Z. Wahrscheinlichkeitstheorie verw. Geb. 12, 185 – 222 (1969)

Information Sources with Different Cost Scales and the Principle of Conservation of Entropy

I. Csiszár, G. Katona and G. Tusnády *

Summary. The aim of this paper is to provide a mathematically rigorous and sufficiently general treatment of the basic information-theoretic problems concerning sources with symbols of different costs and noiseless coding in a general sense. The main new concepts defined in this paper are the entropy rate (entropy per unit cost) of a source with respect to a stochastic cost scale and the encoding (in particular decodable encoding) of a source in a general sense. On the basis of these concepts, we prove some general theorems on the relation of entropy rates with respect to different cost scales and on the effect of encoding to the entropy rate. In particular, the "principle of conservation of entropy" and the "noiseless coding theorem" are proved under very general conditions.

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

* The main results of this research have been reported at the Second Congress of Bulgarian Mathematicians, Varna, August 1967, at the International Symposium on Information Theory, San Remo, September 1967 and at the Colloquium on Information Theory, Debrecen, September 1967.

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) H/L where H is the entropy (per symbol) of the original message and L 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 Sidel'nikov [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.

In Section 1, some notational conventions are introduced and, for the sake of later reference, some simple lemmas are established (without any claim of novelty).

In Section 2, we define entropy rate with respect to an arbitrary cost scale (Definition 2.2) and we prove the entropy rate comparison theorem (Theorem 2.2). This theorem can be considered as a general solution of problem (i), as one sees most clearly from its specialisations (Theorems 2.3 and 2.4). It is also shown, how the entropy rate with respect to an arbitrary cost scale is bounded by the channel capacity (Theorem 2.5).

In Section 3, we define encoding (and, in particular, decodable encoding) in a general sense (Definitions 3.1, 3.2 and 3.3). Applying a strengthened version of the entropy rate comparison theorem, we prove a general form of the principle of conservation of entropy (Theorem 3.3); this result and, in particular, its specialisation Theorem 3.4 provides a solution of problem (iii). We also prove a general version of the noiseless coding theorem (see problem (ii)) even for such cases where the principle of conservation of entropy does not hold (Theorem 3.5).

Both Sections 2 and 3 contains several examples; some of them are illustrations of our general concepts and results showing their relation to familiar ones, while the others show the limitations of eventual further generalizations.

Section 4 contains some comments and unsolved problems.

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

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

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

$$H(\xi|\eta) = \sum_y P(\eta = y) H_{\{\eta = y\}}(\xi) = -\sum_{x,y} P(\eta = y) P(\xi = x|\eta = y) \log_2 P(\xi = x|\eta = y) \tag{1.2}$$

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

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

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

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

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

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

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

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*

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

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

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

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

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*

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

and

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

Proof of Lemma 1.1. The triangle inequality (1.9) follows from

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

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

$$|H(\xi_1|\eta)-H(\xi_2|\eta)|\leqq d(\xi_1,\xi_2), \quad |H(\xi|\eta_1)-H(\xi|\eta_2)|\leqq d(\eta_1,\eta_2) \tag{1.16}$$

whence (1.11) directly follows:

$$\begin{aligned}|H(\xi_1|\eta_1)-H(\xi_2|\eta_2)|&\leqq|H(\xi_1|\eta_1)-H(\xi_2|\eta_1)|+|H(\xi_2|\eta_1)-H(\xi_2|\eta_2)|\\&\leqq d(\xi_1,\xi_2)+d(\eta_1,\eta_2);\end{aligned}$$

at last, from (1.4), (1.5) and the second inequality of (1.16) we get

$$\begin{aligned}|I(\xi_1,\eta_1)-I(\xi_2,\eta_2)| &= |H(\xi_1)-H(\eta_2)-H(\xi_1|\eta_1)+H(\eta_2|\xi_2)|\\ &= |H(\xi_1|\eta_2)-H(\eta_2|\xi_1)-H(\xi_1|\eta_1)+H(\eta_2|\xi_2)|\\ &\leqq d(\eta_1,\eta_2)+d(\xi_1,\xi_2)\end{aligned}$$

i.e. (1.12).

Proof of Lemma 1.2. Set $p_k=P(\xi=k)$ and $q_k=\frac{1}{3}2^{-|k|}$ $(k=0,\pm1,\dots)$. Then $\{q_k\}$ is a probability distribution and the well-known inequality

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

gives rise to (1.13).

(1.14) can be proved in the same way, with the choice

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

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

Lemma 1.3. *Let $\xi(t)$, $t\geqq0$ 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) *$E|\xi(t)|\to E|\xi|<\infty$ $(t\to+\infty)$*

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*

$$\overline{\lim_{t\to\infty}} \int_{|\xi(t)|\geqq K} |\xi(t)|\,P(d\omega)\to 0 \qquad as\ K\to+\infty.$$

Remark 1.2. If there exists $t_0\geqq0$ such that for every finite $t_1>t_0$ the RV's $\xi(t)$, $t_0\leqq t\leqq t_1$ are u.i. then, obviously, condition (a) is equivalent to saying that $\xi(t)$ is u.i. for $t\geqq t_0$.

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

$$E|\xi(t)|-E|\xi|=\int_{A(t,K)}(|\xi(t)|-|\xi|)\,P(d\omega)+\int_{\bar A(t,K)}|\xi(t)|\,P(d\omega)-\int_{\bar A(t,K)}|\xi|\,P(d\omega)$$

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

$$\overline{\lim_{t\to\infty}}\int_{|\xi(t)|\geqq 2K}|\xi(t)|\,P(d\omega)\leqq\overline{\lim_{t\to\infty}}\int_{\bar A(t,K)}|\xi|\,P(d\omega)=\int_{|\xi|\geqq K}|\xi|\,P(d\omega)$$

(we used that $\{\omega\colon|\xi(t)|\geqq 2K\}\subset\bar A(t,K)$) 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 $\mathfrak{U}(X)$ and $\tilde{\mathfrak{U}}(X)$ 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 $\mathfrak{U}(X)$. The set X will be called an alphabet iff identical sequences from $\mathfrak{U}(X) \cup \tilde{\mathfrak{U}}(X)$ 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 $\mathfrak{U}(X)$ and $\tilde{\mathfrak{U}}(X)$ are disjoint. This condition means only that the elements of X are really "elementary", excluding e.g. the possibility of $x_1=a$, $x_2=b$, $x_3=a\,b$ or $x_1=a$, $x_2=b\,c$, $x_3=a\,b$, $x_4=c$ etc.

For each alphabet X, $\mathfrak{U}(X)$ is a free semigroup with respect to juxtaposition; the zero element is the void sequence u_0. Observe, however, that our concept of alphabet means somewhat more than this; e.g. for $x_1=0$, $x_2=0\,1$, $x_3=1\,1$ the set $X=\{x_1, x_2, x_3\}$ is not an alphabet in our sense (as e.g. $x_1 x_3 x_3 x_3 \dots = x_2 x_3 x_3 \dots$), though $\mathfrak{U}(X)$ is obviously a free semigroup.

The elements of an alphabet will be referred to as letters. If X is an alphabet and $u = x_{i_1} \dots x_{i_m} \in \mathfrak{U}(X)$ then m, the length[1] of the sequence u, is uniquely determined by u; it will be denoted by $\|u\|$. Of course, we set $\|u_0\|=0$ and for $u \in \tilde{\mathfrak{U}}(X)$ we set $\|u\| = +\infty$. For $u, v \in \mathfrak{U}(X) \cup \tilde{\mathfrak{U}}(X)$ we shall write $u \prec v$ iff $\|u\| \leqq \|v\|$ and the sequence of the first $m = \|u\|$ letters of v is identical with u. Obviously, $\prec$ is a partial order on $\mathfrak{U}(X) \cup \tilde{\mathfrak{U}}(X)$. Subsets of $\tilde{\mathfrak{U}}(X)$ of form $C(u) = \{\tilde{u}\colon \tilde{u} \succ u\}$ $(u \in \mathfrak{U}(X))$ will be referred to as cylinder sets; the smallest σ-algebra of subsets of $\tilde{\mathfrak{U}}(X)$ containing all cylinder sets will be denoted by $\mathfrak{B}$.

(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

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

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

$$\xi(k; n) = \begin{cases} \xi(k) \dots \xi(n) & \text{if } k \leqq n \\ \text{the void sequence } u_0 & \text{if } k > n \end{cases}$$

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

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

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

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

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

$\eta(t)$ $(0 \leqq t < +\infty)$ $\eta(t) = \xi(1; \nu(t))$

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

[1] In this paper, the term "length" will be used also in another sense, cf. Example 2.2. The notation $\|u\|$, however, will always mean the number of letters in u, while other types of length will be denoted by $l(u)$.

Processes $\mathfrak{Z}$ will be thought of as associated to processes $\mathfrak{X}$ (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 nondecreasing sequence of nonnegative RV's $\tau(n)$ $(n=1, 2, \ldots)$ with $\tau(n) \xrightarrow{\text{a.s.}} +\infty$ defines a sequence $\mathfrak{Z}$ and so does each family of IRV's $\nu(t) \geqq 0$ $(0 \leqq t < +\infty)$ 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 $\nu(t)/t$ is u.i. for $t \to \infty$ (or, equivalently[2], for $t \geqq t_0 > 0$).

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 $\tau(n)$ representes 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.

Remark 2.1. A source $\mathfrak{X}$ can be interpreted also as a measurable mapping of $(\Omega, \mathfrak{F})$ into $(\tilde{\mathfrak{U}}(X), \mathfrak{B})$; it maps also the probability measure P defined on $\mathfrak{F}$ into a probability measure $P_{\mathfrak{X}}$ defined on $\mathfrak{B}$. Often a source is defined as the triple $(\tilde{\mathfrak{U}}(X), \mathfrak{B}, P_{\mathfrak{X}})$ or its analogue with doubly infinite sequences, as e.g. in [9]. This definition has, however, the shortcoming that it does not leave room for RV's not uniquely determined by the message sequence. Thus, for our purposes, the mentioned definition would be adequate only in case of "intrinsic" cost scales, cf. Example 2.2 below.

Example 2.1. The simplest cost scale is defined by

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

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

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

$$\zeta(n)=l(\xi(1;n))-l(\xi(1;n-1)), \qquad \tau(n)=l(\xi(1;n)) \qquad (n=1, 2, \ldots) \tag{2.2}$$

[2] As the sample functions of $\nu(t)$ are non-decreasing, $\nu(t)/t$ is u.i. in every finite interval (t_0, t_1) $(0 < t_0 < t_1)$, provided that $E\,\nu(t_1) < +\infty$. Hence follows at once that $\nu(t)/t$ is u.i. for $t \to \infty$ iff it is u.i. for $t \geqq t_0$; cf. Remark 1.2.

defines a cost scale $\mathfrak{Z}$. Cost scales of this kind will be called *strictly intrinsic* (in general, we shall say that $\mathfrak{Z}$ is an *intrinsic* cost scale for $\mathfrak{X}$ iff each $\zeta(n)$ is uniquely determined by the infinite message sequence $\xi(1)\,\xi(2)\ldots$). In particular, if

$$l(u)=\sum_{j=1}^{n} l(x_{i_j}) \qquad (u=x_{i_1}\ldots x_{i_n}), \tag{2.3}$$

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

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

may be called a *memoryless* intrinsic cost scale; observe, that $l(x)\equiv 1$ gives rise to the counting scale $\mathfrak{C}$. E.g. $l(x)$ may be the length or duration of the symbol $x\in X$. Then, if the symbols are emitted consecutively, without intervals,

$$\tau(n)=\sum_{k=1}^{n} l(\xi(k))$$

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 external 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 2.2. A cost scale $\mathfrak{Z}$ is trivially regular if $\nu(t)/t$ is uniformly bounded; e.g. a strictly intrinsic cost scale (cf. Example 2.2) is surely regular if $l(u)/\|u\|$ is bounded away from 0 $(u\neq u_0)$. For cost scales $\mathfrak{Z}$ with the property that there exists a RV γ such that $\nu(t)/t \xrightarrow{P} \gamma$, the necessary and sufficient condition of regularity consists in

$$\frac{E\,\nu(t)}{t}\to E\,\gamma<\infty$$

on account of Lemma 1.3 (observe that if $\mathfrak{Z}$ is regular, i.e. $\nu(t)/t$ is u.i. for $t\to\infty$ then necessarily $E\,\gamma<\infty$).

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 $\mathfrak{U}(X)$, 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 with respect to the cost scale $\mathfrak{Z}$ is the limit

$$H(\mathfrak{X}\|\mathfrak{Z})=\lim_{t\to\infty}\frac{1}{t}H\big(\eta(t)\big) \tag{2.5}$$

provided that it exists. If the limit does not exist, we shall denote the limsup and liminf of $(1/t)\,H\big(\eta(t)\big)$ 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. $\mathfrak{Z}=\mathfrak{C}$ (cf. Example 2.1) then $\eta(t)=\xi(1;[t])$ and Definition 2.2 reduces to the usual definition of entropy per symbol

$$H(\mathfrak{X})=\lim_{n\to\infty}\frac{1}{n}H\big(\xi(1;n)\big)=H(\mathfrak{X}\|\mathfrak{C}). \tag{2.6}$$

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 $\mathfrak{Z}$. For certain purposes also the other alternative, i.e. that both $\mathfrak{X}$ and $\mathfrak{Z}$ are information-carrying, may be of interest; this, however, would involve considering the pair $\xi\big(1,\nu(t)\big)$, $\zeta\big(1;\nu(t)\big)$ $\big($with $\zeta\big(1;\nu(t)\big)=\zeta(1),\ldots,\zeta\big(\nu(t)\big)\big)$ rather than $\xi\big(1;\nu(t)\big)$ itself, leading to entropies of possibly continuous distributions, a problem we do not want to tackle in this paper. Of course, if $\mathfrak{Z}$ is a strictly intrinsic cost scale (cf. Example 2.2) there is no difference between the two approaches.

