CDMTCS
Research
Report
Series

Entropic Measures, Markov Information Sources and Complexity

Cristian S. Calude and
Monica Dumitrescu
University of Auckland, New Zealand
Bucharest University, Romania

CDMTCS-150
February 2001

Centre for Discrete Mathematics and
Theoretical Computer Science

Entropic Measures, Markov Information Sources and Complexity

Cristian S. Calude
Department of Computer Science
University of Auckland
Private Bag 92019
Auckland
New Zealand
cristian@cs.auckland.ac.nz

Monica Dumitrescu
Faculty of Mathematics
University of Bucharest
Str. Academiei 14
R-70109 Bucharest
Romania
mdumi@pro.math.unibuc.ro

February 2001

Abstract

The concept of entropy plays a major part in communication theory. The Shannon entropy is a measure of uncertainty with respect to *a priori* probability distribution. In algorithmic information theory the information content of a message is measured in terms of the size in bits of the smallest program for computing that message. This paper discusses the classical entropy and entropy rate for discrete or continuous Markov sources, with finite or continuous alphabets, and their relations to program-size complexity and algorithmic probability. The accent is on ideas, constructions and results; no proofs will be given.

1 Introduction

In the classical theory of information the entropy is a measure of uncertainty contained in a stochastic system which can be described through a probability distribution. It does not allow one to call a particular outcome random, except in an intuitive, heuristic sense; it gives no explicit definition of "randomness", which is considered implicitly, by means of probability fields, random variables and stochastic processes.

In algorithmic information theory information is measured in terms of program-size complexity of self-delimited programs and algorithmic probabilities. The information-theoretic complexity of an object is given by the size in bits of the smallest program for computing that object, i.e. its program-size complexity; see, for example, [8, 3, 9, 10, 11]. Algorithmic information theory offers an algorithmic way to define the notions of random (finite) string and random (infinite) sequence.

This paper discusses the classical entropy and entropy rate for discrete or continuous Markov sources, with finite or continuous alphabets, and their relations to program-size complexity and algorithmic probability. We will concentrate on ideas, constructions and results; no proofs will be given.

2 Notation

An information source produces a message or sequences of messages to be transmitted through a communication channel. Messages can be generated either continuously or at discrete moments of time, and the alphabet A of the source can be either finite or an arbitrary subset of real numbers. Let us denote by T the time set for broadcasting; for example, T can be the set of integers or a set of real numbers.

An *information source* is an infinite probability space $\left(A^T, \mathcal{K}^{\mathcal{T}}, \mu\right)$, and its output consists of a stochastic process $\{X_t, t \in T\}$ with the time parameter $t \in T$, the state space A and the probability distribution μ.

A *Markov information source* satisfies the condition

$$\Pr\left(X_t \in B \mid X_u, u \leq s\right) = \Pr\left(X_t \in B \mid X_s\right),$$

for every $s < t$ and every Borel set B, where $\Pr(X_t \in B \mid Y)$ denotes the conditional probability of $\{X_t \in B\}$ given Y.

An information source is called *stationary* if the distribution μ is shift invariant; that is, the distribution of $(X_{t_1+s}, \ldots, X_{t_n+s})$ is independent of s for any positive integer n and $t_1, \ldots, t_n \in T$. For more details we refer to [14, 19].

By $\mathbf{N}, \mathbf{Z}, \mathbf{Q}$ and $\mathbf{R}$, we denote the sets of nonnegative integers, integers, rationals and reals, respectively. By log we denote the base 2 logarithm; exp denotes the exponential function. The set of all strings over the finite alphabet A is denoted by A^*. The length of a string s is denoted by $|s|$; by A^n we denote the set of all strings of length n. A string s is a prefix of a string t ($s \subseteq t$) if there is a string $r \in A^*$ such that $sr = t$. A subset S of A^* is *prefix-free* if whenever s and t are in S and $s \subseteq t$, then $s = t$. For example, the set $\{1^i 0 \mid i \geq 0\}$ is prefix-free.

We shall employ a special model of deterministic Turing machine, namely *self-delimiting Turing machines* or *(Chaitin) machines* (simply, *machines*): these are Turing machines (transforming binary strings into binary strings) having prefix-free domains. Note that every prefix-free computably enumerable set of strings is the domain of some machine. We refer to [25, 3] for more about Turing machines, computable sets and functions, computably enumerable (c.e.) sets.

