Proceedings of the Colloquium on Information Theory
organized by the Bolyai Mathematical Society, Debrecen, (Hungary), 1967

Information sources with different cost scales and the principle of conservation of entropy

by

I. Csiszár, G. Katona and G. Tusnády

Math. Inst. of the Hung. Acad. of Sci, Budapest, Hungary

INTRODUCTION

Despite of the vast progress information theory has made in the last decade, some problems important from the point of view of the very foundations - to the authors' knowledge - still lack a rigorous and sufficiently general exposition. In this paper we attempt to fill some of these gaps, concerning problems of the following type (precise definitions will be given in Section 2):

(i) If the message symbols produced by an information source are of different cost, the entropy per unit cost can be defined either as the limit of the entropy of the message sequence of cumulative cost t divided by t, or as the entropy per symbol (limit of the entropy of the first n symbols divided by n) divided by the average symbol cost. (Most frequently, the cost of a symbol is its duration and entropy per unit cost is entropy per second.) Dating back to Shannon's fundamental paper [16], in general the second definition is adopted (see also [12], [15], etc.) but in heuristic reasonings it is often implicitly assumed to be equivalent to the first one. In the literature consideration is usually restricted to the simplest case that each symbol of the source alphabet has some fixed cost (duration), but no proof of the equivalence of the two possible definitions of entropy per second seems to have been published even for that case. It should be noted that also the general case has considerable interest, in particular if one looks at sources producing message symbols at random times - according to some point process - and "cost" is interpreted as the length of the time interval between two subsequent symbols.

(ii) For the interpretation of entropy as the measure of the amount of information the so-called noiseless coding theorem is of basic importance. It asserts, intuitively, that the greatest lower bound of the average number L of code characters per symbol needed to encode in a uniquely decipherable way the output of a source of entropy rate H equals $\frac{H}{\log_2 s'}$, where s' is the size of the coding alphabet. The "noiseless coding theorem" is usually stated and proved, however, for rather special codes only, namely for those defined by a fixed assignment of sequences of code symbols to the letters of the source alphabet, or to sequences("blocks") of fixed length of letters of the source alphabet (see e.g. Feinstein [9], Ash [1], etc.). On the other hand, the theorem is expected to be true for "all conceivable" codes and in order that it be really valuable from the "foundations" point of view it should be proved for "arbitrary" codes, including blockwise encodings with variable block length and the code mapping varying from block to block, in dependence on the previously encoded message symbols. The strongest results known in this direction are apparently those of Billingsley [2]. Moreover, if the message symbols or the code characters or both are of different cost, if H is interpreted as the entropy per unit cost, L as the average cost dilation due to the encoding, and $\log_2 s'$ is replaced by the "capacity of the noiseless channel" as defined by Shannon [16], one may infer that the statement still remains valid. In fact - for the case of fixed symbol costs - this statement occurs already in [16], but, to the authors' knowledge, no exact proof has been published so far, except for special (Markovian) sources.

(iii) It is "intuitively clear" that uniquely decipherable coding gives rise to a message of entropy (per symbol) $\frac{H}{L}$ where H is the entropy (per symbol) of the original message and is the average number of code characters per message symbol. A rigorous proof of this assertion for finite--state Markovian sources and encodings performed by finite-state transducers has been given by Sidelnikov [17], and for arbitrary sources and simple letter codes by two of the present authors [10]. In case of symbols of different cost, a similar relation is to be expected for the entropies per unit cost. In this direction there seems nothing to have been published.

The problems listed under (i), (ii) and (iii) are very closely related to each other. As a main tool for dealing with them we introduce the concept of entropy rate with respect to a stochastic cost scale and establish a theorem on the relation of entropy rates with respect to different cost scales under general conditions. Applying this result, we obtain an apparently satisfactory solution of problems (i)-(iii), for sources with finite alphabet; in particular we prove the "principle of conservation of entropy" for a very wide class of encoding procedures. Our method is straightforward and follows closely intuition. Our aim was to make familiar heuristic reasonings rigorous rather than to replace them by ad hoc non-information-theoretic arguments, in this respect, even for the particular cases of our results that have been proved earlier, our proofs seem preferable to the existing ones.

§ 1. PRELIMINARIES

(A) Throughout this paper, the terms "random variable", "discrete random variable" (= random variable with finite or countable state space), "integer valued random variable", "almost surely" (= with probability one), "uniformly integrable" and "if and only if" will be abbreviated as RV, DRV, IRV, a. s., u. i. and iff, respectively.

All RV's will be assumed to be defined on the same probability space $(\Omega, \mathcal{F}, P)$. RV's will be denoted by greek letters, omitting, as a rule, the argument ω. Except Ω and ω (typical element of Ω), all greek letters occurring in this paper denote RV's. In case of families of RV's (= stochastic processes) we shall write the parameter as an argument rather than as an index; thus a typical element of a sequence of RV's will be denoted by $\xi(n)$ rather than by ξ_n (of course, $\xi(n)$ means really $\xi(n,\omega)$).

If $A \in \mathcal{F}$, $P(A) > 0$, symbols with subscript A will refer to the probability measure $P_A(\cdot) = P(\cdot \mid A)$; if $P(A) = 0$, such symbols will be meant to be 0. E.g., $E_{\{a \le \xi \le b\}}(\eta)$ means $E(\eta \mid a \le \xi \le b)$ if $P(a \le \xi \le b) > 0$ and 0 otherwise.

(B) By entropy (conditional entropy) of DRV's we shall always mean entropy in the sense of Shannon:

$$H(\xi) = -\sum_x P(\xi = x) \log_2 P(\xi = x) \tag{1.1}$$

$$\begin{aligned} H(\xi \mid \eta) &= \sum_y P(\eta = y) H_{\{\eta = y\}}(\xi) = \\ &= -\sum_{x,y} P(\eta = y) P(\xi = x \mid \eta = y) \log_2 P(\xi = x \mid \eta = y) \end{aligned} \tag{1.2}$$

where x and y range over the state space of ξ and η, respectively (eventual undefined terms in (1. 1) and (1. 2) are considered as zeros).

We shall need also the concept of information distance (of DRV's)

$$d(\xi,\eta) = H(\xi|\eta) + H(\eta|\xi) \tag{1.3}$$

and the mutual information (of DRV's with $d(\xi,\eta) < \infty$)

$$I(\xi,\eta) = H(\xi) - H(\xi|\eta) = H(\eta) - H(\eta|\xi) . \tag{1.4}$$

The equality $H(\xi) - H(\xi|\eta) = H(\eta) - H(\eta|\xi)$ follows from (1.5) below, if $d(\xi,\eta) < \infty$. For the purposes of this paper, we need not define $I(\xi,\eta)$ if $d(\xi,\eta) = \infty$.

The well-known basic identities and inequalities concerning entropies and conditional entropies (due essentially to Shannon [16], see also e.g. [1], [9]) such as

$$H(\xi,\eta) = H(\xi) + H(\eta|\xi) = H(\eta) + H(\xi|\eta) \tag{1.5}$$

$$H(\xi,\eta|\zeta) = H(\xi|\zeta) + H(\eta|\xi,\zeta) = H(\eta|\zeta) + H(\xi|\eta,\zeta) \tag{1.5'}$$

$$0 \le H(\xi|\eta,\zeta) \le H(\xi|\eta) \le H(\xi) \tag{1.6}$$

$$0 \le I(\xi,\eta) \le \max(H(\xi), H(\eta)) \tag{1.7}$$

$$0 \le H(\xi) \le \log_2 \{\text{number of possible values of } \xi\} \tag{1.8}$$

$$0 \le H(\xi|\eta) \le E \log_2 \{\text{number of possible values of } \xi \text{ given } \eta\} \tag{1.8'}$$

will be used freely, without any further reference. We shall need also some other simple but somewhat less standard inequalities summarized in the following lemmas:

LEMMA 1.1. We have for arbitrary DRV's

$$d(\xi,\eta)+d(\eta,\zeta)\geq d(\xi,\zeta) \tag{1.9}$$

$$|H(\xi)-H(\eta)|\leq d(\xi,\eta) \tag{1.10}$$

$$|H(\xi_1|\eta_1)-H(\xi_2|\eta_2)|\leq d(\xi_1,\xi_2)+d(\eta_1,\eta_2) \tag{1.11}$$

$$|I(\xi_1;\eta_1)-I(\xi_2,\eta_2)|\leq d(\xi_1,\xi_2)+d(\eta_1,\eta_2) \tag{1.12}$$

provided (in (1.10)-(1.12)) that the left hand side is meaningful.