Adopting the viewpoint that all relevant information is carried by the process $\mathfrak{X}$, the quantity

$$H^*(\mathfrak{X}\|\mathfrak{Z})=\lim_{t\to\infty}\frac{1}{t}H\big(\eta(t)\,|\,\nu(t)\big) \tag{2.7}$$

might seem a better measure of entropy rate than (2.5). E.g. if $\mathfrak{X}$ is a trivial source emitting a sequence of identical symbols and thus producing no information at all, the quantity (2.5) may happen to be positive while (2.6) obviously equals 0. However, in view of the simple relation

$$H(\mathfrak{X}\|\mathfrak{Z})=H^*(\mathfrak{X}\|\mathfrak{Z})+\lim_{t\to\infty}\frac{1}{t}H\big(\nu(t)\big) \tag{2.8}$$

between H and H^* $\big($which is an immediate consequence of $H(\eta)=H(\eta,\nu)=H(\eta\,|\,\nu)+H(\nu)\big)$, we loose nothing by adopting as a definition of entropy rate the more convenient (2.5) rather than (2.7). Moreover, in practically interesting cases we always have $H\big(\nu(t)\big)=o(t)$ (see Lemma 2.1 below) thus actually $H(\mathfrak{X}\|\mathfrak{Z})=H^*(\mathfrak{X}\|\mathfrak{Z})$.

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

Lemma 2.1. *If for some $a>0$, the a'th moment of $\nu(t)$ exists and it is $\exp\{o(t)\}$ then*

$$H\big(\nu(t)\big)=o(t) \qquad (t\to\infty). \tag{2.9}$$

In particular, (2.9) *holds for every regular cost scale* $\mathfrak{Z}$.

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

$$d(v_1, v_2) \leqq 2E\,|v_1 - v_2| + 2\log_2 3; \tag{2.10}$$

thus if

$$\frac{v_1(t) - v_2(t)}{t} \xrightarrow{L_1} 0,$$

then (2.9) *holds or does not hold simultaneously for* $\mathfrak{Z}_1$ *and* $\mathfrak{Z}_2$.

Proof. As

$$E\log_2(v+1) = \frac{1}{a} E\log_2(v+1)^a \leqq \frac{1}{a}\log_2 E(v+1)^a = o(t)$$

if $E\,v^a = \exp\{o(t)\}$, the first statement is an immediate consequence of Lemma 1.2; if $\mathfrak{Z}$ is regular i.e. if $v(t)/t$ is u.i. for $t \geqq t_0$ then $E\,v(t) = O(t)$ and all the more $E\,v(t) = \exp\{o(t)\}$. Furthermore, as

$$d(v_1, v_2) = H(v_1 | v_2) + H(v_2 | v_1) = H(v_1 - v_2 | v_2) + H(v_1 - v_2 | v_1) \leqq 2H(v_1 - v_2),$$

the inequality (2.10) is a consequence of Lemma 1.2. The last assertion follows from (2.10) and Lemma 1.1.

Example 2.5. Let the possible values of the costs $\zeta(n)$ be 0 and 1; for a binary sequence u of n digits set

$$P(\zeta(1), \dots, \zeta(n) = u) = \begin{cases} 2^{-n} & \text{if the first digit of } u \text{ is } 1 \\ \dfrac{a}{k\log^2 k}\, 2^{-(n-k)} & \text{if the first } k-1 \text{ digits of } u \text{ are } 0 \\ & \text{and the } k\text{'th digit of } u \text{ is } 1 \ (2 \leqq k \leqq n) \\ \displaystyle\sum_{k=n+1}^{\infty} \frac{a}{k\log^2 k} & \text{if } u \text{ consists of } n \text{ zeros} \end{cases}$$

where

$$a = \frac{1}{2}\left(\sum_{k=2}^{\infty} \frac{1}{k\log^2 k}\right)^{-1}.$$

Clearly, the joint distributions of the $\zeta(n)$'s have been defined in a consistent way. It is easily seen that in the sequence $\zeta(1), \zeta(2), \dots$ a 1 appears a.s. and under the condition that $\zeta(n)$ is the first 1, the RV's $\zeta(n+1), \zeta(n+2), \dots$ are independent and take on their values with probabilities $\frac{1}{2} - \frac{1}{2}$; hence, in particular,

$$\frac{\tau(n)}{n} \xrightarrow{\text{a.s.}} \frac{1}{2}.$$

This cost scale $\mathfrak{Z}$ is not regular. As we have $v(t) = k + [t]$ iff there are exactly $[t]$ 1's in the cost sequence $\zeta(1), \dots, \zeta(k + [t])$ and if $\zeta(k + [t] + 1) = 1$, we have (if $k \geqq 1$)

$$P(v(t) = k + [t]) \geqq P(\xi(1) = \dots = \xi(k) = 0;\ \xi(k+1) = \dots = \xi(k + [t] + 1) = 1)$$

$$= \frac{a}{(k+1)\log^2(k+1)}\, 2^{-[t]},$$

thus $H(v(t))=\infty$ for all $t\geqq 0$. This pathological property remains unaffected if we replace the zero costs by small positive ones (depending on n), provided that they tend to zero rapidly enough as $n\to\infty$.

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[3] (2.4), when interpreting $l(x)$ 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 $H(\mathfrak{X})/L$ is implicitely replaced by our $H(\mathfrak{X}\|\mathfrak{Z})$. Our results will, in particular, provide a justification for such reasonings under general conditions and show their limitations, too (cf. Theorem 2.4 and Remark 2.5).

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

$$\begin{aligned} &|r|^+=\max(0,r), && |r|^-=-\min(0,r), && |r|=|r|^++|r|^- \\ &|r|_K^+=\min(|r|^+,K), && |r|_K^-=\min(|r|^-,K), && |r|_K=\min(|r|,K). \end{aligned} \tag{2.11}$$

The following estimates will play a fundamental role in the sequel.

Theorem 2.1. *Let $\mathfrak{X}$ be a source with alphabet X of size s, and let $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ be two different cost scales for $\mathfrak{X}$. Then*[4]

$$H(\eta_1|\eta_2)\leqq H(v_1|v_2)+\log_2 s\cdot E|v_1-v_2|^+ \tag{2.12}$$

and also, if $A\in\mathfrak{F}$ is such that on $\bar{A}=\Omega\setminus A$ we have $v_1-v_2\leqq Kt$ (where $K>0$ is arbitrary)

$$H(\eta_1|\eta_2)\leqq 1+P(\bar{A})\,H_{\bar{A}}(v_1|v_2)+\log_2 s\cdot E|v_1-v_2|_{Kt}+P(A)\,H_A(\eta_1). \tag{2.13}$$

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

$$H(\eta_1|\eta_2)=H(v_1,\eta_1|v_2,\eta_2)\leqq H(v_1|v_2)+H(\eta_1|v_1,v_2,\eta). \tag{2.14}$$

As for given v_1, v_2 and $\eta_2=\xi(1;v_2)$ the number of possible "values" of $\eta_1=\xi(1;v_1)$ is at most $s^{|v_1-v_2|^+}$, the last term in (2.14) is $\leqq E(\log_2 s^{|v_1-v_2|^+})=\log_2 s\cdot E|v_1-v_2|^+$, proving (2.12). We may also write (setting $\alpha=1$ if $\omega\in A$ and $\alpha=0$ otherwise),

$$H(\eta_1|\eta_2)\leqq H(\alpha,\eta_1|\eta_2)=H(\alpha)+P(\bar{A})\,H_{\bar{A}}(\eta_1|\eta_2)+P(A)\,H_A(\eta_1|\eta_2);$$

hence, applying (2.12) to $H_{\bar{A}}(\eta_1|\eta_2)$ and taking into account

$$P(\bar{A})\,E_{\bar{A}}\,|v_1-v_2|^+\leqq E\,|v_1-v_2|^+_{Kt} \tag{2.15}$$

and using the obvious inequalities $H(\alpha)\leqq 1$, $H_A(\eta_1|\eta_2)\leqq H_A(\eta_1)$ we obtain (2.13).

[3] More general cost scales do not seem to have been considered in the literature.

[4] Actually, the assumption that v_i equals $v_i(t)$ corresponding to $\mathfrak{Z}_i$ $(i=1,2)$ is nowhere used in the proof; thus the estimates (2.12) and (2.13) hold for arbitrary IRV's v_1 and v_2 and $\eta_i=\xi(1;v_i)$ $(i=1,2)$.

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 $c\,\zeta(n)$ $(n=1,2,\ldots)$. Let us denote the τ's and v's corresponding to the cost scale $c\,\mathfrak{Z}$ by τ^c and v^c, respectively:

$$\tau^c(n)=c\,\tau(n),\quad n=1,2,\ldots;\qquad v^c(t)=v\left(\frac{t}{c}\right),\quad 0\leqq t<\infty.$$

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

$$\mathfrak{Z}_1\prec\mathfrak{Z}_2 \quad\text{iff}\quad \frac{1}{t}|v_2(t)-v_1(t)|^+\xrightarrow{L_1}0, \tag{2.16}$$

$$\mathfrak{Z}_1\sim\mathfrak{Z}_2 \quad\text{iff}\quad \frac{1}{t}(v_1(t)-v_2(t))\xrightarrow{L_1}0. \tag{2.17}$$

If $\mathfrak{Z}_1\prec\mathfrak{Z}_2$ or $\mathfrak{Z}_1\sim\mathfrak{Z}_2$, we say that $\mathfrak{Z}_1$ *preceeds* $\mathfrak{Z}_2$ or $\mathfrak{Z}_1$ is *equivalent* to $\mathfrak{Z}_2$, respectively.

If in (2.16) and (2.17) instead of L_1-convergence only convergence in probability is required, we shall say that $\mathfrak{Z}_1$ weakly preceeds $\mathfrak{Z}_2$ or $\mathfrak{Z}_1$ is weakly equivalent to $\mathfrak{Z}_2$:

$$\mathfrak{Z}_1\overset{w}{\prec}\mathfrak{Z}_2 \quad\text{iff}\quad \frac{1}{t}|v_2(t)-v_1(t)|^+\xrightarrow{P}0, \tag{2.16'}$$

$$\mathfrak{Z}_1\overset{w}{\sim}\mathfrak{Z}_2 \quad\text{iff}\quad \frac{1}{t}(v_1(t)-v_2(t))\xrightarrow{P}0. \tag{2.17'}$$

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

$$\frac{1}{t}(v_1(t)-v_2^c(t))=\frac{1}{t}\left(v_1(t)-v_2\left(\frac{t}{c}\right)\right)\xrightarrow{L_1}0 \tag{2.18}$$

or

$$\frac{1}{t}\left(v_1(t)-v_2\left(\frac{t}{c}\right)\right)\xrightarrow{P}0. \tag{2.18'}$$

Of course, in case of regular cost scales, the replacement of L_1-convergence by stochastic convergence in the above definitions makes no difference. Thus the "weak" concepts defined above provide a real generalisation only for non-regular cost scales.

Intuitively, $\mathfrak{Z}_1\prec\mathfrak{Z}_2$ means that "in general" $\zeta_1(n)\leqq\zeta_2(n)$ i.e. the "cost" of one symbol is for $\mathfrak{Z}_1$ "smaller" than it is for $\mathfrak{Z}_2$. $\mathfrak{Z}_1\sim\mathfrak{Z}_2$ means that the cost of one symbol is essentially the same for both scales. We think that these loose statements do have some heuristic value but the reader may prefer to disregard them completely.

Clearly, $\prec$ and $\overset{w}{\prec}$ are partial orders and $\sim$ and $\overset{w}{\sim}$ are equivalence relations for cost scales. $c<1$ and $c>1$ imply $c\,\mathfrak{Z}\prec\mathfrak{Z}$ and $\mathfrak{Z}\prec c\,\mathfrak{Z}$, respectively. Quasi-equivalence, too, is an equivalence relation; 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 (weakly) quasi-equivalent cost scales, there obviously holds

$$c_{12}\,c_{23}=c_{13}, \qquad c_{21}=\frac{1}{c_{12}}. \tag{2.19}$$

Theorem 2.2. *Let $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ be two cost scales for a source $\mathfrak{X}$ with finite alphabet X, and let the entropy rates $H(\mathfrak{X}\|\mathfrak{Z}_1)$ and $H(\mathfrak{X}\|\mathfrak{Z}_2)$ exist.*

(a) *If $\mathfrak{Z}_1 \prec \mathfrak{Z}_2$ and $\frac{1}{t}H(\nu_2(t))\to 0$ $(t\to\infty)$ then*

$$H(\mathfrak{X}\|\mathfrak{Z}_1)\geqq H(\mathfrak{X}\|\mathfrak{Z}_2). \tag{2.20}$$

(b) *$\mathfrak{Z}_1\sim\mathfrak{Z}_2$ implies*

$$H(\mathfrak{X}\|\mathfrak{Z}_1)= H(\mathfrak{X}\|\mathfrak{Z}_2). \tag{2.21}$$

(c) *If $\mathfrak{Z}_1\prec c\,\mathfrak{Z}_2$ or $\mathfrak{Z}_1\sim c\,\mathfrak{Z}_2$, the right hand sides of* (2.20) *or* (2.21), *respectively, should be divided by c.*

In particular, if $\mathfrak{Z}_1$ is quasi-equivalent to $\mathfrak{Z}_2$ with quotient $c_{12}=c>0$ then

$$H(\mathfrak{X}\|\mathfrak{Z}_1)=\frac{1}{c}\,H(\mathfrak{X}\|\mathfrak{Z}_2). \tag{2.22}$$

If the entropy rates in question do not exist, all assertions remain true both for the lower and upper entropy rates. In particular, for equivalent (or quasi-equivalent) cost scales, if the entropy rate with respect to either of them exists, it exists also with respect to the other and (2.21) *(or* (2.22)*) holds.*

Proof. The assertions follow easily from Theorem 2.1. In fact, as

$$H(\eta_2)\leqq H(\eta_1,\eta_2)=H(\eta_1)+H(\eta_2|\eta_1),$$

and as $H(\nu_2|\nu_1)\leqq H(\nu_2)$, the estimate (2.12) (interchanging the role of the indices 1 and 2; recall that ν_i and η_i are abbreviations for $\nu_i(t)$ and $\eta_i(t)$, $i=1,2$) and (2.16) imply $H(\eta_2)\leqq H(\eta_1)+H(\nu_2)+o(t)$ $(t\to\infty)$. This, by Definition 2.2 and the assumption $(1/t)\,H(\nu_2)\to 0$ means just (2.20). Furthermore, (2.12) implies also

$$d(\eta_1,\eta_2)\leqq d(\nu_1,\nu_2)+\log_2 s\cdot E\,|\nu_1-\nu_2|. \tag{2.23}$$

Hence, using (1.10) and (2.10), we obtain

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

thus $\mathfrak{Z}_1\sim\mathfrak{Z}_2$ implies, in fact, (2.21). To prove (c) we have only to observe that for any cost scale $\mathfrak{Z}$ and any $c>0$

$$H(\mathfrak{X}\|c\,\mathfrak{Z})=\lim_{t\to\infty}\frac{H(\eta^c(t))}{t}=\lim_{t'\to\infty}\frac{H(\eta(t'))}{c\,t'}=\frac{1}{c}\,H(\mathfrak{X}\|\mathfrak{Z}) \tag{2.24}$$

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

The last statement of Theorem 2.2 is obvious from the proof of assertions (a)–(c).

