right) - 2(\mathbb{E}Y)^2 + (\mathbb{E}Y)^2 \\ &= \mathbb{E}\left(Y^2\right) - (\mathbb{E}Y)^2 = g''(1) + g'(1) - (g'(1))^2. \end{aligned}$$

This insight helps us to compute expectation and variance of the size of the n'th generation of our branching process. ■

Theorem 18 *If $\mu = \mathbb{E}X_1$ and $\sigma^2 = V(X_1)$ are the expectation and the variance of the number of offspring then the expectation and variance of the n'th generation are*

$$E(X_n) = \mu^n, V(X_n) = \sigma^2\mu^{n-1}\left(1 + \mu + \mu^2 + ... + \mu^{n-1}\right).$$

Proof. We have $\mu = g'(1)$ and $\sigma^2 = g''(1) + g'(1) - (g'(1))$ and, of course, the formula is true. The general case will be proved by induction. Suppose we have shown that the formula holds true for some particular n. From $g_{n+1}(z) = g(g_n(z))$ we find that

$$g'_{n+1}(z) = g'(g_n(z))(g'_n(z))$$

and

$$g''_{n+1}(z) = g''(g_n(z))(g'_n(z))^2 + g'(g_n(z))g''_n(z).$$

Hence

$$\mathbb{E}(X_{n+1}) = g'_{n+1}(1) = g'(1)g'_n(1) = \mu\mathbb{E}X_n = \mu\mu^n = \mu^{n+1}.$$

On the other hand

$$\begin{aligned} V(X_{n+1}) &= g''_{n+1}(1) + \mu^{n+1} - \mu^{2m+2} \\ &= g''(1)\mu^{2n} + \mu g''_n(1) + \mu^{n+1} - \mu^{2n+2} \\ &= \left(\sigma^2 - \mu + \mu^2\right)\mu^{2n} + \mu\left(Var(X_n) - \mu^n + \mu^{2n}\right) + \mu^{n+1} - \mu^{2n+2} \\ &= \sigma^2\mu^{2n} + \mu V(X_n) \end{aligned}$$

from which the inductive step follows. ■

Now the picture becomes clearer. When $\mu < 1$, then both $E(X_n)$ and $V(X_n)$ converge to zero geometrically fast. When $\mu = 1$ also $E(X_n) = 1$ for all n but $V(X_n) = n\sigma^2$ diverges to infintiy. For $\mu > 1$ both, expectaction and variance go to infintiy geomatrically fast.

The following simple criterion for extinction might therefore have been expected.

Theorem 19 *Let x_n be the probability of extinction by time n. Then*

$$x_{n+1} = g(x_n)$$

and

$$x_n \to x$$

where x is the smallest non-negative root of

$$z = g(z)$$

and is the probability that the process ever becomes extinct. Moreover

$$x = 1 \Leftrightarrow \mu \leq 1.$$

Proof. : By defintion

$$x_n = P\left(X_n = 0\right) = g_n(0).$$

Thus

$$x_{n+1} = g_{n+1}(0) = g\left(g_n\left(0\right)\right) = g\left(x_n\right).$$

Whenever $X_n = 0$, then $X_{n+1} = 0$, so plainly

$$x_n \leq x_{n+1};$$

since x_n is a probability, the sequence (X_n) is bounded above and hence convergent to some $x \leq 1$. But $g(z) = \sum_{k=0}^{\infty} p_k z^k$ is continuous on $0 \leq z \leq 1$ and so the relation $x_{n+1} = g\left(x_n\right)$ shows that

$$x = g\left(x\right).$$

We now prove that the solution we seek is the smallest root of this equation in $[0, 1]$. Let $\alpha = g\left(\alpha\right)$ be any root with $\alpha \geq 1$. Then

$$x_1 = g(0) \leq g\left(\alpha\right) = \alpha$$

since g is increasing. Inductively, suppose $x_n \leq \alpha$; then

$$x_{n+1} = g\left(x_n\right) \leq g\left(\alpha\right) = \alpha$$

as required. Since $x_n \leq \alpha$ for all n, so

$$x = \lim\left(x_n\right) \leq \alpha$$

and the solution we seek is indeed the smallest non - negative - root.

Finally we show $x = 1 \Leftrightarrow \mu \leq 1$. The easiest case is $p_0 = 0$. Here extinction is impossible, i.e. $x = 0$ and, snce $p_1 < 1$, it is obvious that $\mu > 1$. So we may assume that $g(0) = p_0 > 0$, and we know that

$$g'(z) = \sum_{k=0}^{\infty} k p_k z^{k-1}$$

is increasing when $z \geq 0$ with $\mu = g'(1)$. Suppose $\mu > 1$, then since $g(1) = \sum p_k z^k \mid_{z=1} = \sum p_k = 1$ when $z < 1$ we must have $g(z) < z$. As $g(0) > 0$ the Intermediate Value Theorem shows there is some $x, 0 < x < 1$. Since $g(1) = 1$, and g' is increasing over $(x, 1)$, the chord joining $(x, g(x))$ to $(1, g(1))$ lies entirely above the graph of g, hence

$$\mu = g'(1) > 1.$$

The above theorem illustrates why we excluded the case when $p_1 = 1$. For $\mu = 1$, yet extinction is impossible, so the statement of the theorem would need a caveat. The above theorem shows a dichotomie: If the expected number of offspring is at most one the process dies out, if this expectation is larger than one, the process survives forever with positive probability. ■