REMARK 1.1. This lemma means that the information-distance is a metric in the space of DRV's with finite entropy and the different information quantities are (uniformly) continuous functions with respect to it.

LEMMA 1.2. If ξ is an IRV with finite expectation then

$$H(\xi)\leq E|\xi|+\log_2 3 \tag{1.13}$$

and

$$H(\xi)\leq E\log_2(|\xi|+1)+\log_2\left(\frac{\pi^2}{3}-1\right). \tag{1.14}$$

PROOF OF LEMMA 1.1. The triangle inequality (1.9) follows from

$$H(\xi|\eta)+H(\eta|\xi)\geq H(\xi|\eta,\zeta)+H(\eta|\xi)=H(\xi,\eta|\zeta)\geq H(\xi|\zeta) \tag{1.15}$$

and the corresponding inequality obtained by changing the role of ξ and ζ. (1.10) is an immediate consequence of (1.5) and (1.3). By obvious substitutions, (1.15) gives rise also to

$$\begin{aligned} &|H(\xi_1|\eta)-H(\xi_2|\eta)\leq d(\xi_1,\xi_2)\\ &|H(\xi|\eta_1)-H(\xi|\eta_2)\leq d(\eta_1,\eta_2)\end{aligned} \tag{1.16}$$

whence (1. 11) directly follows:

$$|H(\xi_1|\eta_1)-H(\xi_2|\eta_2)| \le |H(\xi_1|\eta_1)-H(\xi_2|\eta_1)| +$$

$$+|H(\xi_2|\eta_1)-H(\xi_2|\eta_2)| \le d(\xi_1,\xi_2)+d(\eta_1,\eta_2)$$

at last, from (1. 14), (1. 5) and the second inequality of (1. 16) we get

$$|I(\xi_1,\eta_1)-I(\xi_2,\eta_2)| = |H(\xi_1)-H(\eta_2)-H(\xi_1|\eta_1)+H(\eta_2|\xi_2)| =$$

$$= H(\xi_1|\eta_2)-H(\eta_2|\xi_1)-H(\xi_1|\eta_1)+H(\eta_2|\xi_2)| \le d(\eta_1,\eta_2)+d(\xi_1,\xi_2)$$

i. e. (1. 12).

PROOF OF LEMMA 1. 2. Set $p_k = P(\xi=k)$ and $q_k = \frac{1}{3}\,2^{-|k|}$ $(k=0,\pm1,\pm2,\dots)$. Then $\{q_k\}$ is a probability distribution and the well-known inequality

$$\sum p_k \log_2 \frac{p_k}{q_k} \ge 0$$

gives rise to (1. 13).

(1. 14) can be proved in the same way, with the choice

$$q_k = \frac{c}{(|k|+1)^2} \qquad (k=0,\pm1,\pm2,\dots) \qquad c=\left(\frac{\pi^2}{3}-1\right)^{-1} .$$

(C) We shall have to do with three types of convergence of RV' s: convergence in probability (or stochastic concergence), almost sure convergence (convergence with probability one) and convergence in L_1 -norm. They will be denoted by $\xrightarrow{P}$, $\xrightarrow{a.s.}$ and $\xrightarrow{L_1}$, respectively.

LEMMA 1. 3. Let $\xi(t)$, $t\ge0$ be a family of RV' s and let $\xi(t)\xrightarrow{P}\xi$, $(t\to+\infty)$. Then the conditions

a) $\xi(t)$ is uniformly integrable (u. i.) for $t\to\infty$ and

b) $E|\xi(t)| \to E|\xi| < \infty$ $\qquad (t\to\infty)$

are equivalent and imply $\xi(t)\xrightarrow{L_1}\xi$ $(t\to\infty)$; and conversely, $\xi(t)\xrightarrow{L_1}\xi\in L_1$ implies $\xi(t)\xrightarrow{P}\xi$ and both (a) and (b).

Here condition (a) means that $\overline{\lim\limits_{t\to\infty}} \int\limits_{|\xi(t)|\ge k} |\xi(t)|\,P(d\omega) \to 0$ as $k \to +\infty$.

REMARK 1.2. If there exists $t_0 \ge 0$ such that for every finite $t_1 > t_0$ the RV's $\xi(t)$, $t_0 \le t \le t_1$ are u.i. then, obviously, condition (a) is equivalent to saying that $\xi(t)$ is u.i. for $t \ge t_0$.

PROOF. From $\xi(t) \xrightarrow{L_1} \xi \in L_1$ obviously follows both $\xi(t) \xrightarrow{P} \xi$ and (b). These, in turn, imply, on account of

$$E|\xi(t)| - E|\xi| =$$

$$= \int\limits_{A(t,k)} (|\xi(t)| - |\xi|)\,P(d\omega) + \int\limits_{\bar{A}(t,k)} |\xi(t)|\,P(d\omega) - \int\limits_{\bar{A}(t,k)} |\xi|\,P(d\omega)$$

(with $A(t,k) = \{\omega : |\xi(t) - \xi| < k,\ |\xi| < k\}$, $\bar{A}(t,k) = \Omega \setminus A(t,k)$, $k > 0$ fixed) the relation

$$\overline{\lim_{t\to\infty}} \int\limits_{|\xi(t)|\ge 2k} |\xi(t)|\,P(d\omega) \le \overline{\lim_{t\to\infty}} \int\limits_{\bar{A}(t,k)} |\xi|\,P(d\omega) = \int\limits_{|\xi|>k} |\xi|\,P(d\omega)$$

(we used that $\{\omega : |\xi(t)| \ge 2k\} \subset \bar{A}(t,k)$) whence (a) directly follows. Finally, $\xi(t) \xrightarrow{P} \xi$ and (a) obviously imply $\xi(t) \xrightarrow{L_1} \xi$ completing the proof.

(D) If $X = \{x_1, \dots, x_s\}$ is an arbitrary finite set, we denote by $U(X)$ and $\tilde{U}(X)$ the set of all finite and infinite sequences, respectively, of elements of X . Here "sequence" means simply juxtaposition of elements without commas. The void sequence u_0 will be also considered to belong to $U(X)$. The set X will be called an alphabet iff identical sequences from $U(X) \cup \tilde{U}(X)$ are elementwise identical, too, i.e. $x_{i_1}x_{i_2}\dots x_{i_m} = x_{j_1}x_{j_2}\dots x_{j_n}$ implies $n = m$ and $i_k = j_k$ $(k = 1, 2, \dots, n)$, $x_{i_1}x_{i_2}\dots = x_{j_1}x_{j_2}\dots$ implies $i_k = j_k$ $(k = 1, 2, \dots)$ and $U(X)$ and $\tilde{U}(X)$ are disjoint. This condition means only that the elements of X are really "elementary", excluding e.g. the possibility of $x_1 = a$, $x_2 = b$, $x_3 = ab$ or $x_1 = a$, $x_2 = bc$, $x_3 = ab$, $x_4 = c$ etc.

The elements of an alphabet will be referred to as letters. If X is an alphabet and $u = x_{i_1} x_{i_2} \cdots x_{i_m} \in \mathcal{U}(X)$ then m, the length of the sequence u, is uniquely determined by u; it will be denoted by $\|u\|$. Of course, we set $\|u_0\| = 0$ and for $u \in \tilde{\mathcal{U}}(X)$ we set $\|u\| = +\infty$. For $u, v \in \mathcal{U}(X) \cup \tilde{\mathcal{U}}(X)$ we shall write $u \prec v$ iff $\|u\| \leq \|v\|$ and the sequence of the first $m = \|u\|$ letters of v is identical with u. Obviously, $\prec$ is a partial order on $\mathcal{U}(X) \cup \tilde{\mathcal{U}}(X)$. Subsets of $\tilde{\mathcal{U}}(X)$ of form $c(u) = \{\tilde{u} : \tilde{u} \succ u\}$ $(u \in \mathcal{U}(X))$ will be referred to as cylinder sets; the smallest σ-algebra of subsets of $\tilde{\mathcal{U}}(X)$ containing all cylinder sets will be denoted by $\mathcal{B}$.

