Quasi-Questionnaires, Codes and Huffman’s Length

C. F. Picard

New concepts are defined, in particular the quasi-question or vertex with an outgoing arc of zero probability. A quasi-questionnaire is a probabilistic homogeneous (rooted) tree with quasi-questions.

It is shown that every instantaneous code is a quasi-questionnaire with precise restrictive conditions; it may also be a questionnaire, without an arc of zero probability.

Also, an approximation is given — without use of the classical construction — of the average length of Huffman’s code with a given alphabet and given probabilities of code-words.

INTRODUCTION

A questionnaire is a graph with the set $X$ of vertices having the partition $Q \cup E$ such that:

- $Q$, the set of questions, is formed by the vertices which are origins of at least two arcs; there is one and only one vertex $x_0 \in Q$ which is terminal extremity of no arc: it is the root.
- $E$ is formed by the set of terminal vertices, called events.
- If there is one and only one path from $x_0$ to all the other vertices, then the graph is an arborescence (or rooted tree) and there enters exactly one arc in every vertex, but $x_0$: it is always the case in this paper.
- There exists a mapping $P : i \rightarrow p(i)$ from $X$ on the interval $[0, 1]$ such that

$$\sum_{e \in E} p(e) = 1, \quad p(i) = \sum_{j \in I_i} p(j) \quad \text{for all} \quad i \in Q,$$

where $I_i$ is the set of successors of $i$; then $p(x_0) = 1$.
- $|I_i| > 1$ for all $i \in Q$.

The outward degree of every question is called the basis of the question and it is written $a_i$ (or $a$ if possible). If all the questions have same basis, $a$, the questionnaire is called homogeneous. There is a compatibility relation between the number of
If this relation is true, then the questionnaire is strictly homogeneous; else there is a question with $\beta + 1$ for basis, where $\beta$ is the rest of the integer division $(N - 1) \div (a - 1)$; $\beta$ is then strictly less than $a - 1$; all the other questions have same basis $a$ and the questionnaire is called homogeneous in the wide sense.

If $a = 2$, the questionnaire is always strictly homogeneous and is called a dichotomic one, if $a > 2$, then it is a polychotomic one.

An heterogeneous (rooted-tree) questionnaire is defined in the case where the bases are not the same for all the questions: $\mathcal{A} = \{a_1, \ldots, a_M\}$ is the set of the bases and the compatibility relation is now

$$\sum_{i=1}^{M} (a_i - 1) = N - 1.$$

It is possible to associate a code to a questionnaire by a mapping from the set of events on the set of codewords. At every answer $r_0, r_1, \ldots, r_{a-1}$ to the question $i$ (without an explanation of the kind of answer), we associate a $a$-ary digit $j \in \{0, 1, \ldots, a - 1\}$. The path from $x_0$ to $x$ must be coded with, say, $i$ $a$-ary digits if this path contains $i$ arcs: at every arc, corresponds one digit, the left one for the answer to $x_0$, the right one for $x$. In doing this, we get a word with $i$ digits and this is the codeword of the code associated to the questionnaire. Because the questionnaire is a rooted-tree, it is possible to code all the events with other words and the code is decipherable and instantaneous. Then Huffman optimal coding procedure is able to give the algorithm for building an optimal homogeneous questionnaire. The case of heterogeneous questionnaire needs to use sometimes an alphabet with $a_i$ letters (0 to $a_i-1$) for the question of basis $a_i$, and we gave already a generalization of Huffman's algorithm to the heterogeneous case. From the coding point of view, it is for example the use of letters and digits to code an event.

In fact the root operates a partition of $E$ in $a_0$ subsets and every other question $i$ operates a partition of a subset of $E$ in $a_i$ subsets. If all the successors of a question $i$ are events $e_1, \ldots, e_{a_i}$ then $i$ operates the final partition of a subset of size $a$.

The theory gives rules for building algorithms for feasible questionnaires, when some restrictions are done over the partitions of $E$ and of its subsets; for example the unique type of question is to operate an ordinary comparison between two numbers with two (or three) outcomes as $a > b$, $a \leq b$ (or $a > b$, $a = b$, $a < b$).

Some operations are defined over the questionnaires: they allows to build sophisticated graphs and questionnaires with very ordinary ones; they give too questionnaires with one or other extremum property.

The theory of questionnaires leads to use the information theory in view of the evaluation of the information value of a questionnaire. In this paper, we use essentially the concept of routing length i.e. the expectation of the length of a path in a questionnaire and the concept of information transmitted by a questionnaire.
when the set $E$ is given, with its distribution of probabilities; it has been shown that
the information transmitted by a questionnaire is always less than the routing length.

1. QUASI-QUESTIONNAIRES

**Definition 1.** A *quasi-event* is a terminal vertex of a probabilistic tree with a zero
probability.

**Definition 2.** A *quasi-question* is an inner vertex of a homogeneous tree with at
least one terminal vertex of nonzero probability as descendant and one quasi-event
as successor.