Remark 2.3. The intuitive meaning of Theorem 2.2 is clear. In certain cases the conditions of the theorem may be weakened, replacing the relations $\prec$ and $\sim$ by their "weak" analogues. This is the case when some additional conditions ensure that the conditional entropy figuring in the last term of (2.13) is $O(t)$; then (2.13) may be used instead of (2.12) to yield the desired results. We shall encounter such situations in Section 3. Let us also remark, that if in (a) we drop the condition $(1/t)\,H(\nu_2(t))\to 0$, $\mathfrak{Z}_1 \prec \mathfrak{Z}_2$ still implies $H(\mathfrak{X}\|\mathfrak{Z}_1)\geqq H^*(\mathfrak{X}\|\mathfrak{Z}_2)$ (cf. (2.8)), and that assertion (b) holds also for H^* instead of H (cf. (2.8), (2.10) and (1.10)).

Just in view of the "intuitively obvious" character of Theorem 2.2, it is instructive to point out that e.g. $\mathfrak{Z}_1 \overset{w}{\sim} \mathfrak{Z}_2$ is not sufficient, in general, for (2.21) to hold. This follows also from Example 2.5, but the following example seems more suitable.

Example 2.6. Let $\mathfrak{X}$ be a source with alphabet $X=\{0,1,2\}$; let the joint distribution of the $\xi(n)$'s be defined by

$$P(\xi(1;n)=u)=\begin{cases} \dfrac{1}{k(k+1)}\,2^{-(n-k)}\cdot 2^{-(n-l)} & \text{if } u \text{ contains } l \text{ 2's} \qquad (l\geqq 1),\\ & \text{the first at the } k\text{'th place } (1\leqq k\leqq n-l+1)\\ \dfrac{1}{n+1}\,2^{-n} & \text{if } u \text{ consists of 0's and 1's.}\end{cases}$$

Let $l(0)=l(1)=0$, $l(2)=2$, $\zeta(n)=l(\xi(n))$. Then, in the same way as in Example 2.5, we have

$$\frac{\tau(n)}{n}\xrightarrow{\text{a.s.}} 1,$$

and

$$P\left(\nu(t)=k+\left[\frac{t}{2}\right]\right)\geqq P\left(\xi(i)\neq 2,\ i=1,\dots,k;\ \xi(k+1)=\dots=\xi\left(k+\left[\frac{t}{2}\right]+1\right)=1\right)$$

$$=\frac{1}{(k+1)(k+2)}\,2^{-[t/2]},$$

whence we see that $\mathfrak{Z}$ is not regular (actually, $E\,\nu(t)=+\infty, t\geqq 0$). Furthermore, given $\nu(t)=k+[t/2]$, the k letters of $\eta(t)=\xi(1;\nu(t))$, different from 2, may be 0 and 1 independently of each other and with probabilities $\frac{1}{2}-\frac{1}{2}$, thus

$$H_{(\nu(t)=k+[t/2])}(\eta(t))\geqq k,$$

implying $H(\eta(t)|\nu(t))=+\infty$ for all $t\geqq 0$, i.e. $H^*(\mathfrak{X}\|\mathfrak{Z})=+\infty$, and all the more $H(\mathfrak{X}\|\mathfrak{Z})=+\infty$. Observe, that our $\mathfrak{Z}$ is a memoryless intrinsic cost scale weakly equivalent to $\mathfrak{C}$, while (2.21) does not hold for $\mathfrak{Z}_1=\mathfrak{Z}$, $\mathfrak{Z}_2=\mathfrak{C}$. Of course, the zero costs $\zeta(n)=0$ may be replaced by small positive ones, tending rapidly enough to 0 as $n\to\infty$ without any essential change.

In order to apply Theorem 2.2 to concrete problems it will be convenient to establish some simple sufficient or necessary and sufficient conditions of the relations $\prec$, $\sim$ for different cost scales.

Lemma 2.2. *Let $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ be two arbitrary cost scales.*

(a) *Let*

$$\overline{\lim_{t\to\infty}}\, P\left(\frac{v_2(t)}{t} \geqq K\right) \to 0 \qquad \text{for } K\to\infty\,.$$

Then the relations

$$\left|\frac{v_1(t)}{v_2(t)}-1\right|^{-} \xrightarrow{P} 0 \quad \text{and} \quad \frac{v_1(t)}{v_2(t)} \xrightarrow{P} 1$$

are sufficient conditions for $\mathfrak{Z}_1 \overset{w}{\prec} \mathfrak{Z}_2$ and $\mathfrak{Z}_1 \overset{w}{\sim} \mathfrak{Z}_2$, respectively, and if $v_2(t)/t$ is bounded away from 0, *they are necessary, too.*

(b) *$\mathfrak{Z}_1 \prec \mathfrak{Z}_2$ or $\mathfrak{Z}_1 \sim \mathfrak{Z}_2$ holds iff $\mathfrak{Z}_1 \overset{w}{\prec} \mathfrak{Z}_2$ or $\mathfrak{Z}_1 \overset{w}{\sim} \mathfrak{Z}_2$ and, in addition,*

$$\left|\frac{v_2(t)-v_1(t)}{t}\right|^{+} \quad \text{or} \quad \frac{v_1(t)-v_2(t)}{t}$$

is u.i. for $t\to\infty$, respectively. The last conditions are certainly fulfilled if $\mathfrak{Z}_2$ resp. both $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ are regular.

(c) *If $\mathfrak{Z}_2$ is regular and*

$$\frac{v_1(t)}{v_2(r\,t)} \xrightarrow{P} 1, \qquad (r>0),$$

then $\mathfrak{Z}_1 \overset{w}{\sim} c\,\mathfrak{Z}_2$ where $c=1/r$, and also $\mathfrak{Z}_1 \prec c\,\mathfrak{Z}_2$; if $\mathfrak{Z}_1$, too, is regular, then actually $\mathfrak{Z}_1 \overset{w}{\sim} c\,\mathfrak{Z}_2$.

Remark 2.4. If $\mathfrak{Z}_2$ is regular then

$$\overline{\lim_{t\to\infty}}\, P\left(\frac{v_2(t)}{t} \geqq K\right) \to 0 \qquad (K\to\infty)$$

surely holds. If the cost scale $\mathfrak{Z}_2$ is such that $\zeta_2(n) \geqq b > 0$ a.s. ($n=1, 2, \ldots$), then

$$\frac{v_2(t)}{t} \leqq \frac{1}{b},$$

i.e., in this case $v_2(t)/t$ is trivially regular; if $\zeta_2(n) \leqq B$ a.s. ($n=1, 2, \ldots$) then

$$\frac{v_2(t)}{t} \geqq \frac{1}{B}$$

thus in this case $v_2(t)/t$ is bounded away from 0. Observe that in respect of

$$\left|\frac{v_2(t)-v_1(t)}{t}\right|^{+} \quad \text{and} \quad \frac{v_1(t)-v_2(t)}{t}$$

"u.i. for $t\to\infty$" does not necessarily imply "u.i. for $t \geqq t_0$" as these RV's may not be u.i. in finite intervals.

Proof of Lemma 2.2. From

$$\frac{v_1 - v_2}{t} = \frac{v_2}{t}\left(\frac{v_1}{v_2} - 1\right)$$

immediately follows that the relations

$$\left|\frac{v_1(t)}{v_2(t)} - 1\right|^{-} \xrightarrow{P} 0 \quad \text{and} \quad \frac{v_1(t)}{v_2(t)} \xrightarrow{P} 1$$

imply (2.16′) and (2.17′), respectively, provided that

$$\overline{\lim_{t\to\infty}}\, P\left(\frac{v_2(t)}{t} \geqq K\right) \to 0$$

for $K \to \infty$ and that if $v_2(t)/t$ is bounded away from 0, also the converse implications are true, proving assertion (a). The statement (b) is a direct consequence of Definition 2.3 and of Lemma 1.3. At last, (c) follows immediately from (a) and (b), applying them to $v_2^c(t) = v_2(t/c) = v_2(r\,t)$ instead of $v_2(t)$, and using the first assertion of Remark 2.4.

Lemma 2.3. *Let $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ be two cost scales and let one of them have the property*

$$b \leqq \zeta(n) \leqq B \qquad a.s.\ (0 < b < B;\ n = 1, 2, \ldots). \tag{2.25}$$

Then each of the three conditions

$$\frac{v_1(t)}{v_2(r\,t)} \xrightarrow{P} 1; \tag{2.26}$$

$$\frac{\tau_2(v_1(t))}{t} \xrightarrow{P} r > 0; \tag{2.27}$$

$$\frac{\tau_1(v_2(t))}{t} \xrightarrow{P} c > 0 \tag{2.28}$$

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

Proof. According to (2.18′), $\mathfrak{Z}_1 \overset{w}{\sim} (1/r)\,\mathfrak{Z}_2$ means

$$\frac{v_1(t) - v_2(r\,t)}{t} \xrightarrow{P} 0. \tag{2.29}$$

Without any loss of generality, let e.g. $\mathfrak{Z}_2$ have the property (2.25), i.e. $0 < b \leqq \zeta_2(n) \leqq B$; then

$$\left[\frac{r\,t}{B}\right] \leqq v_2(r\,t) \leqq \frac{r\,t}{b}$$

a.s., thus the equivalence of (2.26) and (2.29) is obvious. Furthermore, as by the definition of the τ's and v's the relation $\tau_2(v_1(t)) \leqq y\,t$ is equivalent to $v_1(t) \leqq v_2(y\,t)$ $(0 < y < \infty)$, the relation (2.27) is equivalent to

$$P(v_1(t) \leqq v_2(y\,t)) \to \begin{cases} 0 & \text{if } y < r \\ 1 & \text{if } y > r \end{cases} \qquad (t \to \infty)$$

and this, in turn, is equivalent to (2.29), in view of the assumption $0<b\leqq\zeta_2(n)\leqq B$. Similarly, (2.28) is equivalent to

$$P\big(v_2(t)\leqq v_1(y\,t)\big)\to\begin{cases}0 & \text{if } y<c\\ 1 & \text{if } y>c\end{cases}\qquad (t\to\infty)$$

i.e. to

$$P\big(v_2(y'\,t)\leqq v_1(t)\big)\to\begin{cases}0 & \text{if } y'=\dfrac{1}{y}>\dfrac{1}{c}=r\\ 1 & \text{if } y'<r\end{cases}\qquad (t\to\infty)$$

which, again, is equivalent to (2.29).

The following consequence of Theorem 2.2 and Lemmas 2.2 and 2.3 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 cost scales for $\mathfrak{X}$ such that*

(a) *one of them has the property* (2.25)

(b) *also the other is regular.*

Then either of the three equivalent conditions (2.26)–(2.28) *implies*

$$H(\mathfrak{X}\|\mathfrak{Z}_1)=r\,H(\mathfrak{X}\|\mathfrak{Z}_2)\tag{2.30}$$

(in the sense that if either side exists so does also the other and they are equal). If the condition (b) *is dropped, the relations* (2.26)–(2.28) *still imply*

$$(-1)^i\big(H(\mathfrak{X}\|\mathfrak{Z}_1)-r\,H(\mathfrak{X}\|\mathfrak{Z}_2)\big)\geqq 0\tag{2.31}$$

if $\mathfrak{Z}_i$ has the property (2.25), *provided that both entropy rates exist.*

If the entropy rates in question do not exist, (2.30) *(or* (2.31)*) still holds both for the lower and upper entropy rates.*

The most important cost scales are those quasi-equivalent to the counting scale $\mathfrak{C}$ (cf. Example 2.1). By Definition 1.3, a cost scale $\mathfrak{Z}$ is quasi-equivalent (weakly quasi-equivalent) to the counting scale $\mathfrak{C}$ iff there exists a constant $c>0$ such that

$$\frac{v(t)}{t}\xrightarrow{L_1}\frac{1}{c}\qquad\left(\text{or}\ \ \frac{v(t)}{t}\xrightarrow{P}\frac{1}{c}\right).$$

Of course, all cost scales quasi-equivalent to the counting scale are regular and thus for such cost scales always $H(\mathfrak{X}\|\mathfrak{Z})=H^*(\mathfrak{X}\|\mathfrak{Z})$ (cf. Lemma 2.1).

If for a given cost scale $(1/t)\,v(t)$ converges in probability to a (finite) constant r, this r will be called the symbol rate[5] of the source with respect to the given cost scale.

Similarly, if $(1/n)\,\tau(n)$ converges in probability to a (finite) constant c, this c will be called the average symbol cost (with respect to the given cost scale).

[5] The term is really suitable only for the case where "cost" is time, when r represents the average number of symbols per second; for the sake of brevity, however, we adopt it for the general case, too.