(E) For the rest of this paper, it will be convenient to restrict the use of certain letters, attaching them some specific meanings in a consistent way. Our notational conventions will be the following ones:

X : finite alphabet

$\xi(n)$ $(n = 1, 2, \ldots)$ sequence of DRV's with common state space X

$\mathfrak{X}$ abbreviation for the above sequence

$$\xi(k,n) = \begin{cases} \xi(k), \xi(k+1), \ldots, \xi(n) & \text{if} \quad k \leq n \\ \text{the void sequence } u_0 & \text{if} \quad k > n \end{cases}$$

$\zeta(n)$ $(n = 1, 2, \ldots)$ sequence of nonnegative real RV's, such that

$$\sum_{k=1}^{n} \zeta(n) \xrightarrow{a.s.} +\infty \qquad \text{as} \qquad n \to \infty .$$

$\mathfrak{Z}$ abbreviation for the above sequence

$$\tau(n) \quad (n = 1, 2, \ldots) : \quad \tau(n) = \sum_{k=1}^{n} \zeta(k) \qquad (\tau(0) = 0)$$

$\nu(t)$ $(0 \leq t < +\infty)$: number of n's with $\tau(n) \leq t$

$$\eta(t) \quad (0 \leq t < +\infty) : \quad \eta(t) = \xi(1, \nu(t))$$

($\nu(t)$ and $\eta(t)$ are well-defined a. s., due to the assumption $\tau(n) \xrightarrow{a.s.} +\infty$).

Processes $\mathfrak{Z}$ will be thought of as associated to processes (cf. Definition 2.1 below). Different $\mathfrak{X}$ processes will be distinguished by dashes; different $\mathfrak{Z}$ processes associated with the same $\mathfrak{X}$ will be dashed in the same way as $\mathfrak{X}$, and they will be distinguished by indices. Given $\mathfrak{X}$ and $\mathfrak{Z}$, the corresponding ξ's, ζ's, τ's, ν's and η's will be given the same dashes and (or) indices as $\mathfrak{X}$ and $\mathfrak{Z}$.

Instead of $\xi(n)$, $\zeta(n)$, $\tau(n)$, $\nu(t)$ and $\eta(t)$ we shall often write simply ξ, ζ, τ, ν and η, if omitting the argument does not cause ambiguity.

Observe that $\mathfrak{Z}$ is uniquely defined both by the RV's τ and ν; each non-decreasing sequence of nonnegative RV's $\tau(n)$ $(n=1,2,\ldots)$ with $\tau(n) \xrightarrow{a.s.} +\infty$ defines a sequence and so does each family of IRV's $\nu(t) \geq 0$ $(0 \leq t < +\infty)$ with right-continuous sample functions tending to infinity as $t \to \infty$.

§ 2. INFORMATION SOURCES WITH DIFFERENT COST SCALES; COMPARISON OF THE CORRESPONDING ENTROPY RATES

DEFINITION 2.1. An information source $\mathfrak{X}$ with finite alphabet X is a sequence of DRV's $\xi(n)$ $(n=1,2,\ldots)$ having the finite alphabet X as common state space. A cost scale $\mathfrak{Z}$ is a point process on $[0, \infty)$ described in terms of ζ's, τ's and ν's, see Section 1 (E). A cost scale $\mathfrak{Z}$ will be called regular if $\frac{\nu(t)}{t}$ is u.i. for $t \to \infty$.

Intuitively, $\xi(n)$ represents the n'th message symbol emitted by the source, $\zeta(n)$ its cost, $\tau(n)$ the cumulative cost of the first n message symbols and $\nu(t)$ the number of message symbols with cumulative cost just not exceeding t. E.G. the "cost" may be time as in examples 2.2 and 2.3 below; then $\tau(n)$ represents the epoch at which the emission of the n'th message symbol terminates and $\nu(t)$ is the number of message symbols emitted up to the epoch t.

EXAMPLE 2.1. The simplest cost scale is defined by

$$(2.1) \qquad \zeta(n)=1, \quad \tau(n)=n \quad (n=1,2,\ldots), \quad \nu(t)=[t] \quad (0 \leq t < \infty).$$

This cost scale will be referred to as the counting scale $\mathfrak{C}$.

EXAMPLE 2.2. Let $\ell(u)$ be a nonnegative valued function on $\mathcal{U}(X)$ such that $\ell(u_0)=0$ and $u \prec v$ implies $\ell(u) \leq \ell(v)$. Then, for any source $\mathfrak{X}$ with alphabet X,

$$\zeta(n) = \ell(\xi(1,n)) - \ell(\xi(1,n-1)), \quad \tau(n) = \ell(\xi(1,n)) \qquad (n = 1,2,\dots) \tag{2.2}$$

defines a cost scale $\mathfrak{Z}$. Cost scales of this type will be called <u>strictly intrinsic</u>.

In particular, if

$$\ell(u) = \sum_{j=1}^{n} \ell(x_{i_j}) \qquad (u = x_{i_1} x_{i_2} \dots x_{i_n}) \tag{2.3}$$

the corresponding cost scale $\mathfrak{Z}$ defined by

$$\zeta(n) = \ell(\xi(n)), \quad \tau(n) = \sum_{k=1}^{n} \ell(\xi(k)) \qquad (n = 1,2,\dots) \tag{2.4}$$

may be called a <u>memoryless</u> intrinsic cost scale; observe that $\ell(x) \equiv 1$ gives rise to the counting scale $\mathcal{C}$. E.g. $\ell(x)$ may be the length or duration of the symbol $x \in X$. Then, if the symbols are emitted consecutively, without intervals, $\tau(n) = \sum_{k=1}^{n} \ell(\xi(k))$ is the epoch at which the emission of the n'th message symbol terminates.

EXAMPLE 2.3. The cost of transmission may depend on random outer disturbances independent of the symbols to be transmitted; in our model this means that $\mathfrak{X}$ and $\mathfrak{Z}$ are independent stochastic processes. The same holds if the symbols are emitted at random epochs, independent of the symbols themselves and "cost" means time.

EXAMPLE 2.4. A cost scale may be defined by letting $\tau(n)$ denote the number of binary digits needed to encode the first n message symbols when a particular method of encoding is used.

REMARK 2.1. A cost scale $\mathfrak{Z}$ is trivially regular if $\frac{\nu(t)}{t}$ is uniformly bounded; e.g. a strictly intrinsic cost scale (cf. example 2.2) is surely regular if $\frac{\ell(u)}{\|u\|}$ is bounded away from 0 $(u \neq u_0)$.

If $\mathfrak{X}$ is a source, sequences of type $\xi(1,n)$ will be called finite messages of $\mathfrak{X}$ (for the notations cf. § 1, (E)). In particular, $\eta(t) = \xi(1, \nu(t))$ is the message of cumulative cost just not exceeding t (with respect to the cost scale $\mathfrak{Z}$). E.g. if "cost" is time then $\eta(t)$ is the message emitted in the time interval $[0,t]$.

Obviously, $\eta(t)$ is DRV; its possible "values" are finite sequences belonging to $U(X)$, including, possibly, the void sequence u_0

The entropy of $\eta(t) = \xi(1, \nu(t))$ can be considered as the average information content of a message of cost t (with respect to the given cost scale). This suggests the following.

DEFINITION 2. 2. The entropy rate of the source $\mathfrak{X}$ with respect to the cost scale $\mathfrak{Z}$ is the limit

(2. 5) $$H(\mathfrak{X} \| \mathfrak{Z}) = \lim_{t \to \infty} \frac{1}{t} H(\eta(t))$$

provided that it exists. If the limit does not exist, we shall denote the limsup and liminf of $\frac{1}{t} H(\eta(t))$ by $\overline{H}(\mathfrak{X} \| \mathfrak{Z})$ and $\underline{H}(\mathfrak{X} \| \mathfrak{Z})$, respectively. (The double bar is used in order to avoid confusion with conditional entropy.)