The *program-size complexity* induced by the machine M is $H_M(x) = \min\{|z| \mid M(z) = x\}$, with the convention that the minimum of the empty set is undefined. The *algorithmic probability* of the machine M to produce the output x is

$$P_M(x) = \sum_{M(u)=x} 2^{-|u|}, \tag{1}$$

and the halting probability of M is $\Omega_M = \sum_{x \in A^*} P_M(x)$.

A machine U is *universal* if for every machine M, there is a constant c_M (depending upon M) with the following property: if $M(x)$ halts, then there is an $x' \in A^*$ such that $U(x') = M(x)$ and $|x'| \leq |x| + c_M$; c_M is the simulation constant of M on U. Universal machines can be effectively constructed. See more in [3].

3 Discrete Time Markov Sources

In this section we discuss the entropy of various discrete time Markov sources.

3.1 Finite alphabet stationary sources

Let $\left(A^{\mathbf{Z}}, \mathcal{K}^{\mathbf{Z}}, \mu\right)$ be a *discrete time information source, with a finite alphabet A*. For an n-dimensional outcome $(X_1, \ldots, X_n)$, *Shannon's entropy* is defined by the relation

$$H(X_1, \ldots, X_n) = -\sum_{x_1, \ldots, x_n \in A} \mu(x_1, \ldots, x_n) \log \mu(x_1, \ldots, x_n).$$

In most cases, the entropy $H(X_1, \ldots, X_n)$ diverges as $n \to \infty$, hence, the source has infinitely large entropy. This fact suggests that what is important is not the limit of $H(X_1, \ldots, X_n)$, but its rate of growth. Thus, the *entropy of the source* is defined by

$$\overline{H}(X) = \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n},$$

when the limit exists.

Proposition 3.1 *If the discrete time information source $\left(A^{\mathbf{Z}}, \mathcal{K}^{\mathbf{Z}}, \mu\right)$, with a finite alphabet A is stationary, then the entropy of the source exists and is equal to* $\inf_n \frac{H(X_1, \ldots, X_n)}{n}$.

Proposition 3.2 *Let $\left(A^{\mathbf{Z}}, \mathcal{K}^{\mathbf{Z}}, \mu\right)$ be a discrete time stationary information source, with a finite alphabet A, such that*

$$\mu(x_1, \ldots, x_n) = p(x_1) \prod_{i=1}^{n-1} p(x_{i+1} \mid x_i),$$

and

$$p(x) \geq 0,\ \sum_{x \in A} p(x) = 1,$$

$$p(x' \mid x) \geq 0,\ \sum_{x' \in A} p(x' \mid x) = 1,\ \textit{for every } x \in A.$$

Then its entropy is given by the formula:

$$\overline{H}(X) = H(X_1 \mid X_0) = -\sum_{x \in A} p(x) \sum_{x' \in A} p(x' \mid x) \log p(x' \mid x).$$

The proofs of the above results can be found in [19].

3.2 Finite alphabet non-stationary sources

In many applications, the discrete Markov information source is *not stationary*, but there exists a stationary source which may be associated with it. An important example is a source which produces messages representing a random walk with two absorbing barriers. This model was studied in [16].

Let us suppose that the letters of the alphabet A are simply denoted by $0, 1, \dots, s$. Broadcasting is governed by the parameter θ which gives the probability of a jump from the ith letter to the $(i-1)$th letter. Suppose that the transition matrix $\mathbf{P}$ of the associated Markov chain has the elements

$$p_\theta(0 \mid 0) = p_\theta(s \mid s) = 1, \quad \text{for every } \theta, \tag{2}$$

$$p_\theta(i-1 \mid i) = \theta,\ p_\theta(i+1 \mid i) = 1-\theta,\ i = 1, \dots, s-1. \tag{3}$$

This means that the states 0 and s are absorbing (i.e. once one of these states is reached it is not possible to move to any other state), while $1, \dots, s-1$ are transient (i.e. the probability that the process returns into one of these states after a finite period of time is less than 1).