In view of Lemma 2.3 (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 ($c=1/r$) and these are necessary and sufficient conditions of $\mathfrak{Z}\overset{w}{\sim}c\,\mathfrak{C}$, too. In order that also $\mathfrak{Z}\sim c\,\mathfrak{C}$ hold i.e. that $\mathfrak{Z}$ be quasi-equivalent to $\mathfrak{C}$, the regularity of $\mathfrak{Z}$, or, equivalently, the relation

$$\frac{E\,\nu(t)}{t}\to r$$

(cf. Remark 2.2) is necessary and sufficient. 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 cost scale $\mathfrak{Z}$ a positive symbol rate r or, equivalently, a positive symbol cost c exists ($c\,r=1$), then*[6],

$$H(\mathfrak{X}\|\mathfrak{Z})\geqq r\,H(\mathfrak{X})=\frac{1}{c}\,H(\mathfrak{X}); \tag{2.32}$$

if $\mathfrak{Z}$ is regular, i.e., if $\nu(t)/t$ is u.i. for $t\geqq t_0$ or, equivalently, if also

$$\frac{E\,\nu(t)}{t}\to r$$

holds, we have the equality

$$H(\mathfrak{X}\|\mathfrak{Z})=r\,H(\mathfrak{X})=\frac{1}{c}\,H(\mathfrak{X}) \tag{2.33}$$

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

Remark 2.5. If $\mathfrak{Z}$ is a cost scale of type (2.4) (cf. footnote 3) and the source $\mathfrak{X}$ is such that

$$\frac{1}{n}\sum_{k=1}^{n} l(\xi(k))\xrightarrow{P} L>0 \tag{2.34}$$

(e.g. if $\mathfrak{X}$ is a stationary ergodic source, i.e. if $\xi(1)$, $\xi(2)$, ... is a stationary ergodic sequence of DRV's), "entropy per unit cost" is often defined as the ratio $H(\mathfrak{X})/L$. By Theorem 2.4, this definition is equivalent to (2.5), provided that

$$\frac{E\,\nu(t)}{t}\to\frac{1}{L}\qquad (t\to\infty).$$

This is the case, in particular, if $l(x)>0$ 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). As Example 2.6 shows, the strict inequality in (2.32) can obtain even for memoryless intrinsic cost scales, if there are no restrictions on the source $\mathfrak{X}$. For stationary ergodic sources this is probably impossible; actually, it seems most likely that any stationary ergodic sequence $\zeta(1)$, $\zeta(2)$, ... defines a regular cost scale, but we did not succeed in proving this. Observe that

[6] If the entropy rates do not exist, the result holds both for the lower and upper entropy rates.

for more general cost scales (even for strictly intrinsic ones) the pathological phenomenon of Example 2.6 can occur also for strictly positive $\zeta(n)$'s, as indicated in the same example.

Example 2.7. Let $\mathfrak{X}$ be an arbitrary source with finite entropy per symbol $H(\mathfrak{X})=H$ (i.e., H is the entropy rate with respect to the counting scale $\mathfrak{C}$). Consider an arbitrary cost scale $\mathfrak{Z}$ which is stochastically independent of the process $\mathfrak{X}$. Then

$$H(\eta)-H(\nu)=H(\eta\,|\,\nu)=H\big(\xi(1;\nu)\,|\,\nu\big)=\sum_{n=0}^{\infty} P(\nu=n)\, H_{\{\nu=n\}}\big(\xi(1;\nu)\big)$$
$$=\sum_{n=0}^{\infty} P(\nu=n)\, H\big(\xi(1;n)\big),$$

thus, using the assumption $H\big(\xi(1;n)\big)=n\big(H+o(1)\big)$ $(n\to\infty)$ and the fact $\nu(t)\xrightarrow{\text{a.s.}}+\infty$ we obtain

$$H\big(\eta(t)\big)=H\big(\nu(t)\big)+E\,\nu(t)\big(H+o(1)\big) \qquad (t\to\infty).$$

This means that in this special case $\dfrac{E\,\nu(t)}{t}\to r<\infty$ (when, according to Lemma 2.1, $H(\nu(t))=o(t)$) implies $H(\mathfrak{X}\,\|\,\mathfrak{Z})=r\,H(\mathfrak{X})$, and also that if $(1/t)\,E\,\nu(t)\to+\infty$ then $H(\mathfrak{X}\,\|\,\mathfrak{Z})=+\infty$.

Of course, the relation

$$\frac{E\,\nu(t)}{t}\to r$$

is, in general, by no means sufficient for $H(\mathfrak{X}\,\|\,\mathfrak{Z})=r\,H(\mathfrak{X})$, as e.g. the following example shows.

Example 2.8. Let $\mathfrak{X}$ be an arbitrary source (with finite alphabet) with regular cost scale $\mathfrak{Z}$, such that

$$\frac{\tau(n)}{n}\xrightarrow{P}\gamma \qquad (t\to\infty)$$

where γ is a DRV taking on the (positive) values $c_1, c_2, \ldots, c_q$ with probabilities $p_1, p_2, \ldots, p_q$. We have, using (1.1), (1.2) and (1.4)

$$H\big(\eta(t)\big)=H\big(\eta(t)\,|\,\gamma\big)+I\big(\eta(t),\gamma\big)=\sum_{i=1}^{q} p_i\, H_{\{\gamma=c_i\}}\big(\eta(t)\big)+I\big(\eta(t),\gamma\big). \tag{2.35}$$

Assume now that

$$H_i=\lim_{n\to\infty}\frac{1}{n}\, H_{\{\gamma=c_i\}}\big(\xi(1;n)\big)$$

exist for $i=1,\ldots,r$. Then Theorem 2.4 applies for the conditional probability measures[7] $P_i(\cdot)=P(\cdot\,|\,\gamma=c_i)$ $(i=1,\ldots,q)$, yielding

$$\lim_{t\to\infty}\frac{1}{t}\, H_{\{\gamma=c_i\}}\big(\eta(t)\big)=\frac{H_i}{c_i} \qquad (i=1,\ldots,q),$$

[7] If $\mathfrak{Z}$ is regular i.e. $\nu(t)/t$ is u.i. for $t\geqq t_0$, it is u.i. also with respect to the conditional probability measures $P_i(\cdot)$.

whence

$$H(\mathfrak{X}\|\mathfrak{Z})=\lim_{t\to\infty}\frac{1}{t}H(\eta(t))=\sum_{i=1}^{q}p_i\frac{H_i}{c_i},$$

as, in (2.35), the term $I(\eta(t),\gamma)$ is bounded by $H(\gamma)$. By the same argument

$$H(\mathfrak{X})=H(\mathfrak{X}\|\mathfrak{C})=\lim_{n\to\infty}\frac{1}{n}H(\xi(1;n))=\sum_{i=1}^{q}p_i H_i.$$

Let us denote by $H_\gamma(\mathfrak{X}\|\mathfrak{Z})$ and $H_\gamma(\mathfrak{X})$ the random variables taking on the values

$$\lim_{t\to\infty}\frac{1}{t}H_{\{\gamma=c_i\}}(\eta(t))=\frac{H_i}{c_i}$$

and H_i, respectively, if $\gamma=c_i$. Then our result may be written as

$$H(\mathfrak{X}\|\mathfrak{Z})=E\left(\frac{1}{\gamma}H_\gamma(\mathfrak{X})\right),\qquad H(\mathfrak{X})=E(\gamma H_\gamma(\mathfrak{X}\|\mathfrak{Z})). \tag{2.36}$$

Hence we see, in particular, that the relation

$$\frac{E\,\nu(t)}{t}\to r\qquad\left(\text{in our case } r=E\,\frac{1}{\gamma}=\sum_{i=1}^{q}\frac{p_i}{c_i}\right)$$

in general does not imply $H(\mathfrak{X}\|\mathfrak{Z})=rH(\mathfrak{X})$; neither does

$$\frac{E\,\tau(n)}{n}\to c\qquad\left(\text{in our case, } c=E\gamma=\sum_{i=1}^{q}p_i c_i,\text{ provided that }\frac{\tau(n)}{n}\text{ is u.i.}\right)$$

imply

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

This fact is not due to the eventual inadequate definition of entropy rate; indeed, as $\mathfrak{Z}$ is regular, $H(\mathfrak{X}\|\mathfrak{Z})=H^*(\mathfrak{X}\|\mathfrak{Z})$.

Remark 2.6. If

$$\frac{\tau(n)}{n}\xrightarrow{P}\gamma$$

where γ is an arbitrary RV, the formulas (2.36) cannot be expected to remain true, in general. E.g. if $X=\{0,1\}$ and

$$\zeta(n)=\sum_{k=1}^{n-1}\frac{\xi(k)}{2^k}+n\frac{\xi(n)}{2^n}\quad\text{then}\quad\frac{\tau(n)}{n}=\sum_{k=1}^{n}\frac{\xi(k)}{2^k}\quad\text{and}\quad\gamma=\sum_{k=1}^{\infty}\frac{\xi(k)}{2^k}$$

contains full information on the whole message, thus both $H_\gamma(\mathfrak{X})$ and $H_\gamma(\mathfrak{X}\|\mathfrak{Z})$ are identically 0. It can be shown, under certain regularity conditions, that the first (second) equality in (2.36) holds iff $I(\eta(t), \gamma)=o(t)$ $\left(I(\xi(1; n), \gamma)=o(n)\right)$. Here we do not enter the problem of giving sufficient conditions for these last relations.

Example 2.9. 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; 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 $u=x_{i_1} x_{i_2} \ldots x_{i_n} \in \mathfrak{U}(X)$ is transmissible iff $x_{i_k} \in X(a_{i_{k-1}})$ $(k=1, \ldots, n)$ where a_{i_k} is defined recursively by $a_{i_k}=G(x_{i_k}, a_{i_{k-1}})$ $(k=1, \ldots, n)$; denote the set of all transmissible sequences by $\mathfrak{U}_0$. Let $l(x, a) \geqq 0$ $(a \in A,\ x \in X(a))$ be the cost of transmission of x at the state a; then

$$l(u)=\sum_{k=1}^{n} l(x_{i_k}, a_{i_{k-1}})$$

represents the cost of transmission of the sequence $u=x_{i_1} \ldots x_{i_n} \in \mathfrak{U}_0$. We make the usual assumptions that for each pair of states $a', a'' \in A$ there exists $u=x_{i_1} \ldots x_{i_k} \ldots x_{i_n}$ $u \in \mathfrak{U}_0$ such that $a_{i_k}=a'$, $a_{i_n}=a''$, and that for all $u \in \mathfrak{U}_0$ with $\|u\| \geqq m$ (say) $l(u)>0$. Let $\mathfrak{X}$ be a source transmissible by the channel (i.e. $\xi(1) \ldots \xi(n) \in \mathfrak{U}_0$ a.s. $n=1, 2, \ldots$) and let the (regular) cost scale $\mathfrak{Z}$ be defined by (2.2), with the present $l(u)$. Let $N(t)$ denote the number of different sequences $u \in \mathfrak{U}_0$ with $l(u) \leqq t$ and such that $l(u\,x)>t$ for some $x \in X$, with $u\,x \in \mathfrak{U}_0$; we define the channel capacity by

$$C=\varlimsup_{t \to \infty} \frac{\log_2 N(t)}{t}. \tag{2.37}$$

Then $\eta(t)=\xi(1; v(t))$ has at most $N(t)$ possible values, implying $H(\eta(t)) \leqq \log_2 N(t)$, $\bar{H}(\mathfrak{X} \| \mathfrak{Z}) \leqq C$; thus if

$$\frac{\tau(n)}{n}=\frac{1}{n} l(\xi(1; n)) \xrightarrow{P} L,$$

we have, according to (2.33),

$$\frac{1}{L} \bar{H}(\mathfrak{X}) \leqq C. \tag{2.38}$$

It may be shown that in (2.37) actually the limit exists; its value has been calculated by Shannon[8] [16]. He obtained $C=\log_2 w_0$, where w_0 is the greatest

[8] Shannon, apparently having commensurable symbol costs in mind, considered the number of sequences $u \in \mathfrak{U}_0$ such that exactly $l(u)=t$; the present interpretation of $N(t)$ is more adequate and it gives rise to the same system of difference equations as the original one. A rigorous proof of Shannon's capacity formula $\lim\limits_{t \to \infty} \dfrac{\log_2 N(t)}{t}=\log_2 w_0$ (with w_0 defined as above and with an unessential difference in the definition of $N(t)$) has been given by Ljubič [13]; he has pointed out, too, that with Shannon's original interpretation of $N(t)$, this relation holds only for the upper limit. For the case of $l(u)=\|u\|$ see also [14] (with another terminology) and for the case of no different states see also [12].

positive root of the equation

$$\mathrm{Det}\left[\sum_{k=1}^{s} w^{-l_{ij}^{k}} - d_{ij}\right] = 0; \tag{2.39}$$

here $l_{ij}^{k} = l(x_k, a_i)$ if $x_k \in X(a_i)$, $G(x_k, a_i) = a_j$ and 0 otherwise and $d_{ij} = 1$ if $i = j$ and 0 otherwise.

The inequality (2.38), 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 (2.38) concern stationary Markovian sources only[9], and they boil down to the formal verification of the inequality between the algebraic expressions representing the both sides. (A simple non-computational proof of this kind is given in [7]; a computational approach is presented in [15].)

As a matter of fact, (2.38) is a much weaker result than (2.33) and an inequality of this type can be proved under more general conditions. We conclude this section by a generalization of inequality (2.38) for entropy rates $H(\mathfrak{X} \| \mathfrak{Z})$ instead of $H(\mathfrak{X})$ and to the case that an "average symbol cost" L exists only in an expectation sense (observe that in the latter case, as we have seen in Example 2.8, no equality of type (2.33) holds).