If the cost of each symbol is unity i.e. $\mathfrak{Z} = \mathcal{C}$ (cf. example 2. 1) then $\eta(t) = \xi(1, [t])$ reduces to the usual definition of entropy per symbol

(2. 6) $$H(\mathfrak{X}) = H(\mathfrak{X} \| \mathcal{C}) = \lim_{n \to \infty} \frac{1}{n} H(\xi(1,n)) .$$

The idea underlying definition 2. 2 is that the relevant information is carried by the message symbols i.e. by the process $\mathfrak{X}$ and not by the process $\mathfrak{Z}$.

In the sequel we shall omit the arguments t where doing so does not cause ambiguity.

LEMMA 2. 1. For every regular cost scale

(2. 7) $$H(\nu(t)) = o(t) \qquad (t \to \infty)$$

holds. Furthermore, for two arbitrary cost scales $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$

(2. 8) $$d(\nu_1, \nu_2) \leq 2E|\nu_1 - \nu_2| + 2 \log_2 3$$

thus if $\frac{\nu_1(t)-\nu_2(t)}{t} \xrightarrow{L_1} 0$, then (2. 7) holds or does not hold simultaneously for $\mathcal{Z}_1$ and $\mathcal{Z}_2$.

PROOF. If $\mathcal{Z}$ is regular i. e. if $\frac{\nu(t)}{t}$ is u. i. for $t \to \infty$ then $E(\nu(t)) = O(t)$ and $E \log_2(\nu+1) \le \log_2 E(\nu+1) = \log_2 O(t) = o(t)$ holds. Furthermore, as

$$d(\nu_1, \nu_2) = H(\nu_1|\nu_2) + H(\nu_2|\nu_1) = H(\nu_1-\nu_2|\nu_2) + H(\nu_2-\nu_1|\nu_1) \le 2H(\nu_1-\nu_2)$$

the inequality (2. 8) is a consequence of lemma 1. 2. The last assertion follows from (2. 8) and lemma 1. 1 ((1. 10)).

In this section we shall be interested in the relation of entropy rates with respect to different cost scales of the same information source. Let us remark that for memoryless intrinsic cost scales (2. 4), when interpreting $\ell(x)$ as the length or duration of the symbol $x \in X$, "Entropy per second" is commonly defined (cf. [16], [12], etc.) as entropy per symbol (2. 6) divided by the "average symbol length" L "and not as in definition 2. 2; in some reasonings, however, this $\frac{H(\mathcal{X})}{L}$ is implicitely replaced by our $H(\mathcal{X}\|\mathcal{Z})$. Our results will, in particular, provide a justification for such reasonings under general conditions (cf. theorem 2. 4).

For an arbitrary real number r and $k > 0$ we set

$$|r|^+ = \max(0,r), \quad |r|^- = -\min(0,r), \quad |r| = |r|^+ + |r|^-$$

(2. 9)

$$|r|_k^+ = \min(|r|^+, k), \quad |r|_k^- = \min(|r|^-, k), \quad |r|_k = \min(|r|, k).$$

The following estimates will play fundamental role in the sequel.

THEOREM 2. 1. Let $\mathcal{X}$ be a source with alphabet X of size s and let $\mathcal{Z}_1$ and $\mathcal{Z}_2$ be two different cost scales for $\mathcal{X}$. Then

(2. 10) $$H(\eta_1|\eta_2) \le H(\nu_1|\nu_2) + \log_2 s \, E|\nu_1-\nu_2|^+$$

and also, if $A \in \mathcal{F}$ is such that on $\bar{A} = \Omega \setminus A$ we have $\nu_1 - \nu_2 \le kt$ (where $k > 0$ is arbitrary)

(2.11) $H(\eta_1|\eta_2) \leq 1+P(\bar{A})H_{\bar{A}}(\nu_1|\nu_2)+\log_2 s E|\nu_1-\nu_2|_{kt}^{+} + P(A)H_A(\eta_1)$.

PROOF. As $\nu_i = \|\eta_i\|$ is uniquely determined by $\eta_i = \xi(1,\nu_i)$ $(i=1,2)$, we have

(2.12) $H(\eta_1|\eta_2) = H(\nu_1,\eta_1|\nu_2,\eta_2) \leq H(\nu_1|\nu_2)+H(\eta_1|\nu_1,\nu_2,\eta_2)$.

As for given ν_1,ν_2 and $\eta_2=\xi(1,\nu_2)$ the number of possible "values" of $\eta_1=\xi(1,\nu_1)$ is at most $s^{|\nu_1-\nu_2|^+}$, the last term in (2.12) is $\leq E(\log_2 s^{|\nu_1-\nu_2|^+}) =$ $= \log_2 s E|\nu_1-\nu_2|^+$ proving (2.10). We may also write (setting $\alpha=1$ if $\omega \in A$ and $\alpha=0$ otherwise), $H(\eta_1|\eta_2) \leq H(\alpha,\eta_1|\eta_2) = H(\alpha)+P(\bar{A})H_{\bar{A}}(\eta_1|\eta_2)+$ $+P(A)H_A(\eta_1|\eta_2)$ hence, applying (2.10) to $H_{\bar{A}}(\eta_1|\eta_2)$ and taking into account

(2.13) $$P(\bar{A})E_{\bar{A}}|\nu_1-\nu_2|^+ \leq E|\nu_1-\nu_2|_{kt}^{+}$$

and using the obvious inequalities $H(\alpha)\leq 1$, $H_A(\eta_1|\eta_2) \leq H_A(\eta_1)$ we obtain (2.11).

In order that the implications of theorem 2.1 can be formulated concisely, we introduce some definitions concerning cost scales.

If $\mathfrak{Z}$ is an arbitrary cost scale and $c>0$, we may define the cost scale $c\mathfrak{Z}$ as the sequence of RV's $c\zeta(n)$ $(n=1,2,\dots)$. Let us denote the τ's and ν's corresponding to the cost scale $c\mathfrak{Z}$ by τ^c and ν^c , respectively: $\tau^c(n) = c\tau(n)$, $n=1,2,\dots$; $\nu^c(t) = \nu(\frac{t}{c})$, $0 \leq t < \infty$.

DEFINITION 2.3. For two cost scales $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ we write

(2.14) $$\mathfrak{Z}_1 \sim \mathfrak{Z}_2 \qquad \text{iff} \qquad \frac{1}{t}(\nu_1(t)-\nu_2(t)) \xrightarrow{L_1} 0 .$$

If $\mathfrak{Z}_1 \sim \mathfrak{Z}_2$, we say that $\mathfrak{Z}_1$ is equivalent to $\mathfrak{Z}_2$.

The cost scales $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ will be said to be quasi-equivalent with quotient $c_{12}=c>0$ if $\mathfrak{Z}_1 \sim c\mathfrak{Z}_2$, i.e. if

(2.15) $$\frac{1}{t}(\nu_1(t)-\nu_2^c(t)) = \frac{1}{t}(\nu_1(t)-\nu_2(\tfrac{t}{c})) \xrightarrow{L_1} 0 \ .$$

Of course, in case of regular cost scales, the replacement of L_1 -convergence by stochastic convergence in the above definitions (in (2.14) and (2.15)) makes no difference.

Intuitively, $\mathfrak{Z}_1 \sim \mathfrak{Z}_2$ means that the cost of one symbol is essentially the same for both scales.

The equivalence and the quasi-equivalence are equivalence relations, and the quotient c_{12} is uniquely determined by $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$, except for the trivial case $E\nu_i(t) = o(t)$ $(i=1,2)$. If $\mathfrak{Z}_1, \mathfrak{Z}_2$ and $\mathfrak{Z}_3$ are quasi-equivalent cost scales, there obviously holds

(2.16) $$c_{12}c_{23} = c_{13}, \qquad c_{21} = \frac{1}{c_{12}} \ .$$

THEOREM 2. Let $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ be two cost scales for a source $\mathfrak{X}$ with finite alphabet X, and let one of the entropy rates $H(\mathfrak{X}\,\|\,\mathfrak{Z}_1)$ and $H(\mathfrak{X}\,\|\,\mathfrak{Z}_2)$ exist. If $\mathfrak{Z}_1$ is quasi-equivalent to $\mathfrak{Z}_2$ with quotient $c_{12} = c > 0$ then

(2.17) $$H(\mathfrak{X}\,\|\,\mathfrak{Z}_1) = \frac{1}{c} H(\mathfrak{X}\,\|\,\mathfrak{Z}_2) \ .$$

If the entropy rates in question do not exist, the assertion remains true both for the lower and upper entropy rates.