**Definition 3.** The *probability of a quasi-question* $x$ is the sum of the probabilities
of the terminal vertices of nonzero probabilities which belong to the descendance of $x$.
If a quasi-question $x$ has only one successor $y$ of nonzero probability, all the other
successors of $x$ are quasi-events and the information given by $x$ is zero because the
conditional probability of $y$ is equal to 1.

A quasi-question is not a question in the usual sense of questionnaires.
To repeat, an event is a terminal vertex of nonzero probability and a *question*
is a vertex with *successors* which are events or questions.

**Definition 4.** A *polychotomic quasi-questionnaire* is a probabilistic homogeneous
tree with the two following properties:
1. the vertices are events, quasi-events, questions or quasi-questions;
2. the sum of the probabilities of the events is 1.

It will be noted that in a quasi-questionnaire* there are two kinds of terminal ver-
tices:

- terminal vertices giving a complete system of events that is the set $E = \{e_i\}$ such
  as $\sum_{e \in E} p(e) = 1$, where $P = \{p(e_i) \mid e_i \in E\}$ is given.
- terminal vertices such as $\forall e_i, p(e_i) = 0$ giving the set $E = \{\bar{e}_i\}$.

This case is very different from the incomplete systems studied by, for example,
Renyi [8] and those used in the sub-questionnaires (Dubail [3]).

In a probabilistic tree, a non-terminal vertex with only quasi-events as successors
will be "reduced" in a quasi-event with reduction of the order of the tree.

In a wider sense, a homogeneous questionnaire, such as the number of events $N$
is $N = a^d + a(a - 1) + \beta$, may be considered as a quasi-questionnaire with
$(N - \beta - 1)/(a - 1)$ questions and a quasi-question with $(a - \beta - 1)$ quasi-events
as outcomes.

* See figure in the annex.
The average path-length of a quasi-questionnaire $K$ is defined by the usual sum:

\[ L = \sum_{e \in E} p(e) r(e) \]

where the rank of the event $e_i$ is $r(e_i)$. That is, the event $e_i$ is connected with the root by a path whose length is $r(e_i)$. Of course, the possible extension of the sum to the quasi-event $\tilde{e}_j$ does not change $L$. $L$ is called routing-length (or ring-length).

**Theorem 1.** The routing-length of a quasi-questionnaire is

\[ L = \sum_{q \in Q} p(q) \]

where $q$ is a question ($q \in Q$) or a quasi-question ($q \in \bar{Q}$).

The equivalence between (1) and (2) is proved in the same way as for rooted-tree questionnaires (Picard [7]).

To a polychotomic questionnaire $K$ constructed on $(a, P)$ we can map an infinity of quasi-questions obtained by substituting $\bar{P}$ to $P$, where $\bar{P}$ is a distribution of the same $N$ nonzero probabilities and of a multiple of $(a - 1)$ zero probabilities.

But reciprocally to a quasi-questionnaire $\bar{K}$ is associated only a couple $(a, P)$ where $P$ is a distribution of $N$ nonzero probabilities.

Let $N$ be the number of quasi-events of $\bar{K}$ and let $\bar{M} > 1$ be the number of quasi-questions.

The attempt is made to do some transfers of terminal vertices by interchanging the probabilities of two vertices, an event $e_i$ of rank $r(e_i)$ and a quasi-event of rank $r(e)$. If $r_e < r(e_i)$ then the ring-length of the new tree is:

\[ L - p(e) [r(e_i) - r_e] \]

This type of transfer must be repeated as often as possible. If during this process a tree is such that there is a vertex $y$ with only quasi-events as descendants, then only the vertex $y$ will be substituted for the subtree issued from $y$, and $y$ will be taken as quasi-event.

At the end $N$ will be reduced by a multiple of $a - 1$ and $\bar{M}$ will also be reduced.

Let $r_M$ be the highest rank of the events; then there is no quasi-event of rank $r_e > r_M$; if not, there would be a vertex of rank $r_M$ of which no descendant would be an event.

If all the quasi-events are of the rank $r_M$, it is possible to do again a transfer between an event and a quasi-event.

Such transfers will not reduce the ring-length but may permit to give a quasi-events as successors to the same vertex. Then the number of quasi-events and quasi-questions will be reduced.

If and only if $N < a - 1$, no possible transfer can reduce $N$.

All transfers performed one after the other make it possible to form a homogeneous
questionnaire $K'$ in the strict sense (if $N' = 0$) or in the wide sense (if $0 < N' < a - 1$); the events $e_i$ of this questionnaire $K'$ have as rank $r'(e_i)$ such that $r'(e_i) < r(e_i) \forall i$.