Theorem 2.5. *Let $\mathfrak{Z}_2$ be a cost scale for the source $\mathfrak{X}$ such that*

$$\frac{\tau_2(n)}{n} \geqq b > 0 \qquad \textit{for } n \geqq n_0 \textit{ (say)}.$$

Let $N(t)$ denote the number of different possible values of $\eta_2(t) = \xi(1; \nu_2(t))$. Let $\mathfrak{Z}_1$ be another cost scale for $\mathfrak{X}$ such that $E\,\tau_2(\nu_1(t)) = O(t)$ $(t \to \infty)$. Then

$$H(\eta_1(t)) \leqq C\,E\,\tau_2(\nu_1(t)) + o(t) \qquad (t \to \infty) \tag{2.40}$$

with

$$C = \overline{\lim_{t \to \infty}} \frac{\log_2 N(t)}{t} \leqq \frac{\log_2 s}{b}. \tag{2.41}$$

In particular, if

$$L = \lim_{t \to \infty} \frac{E\,\tau_2(\nu_1(t))}{t} < \infty \tag{2.42}$$

exists, we have

$$\overline{H}(\mathfrak{X} \| \mathfrak{Z}_1) \leqq L\,C. \tag{2.43}$$

Proof. Let us define a DRV $\kappa(t)$ by

$$\kappa(t) = \frac{k}{K}\,t \quad \text{iff} \quad \frac{k-1}{K}\,t \leqq \tau_2(\nu_1(t)) < \frac{k}{K}\,t \qquad (k = 1, 2, \ldots) \tag{2.44}$$

[9] For the simplest case that there is only one state, when $l(x, a)$ reduces to $l(x)$ and $N(t)$ to the number of sequences $u \in \mathfrak{U}(X)$ with $t - B < l(u) \leqq t$ $(B = \max_{x \in X} l(x))$, (2.39) reduces to $\sum_{x \in X} w^{-l(x)} = 1$. This model has been considered in [5] and [12], and (2.38) has been proved for independent $\xi(1), \xi(2) \ldots$; in this particular case hence one easily obtains a proof for arbitrary $\mathfrak{X}$, too, but in the general case a transition from stationary Markovian sources to arbitrary ones does not seem easy.

where K is a fixed positive number. From (2.44) immediately follows

$$\nu_2\left(\kappa(t)-\frac{t}{K}\right)\leqq \nu_1(t)\leqq \nu_2(\kappa(t)); \tag{2.45}$$

in particular, if $\eta_2(\kappa(t))$ is given, $\eta_1(t)$ is uniquely determined by the value of $\nu_1(t)$, thus

$$H(\eta_1(t))\leqq H(\eta_1(t), \eta_2(\kappa(t))=H(\eta_2(\kappa(t))+H(\nu_1(t)|\eta_2(\kappa(t))). \tag{2.46}$$

Now, using (2.41) and (1.8′), we have for $t\geqq t_0(K)$ (say)

$$H(\eta_2(\kappa(t))|\kappa(t))\leqq E\log_2 N(\kappa(t))\leqq C\left(1+\frac{1}{K}\right)E\kappa(t), \tag{2.47}$$

whence, in view of

$$H(\kappa(t))=H\left(\frac{K}{t}\kappa(t)\right)\leqq E\left(\frac{K}{t}\kappa(t)\right)+\log_2 3$$

(cf. Lemma 1.2) we obtain

$$\begin{aligned} H(\eta_2(\kappa(t)))&\leqq H(\eta_2(\kappa(t)), \kappa(t))=H(\eta_2(\kappa(t))|\kappa(t))+H(\kappa(t))\\ &\leqq C\left(1+\frac{1}{K}\right)E\kappa(t)+\frac{K}{t}E\kappa(t)+\log_2 3 \qquad (t\geqq t_0(K)); \end{aligned} \tag{2.48}$$

furthermore, the assumptions $\dfrac{\tau_2(n)}{n}\geqq b>0$ $(n\geqq n_0)$ and $E\tau_2(\nu_1(t))=O(t)$ imply $E\nu_1(t)=O(t)$, thus, by Lemma 2.1,

$$H(\nu_1(t))=o(t). \tag{2.49}$$

As, according to (2.44), $\kappa(t)\leqq\tau_2(\nu_1(t))+(t/K)$, (2.46), (2.48) and (2.49) immediately give rise to (2.40), using the assumption $E\tau_2(\nu_1(t))=O(t)$ and that $K>0$ was arbitrary.

The inequality (2.41) is an obvious consequence of the assumption

$$\frac{\tau_2(n)}{n}\geqq b>0, \quad \text{yielding} \quad \nu_2(t)\leqq\frac{t}{b}, \qquad N(t)\leqq\frac{t}{b}s^{t/b}.$$

(2.43) follows from (2.40) and (2.42) by Definition 2.2.

Example 2.10. Let $\mathfrak{U}_0$ be a subset of $\mathfrak{U}(X)$ such that $u\prec v\in\mathfrak{U}_0$ implies $u\in\mathfrak{U}_0$ and that for each $u\in\mathfrak{U}_0$ there exists $x\in X$ with $u\,x\in\mathfrak{U}_0$. Let further $l(u)$ be a non-negative-valued function on $\mathfrak{U}_0$ such that $u\prec v\in\mathfrak{U}_0$ implies

$$l(u)\leqq l(v) \quad \text{and} \quad \frac{l(u)}{\|u\|}\geqq b>0 \quad \text{if} \quad \|u\|\geqq n_0.$$

Let $N(t)$ be the number of different sequences $u \in \mathfrak{U}_0$ such that $l(u) \leqq t$ and $l(u\,x) > t$ for some $x \in X$ with $u\,x \in \mathfrak{U}_0$. One may consider the pair $(\mathfrak{U}_0, l(\cdot))$ as a noiseless channel in a general sense, and

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

may be interpreted as its capacity. Theorem 2.5 implies that if a source $\mathfrak{X}$ is transmissible by the channel (i.e. if $\xi(1; n) \in \mathfrak{U}_0$ a.s., $n = 1, 2, \ldots$) and if $\mathfrak{Z}$ is a cost scale for $\mathfrak{X}$ satisfying $E\,l(\eta(t)) = O(t)$ then

$$H(\eta(t)) \leqq C\,E\,l(\eta(t)) + o(t) \tag{2.50}$$

and if

$$\frac{E\,l(\eta(t))}{t} \to L > 0$$

then

$$\frac{\bar{H}(\mathfrak{X} \| \mathfrak{Z})}{L} \leqq C \leqq \frac{\log_2 s}{b}. \tag{2.51}$$

Of course, the finite-state channels (cf. Example 2.9) deserve the main interest; even for that case (and even for the case of one state only, cf. footnote 9), (2.51) provides a considerable generalization of (2.38).

Remark 2.7. The assumption

$$\frac{\tau_2(n)}{n} \geqq b > 0 \qquad (n \geqq n_0)$$

in Theorem 2.5 has been made in order to exclude the trivial case $C = \infty$ and to ensure $H(\nu_1(t)) = o(t)$. In the second respect, what really needed is only

$$H(\nu_1(t) | \eta_2(\kappa(t))) = o(t)$$

(cf. (2.46)); this relation trivially holds if $\mathfrak{Z}_1$ is any strictly intrinsic cost scale, as then (cf. (2.45)) $\nu_1(t)$ is uniquely determined by $\eta_2(\kappa(t))$. Also, if instead of

$$\frac{\tau_2(n)}{n} \geqq b > 0 \quad \text{we assume only} \quad \frac{\tau_2(n)}{\log_2 n} \geqq b(n) \quad \text{with } b(n) \to \infty \text{ as } n \to \infty,$$

the relation $E\,\tau_2(\nu_1(t)) = O(t)$ still implies $E \log_2 \nu_1(t) = o(t)$ thus $H(\nu_1(t)) = o(t)$ (cf. Lemma 1.2). For quite arbitrary $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$, however, even if

$$L = \lim_{t \to \infty} \frac{E\,\tau_2(\nu_1(t))}{t}$$

exists, (2.40) and (2.43) may not hold in pathological cases. E.g. if $\mathfrak{X}$ is a trivial source such that $\xi(n) = \xi(1)$ a.s. ($n = 1, 2, \ldots$) and $\mathfrak{Z}_1$ is the cost scale of Example 2.5 then $H(\mathfrak{X} \| \mathfrak{Z}_1) = \infty$ while $C = 0$ for any strongly deterministic $\mathfrak{Z}_2$; in particular, $\mathfrak{Z}_2$ can be chosen in such a way that $E\,\tau_2(\nu_1(t)) = o(t)$ thus $L = 0$ (e.g. $\zeta_2(n) = 1$ if $n = 2^{2^k}$ and 0 otherwise).

§ 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

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

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

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

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

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) to be denoted by

$$\bar{H}(\mathfrak{X}, \mathfrak{X}' \| \mathfrak{Z}, \mathfrak{Z}'), \quad \underline{H}(\mathfrak{X}, \mathfrak{X}' \| \mathfrak{Z}, \mathfrak{Z}'), \quad \bar{H}(\mathfrak{X} | \mathfrak{X}' \| \mathfrak{Z}, \mathfrak{Z}'), \quad \underline{H}(\mathfrak{X} | \mathfrak{X}' \| \mathfrak{Z}, \mathfrak{Z}'), \quad \text{etc.}$$

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, remain 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).

We wish to consider encodings in "the most general sense" that can "reasonably" be given to this term. The definition of a "general" code has to conform with the intuitive requirement that the encoding proceeds "from letter to letter" (in practice, the encoding often proceeds "from block to block"; blockwise encoding, however, may be described by assigning to each but the last letter of any block the void sequence u_0 while the code of the whole block is assigned to the last letter of the block). Thus, an arbitrary (measurable) mapping of $\tilde{\mathfrak{U}}(X)$ into $\tilde{\mathfrak{U}}(X')$ (cf. Section 1, (D)) will not be considered as a code. A possible way of defining a "real" code is to specify for each letter $x \in X$, the rule how a "code word" is assigned to x, in dependence, in general, on the previously encoded letters. In this sense, a code is a mapping g of $X \times \mathfrak{U}(X)$ into $\mathfrak{U}(X')$ assigning to $x \in X$ the (conceivably void) "code word" $g(x|u)$ if the sequence of previously encoded letters is $u \in \mathfrak{U}(X)$. Another possible way is to define a code as a mapping of $\mathfrak{U}(X)$ into $\mathfrak{U}(X')$ which is monotone with respect to the partial order $\prec$. In fact, writing for $u = x_{i_1} \dots x_{i_n}$

$$f(u) = g(x_{i_1} | u_0)\, g(x_{i_2} | x_{i_1}) \dots g(x_{i_1} | x_{i_1} \dots x_{i_{n-1}}), \tag{3.5}$$

the functions f and g uniquely determine each other.

We prefer to adopt the following

Definition 3.1. Let X and X' be finite alphabets. An arbitrary mapping f of a subset D_f of $\mathfrak{U}(X)$ into $\mathfrak{U}(X')$ will be called a *semicode* from X to X' (D_f will be called the domain of the semicode f). A semicode with domain $\mathfrak{U}(X)$ such that $f(u_0)=u_0$ and

$$u \prec v \quad \text{implies} \quad f(u) \prec f(v) \tag{3.6}$$

will be called a *code* from X to X'.

Each code f from X to X' defines a mapping of infinite sequences, too; in fact, for $\tilde{u}=x_{i_1} x_{i_2} \ldots \in \tilde{\mathfrak{U}}(X)$ we may write (cf. (3.5))

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

Observe that Definition 3.1 does not exclude that the "code" $f(\tilde{u})$ of some infinite sequence $\tilde{u}$ be finite. The set of those $\tilde{u} \in \tilde{\mathfrak{U}}(X)$ for which $f(\tilde{u})$ is infinite, will be denoted by $\tilde{D}_f$. More generally, we define for any semicode f

$$\tilde{D}_f=\{\tilde{u}\colon u^n \in D_f,\ f(u^n) \prec f(u^{n+1}),\ n=1,2,\ldots;\ \lim_{n\to\infty} \|f(u^n)\|=\infty\} \tag{3.7}$$

where u^n is the section of length n of $\tilde{u}$, i.e. $u^n=x_{i_1} \ldots x_{i_n}$ if $\tilde{u}=x_{i_1} x_{i_2} \ldots$.

Infinite sequences belonging to $\tilde{D}_f$ can be encoded even if f is only a semicode; in fact, $f(\tilde{u})$ ($\tilde{u} \in \tilde{D}_f$) may be defined in an obvious way. Of course, for an arbitrary semicode f the set $\tilde{D}_f$ may be conceivably void; if $\tilde{D}_f \neq \emptyset$, the restriction of f to $D_f^0=\{u\colon u \in D_f,\ u \prec \tilde{u}$ for some $\tilde{u} \in \tilde{D}_f\}$ certainly satisfies (3.6); if we extend f from D_f^0 to $\mathfrak{U}(X)$ by setting $f^0(u)=f(u^*)$ where u^* is the longest section of u belonging to D_f^0 (or, if no section of u belongs to D_f^0 then $f^0(u)=u_0$) we obtain a code that gives rise to the same $\tilde{D}_f$ and the same mapping of $\tilde{D}_f$ into $\tilde{\mathfrak{U}}(X)$ as the original f.

The number of sequences $\tilde{u} \in \tilde{\mathfrak{U}}(X)$ encodable by a semicode f may be increased if we prescribe a sequence of "block ends". More exactly, for a sequence $\mathfrak{k}$ of integers $0 \leqq k_1 < k_2 < \cdots$ we set

$$\tilde{D}_f^{\mathfrak{k}}=\{\tilde{u}\colon u^{k_n} \in D_f,\ f(u^{k_n}) \prec f(u^{k_{n+1}}),\ n=1,2,\ldots;\ \lim_{n\to\infty} \|f(u^{k_n})\|=\infty\}; \tag{3.8}$$

sequences $\tilde{u} \in \tilde{D}_f^{\mathfrak{k}}$ may be encoded by f, namely as

$$f^{\mathfrak{k}}(\tilde{u})=\lim_{k\to\infty} f(u^{k_n}) \in \tilde{\mathfrak{U}}(X') \qquad (\tilde{u} \in \tilde{D}_f^{\mathfrak{k}}), \tag{3.9}$$

the limit being understood with respect to the partial order $\prec$.

If $\mathfrak{k}$ is the sequence of all positive integers then $D_f^{\mathfrak{k}}=\tilde{D}_f$ for any semicode f; if f is a code than $\tilde{D}_f^{\mathfrak{k}}=\tilde{D}_f$ and $f^{\mathfrak{k}}(\tilde{u})=f(\tilde{u})$ ($\tilde{u} \in \tilde{D}_f$) for any sequence $\mathfrak{k}$. Observe, however, that if f is only a semicode then some infinite sequences $\tilde{u} \in \tilde{\mathfrak{U}}(X)$ may belong to several $\tilde{D}_f^{\mathfrak{k}}$ but $f^{\mathfrak{k}}(\tilde{u})$ may depend on the particular "block end sequence" $\mathfrak{k}$.

Remark 3.1. It is easy to see that $\tilde{D}_f^{\mathfrak{k}} \in \mathfrak{B}$ for any semicode f and any sequence $\mathfrak{k}$; also, the mapping $f^{\mathfrak{k}}\colon=\tilde{D}_f^{\mathfrak{k}} \to \tilde{\mathfrak{U}}(X')$ 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 $f\colon \tilde{D}_f \to \tilde{\mathfrak{U}}(X)$ may be called an *infinite-code*; clearly the concept of an infinite-code is much more restrictive then that of an arbitrary measurable mapping $\tilde{\mathfrak{U}}(X) \to \tilde{\mathfrak{U}}(X')$ (though all practically realizable mappings seem to belong to this class). Similarly if f is a semicode, we may call the system of the mappings $f^{\mathfrak{t}}\colon \tilde{D}_f^{\mathfrak{t}} \to \tilde{\mathfrak{U}}(X)$ an infinite-semicode. Two different codes (semicodes) f_1 and f_2 may give rise to the same infinite-codes (infinite-semicodes). 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 (infinite-semicode) may be identified with the corresponding equivalence class of codes (semicodes) and vice-versa.

Remark 3.2. A different way of defining codes in a general sense has been suggested by Billingsley [2]. His approach is based on infinite sequences and encoding of finite sequences does not enter to his definition. Instead of this, a continuity condition is imposed in order to bring the general definition closer to real possibilities; it may be seen that each code in the sense of Billingsley is an infinite-code in our sense (cf. Remark 3.1), with $\tilde{D}_f = \tilde{\mathfrak{U}}(X)$.