PROOF. (2.10) implies

(2.18) $$d(\eta_1, \eta_2^c) \le d(\nu_1, \nu_2^c) + \log_2 s\, E|\nu_1 - \nu_2^c| \ .$$

Hence, using (1.10) and (2.8), we obtain

$$|H(\eta_1) - H(\eta_2^c)| \le (2 + \log_2 s)\, E|\nu_1 - \nu_2^c| + 2\log_2 3$$

what results by definition the desired statement

$$H(\mathfrak{X} \| \mathfrak{Z}_1) = \lim_{t\to\infty} \frac{1}{t} H(\eta_1(t)) = \lim_{t\to\infty} \frac{1}{t} H(\eta_2^c(t)) =$$

$$= \lim_{t\to\infty} \frac{1}{ct} H(\eta_2(t)) = \frac{1}{c} H(\mathfrak{X} \| \mathfrak{Z}_2)$$

(where $\eta^c(t) = \xi(1, \nu^c(t)) = \xi\left(1, \nu\left(\frac{t}{c}\right)\right) = \eta\left(\frac{t}{c}\right)$).

In order to apply theorem 2.2 to concrete problems it will be convenient to establish some simple sufficient conditions of the relation $\sim$ for different cost scales.

LEMMA 2.2. Let $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ be two regular cost scales and let one of them have the property

(2.19) $\qquad b \le \zeta(n) \le B \qquad$ a.s. $\qquad (0 < b < B;\ n = 1, 2, \dots)$.

Then each of the conditions

(2.20) $$\frac{\nu_1(t)}{\nu_2(rt)} \xrightarrow{P} 1$$

(2.21) $$\frac{1}{t} \tau_2(\nu_1(t)) \xrightarrow{P} r > 0$$

(2.22) $$\frac{1}{t} \tau_1(\nu_2(t)) \xrightarrow{P} c > 0$$

is equivalent to $\mathfrak{Z}_1 \sim c\,\mathfrak{Z}_2$, where $c = \frac{1}{r}$.

PROOF. For regular cost scales $\mathfrak{Z}_1 \sim \frac{1}{r}\mathfrak{Z}_2$ means

(2.23) $$\frac{1}{t}(\nu_1(t) - \nu_2(rt)) \xrightarrow{P} 0 .$$

Without any loss of generality, let e.g. $\mathfrak{Z}_2$ have the property (2.19), i.e. $0 < b \le \zeta_2(n) \le B$; then $\left[\frac{rt}{B}\right] \le \nu_2(rt) \le \frac{rt}{b}$ a.s., thus the equivalence of (2.20) and (2.23) is obvious. Furthermore, as by the definition of the τ's and ν's the relation $\tau_2(\nu_1(t)) \le yt$ is equivalent to $\nu_1(t) \le \nu_2(yt)$ $(0 < y < \infty)$, the relation (2.21) is equivalent to

$$P(\nu_1(t) \le \nu_2(yt)) \longrightarrow \begin{cases} 0 & \text{if} \quad y < r \\ & \qquad\qquad (t \to \infty) \\ 1 & \text{if} \quad y > r \end{cases}$$

and this, in turn, is equivalent to (2.23), in view of the assumption $0 < b \le \zeta_2(n) \le B$. Similarly, (2.28) is equivalent to

$$P(\nu_2(t) \le \nu_1(yt)) \longrightarrow \begin{cases} 0 & \text{if} \quad y < c \\ & \qquad\qquad (t \to \infty) \\ 1 & \text{if} \quad y > c \end{cases}$$

i.e. to

$$P(\nu_2(y't) \le \nu_1(t)) \longrightarrow \begin{cases} 0 & \text{if} \quad y' = \frac{1}{y} > \frac{1}{c} = r \\ & \qquad\qquad (t \to \infty) \\ 1 & \text{if} \quad y' < r \end{cases}$$

which, again, is equivalent to (2.23).

The following consequence of theorem 2.2 and lemma 2.2 is worth being formulated as a new theorem.

THEOREM 2.3. Let $\mathfrak{X}$ be a source with finite alphabet and let $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ be two regular cost scales for $\mathfrak{X}$ such that one of them has the property (2.19). Then either of the three equivalent conditions (2.20)-(2.22) implies

$$H(\mathfrak{X} \,\|\, \mathfrak{Z}_1) = r H(\mathfrak{X} \,\|\, \mathfrak{Z}_2) \tag{2.24}$$

(in the sense that if either side exists so does also the other and they are equal). If the entropy rates in question do not exist, (2.24) still holds both for the lower and upper entropy rates.

The most important cost scales are those quasi-equivalent to the counting scale $\mathcal{C}$ (cf. example 2.1). By definition 1.3, a cost scale $\mathfrak{Z}$ is quasi-equivalent to the counting scale $\mathcal{C}$ iff there exists a constant $c > 0$ such that $\frac{\nu(t)}{t} \xrightarrow{L_1} \frac{1}{c}$. Of course, all cost scales quasi-equivalent to the counting scale are regular (cf. Lemma 2.1).

If for a given cost scale $\frac{\nu(t)}{t}$ converges in probability to a (finite) constant r, this r will be called the symbol rate of the source with respect to the given cost scale. Similarly, if $\frac{\tau(n)}{n}$ converges in probability to a (finite) constant c, this c will be called the average symbol cost (with respect to the given cost scale).

In view of lemma 2.2 (with $\mathfrak{Z}_1 = \mathfrak{Z}$, $\mathfrak{Z}_2 = \mathfrak{C}$) a symbol rate $r > 0$ exists iff an average symbol cost $c > 0$ exists $(c = \frac{1}{r})$ and these are also necessary and sufficient conditions of $\mathfrak{Z} \sim c\mathfrak{C}$ for regular $\mathfrak{Z}$. Thus we obtain, as a particular case of theorem 2.3.

THEOREM 2.4. If for a source $\mathfrak{X}$ with finite alphabet X and with a regular cost scale $\mathfrak{Z}$ a positive symbol rate r or, equivalently, a positive symbol cost c exists $(cr = 1)$ then

$$H(\mathfrak{X} \| \mathfrak{Z}) = rH(\mathfrak{X}) = \frac{1}{c} H(\mathfrak{X}) \tag{2.25}$$

if either of $H(\mathfrak{X})$ and $H(\mathfrak{X} \| \mathfrak{Z})$ exists.

REMARK 2.2. If $\mathfrak{Z}$ is a cost scale of (2.4) (more general cost scales do not seem to have been considered in the literature) and the source $\mathfrak{X}$ is a stationary ergodic source (i.e. $\xi(1), \xi(2), \dots$ is a stationary ergodic sequence of DRV's), then

$$\frac{1}{n} \sum_{k=1}^{n} \ell(\xi(k)) \xrightarrow{P} L > 0 . \tag{2.26}$$

In this case "entropy per unit cost" is often defined as the ratio $\frac{H(\mathfrak{X})}{L}$. By theorem 2.4 this definition is equivalent to (2.5), provided that $\mathfrak{Z}$ is regular. This is the case, in particular, if $\ell(x) > 0$ for all $x \in X$, or if the message symbols $\xi(n)$ are independent and identically distributed (in the latter case, $L = E\zeta(1)$ by the law of large numbers and $\frac{E(\nu(t))}{t} \to \frac{1}{E\zeta(1)}$ by the renewal theorem, and this ensures the regularity by lemma 1.3.).