Consequently the ring-length $L'$ of the questionnaire $K'$ is

\[ L' \leq L. \]

Furthermore, if a quasi-event has a rank $r_h < r_M$ then the inequality yields:

\[ L' < L. \]

If all the quasi-events are of maximal rank then

1. if $K$ has $N < a - 1$ quasi-events then, at the most, $N - 1$ transfers will permit to form a homogeneous questionnaire, in the wide sense, with a ring-length

\[ L' = L; \]

2. if $K$ has $N > a - 1$ quasi-events, then some transfers will permit to form at least a non-terminal vertex with no more than one outgoing event.

If there is no outgoing event this vertex will be replaced by a quasi-event of rank less than $r_M$ in such a way that a new transfer will permit to reduce the ring-length; if not, the following case will occur.

Every quasi-question of rank $r$ with only one outgoing event $e_i$ may be replaced by this event; consequently, there is a reduction of ring-length of

\[ (r(e_i) - r) p(e_i) \]

Thus, the following result has been obtained:

**Theorem 2.** Every quasi-questionnaire $K$ constructed on $(a, P)$, with a routing-length $L$, and with $N > a - 1$ quasi-events may be reduced to a questionnaire $K'$ constructed on $(a, P)$ with a routing-length $L' < L$.

We note that the restriction $N > a - 1$ follows from the fact that if $N = a^k + o(a - 1) + \beta$ then the same rooted tree $K$ may be considered both as a questionnaire and as a quasi-questionnaire.

2. **INSTANTANEOUS CODES AND QUASI-QUESTIONNAIRES**

Some connexions have already been noted between coding and questionnaire (see, for example, Picard [7] § i - 5 and Césari [2]).

The couple $(a, P)$ is the base of the questions and the distribution of the probabilities of the complete system of events $E$ (in the sense of questionnaire) or the number of the letters of the alphabet and the probabilities of the code-words which still form a complete system of events $E$ (in the sense of coding).
The words of an instantaneous code are those of the terminal vertices of a probabilistic rooted-tree which have nonzero probability.

The problem of instantaneous coding is to determine the ordered sequence of the arcs of the path connecting the root to the code-words.

This coding is also used in the Theory of Questionnaires to locate the events. But in this last theory the questions or inner vertices are also to be noted; since the characters used to code a question, i.e. to locate its position on the rooted-tree, are the prefix of the characters of the terminal vertices, we must, to decode them, know the rank of the vertices ([7], § 1.2.1.).

An instantaneous code is a quasi-questionnaire because definition 4 holds.

Let $A$ be a strict homogeneous rooted-tree. Every vertex of $A$ has either no successor or a successor.

Let $E$ be the set of the terminal vertices with $N$ elements, $E = \{e_i | i = 1, 2, \ldots, N\}$ and let $r(e_i)$ be their ranks.

Picard [7] and Ash [1], for example, have proved that it is possible to construct an optimal questionnaire on $A$, without “discrepancy” $K_0$, using an information equal to the ring-length (called also the absolutely optimal code) assigning the probability $p(e_i) = a^{-r(e_i)}$ to the vertices of rank $r(e_i)$.

The calculation step by step (from the events to the root) of the probabilities of the questions shows that all the vertices of $K_0$ have a probability connected to the rank by: $p(x) = a^{-r(x)}$. Indeed, a question $x$ of rank $r$ has for successors: the vertices (events or questions) of rank $r + 1$ and of probability $a^{-r(x)}$ thus $p(x) = \sum a^{-r(x+1)} = a^{-r}$. In particular the root $x_0$ is such that $p(x_0) = a^0 = 1$. From $p(x_0) = \sum a^{-r(e_i)}$ it follows

\begin{equation}
\sum_{e_i \in E} a^{-r(e_i)} = 1
\end{equation}

The property (6) holds for every questionnaire constructed on $A$.

Let then $E$ be the partition $E = E' \cup E'$ of the terminal vertices of $A$ and let $q(e)$ be the probabilities of an instantaneous code $K$, such that:

\[ \sum_{e_i \in E} q(e_i) = 1 \quad \text{and} \quad \left\{ \begin{array}{l}
(\forall e \in E) q(e) = 0 , \\
(\forall e_i \in E') q(e_i) = 0
\end{array} \right. \]

the property (6) implies:

\begin{equation}
\sum_{e_i \in E} a^{-r(e_i)} + \sum_{e_i \in E'} a^{-r(e_i)} = 1
\end{equation}

thus,

\begin{equation}
\sum_{e_i \in E'} a^{-r(e_i)} < 1
\end{equation}

and nevertheless $E'$ is a complete system of events.
E is the set of quasi-events.

The quasi-questionnaire is a questionnaire if and only if \( E = \emptyset \) so that the following theorem is proved:

**Theorem 3.** *An instantaneous code constructed on a complete system of events \( E = \{e_i\} \) is a strictly homogeneous questionnaire, if and only if:*

\[
\sum_{e \in E} a^{-r(e)} = 1.
\]

The existence of an instantaneous code for which the rank of every codeword \( e_i \) is the integer immediately greater than the logarithm (of base \( a \)) of the inverse of the probability \( p(e_i) \) has been indicated by Shannon ([9] § 9) and proved by Feinstein [4].