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

$$0,\ \|f(x_{i_1})\|, \dots, \|f(x_{i_1} \dots x_{i_m})\|, \dots, \|f(x_{i_1} \dots x_{i_p})\| \tag{3.10}$$

and $k_0 = 0$. For $\tilde{u} \in \tilde{\mathfrak{U}}(X)$ 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 $n = 0, 1, 2, \dots$.

Each $u \in \mathfrak{U}(X)$ may be partitioned into blocks $x_{i_j} \dots x_{i_l}$ with $j = k_{n-1} + 1$, $l = k_n$ and, conceivably, an "incomplete" block $x_{i_j} \dots x_{i_p}$, where $j - 1$ is the last point of increase of the sequence (3.10). The n'th block $x_{i_j} \dots x_{i_l}$ $(j = k_{n-1} + 1,\ l = k_n)$ is encoded by the nonvoid sequence $g(x_{i_l} | x_{i_1} \dots x_{i_{l-1}})$ (cf. (3.5)), and the last incomplete block (if any) is not encoded (i.e., it is encoded by the void sequence u_0). This interpretation suggests to define the essential domain of f by

$$D_f^* = \{u = x_{i_1} \dots x_{i_p} \colon p = k_n(u) \text{ for some } n \geqq 1\} \cup \{u_0\}, \tag{3.11}$$

i.e. as the set of all $u \in \mathfrak{U}(X)$ consisting of complete blocks, including the void sequence u_0.

Example 3.1. If, in (3.5), $g(x|u) = g(x)$, where $g(x)$ is some fixed mapping of X into $\mathfrak{U}(X') \setminus \{u_0\}$, the resulting code will be called a *simple letter code*.

In this case we have

$$\begin{aligned} f(u) &= g(x_{i_1})\, g(x_{i_2}) \dots g(x_{i_n}) && (u = x_{i_1} \dots x_{i_n} \in \mathfrak{U}(X)) \\ f(\tilde{u}) &= g(x_{i_1})\, g(x_{i_2}) \dots && (\tilde{u} = x_{i_1} x_{i_2} \dots \in \tilde{D}_f = \tilde{\mathfrak{U}}(X)) \end{aligned} \tag{3.12}$$

and

$$D_f^* = \mathfrak{U}(X).$$

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 $\mathfrak{U}(X')$ and A, respectively.

Let an initial state $a_0 \in A$ be specified, set $f(u_0) = u_0$ and for $u = x_{i_1} \dots x_{i_n}$, $n > 0$ set

$$f(u) = F(x_{i_1}, a_0)\, F(x_{i_2}, a_1) \dots F(x_{i_n}, a_{n-1}) \tag{3.13}$$

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*[10] and f will be referred to as the code generated by this finite-state transducer. Observe that if we drop the condition that the set of states is finite then each code f can be generated in this way, setting $A = \mathfrak{U}(X)$, $a_0 = u_0$, and $F(x, a) = g(x|a)$ $G(x, a) = a\,x$ for $a \in \mathfrak{U}(X)$. A simple letter code (Example 3.1) may be thought of as a code generated by a trivial finite-state transducer, having only one state.

Definition 3.2. Let $\mathfrak{X}$ be a source with finite alphabet X and let f be a semicode from X to X'. We say that $\mathfrak{X}$ is encodable by f, with respect to a cost scale $\mathfrak{Z}_1$, if $\xi(1)\,\xi(2)\ldots \in \tilde{D}_f^{(\mathfrak{n})}$ a.s. where the (random) sequence $\mathfrak{n}$ consists of the values taken on by $\nu_1(t)$ $(0 \leqq t < +\infty)$. The encoding results in a new source $\mathfrak{X}' = f(\mathfrak{X})$ defined by $\xi'(1), \xi'(2), \ldots;$ $\xi'(1)\,\xi'(2)\ldots = f^{(\mathfrak{n})}(\xi(1)\,\xi(2)\ldots)$. We also define the mapped cost scale $\mathfrak{Z}_1' = f(\mathfrak{Z}_1)$ by

$$\nu_1'(t) = \| f(\eta_1(t)) \| \tag{3.14}$$

or, equivalently, by

$$\tau_1'(n) = t \quad \text{iff} \quad \| f(\eta_1(t')) \| < n \leqq \| f(\eta_1(t)) \| \qquad \text{for all } t' < t. \tag{3.15}$$

According to (3.8) and (3.9), if $\mathfrak{X}$ is encodable by f with respect to $\mathfrak{Z}_1$, all RV's $\xi'(k)$, $\nu_1'(t)$ and $\tau_1'(n)$ are well-defined a.s. and the latter ones do define a cost scale for $\mathfrak{X}'$. (Of course, both $f(\mathfrak{X})$ and $f(\mathfrak{Z}_1)$ depend, in general, both on $\mathfrak{X}$ and $\mathfrak{Z}_1$; we think, however, that our notation will not cause ambiguities.)

If f is not only a semicode but a code, when saying "$\mathfrak{X}$ is encodable by f" we need not mention the cost scale, as for codes $\tilde{D}_f^{(\mathfrak{t})} = \tilde{D}_f$ does not depend on $\mathfrak{t}$. As a point of fact, $\mathfrak{X}$ is encodable by a code f iff the IRV's

$$\kappa(n) = k_n(\xi(1)\,\xi(2)\ldots) \tag{3.16}$$

(k_n has been defined in connection with (3.10)) are well-defined a.s. for $n = 1, 2, \ldots$. Furthermore, if $\zeta(n) > 0$ a.s. $(n = 1, 2, \ldots)$, then, in Definition 3.2, $\mathfrak{n}$ is the sequence of all nonnegative integers and encodability does not depend on the cost scale, even for semicodes; in this case, in fact, one may restrict attention to encodings by codes. If, however, some $\zeta(n)$'s may be zero with positive probability, the sequence $\mathfrak{n}$ becomes really a random one and semicodes offer a greater generality then codes. For the following, it is essential to allow this possibility, too; in particular, for mapped cost scales of type (3.14), (3.15), $\nu'(t)$ has, as a rule, jumps greater than 1, i.e. for such cost scales there exist symbols of zero cost (cf. Remark 3.5 below).

If a source $\mathfrak{X}$ is encodable by a semicode f with respect to a cost scale $\mathfrak{Z}_1$, and we set $\mathfrak{X}' = f(\mathfrak{X})$, $\mathfrak{Z}_1' = f(\mathfrak{Z}_1)$. (3.14) implies $\eta_1'(t) = \xi'(1; \nu_1'(t)) = f(\eta_1(t))$ and thus $H(\eta_1'(t)) \leqq H(\eta_1(t))$, i.e.[11]

$$H(\mathfrak{X}' \| \mathfrak{Z}_1') \leqq H(\mathfrak{X} \| \mathfrak{Z}_1). \tag{3.17}$$

[10] Also the term "finite state automaton" has been used, but we prefer to avoid it, since the set of the possible outputs (in one tact) is no alphabet, in general.

[11] If the entropy rates in question do not exist, the inequality holds both for the lower and upper entropy rates.

If $\mathfrak{Z}'$ is a given cost scale for $\mathfrak{X}'$ and there exists $c'>0$ such that $\mathfrak{Z}_1'=f(\mathfrak{Z}_1)\prec c'\,\mathfrak{Z}'$ (cf. Definition 2.3) then, in view of Theorem 2.2, (3.17) implies[11]

$$H(\mathfrak{X}'\|\mathfrak{Z}')\leqq c'\,H(\mathfrak{X}\|\mathfrak{Z}_1) \tag{3.18}$$

provided that $H\big(v'(t)\big)=o(t)$ or else that $f(\mathfrak{Z}_1)\sim c'\,\mathfrak{Z}'$.

Remark 3.3. If a source $\mathfrak{X}$ is encodable by a semicode (or code) f with respect to a regular cost scale $\mathfrak{Z}_1$, 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 consists in the boundedness of $\|f(u)\|/\|u\|$ for $u\in D_f\,(u\neq u_0)$. In particular, for codes generated by finite-state transducers (cf. Example 3.2), this condition is trivially fulfilled.

Though the regularity of a cost scale $\mathfrak{Z}$ may get lost in the transition to $f(\mathfrak{Z})$, an important property of regular cost scales remains preserved, as the following partial sharpening of Theorem 2.2 shows.

Theorem 3.2. *Let $\mathfrak{Z}_1'$ and $\mathfrak{Z}_2'$ be two cost scales for a source $\mathfrak{X}'$; let there exist sources $\mathfrak{X}_i$, cost scales $\mathfrak{Z}_i$ and semicodes f_i $(i=1,2)$ such that $\mathfrak{X}_i$ is encodable by f_i with respect to $\mathfrak{Z}_i$ and $f_i(\mathfrak{X}_i)=\mathfrak{X}'$, $f_i(\mathfrak{Z}_i)=\mathfrak{Z}_i'$ $(i=1,2)$. Then, if $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ are regular and $\mathfrak{Z}_1'$ and $\mathfrak{Z}_2'$ are weakly quasi-equivalent, i.e. $\mathfrak{Z}_1'\overset{w}{\simeq} c'\,\mathfrak{Z}_2'$ for some $c'>0$ then*

$$H(\mathfrak{X}'\|\mathfrak{Z}_1')=\frac{1}{c'}\,H(\mathfrak{X}'\|\mathfrak{Z}_2') \tag{3.19}$$

in the sense that if either side exists so does also the other and they are equal; if the entropy rates do not exist, the equality holds both for the lower and upper entropy rates.

Proof. In view of (2.24) and the obvious relation $f_2(c'\,\mathfrak{Z}_2)=c'\,f_2(\mathfrak{Z}_2)$, we may assume $c'=1$, without any loss of generality. Applying Theorem 2.1 with $A=\{\omega\colon |v_1'-v_2'|>t\}$, we obtain (the arguments t being omitted)

$$H(\eta_1'|\eta_2')\leqq 1+P(\bar{A})\,H_{\bar{A}}(v_1'|v_2')+\log_2 s'\,E\,|v_1'-v_2'|_t^+ +P(A)\,H_A(\eta_1'), \tag{3.20}$$

where s' denotes the size of the alphabet X' of $\mathfrak{X}'$. Here

$$H_{\bar{A}}(v_1'|v_2')\leqq\log_2(2t+1) \tag{3.21}$$

as, if A obtains, the number of possible values of v_1' given v_2' is $\leqq 2t+1$. On account of

$$\mathfrak{Z}_1'\sim\mathfrak{Z}_2',\quad \text{i.e.}\quad \frac{v_1'-v_2'}{t}\xrightarrow{P}0,$$

we also have

$$E\,|v_1'-v_2'|_t^+=o(t). \tag{3.22}$$

To prove that also the last term of (3.20) is $o(t)$, let M be a fixed positive integer, and set $B_M=\{\omega\colon v_1\leqq M\,t\}$, $B_k=\{\omega\colon (k-1)\,t<v_1\leqq k\,t\}$, $k=M+1, M+2, \dots$. Then, if B_k obtains $(k=M, M+1, \dots)$, $\eta_1=\xi(1;v_1)$, and thus $\eta_1'=f_1(\eta_1)$, too, can take on at most

$$\sum_{i=0}^{[kt]} s_1^i<s_1^{kt+1}$$

different values, where s_1 is the size of the alphabet of the source $\mathfrak{X}_1$; thus, introducing the IRV β by setting $\beta=k$ if $\omega\in B_k$, we may write

$$
\begin{aligned}
H_A(\eta'_1)&\leqq H_A(\eta'_1,\beta)=H_A(\beta)+H_A(\eta'_1|\beta)=H_A(\beta)+\sum_{k=M}^{\infty}P(B_k|A)\,H_{A\cap B_k}(\eta'_1)\\
&\leqq H_A(\beta)+\log_2 s_1\sum_{k=M}^{\infty}P(B_k|A)(k\,t+1) \qquad (3.23)\\
&=H_A(\beta)+\log_2 s_1\left(1+P(B_M|A)\,M\,t+\sum_{k=M+1}^{\infty}P(B_M|A)\,k\,t\right)
\end{aligned}
$$

implying

$$P(A)\,H_A(\eta'_1)\leqq P(A)\,H_A(\beta)+P(A)\log_2 s_1(1+M\,t)+\log_2 s_1\int_{v_1>M\,t}v_1\,P(d\omega). \qquad (3.24)$$

Here, as $\mathfrak{Z}_1$ is regular and β is a function of v_1, $P(A)\,H_A(\beta)\leqq H(\beta)\leqq H(v_1)=o(t)$. By the assumption $\mathfrak{Z}'_1\overset{w}{\simeq}\mathfrak{Z}'_2$ we have $P(A)\to 0$ thus also the second term on the right hand side of (3.24) is $o(t)$ for any fixed M. Finally, as $\mathfrak{Z}_1$ is regular i.e. v_1/t is u.i. for $t\geqq t_0$, the last term divided by t can be made arbitrarily small for $t\geqq t_0$, if M is large enough. Hence we see that $P(A)\,H_A(\eta'_1)=o(t)$ for $t\to\infty$, that, together with (3.21) and (3.22), gives rise to

$$H(\eta'_1|\eta'_2)=o(t) \qquad (t\to\infty). \qquad (3.25)$$

We obtain, in the same way

$$H(\eta'_2|\eta'_1)=o(t) \qquad (t\to\infty), \qquad (3.26)$$

too, whence, by inequality (1.10) of Lemma 1.1,

$$H(\eta'_1)-H(\eta'_2)=o(t) \qquad (3.27)$$

completing the proof of Theorem 3.2.

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 $\mathfrak{X}$ be encodable by a semicode f from X to X', with respect to some cost scale $\mathfrak{Z}_1$. The encoding $\mathfrak{X}\to\mathfrak{X}'=f(\mathfrak{X})$ will be said to be *decodable* if there exists a semicode f' from X' to X such that $\mathfrak{X}'$ is encodable by f' with respect to $\mathfrak{Z}'_1=f(\mathfrak{Z}_1)$ yielding $f'(\mathfrak{X}')=\mathfrak{X}$ a.s. If, in addition, f' can be chosen in such a way that the cost scale $\mathfrak{Z}_2=f'(\mathfrak{Z}'_1)$ is (weakly) equivalent to $\mathfrak{Z}_1$, the encoding will be said to be (weakly) *properly decodable* (with respect to $\mathfrak{Z}_1$).