EXAMPLE 2.5. Let us be given a finite-state noiseless channel as defined by Shannon [16]; such a channel is specified by the input alphabet X, the set of states A, an assignment of subsets $X(a)$ of X to each $a \in A$ and by a function $G(x, a)$ defined for $a \in A$, $x \in X(a)$ and taking its values in A (X and A are finite sets). In each state a , the channel is capable of transmitting letters from $X(a)$ only and if $x \in X(a)$ is transmitted, the new state will be $a' = G(x, a)$. Let an $a_{i_0} \in A$ be fixed as the initial state;

a sequence $u = x_{i_1} x_{i_2} \cdots x_{i_n} \in U(X)$ is transmissible iff $x_{i_k} \in X(a_{i_{k-1}})$ $(k=1,\dots,n)$ where a_{i_k} is defined recursively by $a_{i_k} = G(x_{i_k}, a_{i_{k-1}})$ $(k=1,2,\dots,n)$ denote the set of all transmissible sequences by U_0 . Let $\ell(x, a) \geq 0$, $(a \in A,\ x \in X(a))$ be the cost of transmission of x at the state a ;

then $\ell(u) = \sum_{k=1}^{n} \ell(x_{i_k}, a_{i_{k-1}})$ representes the cost of transmission of the

sequence $u = x_{i_1} x_{i_2} \cdots x_{i_n} \in U_0$. We make the usual assumptions that for each pair of states $a', a'' \in A$ there exists $u = x_{i_1} \dots x_{i_k} \dots x_{i_n} \in U_0$ such that $a_{i_k} = a'$, $a_{i_n} = a''$, and that for all $u \in U_0$ with $\|u\| > m$(say) $\ell(u) > 0$. Let $\mathfrak{X}$ be a source transmissible by the channel (i. e. $\xi(1) \cdots \xi(n) \in U_0$ a. s. $n = 1, 2, \dots$) and let the (regular) cost scale $\mathfrak{Z}$ be defined by (2. 2), with the present $\ell(u)$. Let $N(t)$ denote the number of different sequences $u \in U_0$ with $\ell(u) \leq t$ and such that $\ell(ux) > t$ for some $x \in X$, with $ux \in U_0$; we define the channel capacity by

(2. 27)
$$C = \overline{\lim_{t \to \infty}} \frac{1}{t} \log_2 N(t) .$$

Then $\eta(t) = \xi(1, \nu(t))$ has at most $N(t)$ possible values, implying $H(\eta(t)) = \log_2 N(t)$, $\bar{H}(\mathfrak{X} \| \mathfrak{Z}) \leq C$; thus if $\frac{\tau(n)}{n} = \frac{1}{n} \ell(\xi(1, n)) \xrightarrow{P} L$ we have, according to (2. 25),

(2. 28)
$$\frac{1}{L} \bar{H}(\mathfrak{X}) \leq C .$$

The inequality (2. 28), first appearing in Shannon' s fundamental paper [16] is often considered to be "obvious". However, its familiar "justification" relays on the equivalence of the two possible definitions of entropy per unit cost and can be made rigorous only on the basis of theorem 2. 4. The existing rigorous proofs of (2. 28) concern stationary Markovian sources only ([7], [15]) or the case of one state ([5], [12]).

Let us now consider the inequality (2. 28) in the most general form that still follows from theorem 2. 3.

THEOREM 2.5. Let $\mathfrak{X}$ be a source with finite alphabet and let $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ two regular cost scales for $\mathfrak{X}$ such that one of them has the property (2.19) and

$$\frac{1}{t} \tau_2(\nu_1(t)) \xrightarrow{P} L > 0 .$$

If $N(t)$ denotes the number of different possible values of $\eta_2(t) = \xi(1, \nu_2(t))$ and

$$C = \overline{\lim_{t \to \infty}} \frac{1}{t} \log_2 N(t)$$

then

(2.29)
$$\frac{1}{L} \bar{H}(\mathfrak{X} \| \mathfrak{Z}_1) \leq C .$$

In the original paper (cf. footnote 1) we proved (2.29) under more general conditions which did not ensure the validity of (2.24).

§ 3. THE PRINCIPLE OF CONSERVATION OF ENTROPY

Let $\mathfrak{X}$ and $\mathfrak{X}'$ be two sources with finite alphabets X and X', respectively. Let $\mathfrak{Z}$ be a cost scale for $\mathfrak{X}$ and $\mathfrak{Z}'$ a cost scale for $\mathfrak{X}'$; then we define the rates

(3.1)
$$H(\mathfrak{X}, \mathfrak{X}' \| \mathfrak{Z}, \mathfrak{Z}') = \lim_{t \to \infty} \frac{1}{t} H(\eta(t), \eta'(t))$$

(3.2)
$$H(\mathfrak{X} | \mathfrak{X}' \| \mathfrak{Z}, \mathfrak{Z}') = \lim_{t \to \infty} \frac{1}{t} H(\eta(t) | \eta'(t))$$

(3.3)
$$I(\mathfrak{X}, \mathfrak{X}' \| \mathfrak{Z}, \mathfrak{Z}') = \lim_{t \to \infty} \frac{1}{t} I(\eta(t), \eta'(t))$$

(3.4)
$$d(\mathfrak{X}, \mathfrak{X}' \| \mathfrak{Z}, \mathfrak{Z}') = \lim_{t \to \infty} \frac{1}{t} d(\eta(t), \eta'(t))$$

provided that the limits exist ($\eta(t)$ and $\eta'(t)$ are defined as in § 1, (E); the argument t will be often omitted). If the limits do not exist, one may consider the corresponding upper and lower rates (upper and lower limits).

As an immediate consequence of theorem 2. 1, lemma 2. 1 and lemma 1. 1 we have

THEOREM 3. 1. All the rates (3. 1)-(3. 4), as well as the corresponding upper and lower rates, remains unchanged if either of $\mathfrak{Z}$ and $\mathfrak{Z}'$ is replaced by an equivalent cost scale.

In this section we shall apply the results of Section 2 to the case that $\mathfrak{X}'$ is obtained from $\mathfrak{X}$ by encoding (or conversely). For this purpose we define the codes in a very general sense.

DEFINITION 3. 1. Let X and X' be finite alphabets. An arbitrary mapping f of $U(X)$ into $U(X')$ such that $f(u_0) = u_0$ and

$$(3.5) \qquad u \prec v \qquad \text{implies} \qquad f(u) \prec f(v)$$

will be called a <u>code</u> from X to X' .

The code of a sequence $u = x_{i_1} \cdots x_{i_n}$ we can write in the form

$$(3.6) \qquad f(u) = g(x_{i_1} | u_0)\, g(x_{i_2} | x_{i_1}) \cdots g(x_{i_n} | x_{i_1} \cdots x_{i_{n-1}}) .$$

Each code f from X to X' defines a mapping of infinite sequences, too; in fact, for $\tilde{u} = x_{i_1} x_{i_2} \ldots \in \tilde{U}(X)$ we may write $f(\tilde{u}) = g(x_{i_1} | u_0)\, g(x_{i_2} | x_{i_1}) \ldots g(x_{i_n} | x_{i_1} \ldots x_{i_{n-1}}) \ldots$. Observe that the definition 3. 1 does not exclude that the "code $f(\tilde{u})$ of some infinite sequence $\tilde{u}$ be finite. The set of those $\tilde{u} \in \tilde{U}(X)$ for which $f(\tilde{u})$ is infinite, will be denoted by $\tilde{D}_f$.

REMARK 3. 1. It is easy to see that $\tilde{D}_f \in \mathfrak{B}$ for any code f and also, the mapping $f : \tilde{D}_f \to \tilde{U}(X')$ is measurable with respect to the σ-algebras $\mathfrak{B}$ and $\mathfrak{B}'$ spanned by the cylinder sets. If f is a code, the mapping $f : \tilde{D}_f \to \tilde{U}(X')$ may be called an infinite-code; clearly, the concept of an infinite-code is much more restrictive than that of an arbitrary measurable mapping $\tilde{U}(X) \to \tilde{U}(X')$ (though all practically realizable mappings seem to belong to this class). Two different codes f_1 and f_2 may give rise to the same infinite-code. In this case we shall say that f_1 and f_2 are equivalent and write $f_1 \sim f_2$. Of course, each infinite-code may be identified with the corresponding equivalence class of codes and vice-versa.