**Noiseless Coding Theorem.** There exists an instantaneous code \( C \) such that

\[
\log \frac{1}{p(e_i)} \leq r(e_i) < \log \frac{1}{p(e_i)} + 1
\]

for every code word \( e_i \).

In such a code, the word \( e_i \) is coded by a word of \( r(e_i) \) characters.

The multiplication of each member of (9) by \( p(e_i) \), and then the summation over the elements \( e_i \) of the set \( E \) of code-words \( C \), lead to the double inequality:

\[
I_n(E) \leq L(C) < I_s(E) + 1
\]

where \( I_n(E) \) is the Shannon's information of code \( C \), expressed with logarithms of base \( a \) and \( L(C) \) the ring-length of the quasi-questionnaire \( C \). One could also write:

\[
r(e_i) = \left\lfloor \log \frac{1}{p(e_i)} \right\rfloor + 1
\]

and

\[
L(C) = \sum_{e \in E} p(e) \left( \left\lfloor \log \frac{1}{p(e)} \right\rfloor + 1 \right)
\]

where \( [x] \) is the greatest integer strictly less than \( x \).

Let then \( C \) be a Feinstein's code of which the ranks of the words are determined by (9).

If \( \log 1/p(e_i) \) is integer for every \( e_i \) then \( r(e_i) = \log 1/p(e_i) \) (\( \forall i \)) and therefore \( \sum_{e \in E} a^{-r(e)} = 1 \), thus \( C \) is a questionnaire.

If \( \log 1/p(e_i) \) is not integer for at least one word \( e_i \), then, according to (9) \( r(e_i) > \log 1/p(e_i) \), that is

\[
p(e_i) > a^{-r(e_i)}.
\]
Since for every word of $C$, $p(e) \geq a^{-r(e)}$, if (13) holds for at least one word, then
\[
\sum_{e \in E} a^{-r(e)} = \sum_{e \in E} p(e) \quad \text{and, thus} \quad \sum_{e \in E} a^{-r(e)} < 1 ;
\]
and so, according to Theorem 3, $C$ is a quasi-questionnaire.

Theorem 4. A code $C$ satisfying the inequalities (9) of the Noiseless Coding Theorem is a strictly homogeneous questionnaire if and only if
\[
r(e) = \log \frac{1}{p(e)} \quad (\forall i) .
\]

The homogeneous questionnaires in the wide sense have in fact a property similar to the homogeneous questionnaires in the strict sense.

The sub-tree with one question of base $\beta + 1 < a$ may admit some events of the same probability of the form $a^{-r(e)}$.

Then if each of the $N - (\beta + 1)$ other events $e_i$ forming the subset $E_i \subset E$ has a probability of the form $a^{-r(e)}$, then the quasi-questionnaire is a homogeneous questionnaire in the wide sense:

Corollary. A code $C$ satisfying the inequalities (9) of the Noiseless Coding Theorem is a homogeneous questionnaire in the wide sense, if and only if, the following conditions hold:

1. $N - (\beta + 1)$ is a multiple of $a - 1$ and $\beta + 1 < a$,
2. $N - (\beta + 1)$ events have a probability of the form $a^{-r(e)}$,
3. $(\beta + 1)$ events have a probability of the form $a^{-r((\beta + 1))}$,
4. $s = \sup_{e \in E_i} r(e_i)$ where $E_i$ is the set of $N - (\beta + 1)$ events of type 2.

Otherwise $C$ is a quasi-questionnaire.

The last condition places the usual restriction on the optimal homogeneous questionnaires in the wide sense: the question of base $\beta + 1$ has as outcome the events of smallest probabilities.

3. OPTIMAL QUESTIONNAIRES AND QUASI QUESTIONNAIRES

The questionnaires of the minimal ring-length $K_H$ or optimal questionnaires, may be constructed — within one equivalence — by Huffman's algorithm (1952).

This algorithm allows the evaluation of the rank of the events, thus the determination of the ring-length $L_H$, but, to our knowledge, there does not exist a formula which gives $L_H$ directly without using this algorithm.

Now let us study the rank of the vertices of the optimal questionnaires,
Let us suppose that the ranks of the events of a questionnaire $K_H$ are at least equal to $k_0$. For every rank $r \leq k_0$, there are exactly $a^r$ vertices and the sum of the probabilities of the vertices of every rank ($0 \leq r \leq k_0$) is 1.

Let $X_r$ be the set of the vertices of rank $r$.

1. $r \leq k_0$ then it follows:

\begin{equation}
|X_r| = a^r
\end{equation}

and

\begin{equation}
\sum_{x \in X_r} p(x) = 1
\end{equation}

Furthermore, if $K_H$ is an optimal questionnaire:

\begin{equation}
x \in X_r \text{ and } y \in X_{r+1} \Rightarrow p(x) \geq p(y)
\end{equation}

and for the predecessor $x_0$ of $y$, $(\forall y \in \Gamma x_0)$:

\begin{equation}
p(x_0) > p(y)
\end{equation}

However, these properties are not sufficient to characterize an optimal questionnaire.