Remark 3.4. The last conditions seem, at first sight, a bit artificial; however, as we are going to see, only proper decodability (or its weak version) is adequate for our purposes. Observe that the conditions $f'(\mathfrak{Z}'_1)\sim\mathfrak{Z}_1$ and $f'(\mathfrak{Z}'_1)\overset{w}{\simeq}\mathfrak{Z}_1$ mean, in view of (3.14) and Definition 2.3,

$$\frac{1}{t}\left(\|f'(f(\eta_1(t)))\|-v_1(t)\right)\xrightarrow{L_1}0 \quad\text{and}\quad \frac{1}{t}\left(\|f'(f(\eta_1(t)))\|-v_1(t)\right)\xrightarrow{P}0,$$

respectively.

Example 3.3. Let f be a code from X to X' and assume that there exists a code f' from X' to X such that $\tilde{D}_{f'} \supset f(\tilde{D}_f)$ and $f'(f(\tilde{u})) = \tilde{u}$ for each $\tilde{u} \in \tilde{D}_f$. Such codes may be called *infinite-decodable* (in accordance with the concept of infinite-code, cf. Remark 3.2). Obviously, for any source $\mathfrak{X}$ encodable by such an f, the encoding $\mathfrak{X} \to f(\mathfrak{X})$ is decodable (but not necessarily properly decodable). E.g. if $f(u)$ is defined as the sequence of the first $[\frac{1}{2}\,\|u\|]$ letters of u, then f is infinite-decodable but not properly decodable.

Example 3.4. Let f be a code from X to X', and let D_f^* be the essential domain of f, see (3.11). The code f will be called *wide sense finite-decodable* if $u, v \in D_f^*$, $u \neq v$ implies $f(u) \neq f(v)$. For any source $\mathfrak{X}$ encodable by a wide sense finite-decodable code the encoding $\mathfrak{X} \to \mathfrak{X}' = f(\mathfrak{X})$ is decodable in the sense of Definition 3.3, with respect to any cost scale $\mathfrak{Z}_1$; an appropriate semicode f' from X' to X is defined by

$$D'_f = f(\mathfrak{U}(X)) = f(D_f^*), \qquad f'(u') = u \quad \text{iff} \quad f(u) = u', \qquad u \in D_f^*.$$

In order to exhibit a condition of proper decodability, let us introduce, for an arbitrary sequence $\mathfrak{K}$ of IRV's $0 = \kappa(0) \leqq \kappa(1) \leqq \kappa(2) \leqq \cdots$ and an arbitrary cost scale $\mathfrak{Z}$ the notation $\mathfrak{Z}|\mathfrak{K}$, where $\mathfrak{Z}^* = \mathfrak{Z}|\mathfrak{K}$ is defined by

$$\zeta^*(n) = \begin{cases} \sum\limits_{i=\kappa(k-1)+1}^{\kappa(k)} \zeta(i) & \text{if } n = \kappa(k) > \kappa(k-1) \text{ for some } k \geqq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.28}$$

Using this notation, letting $\mathfrak{K}$ be the sequence defined by (3.16), we may write

$$\mathfrak{Z}_2 = f'(\mathfrak{Z}'_1) = \mathfrak{Z}_1|\mathfrak{K}. \tag{3.29}$$

Thus the encoding by a wide-sense finite-decodable code f is properly decodable (weakly properly decodable) if $\mathfrak{Z}_1 \sim \mathfrak{Z}_1|\mathfrak{K}$ (or $\mathfrak{Z}_1 \overset{w}{\sim} \mathfrak{Z}_1|\mathfrak{K}$). In particular, if $\mathfrak{X}$ is encodable by f (i.e., the RV's $\kappa(n)$ are well-defined for $n = 1, 2, \ldots$) and $\mathfrak{Z}$ is an arbitrary cost scale, the encoding is surely properly decodable with respect to $\mathfrak{Z}_1 = \mathfrak{Z}|\mathfrak{K}$. Observe, too, that a simple sufficient condition of $\mathfrak{Z}_1 \sim \mathfrak{Z}_2 = \mathfrak{Z}_1|\mathfrak{K}$ consists in the uniform boundedness of the "block lengths" $\kappa(n) - \kappa(n-1)$ $(n = 1, 2, \ldots)$.

Example 3.5. Let the code f from X to X' have the property that $u \neq v$ implies $f(u) \neq f(v)$ for every $u, v \in \mathfrak{U}(X)$; such codes may be called *strict sense finite-decodable*. A code f is strict sense finite-decodable iff it is wide sense finite-decodable and, in addition, no $g(x_{i_k}|x_{i_1} \ldots x_{i_{k-1}})$ equals the void sequence u_0. For such codes $D_f^* = \mathfrak{U}(X)$, $\tilde{D}_f = \tilde{\mathfrak{U}}(X)$, and $\mathfrak{Z}_2 = f'(f(\mathfrak{Z}_1))$ is identical to $\mathfrak{Z}_1$, for arbitrary $\mathfrak{Z}_1$. This means, that each strict sense finite-decodable code defines a properly decodable encoding, for any source $\mathfrak{X}$ with alphabet X and with respect to any cost scale $\mathfrak{Z}$. For simple letter codes (cf. Example 3.1)) finite-decodability (in both senses) is identical to the usual concept of unique decipherability (see e.g. [9]).

Observe that strict sense finite-decodability does not imply infinite-decodability, as already the example of the simple letter code

$$X = \{0, 1, 2\}, \qquad X' = \{0, 1\}, \qquad g(0) = 0, \qquad g(1) = 01, \qquad g(2) = 11$$

shows. (E.g. for $\tilde{u}=022\cdots \neq \tilde{v}=122$ we have $f(\tilde{u})=f(\tilde{v})=01111\ldots$.) On the other hand, a code f may be infinite-decodable and give rise to a properly decodable encoding without being wide sense finite-decodable. A simple counterexample is the code defined by delating the last letter of each $u\in\mathfrak{U}(X)$ (for $u=u_0$ we set $f(u)=u_0$); for simple letter codes, however, infinite-decodability does imply finite-decodability.

Remark 3.5. The fact that finite-decodability does not imply infinite-decodability means that it is essential in Definition 3.3 to allow for the "decoding" f' also semicodes, even if f itself is a code. The above simple example shows that it can happen already for uniquely decipherable simple letter codes that as "decoding" only a semicode f' may be found, and, in order that $\mathfrak{X}'=f(\mathfrak{X})$ be encodable by this f' (to obtain $\mathfrak{X}=f'(\mathfrak{X}')$) a particular cost scale – defined through f – has to be chosen.

Theorem 3.3. *Let $\mathfrak{X}$ be a source with a cost scale $\mathfrak{Z}$. Let $\mathfrak{X}$ be encodable by a semicode f with respect to some cost scale $\mathfrak{Z}_1\sim\mathfrak{Z}$. Let 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 encoding $\mathfrak{X}\to\mathfrak{X}'$ is properly decodable (with respect to $\mathfrak{Z}_1$) 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*[12]

$$H(\mathfrak{X}'\|\mathfrak{Z}')=c'\,H(\mathfrak{X}\|\mathfrak{Z}). \tag{3.30}$$

If $\mathfrak{Z}$ and $\mathfrak{Z}'$ are regular, the conditions of proper decodability and of quasi-equivalence of $\mathfrak{Z}'_1$ and $\mathfrak{Z}'$ may be replaced by the corresponding weak concepts.

Proof. If the encoding $\mathfrak{X}\to\mathfrak{X}'$ is decodable, (3.17) may be applied with interchanged $\mathfrak{X}$ and $\mathfrak{X}'$ to obtain for $\mathfrak{Z}_2=f'(\mathfrak{Z}'_1)$

$$H(\mathfrak{X}\|\mathfrak{Z}_2)\leqq H(\mathfrak{X}'\|\mathfrak{Z}'_1). \tag{3.31}$$

In case of proper decodability i.e. $\mathfrak{Z}_2\sim\mathfrak{Z}_1$ we have $H(\mathfrak{X}\|\mathfrak{Z}_2)=H(\mathfrak{X}\|\mathfrak{Z}_1)$, by Theorem 2.2, thus (3.17) and (3.31) imply

$$H(\mathfrak{X}'\|\mathfrak{Z}'_1)=H(\mathfrak{X}\|\mathfrak{Z}_1); \tag{3.32}$$

If $\mathfrak{Z}$ (and thus $\mathfrak{Z}_1\sim\mathfrak{Z}$, too) is regular, already $\mathfrak{Z}_2\overset{w}{\sim}\mathfrak{Z}_1$ is sufficient for $H(\mathfrak{X}\|\mathfrak{Z}_2)=H(\mathfrak{X}\|\mathfrak{Z}_1)$, on account of Theorem 3.2 (observe that $\mathfrak{Z}_2=f'(f(\mathfrak{Z}_1))$ and that $\mathfrak{Z}_1$ may be considered as the map of itself for the identity code $f(u)=u$).

Furthermore, $\mathfrak{Z}'_1\sim c'\,\mathfrak{Z}'$ implies, on account of Theorem 2.2,

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

If $\mathfrak{Z}_1\sim\mathfrak{Z}$ and $\mathfrak{Z}'$ are regular, the same equality follows already from $\mathfrak{Z}'_1\overset{w}{\sim}c'\,\mathfrak{Z}'$, in view of Theorem 3.2. In both cases, (3.32) gives rise to (3.30), and Theorem 3.3 is proved.

[12] In the sense that if either entropy rate exists so does also the other and the equality holds; if the entropy rates do not exist, the equality holds both for the lower and upper entropy rates.

The intuitive meaning of Theorem 3.3 is clear. (3.30) represents the "*principle of conservation of entropy*", i.e. that (properly) decodable encoding does not change the entropy rate apart from a factor representing the quotient of average costs after and before encoding. The inequality (3.18) means that in the non-decodable case some information may get lost. Observe, that proper decodability is essential in order that the equality (3.30) hold. E.g. if $f(u)$ consists of the first $[n/2]$ letters of u if $\|u\|=n$, the code f is obviously infinite-decodable, but if $\mathfrak{X}$ is any source and $\mathfrak{Z}=\mathfrak{Z}'=\mathfrak{C}$, we have $c'=2$, while the encoded source is identical with the original one. Thus $H(\mathfrak{X}')=H(\mathfrak{X})$.

As a point of fact, $\mathfrak{X}'=f(\mathfrak{X})$ always depends merely on the equivalence class or infinite-code (infinite-semicode) defined by f (cf. Remark 3.1); thus if to a given f an $f_1 \sim f$ exists with $c_1' < c'$, the inequality (3.18) can always be improved by replacing c' by c_1'.

In order to exhibit a "principle of conservation of entropy", a comparison of the cost scales of the original and the secondary sources is needed, and to this end one of the two cost scales has to be "lifted" to the other process. Theorem 3.3 refers to comparison of costs in the secondary process. Also the other alternative would be possible, though, as we are going to see, it seems a bit less convenient. Let us consider only the case that f is a code; let $\mathfrak{Z}$ and $\mathfrak{Z}'$ be cost scales for $\mathfrak{X}$ and $\mathfrak{X}'=f(\mathfrak{X})$, respectively, and define the "inverse image" $\mathfrak{Z}_1=f^{-1}(\mathfrak{Z}')$ of $\mathfrak{Z}'$ by assigning to each symbol $\xi(n)$ the cost of its image, i.e.

$$\zeta_1(n)=\sum_{k=\kappa'(n-1)+1}^{\kappa'(n)} \zeta'(k), \qquad \kappa'(n)=\|f(\xi(1;n))\|. \tag{3.33}$$

Then we clearly have (cf. (3.28))

$$f^{-1}(f(\mathfrak{Z}))=\mathfrak{Z}|\mathfrak{K}, \qquad f(f^{-1}(\mathfrak{Z}'))=\mathfrak{Z}'|\mathfrak{K}' \tag{3.34}$$

with the $\mathfrak{K}$ and $\mathfrak{K}'$ defined by (3.16) and (3.33), respectively. Thus, in particular, (3.17) holds for $\mathfrak{Z}_1=f^{-1}(\mathfrak{Z}')$ and $\mathfrak{Z}_1'=f(\mathfrak{Z}_1)=\mathfrak{Z}'|\mathfrak{K}'$. If the encoding is properly decodable with respect to $\mathfrak{Z}_1=f^{-1}(\mathfrak{Z}')$, then the equality holds, and thus, if $\mathfrak{Z}_1$ is quasi-equivalent to $\mathfrak{Z}$, i.e. $\mathfrak{Z}_1 \sim c\,\mathfrak{Z}$ for some $c>0$ and if $\mathfrak{Z}'|\mathfrak{K}' \sim \mathfrak{Z}'$, we obtain, using Theorem 2.2,

$$H(\mathfrak{X}'\|\mathfrak{Z}')=\frac{1}{c}\,H(\mathfrak{X}\|\mathfrak{Z}). \tag{3.35}$$

The following corollary of Theorem 3.3 is worth formulating as a new theorem.

Theorem 3.4. *Let $\mathfrak{X}$ be a source with a given regular cost scale $\mathfrak{Z}$; let $\mathfrak{X}$ be encodable by a code (or semicode) f with respect to $\mathfrak{Z}$. Let $\mathfrak{Z}'$ be a cost scale for $\mathfrak{X}'=f(\mathfrak{X})$ such that*

$$b \leqq \zeta'(n) \leqq B \qquad a.s. \ (n=1,2,\ldots;\ 0<b<B). \tag{3.36}$$

Then, if

$$\frac{\tau'(\|f(\eta(t))\|)}{t} \xrightarrow{P} r>0, \tag{3.37}$$

we have

$$H(\mathfrak{X}'\|\mathfrak{Z}') \leqq \frac{H(\mathfrak{X}\|\mathfrak{Z})}{r}. \tag{3.38}$$

If, in addition, also the decodability condition

(D) *there exists a semicode f' such that $f'(f(\eta(t)))$ is defined a.s. for all $t \geqq 0$, $f'(f(\eta(t_1))) \prec f'(f(\eta(t_2))) \prec \eta(t_2)$ a.s. for $0 \leqq t_1 \leqq t_2$ and such that*

$$\frac{1}{t}\left(v(t) - \|f'(f(\eta(t)))\|\right) \xrightarrow{P} 0 \tag{3.39}$$

is fulfilled, in (3.28) *the equality holds, i.e.*

$$H(\mathfrak{X}' \| \mathfrak{Z}') = \frac{H(\mathfrak{X} \| \mathfrak{Z})}{r}. \tag{3.40}$$

Proof. According to Lemma 2.3, (3.37) is equivalent to $f(\mathfrak{Z}) \overset{w}{\succsim} (1/r)\,\mathfrak{Z}'$. Thus (3.38) is a particular case of (3.18), as $f(\mathfrak{Z}) \overset{w}{\succsim} (1/r)\,\mathfrak{Z}'$ implies $f(\mathfrak{Z}) \prec (1/r)\,\mathfrak{Z}'$ if $\mathfrak{Z}'$ is regular (here the regularity of $\mathfrak{Z}$ is not needed). The condition (D) ensures that the encoding $\mathfrak{X} \to \mathfrak{X}'$ is weakly properly decodable with respect to $\mathfrak{Z}$, thus Theorem 3.3 applies to yield the identity (3.40).