If we consider a permutation of letters of the alphabet, say $0, s, 1, 2, \dots, s-2, s-1$, then the transition matrix $\mathbf{P}$ is of the form

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{p}(\theta) & \mathbf{Q}(\theta) \end{pmatrix},$$

where $\mathbf{I}$ is the 2×2 identity matrix, $\mathbf{0}$ is a $2 \times (s-1)$ matrix of zeros, $\mathbf{p}(\theta)$ is a $(s-1) \times 2$ matrix and

$$\mathbf{Q}(\theta) = \begin{pmatrix} 0 & 1-\theta & 0 & \dots & 0 & 0 & 0 \\ \theta & 0 & 1-\theta & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \theta & 0 & 1-\theta \\ 0 & 0 & 0 & \dots & 0 & \theta & 0 \end{pmatrix}.$$

Notice that $\mathbf{Q}(\theta)$ is not a stochastic matrix.

Let $(X_1, \dots, X_n)$ be the output of the source for n consecutive moments of time, and let us assume that the absorption has not taken place (i.e. $X_n \neq 0, s$). When this assumption is true for large n, one says that "the absorption has not taken place and will not take place for a long time". Conditional on this fact, one can associate a stationary Markov source, with alphabet $A' = \{1, \dots, s-1\}$, which gives the conditional broadcast of the initial source. According to well known properties of absorbing Markov chains (see [1]), the elements which define the probability distribution for this new source are constructed as follows.

- The stationary distribution on A' is

  $$\pi_\theta(j) = v_j w_j,\ j \in A',$$

  where $v = (v_1, .., v_{s-1})'$ and $w = (w_1, \dots, w_{s-1})'$ are the left and right eigenvectors of the matrix $\mathbf{Q}(\theta)$ corresponding to the largest eigenvalue $\lambda_1(\theta)$, such that

  $$\sum_{j=1}^{s-1} v_j = 1,\ \sum_{j=1}^{s-1} v_j w_j = 1.$$

- The transition matrix of the associated stationary Markov source, denoted $R(\theta)$, has the elements

$$r_\theta(j \mid i) = \frac{1}{\lambda_1(\theta)} p_\theta(j \mid i) \frac{w_j}{w_i},\ i, j = 1, \ldots, s-1.$$

Theorem 3.3 *Let $(A^{\mathbf{Z}}, \mathcal{K}^{\mathbf{Z}}, \mu)$ be a Markov information source with alphabet $A = \{0, 1, \ldots, s\}$ and transition matrix given by (2) and (3). Under the assumption that the absorption has not taken place and will not take place for a long time, the entropy of the stationary associated source is*

$$\overline{H}(X) = -\sum_{i=1}^{s-1} \pi(i) \sum_{j=1^{s-1}} r(j \mid i) \log r(j \mid i),$$

where the stationary distribution $\{\pi_\theta(j), j = 1, \ldots, s-1\}$ and the transition matrix are independent of θ:

$$\pi(j) = \left(\sum_{i=1}^{s-1} \sin^2 \frac{i\pi}{s}\right)^{-1} \sin^2 \frac{j\pi}{s},\ j = 1, \ldots, s-1,$$

$$r(j \mid i) = \begin{cases} 1, & j = 2, \quad i = 1, \\ \sin \frac{(i-1)\pi}{s} \left(2 \cos \frac{\pi}{s} \sin \frac{i\pi}{s}\right)^{-1}, & j = i-1, \quad i = 2, \ldots, s-2, \\ \sin \frac{(i+1)\pi}{s} \left(2 \cos \frac{\pi}{s} \sin \frac{i\pi}{s}\right)^{-1}, & j = i+1, \quad i = 2, \ldots, s-2, \\ 1, & j = s-2, \quad i = s-1, \\ 0, & \text{otherwise.} \end{cases}$$

3.3 Infinite alphabet sources

We discuss now the case of a *Markov information source with discrete time and alphabet $A = \mathbf{R}$*. The Shannon's entropy is replaced by an entropic measure which takes into account the continuous character of the measure μ.

We assume that the vector $(X_1, \ldots, X_n)$ has a probability density $f(x_1, \ldots, x_n)$ with respect to the Lebesgue measure. Then, the *Boltzmann entropy* is

$$h(X_1, \ldots, X_n) = -\int_{R^n} f(x_1, \ldots, x_n) \log f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n, \tag{4}$$

provided the integral exists.

In contrast with Shannon's entropy, $h(X_1, \ldots, X_n)$ itself does not work as a measure of uncertainty. However, it is well known that the difference $h(X_1, \ldots, X_n) - h(X_1', \ldots, X_n')$ of the entropies indicates the difference of uncertainties of $(X_1, \ldots, X_n)$ and $(X_1', \ldots, X_n')$, see [20]. This is an important difference between the continuous entropy and the discrete one: in the discrete case the entropy measures the uncertainty in an absolute way, while in the continuous case the measurement is only relative. Note also that the discrete entropy is always non-negative while the continuous one can be negative.