If, in $K_H$, it was possible to find a vertex $y$ of rank $r \leq k_0 + 1$ such that

\[ p(y) = \frac{1}{a^{r-1}} , \]

then for all the rank $r - 1$ we would have, according to (14), (16) and (17):

\[ \sum_{x \in X_{r-1}} p(x) > a^{r-1} p(y) \geq 1 , \]

which is in contradiction with (15).

Thus:

\begin{equation}
r(x) \leq k_0 + 1 \Rightarrow p(x) < \frac{1}{a^{(r+1)-1}}
\end{equation}

so

\[ r(x) < \log \frac{1}{p(x)} + 1 , \]

which is the second inequality of (9), extended to the questions of $K_H$.

2. $r > k_0 + 1$ with $K_H$ being always an optimal questionnaire.

If for every vertex of rank $r - 1$ we have $p(x_{r-1}) \leq a^{-(r+1)}$, is it possible to have a vertex of rank $r + 1$ such that $p(x_{r+1}) \geq a^{-(r+1)}$?
In this case, all the vertices of rank \( r \) would have the probability \( p(x_r) \geq a^{-r} \), and for at least a question of rank \( r : p(q_r) > a^{-r} \); and the questions of rank \( r - 1 \) would have the probability \( p(q_{r-1}) \geq a^{-r+1} \) and for at least a question of rank \( r - 1 \) \( p(q_{r-1}) > a^{-r+1} \); which leads to a contradiction. Therefore:

**Theorem 5.** In an optimal questionnaire in which the events are of rank at least equal to \( k_0 \), the probability of every vertex \( x \) of rank \( r(x) = k_0 + 1 \) is such that \( p(x) < a^{-k_0+1} \); if for \( r > k_0 + 1 \) every vertex of rank \( r - 1 \) is such that \( p(x_{r-1}) \leq a^{-r+1} \) then all vertices of rank \( r + 1 \) have a probability \( p(x_{r+1}) < a^{-r} \).

A necessary condition of optimization such that: “No vertex of rank \( r \) has a probability greater than the sum of the probabilities of a other vertices of the same rank” \([7]\) leads naturally to theorem 5, but not to stronger inequalities when \( r > k_0 + 1 \).

Theorem 5 leads to the comparison of the ranks of the events in Huffman’s and Feinstein’s codes. Let \( r_d(e) \) and \( r_H(e) \) be the respective ranks of the same event with the same probability \( p(e) \) and let \( \inf_{e \in E} r_H(e) = k_0 \).

Then:

\[
\begin{align*}
\text{If } r_H(e) = k_0 & \Rightarrow p(e) < a^{-k_0+1} \quad \text{and} \quad r_d(e) \geq k_0 , \\
r_d(e) = k_0 + 1 & \Rightarrow p(e) < a^{-k_0} \quad \text{and} \quad r_d(e) \geq k_0 + 1 , \\
r_H(e) = k_0 + 2 & \Rightarrow p(e) < a^{-k_0} \quad \text{and} \quad r_d(e) \geq k_0 + 1 , \\
r_H(e) = k_0 + 3 & \Rightarrow p(e) < a^{-k_0-1} \quad \text{and} \quad r_d(e) \geq k_0 + 2 ,
\end{align*}
\]

and also

\[
\begin{align*}
r_H(e) = k_0 + 2h & \Rightarrow p(e) < a^{-k_0-h+1} \quad \text{and} \quad r_d(e) \geq k_0 + h , \\
r_H(e) = k_0 + 2h + 1 & \Rightarrow p(e) < a^{-k_0-h} \quad \text{and} \quad r_d(e) \geq k_0 + h + 1 ,
\end{align*}
\]

that is,

\[
\begin{align*}
r_H(e) = k_0 + 2h & \Rightarrow r_d(e) \geq r_H(e) - h , \\
r_H(e) = k_0 + 2h + 1 & \Rightarrow r_d(e) \geq r_H(e) - h .
\end{align*}
\]

Let then \( e_1 \) be the event with the greatest probability

\[ p(e_1) = \sup_{e \in E} p(e) ; \]

The rank of \( e_1 \) is \( k_0 \), that is, \( r_d(e_1) = k_0 \) and furthermore

\[ r_d(e_1) = \left[ \log \frac{1}{p(e_1)} \right] + 1 . \]
The preceding inequalities lead to

\[ k_0 \leq \left\lfloor \log \frac{1}{p(e_1)} \right\rfloor + 1. \]

Furthermore, let us suppose:

\[ k_0 \leq \left\lfloor \log \frac{1}{p(e_1)} \right\rfloor - 1, \]

that is:

\[ p(e_1) \leq a^{-k_0 - 1}; \]

since the questionnaire is optimal, the vertices of rank \( k_0 + 1 \) have a probability:

\[ p(x_{k_0+1}) \leq p(e_1) \leq a^{-k_0 - 1} \]

and the questions of rank \( k_0 \):

\[ p(q_{k_0}) \leq a^{-k_0} \]

so that

\[ \sum_{x \leq x_{k_0}} p(x) \leq (a^{k_0} - 1) a^{-k_0} + a^{-k_0 - 1} < 1 \]

which is absurd.