If f is a code from X to X' and $u = x_{i_1} x_{i_2} \cdots x_{i_p} \in U(X)$ let $k_n = k_n(u)$ denote the n'th point of increase of the sequence

(3.7) $$0, \| f(x_{i_1}) \|, \ldots, \| f(x_{i_1}, \ldots, x_{i_m}) \|, \ldots, \| f(x_{i_1}, \ldots, x_{i_p}) \|$$

and $k_0 = 0$. For $\tilde{u} \in \tilde{U}(X)$ we define $k_n(\tilde{u})$ in a similar way; in particular, if $\tilde{u} \in \tilde{D}_f$, then $k_n(\tilde{u})$ is well defined for $n = 0, 1, 2, \ldots$

EXAMPLE 3.1. If in (3.6) $g(x|u) = g(x)$, where $g(x)$ is some mapping of X into $U(X') - \{u_0\}$, the resulting code will be called a simple letter code. In this case we have

(3.8)
$$f(u) = g(x_{i_1}) g(x_{i_2}) \cdots g(x_{i_n}) \qquad (u = x_{i_1} \cdots x_{i_n} \in U(X))$$
$$f(\tilde{u}) = g(x_{i_1}) g(x_{i_2}) \cdots \qquad (\tilde{u} = x_{i_1} x_{i_2} \cdots \in \tilde{U}(X)) .$$

EXAMPLE 3.2. Let X and X' be finite alphabets, let A be a finite set to be called the set of states, and let us be given two functions F and G mapping the Cartesian product $X \times A$ into $U(X')$ and A , respectively. Let an initial state $a_0 \in A$ be specified, set $f(u_0) = u_0$ and for $u = x_{i_1} x_{i_2} \cdots x_{i_n}, n > 0$ set

(3.9) $$f(u) = F(x_{i_1}, a_0) F(x_{i_2}, a_1) \cdots F(x_{i_n}, a_{n-1})$$

where the states a_k are defined, recursively by $a_k = G(x_{i_k}, a_{k-1})$ Then f is a code in the sense of definition 3.1; the encoder (X, X', A, a_0, F, G) will be called, following Shannon [16], a finite-state transducer and f will be referred to as the code generated by this finite-state transducer.

DEFINITION 3.2. Let $\mathfrak{X}$ be a source with finite alphabet X and let f be a code from X to X' . We say that $\mathfrak{X}$ is encodable by f if the IRV's

(3.10) $$\varkappa(n) = k_n(\xi(1) \xi(2) \ldots)$$

(k_n has been defined in connection with (3.7)) are well defined a.s. for $n = 1, 2, \ldots$.)

The encoding results in a new source $\mathfrak{X}' = f(\mathfrak{X})$ defined by

$$\xi'(1), \xi'(2), \ldots ; \xi'(1) \xi'(2) \cdots = f(\xi(1) \xi(2) \ldots) .$$

If there is given a cost scale $\mathfrak{Z}_1$ for $\mathfrak{X}$ then we also define the mapped cost scale $\mathfrak{Z}_1' = f(\mathfrak{Z}_1)$ by

$$\nu_1'(t) = \| f(\eta_1(t)) \| \tag{3.11}$$

or, equivalently, by

$$\tau_1'(n) = t \quad \text{iff} \quad \| f(\eta_1(t')) \| < n < \| f(\eta_1(t)) \| \quad \text{for all } t' < t \ . \tag{3.12}$$

REMARK 3.2. If a source $\mathfrak{X}$ is encodable by a code f, the mapped cost scale $f(\mathfrak{Z}_1)$ need not be regular, in general. In order that the regularity of $\mathfrak{Z}_1$ imply that of $f(\mathfrak{Z}_1)$, a simple sufficient condition consits in the boundedness of $\frac{\| f(u) \|}{\| u \|}$ $(u \neq u_0)$. In particular, for codes generated by finite-state transducers (cf. example 3.2), this condition is trivially fulfilled.

In order that a code be of any practical value, it ought to be possible, in some sense, to recover the original message from its encoded form. We adopt the following

DEFINITION 3.3. Let $\mathfrak{X}$ be a source with finite alphabet X and let f be a code such that $\mathfrak{X}$ encodable by f. The encoding $f: \mathfrak{X} \to \mathfrak{X}'$ will be said to be <u>finite decodable</u> if there exists a natural number d such that to any u' there are at most d u's satisfying $f(u) = u'$.

THEOREM 3.2. Let $\mathfrak{X}$ be a source with a cost scale $\mathfrak{Z}_1$ and let $\mathfrak{X}$ be encodable by a code f. Let further the encoding result in the source $\mathfrak{X}' = f(\mathfrak{X})$ and let $\mathfrak{Z}'$ be a cost scale for $\mathfrak{X}'$. Then if the code $f: \mathfrak{X} \to \mathfrak{X}'$ is finite decodable and if the mapped cost scale $\mathfrak{Z}_1' = f(\mathfrak{Z}_1)$ is quasi--equivalent to $\mathfrak{Z}'$, i.e. $\mathfrak{Z}_1' \sim c'\mathfrak{Z}'$; $c' > 0$, we have

$$H(\mathfrak{X}' \| \mathfrak{Z}') = c' H(\mathfrak{X} \| \mathfrak{Z}_1) \ . \tag{3.13}$$

PROOF. (3.11) implies $\eta_1'(t) = \xi'(1, \nu_1'(t)) = f(\eta_1(t))$ and thus $H(\eta_1'(t)) \leq H(\eta_1(t))$ i.e.

$$H(\mathfrak{X}' \| \mathfrak{Z}_1') \leq H(\mathfrak{X} \| \mathfrak{Z}_1) \ . \tag{3.14}$$

On the other hand, since f is finite decodable,

$$H(\eta_1 | \eta_1') \leq \log_2 d = o(t)$$

holds. Hence, using (1. 5), the inequality

$$H(\mathfrak{X}' \| \mathfrak{Z}_1') \geq H(\mathfrak{X} \| \mathfrak{Z}_1) \tag{3. 15}$$

follows. (3. 14) and (3. 15) give

$$H(\mathfrak{X}' \| \mathfrak{Z}_1') = H(\mathfrak{X} \| \mathfrak{Z}_1) \ . \tag{3. 16}$$

However, $\mathfrak{Z}_1' \sim c'\mathfrak{Z}'$, thus by theorem 2. 2 we obtain

$$H(\mathfrak{X}' \| \mathfrak{Z}_1') = \frac{H(\mathfrak{X}' \| \mathfrak{Z}')}{c'} \ . \tag{3. 17}$$

Finally, (3. 16) and (3. 17) give the desired (3. 13).

The intuitive meaning of theorem 3. 2 is clear. (3. 13) represents the "principle of conservation of entropy", i. e. that finite decodable encoding does not change the entropy rate apart from a factor representing the quotient of average costs after and before encoding. If the coding is not decodable from (3. 14) and (3. 17) we can obtain

$$H(\mathfrak{X}' \| \mathfrak{Z}') \leq c' H(\mathfrak{X} \| \mathfrak{Z}_1)$$

which means that in the non-decodable case some information may get lost.

The following corollary of theorem 3. 2 is worth formulating as a new theorem.

THEOREM 3. 3. Let $\mathfrak{X}$ be a source with a given regular cost scale $\mathfrak{Z}_1$; let $\mathfrak{X}$ be encodable by a finite decodable code f and let $\mathfrak{Z}'$ be a cost scale for $\mathfrak{X}' = f(\mathfrak{X})$ such that

$$b \leq \zeta'(n) \leq B \qquad \text{a. s.} \qquad (n = 1, 2, \dots \ ; \ 0 < b < B) \ . \tag{3. 18}$$

Then if

$$\frac{1}{t} \tau'(\| f(\eta_1(t)) \|) \xrightarrow{P} r > 0 \tag{3. 19}$$

and the code f has the property that $\dfrac{\| f(u) \|}{\| u \|}$ is bounded, we have

$$H(\mathfrak{X}' \| \mathfrak{Z}') = \frac{H(\mathfrak{X} \| \mathfrak{Z})}{r} \ . \tag{3. 20}$$

PROOF. By remark 3.2 the obtained cost scale $\mathfrak{Z}_1' = f(\mathfrak{Z}_1)$ is regular, similarly, $\mathfrak{Z}_1$ also is regular trivially. Using lemma 2.2 we obtain that (3.19) is equivalent to $\mathfrak{Z}_1' \sim \frac{1}{r}\mathfrak{Z}'$. Thus, theorem 3.2 leads to (3.20).

E.g. if $\mathfrak{X}'$ has a memoryless intrinsic cost scale (cf. example 2.2) defined by fixed symbol costs $\ell(x')$ $(x' \in X')$ such that $\zeta'(n) = \ell(\xi'(n))$ or somewhat more generally, if $\mathfrak{X}'$ is to be transmitted by a finite-state noiseless channel (cf. example 2.5), the symbol costs may depend on the "state of the channel", i.e. $\zeta'(n) = \ell(\xi'(n), \alpha(n-1))$ where $\alpha(k)$ represents the state of the channel after the transmission of the k'th message symbol. In these cases the condition (3.18) is trivially fulfilled provided that $\ell(x')$ (or $\ell(x', \alpha)$) is strictly positive.