Most commonly, $\mathfrak{X}'$ has a memoryless intrinsic cost scale (cf. Example 2.2) defined by fixed symbol costs $l(x')$ $(x' \in X')$ so that $\zeta'(n) = l(\xi'(n))$; or, somewhat more generally, if $\mathfrak{X}'$ is to be transmitted by a finite-state noiseless channel (cf. Example 2.9), the symbol costs may depend on the "state of the channel", i.e. $\zeta'(n) = l(\xi'(n), \alpha(n-1))$ where $\alpha(k)$ represents the state of the channel after the transmission of the k'th message symbol. In these cases the condition (3.36) is trivially fulfilled provided that $l(x')$ (or $l(x', a)$) is strictly positive. The most important special case is, of course, that $\mathfrak{Z}'$ is the counting scale $\mathfrak{C}$ or, at least, $\mathfrak{Z}'$ is quasi-equivalent to $\mathfrak{C}$, in which case in (3.38) and (3.40) $H(\mathfrak{X}' \| \mathfrak{Z}')$ may be replaced by $r' H(\mathfrak{X}')$ (if $r' \mathfrak{Z}' \sim \mathfrak{C}$).

Remark 3.6. Theorem 3.4 is perhaps the most impressive form of the "principle of conservation of entropy". As $\tau'(\|f(\eta(t))\|)$ is the cumulative cost of the code of a message of cumulative cost t (i.e., of $\eta(t) = \xi(1; v(t))$), the condition (3.37) requires the existence of an average code cost r per unit message cost, in the sense of convergence in probability. If both $\mathfrak{Z}$ and $\mathfrak{Z}'$ are the counting scale $\mathfrak{C}$, (3.37) reduces to

$$\frac{\|f(\xi(1; n))\|}{n} \xrightarrow{P} r, \tag{3.37'}$$

(3.39) reduces to

$$\frac{1}{n}\|f'(f(\xi(1; n)))\| \xrightarrow{P} 1, \tag{3.39'}$$

and the identity (3.40) becomes

$$H(\mathfrak{X}') = \frac{H(\mathfrak{X})}{r}. \tag{3.40'}$$

This relation, dating back to Shannon [16], has often been regarded as "obvious" but, to the authors' knowledge, it has never been proved in a rigorous way, for arbitrary sources and codes. For the case that f is a simple letter code, a proof of (3.40') appears in [10]. The more general case of codes generated by finite-state transducers (cf. Example 3.2) has been considered by Sidel'nikov[13] [17]; he, however, restricted attention to Markovian sources (though, as he has remarked, some of his results hold in more general cases, too).

Remark 3.7. The decodability condition (D) is not necessary for the equality in (3.38). The role of (proper) decodability was to ensure that in inequality (3.17) the equality sign hold. Of course, this may be the case for non-decodable encodings, too; what really needed is the relation[14]

$$H(\eta_1(t)|\eta_1'(t))=o(t) \qquad (t\to\infty). \tag{3.41}$$

(3.41) surely holds if the number of different possible values of $\eta_1(t)$ yielding the same $\eta_1'(t)=f(\eta_1(t))$ is $\exp\{o(t)\}$; this condition is trivially fulfilled e.g. if f has the property that for any given $u'\in\mathfrak{U}(X')$ there are at most d different $u\in\mathfrak{U}(X)$ with $f(u)=u'$ ($d=1$ means that the code is strict-sense finite-decodable; if $d>1$, the code may still be properly decodable, but it need not be so). E.g. one may require that is $u_1\neq u_2$ and both the first and last letters of u_1 and u_2 coincide then $f(u_1)\neq f(u_2)$ (in this case, with the above notation, $d=s^2$, where s is the size of the alphabet X). Sidel'nikov [17] has imposed just such a condition on the encoding, generated by a finite-state transducer and proved that it implies (3.40'); actually, he has shown this condition to be necessary, too, for the class of sources considered by him. Of course, for arbitrary sources such a simple necessary condition of the equality cannot be hoped for.

We conclude this section by exhibiting a general form of the "noiseless coding theorem". Let us mean by a general noiseless channel with alphabet X' a subset $\mathfrak{U}_0$ of $\mathfrak{U}(X')$ together with a nonnegative function $l(u')$ on $\mathfrak{U}_0$, satisfying the same conditions as in Example 2.10. Let $N(t)$ be the number of different sequences $u'\in\mathfrak{U}_0$ with the property $l(u')\leqq t$, $l(u'\,x')>t$ (for some $x'\in X'$ with $u'\,x'\in\mathfrak{U}_0$) and define the channel capacity by[15]

$$C=\overline{\lim_{t\to\infty}}\,\frac{\log_2 N(t)}{t}. \tag{3.42}$$

Theorem 3.5. *Let $\mathfrak{X}$ be a source with finite alphabet X and let us be given a noiseless channel with alphabet X' of capacity C. Let $\mathfrak{Z}$ be a cost scale for $\mathfrak{X}$ and let f be a semicode from X to X' such that $\mathfrak{X}$ is encodable by f with respect to $\mathfrak{Z}$ in a properly decodable way (or in a weakly properly decodable way, if $\mathfrak{Z}$ is regular) and the resulting source $\mathfrak{X}'=f(\mathfrak{X})$ is transmissible by the given channel (i.e. $\xi'(1;n)\in\mathfrak{U}_0'$ a.s. $n=1,2,\ldots$). Then, if*

$$L=\lim_{t\to\infty}\frac{E\,l(f(\eta(t)))}{t} \tag{3.43}$$

exists, it necessarily satiesfies

$$L\geqq\frac{\bar{H}(\mathfrak{X}\|\mathfrak{Z})}{C}. \tag{3.44}$$

[13] Cf. the next remark.

[14] For decodable encoding, $H(\eta_2(t)|\eta_1'(t))=0$ trivially holds $(\mathfrak{Z}_2=f'(\mathfrak{Z}_1'))$; if $\mathfrak{Z}_2\sim\mathfrak{Z}_1$, this implies (3.41), by Theorem 3.1.

[15] It can be shown under mild regularity conditions, that in (3.42) actually the limit exists; a related problem is discussed in [18].

Proof. Proper decodability implies $H(\mathfrak{X}\|\mathfrak{Z})=H(\mathfrak{X}'\|f(\mathfrak{Z}))$, by Theorem 3.2, and if $\mathfrak{Z}$ is regular, this identity holds in the weakly properly decodable case, too. As for $\mathfrak{Z}_1'=f(\mathfrak{Z})$ we have $\nu_1'(t)=\|f(\eta(t))\|$, $\tau'(\nu_1'(t))=l(f(\eta(t)))$, (3.44) is an immediate consequence of Theorem 2.5.

Remark 3.8. For channels with all conceivable sequences transmissible (i.e. $\mathfrak{U}_0=\mathfrak{U}(X')$) and with memoryless intrinsic cost scales, in the special case of simple letter codes and $\mathfrak{Z}=\mathfrak{C}$, (3.44) has been proved in [6] and [12], where also suitable code constructions are presented. For the case of stationary ergodic sources with $\mathfrak{Z}=\mathfrak{C}$ for $\mathfrak{U}_0=\mathfrak{U}(X')$, $l(u')=\|u\|$ and for general codes as defined by him, Billingsley [2] has proved even a stronger theorem than the corresponding particular case of Theorem 3.5. For more general cases, to the authors' knowledge, the assertion is, though "intuitively obvious", as a mathematical theorem new.

For trivial noiseless channels, with all conceivable sequences transmissible (i.e. $\mathfrak{U}_0=\mathfrak{U}(X')$) and all symbols having unit cost (i.e. $l(u')=\|u'\|$ for all $u'\in\mathfrak{U}(X')$), inequality (3.44) reduces to

$$L\geqq\frac{\overline{H}(\mathfrak{X}\|\mathfrak{Z})}{\log_2 s}. \tag{3.44'}$$

In particular, if $s=2$, (3.44') means that it is impossible to encode the source $\mathfrak{X}$ by binary digits in a properly decodable way such that the average number of binary digits used per unit cost be less than the entropy rate of $\mathfrak{X}$ with respect to the cost scale in question, in accordance with the intuitive concept of entropy. In our mind, to have an exact formulation and a general proof of this familiar assertion is essential for the interpretation of entropy as the measure of the amount of information. Observe that the usual "noiseless coding theorem" (the particular case of (3.44') that $\mathfrak{Z}=\mathfrak{C}$ and f is a simple letter code) is less satisfactory in this respect, as it concerns a very special type of encoding only. (Billingsley's results [2] are much more relevant in this respect; in view of the different definitions of "general" encoding, however, there is only a slight overlapping between his results and that of ours.)

Remark 3.9. Of course, if the principle of conservation of entropy (Theorem 3.4) is valid, (3.44) is an immediate consequence of (3.40); in fact, then (3.44) holds with the r of (3.37) in place of L (observe that if $\dfrac{l(f(\eta(t)))}{t}$ is u.i. for $t\to\infty$ then $L=r$; otherwise we have $L>r$).

§ 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 $C\leqq(\log_2 s)/b$). Observe, too, that the theorems involving coding (Theorems 3.3, 3.4, 3.5) can obviously be extended to countable X (provided, in the case of the first two, that X' remains finite), if some additional assumptions (e.g. strict sense finite-decodability) ensure the validity of (3.41). 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.

One could generalize the concept of cost scale permitting $v(t)$ to be any family of nonnegative IRV's (dropping the condition that the sample functions are nondecreasing). Some of our results would remain true for this case, too, but in lack of examples where a need for such a generalization would arise, it does not deserve closer attention.

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 than to the first n ones); as m is constant, this change of viewpoint does not cause any change in the results. Sometimes one considers randomised encodings, too, where the code sequence assigned to a message sequence is not uniquely determined but it depends on chance. The mapped cost scale $f(\mathfrak{Z}_1)$ can be defined in this case, too, but (3.17) need not hold, except if some additional conditions ensure

$$H(\eta_1'(t)|\eta_1(t))=o(t) \tag{3.45}$$

(such a condition would be e.g. that $f(u)$, though not uniquely determined by u, can take on at most d different values, where d does not depend on u). In general, in case of randomised encodings, in Theorems 3.3 and 3.4 the entropy rate $H(\mathfrak{X}'\|\mathfrak{Z}')$ has to be replaced by the mutual information rate $I(\mathfrak{X}', \mathfrak{X}\|\mathfrak{Z}', \mathfrak{Z})$ (cf. (3.3)); of course, if (3.45) is valid, Theorems 3.3 and 3.4 hold in their original form.

As randomised encoding is essentially equivalent to a noisy channel, the above remark indicates how 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 synchronisation errors, investigated by Dobrušin [8]. A closer study of noisy channels with arbitrary cost sclaes, however, though very desirable both from the theoretical and practical points of view, 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 $\dfrac{H(\mathfrak{X}\|\mathfrak{Z})}{C}$ 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. Ash, R.: Information theory. New York: Interscience Publishers 1965.
2. Billingsley, P.: On the coding theorem for the noiseless channel. Ann. math. Statistics **32**, 594 – 601 (1961).
3. Blachman, N. M.: Minimum-cost encoding of information. IRE Trans. Inform. Theory PGIT-3, 139 – 149 (1954).
4. – Minimum-cost transmission of information. Inform. and Control **7**, 508 – 511 (1964).
5. Bloh, E. L.: Generalization of an inequality of information theory to the case of symbols of inequal duration [in Russian]. Probl. Peredači Inform. **5**, 95 – 99 (1960).
6. – Construction of optimal code constituted of elementary symbols of inequal duration [in Russian]. Probl. Peredači Inform. **5**, 100 – 111 (1960).
7. Csiszár, I.: Two remarks on noiseless coding. Inform. and Control **11**, 317 – 322 (1967).
8. Dobrušin, R. L.: Capacity of channels with synchronisation errors. Paper presented at the Colloquium on Information Theory, Debrecen, 1967.
9. Feinstein, A.: Foundations of information theory. New York: McGraw-Hill Book Co. 1958.
10. Katona, G., and G. Tusnády: The principle of conservation of entropy in a noiseless channel. Studia Sci. Math. Hung. **2**, 29 – 35 (1966).
11. Kinney, J. R.: Singular functions associated with Markov chains. Proc. Amer. math. Soc. **9**, 603 – 608 (1958).
12. Krause, R. M.: Channels which transmit letters of inequal duration. Inform. and Control **5**, 13 – 24 (1962).
13. Ljubič, Ju. I.: Remark on the capacity of the discrete noiseless channel [in Russian]. Uspehi mat. Nauk. **17**, 191 – 198 (1962).
14. Parry, W.: Intrinsic Markov chains. Trans. Amer. math. Soc. **112**, 55 – 66 (1964).
15. Radke, C. E.: Necessary and sufficient conditions on conditional probabilities to maximize entropy. Inform. and Control **9**, 279 – 284 (1966).
16. Shannon, C. E.: A mathematical theory of communication. Bell System techn. J. **27**, 379 – 423, 623 – 656 (1948).
17. Sidel'nikov, V. M.: On statistical properties of transformations induced by finite automata [in Russian]. Kibernetika Kiev **6**, 1 – 14 (1965).
18. Zaïdman, R. A.: On the asymptotics of certain sequences encountered in problems of non-Markov random walks and of information theory [in Russian]. Vestnik Leningrad. Univ. **1**, 23 – 33 (1965).

Dr. I. Csiszár
Dr. G. Katona
G. Tusnády
Mathematisches Institut
der Ungarischen Akademie der Wissenschaften
Reáltanoda u. 13 – 15
Budapest V, Ungarn

(Received, March 18, 1968)