Thus

\[ k_0 = \left\lfloor \log \frac{1}{p(e_1)} \right\rfloor + \alpha \]

where \( \alpha = 0 \) or 1 and \( k_0 \) has no other possible value.

Hence the results:

**Theorem 6.** In an optimal questionnaire \( K_H \), the event of greatest probability \( e_1 \) has a rank \( r_H(e_1) = k_0 \) such that:

\[ k_0 = \left\lfloor \log \frac{1}{p(e_1)} \right\rfloor \]

or

\[ k_0 = \left\lfloor \log \frac{1}{p(e_1)} \right\rfloor + 1. \]

For \( h \geq 0 \), the ranks of the events \( r_H(e) \) are connected with the ranks \( r_H(e) \) of Feinstein's code by:

\[ r_H(e) = k_0 + 2h \quad \text{or} \quad r_H(e) = k_0 + 2h + 1 \Rightarrow r_H(e) \geq r_H(e) - h. \]
Applications. 1. Let us consider the two following Huffman's questionnaires

for which $L_H = 2$.

The first one is such that $k_0 = \log_2 \left( \frac{100}{38} \right) + 1 = 2$.
The second one is such that $k_0 = \log_2 \left( \frac{100}{38} \right) = 1$.

2. Let us consider this example given by Y. Cesari to whom I am grateful:

For $p(e_1) = 0.47$ we have

$$r_H(e_1) = k_0 + 2h_1,$$
$$r_H(e_2) = k_0 + 2h_2 + 1,$$
$$r_H(e_3) = k_0 + 2h_3,$$
$$r_H(e_4) = k_0 + 2h_4,$$

with

$h_1 = h_2 = 0$ and $h_3 = h_4 = 1$

we have:

$$r(e_1) > r_H(e_1) - h_1$$
$$r(e_2) > r_H(e_2) - h_2$$
$$r(e_3) = r_H(e_3) - h_3$$
$$r(e_4) > r_H(e_4) - h_4.$$

The calculation of a bound of the routing-length of the optimal questionnaire

is a consequence of:

$$L_H \leq L \leq L(C)$$

where $L_H$ is the routing-length of an optimal questionnaire $K_H$, $L(C)$ the routing-length of Feinstein's code, $L$ — the routing-length of a questionnaire $K'$ deduced from Feinstein's code, these 3 quasi-questionnaires being constructed on the same couple $(a, P)$.

It is possible to bound the preceding inequalities by some informations:

(20) $$I_n(E) \leq L_H \leq L \leq L(C) \leq I_n(E) + 1.$$
From (12) and (20) we have

\[ L_n < \sum_{e \in E} p(e) \left( \log \frac{1}{p(e)} \right) + 1 \]

if at least the probability of an event is not a power of \(1/a\) (from theorems 2 and 4).

Let us suppose now that this condition holds; the code \(C\) is then a quasi-questionnaire and from theorem 2 there exists a questionnaire \(K'\) such that \(L < L(C)\).

The evaluation of (11), according to (21) requires the determination for every event \(e_i\) of

\[ c_i = \left\lfloor \log \frac{1}{p(e_i)} \right\rfloor + 1. \]

Let us then call \(\text{Sup } c_i = c_N\) and

\[ \sum_{e \in E} a^{-c_i} = 1 - \varrho \]

where

\[ \varrho = \sum_{e \in E} a^{-r(e)} \]

\(\varrho\) is the residue corresponding to the quasi-event of \(C\). \(\varrho\) is then sum of \(N = |E|\) powers of \(1/a\) (distinct or not) corresponding to the quasi-questionnaire \(C\).

These powers may be obtained explicitly from (22): we shall use a system of numeration of base \(a\) to express \(\sum_{e \in E} a^{-c_i}\) and \(\varrho\) may be written:

\[ \varrho = \sum_{r_i=1}^N t_i a^{-r_i} \quad \text{with } 0 \leq t_i < a. \]

\(N\) will be then the sum of the nonzero digits:

\[ N = \sum_{i=1}^{c_N} t_i. \]

**Example.** If \(P = \{0,90; 0,10\}\) then \(c_1 = 1\) and \(c_2 = 4\); \(\sum_{e \in E} 2^{-c_i} = 0,1001\) so that \(\varrho = 0,0111\); thus \(N = 3\) and the ranks of the quasi-events are 2, 3 and 4.