REMARK 3.3. Theorem 3.3 is perhaps the most impressive form of the "principle of conservation of entropy". As $\tau'(\| f(\eta_1(t)) \|$ is the cumulative cost of the code of a message of cumulative cost t (i.e. of $\eta_1(t) = \xi_1(1, \nu_1(t))$) the condition (3.19) requires the existence of an average code cost r per unit message cost, in the sense of convergence in probability. If both $\mathfrak{Z}_1$ and $\mathfrak{Z}'$ are the counting scale $\mathfrak{C}$ (3.19) reduces to

$$\frac{1}{n} \| f(\xi(1,n)) \| \xrightarrow{P} r \tag{3.19'}$$

and the identity (3.20) becomes

$$H(\mathfrak{X}') = \frac{H(\mathfrak{X})}{r} \quad . \tag{3.20'}$$

This relation, dating back to Shannon [16], has often been regarded as " obvious" but, to the authors' knowledge, it has never been provec in a rigorous way, for arbitrary sources and codes. For the case that f is a simple letter code, a proof of (3.20') appears in [10]. The more general case of codes generated by finite-state transducers (cf. example 3.2) has been considered by Sidel'nikov [17]; he however, restricted attention to Markovian sources (though, as he has remarked, some of his results hold in more general cases, too).

We conclude this section by exhibiting a general form of the "noiseless coding theorem" as a consequence of theorem 3.3. (In our original paper we prove a more general form which is not such simple consequence of theorem 3.3.)

THEOREM 3.4. Let $\mathfrak{X}$ be a source with a given regular cost scale $\mathfrak{Z}_1$; let $\mathfrak{X}$ be encodable by a finite decodable code f and let $\mathfrak{Z}'$ be a cost scale for $\mathfrak{X}' = f(\mathfrak{X})$ such that

$$b \le \zeta'(n) \le B \qquad \text{a. s.} \quad (n = 1, 2, \cdots ; \; 0 < b < B) .$$

Denote by $N(t)$ the number of different possible values of $\eta'(t) = \xi(1, \nu'(t))$ and put

$$C = \overline{\lim_{t \to \infty}} \frac{\log_2 N(t)}{t} .$$

Then if

$$\frac{1}{t} \tau'(\| f(\eta_1(t)) \|) \xrightarrow{P} L > 0$$

and the code has the property that $\frac{\| f(u) \|}{\| u \|}$ is bounded, we have

(3.21) $$\frac{\bar{H}(\mathfrak{X} \| \mathfrak{Z})}{C} \le L .$$

PROOF. We have to observe only that by definition of C

$$H(\eta'(t)) \le \log_2 N(t) \qquad \text{and} \qquad H(\mathfrak{X}' \| \mathfrak{Z}') \le C$$

hold, thus the statement follows from (3.20).

REMARK 3.4. (3.21) has been proved for the special case if $\mathfrak{Z}_1 = \mathfrak{C}, \mathfrak{Z}'$ is a memoryless intrinsic cost scale and the code is a simple letter code. For the case of stationary ergodic sources with $\mathfrak{Z}_1 = \mathfrak{C}$ for $\mathfrak{Z}' = \mathfrak{C}$ and for general codes defined by him, Billingsley [2] has proved even a stronger theorem than the corresponding particular case of theorem 3.4. For more general cases, to the authors' knowledge, the assertion is, though "intuitively obvious", as a mathematical theorem new.

§ 4. CONCLUDING REMARKS

In this paper, we restricted ourselves to information sources with finite alphabet; the finiteness assumption has been essential for our basic estimations (2. 12) and (2. 13) (theorem 2. 1) and it remains an open problem, under what conditions does the "entropy rate comparison theorem" (theorems 2. 2 and 3. 2) hold for countable alphabets, too. Theorem 2. 5, however, remains unchanged also for countable X (except for the bound $C \leq \frac{\log_2 s}{b}$). Observe. too, that the theorems involving coding (theorems 3. 2, 3. 3, 3. 4) can obviously be extended to countable X (provided, in the case of the first two that X' remains finite). A possible approach to problems concerning countable alphabets in general would be to reduce them to the finite-alphabet case by an appropriate encoding, using the above remark.

As to the generality of the concept of coding used in this paper (definitions 3. 1 and 3. 2) one might make the objection that in some cases the code sequence assigned to a (finite) message sequence may conceivably depend not only on this sequence but on some subsequent letters, too. In all practical cases, however, the encoding is "of finite delay", i. e. the code sequence assigned to the first n letters of the message to be encoded is uniquely determined by these letters and m subsequent ones, where m is fixed. Then one may consider that the code sequence obtained in this way is actually assigned to the first $n+m$ letters of the message sequence (rather to the first n cases); as m is constant, this change of viewpoint does not cause any change in the results.

It would be very desirable, both from the theoretical and practical points of view, to extend our results in order to include the case of information transmission in the presence of noise. E. g., our results may conceivably be useful in the theory of channels with error synchronisation, investigated by Dobrusin [8]. A closer study of noisy channels with arbitrary cost scales, however, is beyond the scope of the present paper.

In connection with the "noiseless coding theorem" (theorem 3. 5), we did not tacle the problem whether the lower bound $\frac{H(\mathfrak{X} \| \mathfrak{Z})}{C}$ of L can be attained (or approximated to any specified degree) by an appropriate encoding. In practically important cases, this question may be answered in the affirmative using familiar methods, though in the most general case there may arise some difficulties. Another problem we did not enter is that of generalizing McMillan's theorem for sources with cost scales; this problem, though of considerable interest, apparently requires different methods than those used in this paper.

REFERENCES

[1] R. ASH: Information Theory: New York, Interscience Publishers 1965.

[2] P. BILLINGSLEY: On the coding theorem for the noiseless channel. Ann. Math. Statist. 32, 594-601 (1961)

[3] N. M. BLACHMAN: Minimum-cost encoding of information. IRE Trans. Information Theory PGIT. 3, 139-149 (1954)

[4] ---: Minimum-cost transmission of information. Information and Control 7, 508-511 (1964)

[5] E. L. BLOH: Generalization of an inequality of information theory to the case of symbols of inequal duration. (In Russian) Problemy Peredaci Informacii 5, 95-99 (1960)

[6] ---: Construction of optimal code constituted of elementary symbols of inequal duration (in Russian). Problemy Peredaci Informacii 5, 100-111 (1960)

[7] I. CSISZÁR: Two remarks on noiseless coding. Information and Control. To appear.

[8] R. L. DOBRUSIN: Capacity of channels with error synchronisation. Paper presented at the Colloquium of Information Theory, Debrecen, 1967.

[9] A. FEINSTEIN: Foundations of Information Theory. New York: McGraw-Hill Book Co. 1958.

[10] G. KATONA and G. TUSNÁDY: The principle of conservation of entropy in a noiseless channel. Studia Sci. Math. Acad. Sci. Hung. 2, 29-35 (1966)

[11] J. R. KINNEY: Singular functions associated with Markov chains. Proc. Amer. Math. Soc. 9, 603-608 (1958)

[12] R. M. KRAUSE: Channels which transmit letters of inequal duration. Information and Control 5, 13-24 (1962)

[13] Ju. I. LJUVIC: Remark on the capacity of the discrete noiseless channel (in Russian). Uspehi Mat. Nauk. 17, 191-198 (1962)

[14] W. PARRY: Intrinsic Markov chains. Trans. Amer. Math. Soc. 112, 55-66 (1964)

[15] C. E. RADKE: Necessary and sufficient conditions on conditional probabilities to maximize entropy. Information and Control 9, 279-284 (1966)

[16] C. E. SHANNON: A mathematical theory of communication. Bell System Techn. J. 27, 379-432, 623-656 (1948)

[17] V. M. SIDEL'NIKOV: On statistical properties of transformations induced by finite automata (in Russian). Kibernetika (Kiev) 6, 1-14 (1965)

[18] R. A. ZAIDMAN: On the assymptotics of certain sequences encountered in problems of non-Markov random walks and of information theory (in Russian). Vestnik Leningrad. Univ. 1, 23-33 (1965).