Also for \(a = 3\)

\[ P = \begin{bmatrix} 32 & 25 & 20 & 15 & 5 & 2 & 1 \\ 100 & 100 & 100 & 100 & 100 & 100 & 100 \end{bmatrix} \]

thus \(c_1 = 2, 2, 2, 2, 3, 4, 5,\) and

\[ \sum_{i=1}^N 3^{-c_i} = 0,11111 = 1 - \varrho \]
Let us call $f_j$ the ranks of the quasi-events, for $j = 1, \ldots, N$. A quasi-questionnaire $K_j$ is deduced from $C$ by a succession of $N$ permutations (at the most) among the events and the quasi-events; each one may be followed by a substitution in case there would be a non-terminal vertex of zero probability.

The maximal rank of an event of $C$ is:

$$c_N = F_N = 1 + \max_{e \in E} \left( \log \frac{1}{p(e)} \right)$$

and corresponds to the event $e_N$ with the smallest probability $p_N$.

If $e_N$ is the only event of rank $c_N$, before a permutation we can replace the quasi-question with outcome $e_N$ by an event of probability $p_N$; the rank of $e_N$ is $c_N - 1$ and the reduction of the ring-length is $p_N$ (recursive operation).

If the difference of rank between $e_N$ and the event of probability immediately higher, $e_{N-1}$ is $c_N - c_N$, we shall have reduced the ring-length by $p_N(c_N - c_N)$ and the number of quasi-events by $(a - 1)(c_N - c_N)$.

Let us suppose then that $c_N = c_N'$ and let $c_N$ be the rank of $e_N$ in $C$.

If there exists a quasi-event of rank $c_j < c_N$, then the permutation of $e_{N-1}$ with this quasi event gives a reduction of $(c_N - c_j) p(e_{N-1})$.

But we can increase this reduction by doing a permutation between the quasi-event of minimal rank $c_0$ and the event of maximum probability among the events of rank higher than $c_0$; then, by other steps ($j$), by substituting a new value to $c_0$ (greater or equal to the previous value) we can repeat the process.

This algorithm may be expressed as follows:

1. Let $c_0(j)$ be the smallest rank of a quasi-event at the step $j$ ($j$ starting at 1); the permutation of a quasi-event of rank $c_0(j)$ with the event of greatest probability, having a rank $c_i$ greater than $c_0(j)$, leads to the reduction of

$$\max_{e_i > c_0(j)} \{ p(e_i) \} \times (c_i - c_0(j)) .$$
2. For the next step \((j + 1)\) we shall substitute \(\tilde{c}_0(j + 1)\) to \(\tilde{c}_0(j)\) where

\[
\tilde{c}_0(j + 1) \geq \tilde{c}_0(j)
\]

the algorithm will stop at step \(t\) such that for every event \(c \leq \tilde{c}_0(t)\).

At the end, one will obtain a quasi-questionnaire \(K_t\) with \(N_t\) quasi-events and a routing-length:

\[
L_t = L(C) - \sum_{j=1}^{t} (c_i - \tilde{c}_0(j)) \max_{c_i > \tilde{c}_0(j)} \{p(e_i)\}
\]

\(C\) and \(K_t\) have the property \(P\):

The events ordered according to the decreasing probabilities have non-decreasing ranks.

This is a necessary property for optimal questionnaires. This property is true for \(K'\). \(K'\) will be deduced from \(K_t\) according to theorem 2.

Furthermore, the routing-length of \(K'\) is less than that of \(K\); every time \((a - 1)\) quasi-events have been suppressed, then the event of probability \(p_i > p_n\) may receive a rank lower.

Thus

\[
L' < L_t - (N_t \% (a - 1)) p_n.
\]

\((a \% b\) means the integer part of the quotient of \(a\) by \(b\).)

This formula may be expressed more accurately

(1) by writing \(N_t\) in a polynomial form

\[
N_t = \sum_{a} (a - 1)^x
\]

and using the successive remainders;

(2) by substituting to \(p_n\) the \((N_t \% (a - 1))\) smallest probabilities among the \(p(e_i)\).

From (24) and (25):

**Theorem 7.** It is possible to find an upper limit to the routing-length of a poly-chotomic questionnaire by

\[
L_u \leq L(C) - \sum_{j=1}^{t} (c_i - \tilde{c}_0(j)) \max_{c_i > \tilde{c}_0(j)} \{p(e_i)\} - p_n(N_t \% (a - 1))
\]

where \(L(C)\) is defined by (12), \(N_t\) is the number of quasi-events of the quasi-questionnaire \(K_t\) of which all the quasi-events are of maximal rank, \(p_n\) is the smallest probability of \(P\).

For the ranks \(c_i, \tilde{c}_0(j)\) and \(t\) see formula (24). This formula is not trivial.
Nevertheless, it gives the determination of the upper limit of $L_H$ without using Huffman’s construction and has the theoretical interest to use the events by their probabilities $p(e)$ without using the questions.

By substituting a questionnaire $K^*$ to the code $C$ without using the questionnaire $K$, we found a more elementary formula:

Since $C$ has $N$ quasi-events, we can deduce from the polynom $N = \sum \beta_n(a - 1)^n$ how many times the rank of an event may reduced by one so that:

$$L'_H \leq L < L(C) - (N \% (a - 1)) p_N .$$

$L'$ may be stated more accurately as in (25), but we are not sure that the questionnaire $K^*$ with the routing-length $L'$ has the property $P$. (26) is more exact than (27).

According to a communication by E. Gilbert (Bell C°), it seems that E. Moore has shown a simpler formula but less precise than (27) (see also Jelinek’s Annex [6])

$$L_H \leq I_0(E) + 1 - 2 p_N \quad \text{for} \quad a = 2 .$$

ANNEX. EXAMPLES OF QUASI-QUESTIONNAIRES AND CODES

1. Let $a = 2, P = \{90/100, 10/100\}$.

For $c_1 = 1, c_2 = 4$ we have found $\tilde{c}_1(1) = 2$. (24) leads to

$$L_1 = I(C) - 2 \times \frac{10}{100} .$$

then $L' = L_H = 1$ and $I(C) = 1, 3$.

2. Let $a = 3, P = \{32/100, 25/100, 20/100, 15/100, 5/100, 2/100, 1/100\}$

$$\begin{align*}
J = 1 & \quad \tilde{c}_0(1) = 1 \quad c_1 = 2 \quad p(e_1) = \frac{32}{100} , \\
J = 2 & \quad \tilde{c}_0(2) = 2 \quad c_2 = 3 \quad p(e_2) = \frac{5}{100} , \\
J = 3 & \quad \tilde{c}_0(3) = 2 \quad c_3 = 4 \quad p(e_3) = \frac{2}{100} , \\
J \text{ Max} = 4 & \quad \tilde{c}_0(4) = 3 \quad c_4 = 5 \quad p(e_4) = \frac{1}{100} ,
\end{align*}$$

$$\begin{align*}
\mathcal{N}_1 = 2 - a - 1 \Rightarrow L' = L(C).
\end{align*}$$
Thus \( L_H = 1.51 \) and \( Z(C) = 2.12 \).

3. Let \( a = 3, P = \{ 22/45, 1/5, 1/15 \text{ (three times), } 1/54 \text{ (six times) } \} \); the ranks are determined by

\[
3^9 > \frac{22}{45} > 3^{-1} > \frac{1}{5} > 3^{-2} > \frac{1}{15} > 3^{-3} > \frac{1}{54} > 3^{-4}
\]

Thus \( c_i = 1, 2, 3, 4 \), we have

\[
3^{-1} + 3^{-2} + 3^{-2} + 6 \cdot 3^{-4} = 3^{-1} + 2 \cdot 3^{-2} + 2 \cdot 3^{-3}
\]

Thus \( \varrho = 3^{-1} + 3^{-3} \) and \( N = 2 \)

\[
L_H \leq L(C) = \left[ \frac{1}{5} \times 1 + \frac{1}{15} \times 1 + \frac{1}{54} \times 2 \right] - \frac{1}{54},
\]

\[
L_H \leq L(C) = \frac{29}{90}.
\]

Thus:

\[
I(C) = 1 + \frac{84}{90} \text{ and } L_H = 1 + \frac{54}{90}
\]

so that the bound is close to \( L_H \).

4. Let \( C \) be the code defined by: \( a = 2 \) and \( P = \{ 0.425/0.250/0.08125 \text{ (four times)} \} \); the ranks of \( C \) are:

\[
r(e_1) = r(e_2) = 2
\]

\[
r(e_3) = \ldots = r(e_6) = 4
\]

Thus

\[
\sum 2^{-r(e_i)} = 0.11
\]

and

\[
\rho = 0.01.
\]

There is only a quasi-event of rank 2. The difference \( I(C) - L' \) is obtained in the following way:

\( e_3 \) will change from rank 4 to rank 2 by permutation, then \( e_4 \) will change from rank 4 to rank 3 since \( N_{e_4} = 1 \).

Thus \( I(C) - L' = 3 \times 0.08125 = 0.24375 \).

(Received April 6, 1970.)
REFERENCES


VÝTAH

Kvazi-dotazníky, kóry a Huffmanova délka

C. F. PICARD

Definuji se nové koncepce, a to zejména kvazi-otázka neboli uzel s vycházející větví o nulové pravděpodobnosti. Kvazi-dotazník je pravděpodobnostní homogenní strom s kvazi-otázkami.

Ukazuje se, že každý okamžitý kód je kvazi-dotazník s přesně vymezenými podmínkami; může být těž dotazníkem bez větví o nulové pravděpodobnosti.

Těž je dána aproximace — aniž by se použilo klasické konstrukce — průměrné délky Huffmanova kódu s danou abecedou a s danými pravděpodobnostmi kódových slov.