The *entropy rate* (or the *per unit time entropy*) of a discrete information source with alphabet $A = \mathbf{R}$ can be defined by

$$\bar{h}(X) = \lim_{n \to \infty} \frac{h(X_1, \ldots, X_n)}{n},$$

when the limit exists. The proof of the following result can be found in [20].

Theorem 3.4 *Suppose that the information source with discrete time and alphabet $A = \mathbf{R}$ is stationary and has finite continuous entropy for every n. Then, the entropy rate $\bar{h}(X)$ exists, and is equal to the conditional entropy of one step "future" X_1 when the "past" $(\ldots, X_{-1}, X_0)$ is known, that is,*

$$\bar{h}(X) = \lim_{n \to \infty} h(X_1 \mid X_0, \ldots, X_{-n}),$$

where

$$h(X_1 \mid X_0, \ldots, X_{-n}) = -\int_{\mathbf{R}^{n+2}} f(x_{-n}, \ldots, x_0, x_1) \log \frac{f(x_{-n}, \ldots, x_0, x_1)}{f(x_{-n}, \ldots, x_0)} dx_{-n} \ldots dx_0 dx_1.$$

Corollary 3.5 *Let* $\left(A^{\mathbf{Z}}, \mathcal{K}^{\mathbf{Z}}, \mu\right)$ *be a Markov, stationary information source, with discrete time and alphabet* $A = \mathbf{R}$*, such that*

$$f\left(x_1, \ldots, x_n\right) = f\left(x_1\right) \prod_{i=1}^{n-1} f\left(x_{i+1} \mid x_i\right),$$

and

$$f\left(x\right) \geq 0, \int\limits_{\mathbf{R}} f\left(x\right) dx = 1,$$

$$f\left(x' \mid x\right) \geq 0, \int\limits_{\mathbf{R}} f\left(x' \mid x\right) dx' = 1, \textit{ for every } x \in \mathbf{R}.$$

Then its entropy rate has the value

$$\bar{h}\left(X\right) = h\left(X_1 \mid X_0\right) = -\int\limits_{\mathbf{R^2}} f\left(x\right) f\left(x' \mid x\right) \log f\left(x' \mid x\right) dx' dx.$$

3.4 Gaussian sources

An information source $\left(A^{\mathbf{Z}}, \mathcal{K}^{\mathbf{Z}}, \mu\right)$ with $A = \mathbf{R}$ is called *Gaussian* if its output $\{X_t, t \in \mathbf{Z}\}$ is a Gaussian process; that is, the joint distribution of $(X_{t_1}, \ldots, X_{t_n})$ is Gaussian for every finite set $\{t_1, \ldots, t_n\} \subset \mathbf{Z}$.

Let us denote by $N\left(n; \mathbf{m_{t_1,\ldots,t_n}}, \mathbf{\Sigma_{t_1,\ldots,t_n}}\right)$ the n-dimensional Gaussian distribution of $\left(X_{t_1}, \ldots, X_{t_n}\right)$. When the source is stationary, the mean vector $\mathbf{m_{t_1,\ldots,t_n}}$ has equal components, and the elements of $\mathbf{\Sigma_{t_1,\ldots,t_n}}$ (covariances) depend only on the time intervals (i.e. *cov* $\left(X_{t_i}, X_{t_j}\right) = \gamma\left(t_i - t_j\right), i, j = 1, \ldots, n$).

Theorem 3.6 *Let* $\left(A^{\mathbf{Z}}, \mathcal{K}^{\mathbf{Z}}, \mu\right)$ *be a stationary information source, with discrete time and alphabet* $A = \mathbf{R}$*, such that* $\left(X_1, \ldots, X_n\right)$ *has an* n*-dimensional Gaussian distribution* $N\left(n; \mathbf{m}, \mathbf{\Sigma}\right)$. *Then the Boltzmann entropy is given by*

$$h\left(X_1, \ldots, X_n\right) = \frac{1}{2} \log\left(\left(2\pi e\right)^n \mid \mathbf{\Sigma} \mid\right).$$

In particular, if X_1 is an one-dimensional Gaussian random variable with distribution $N\left(m_1, \sigma^2\right)$, then $h\left(X_1\right) = \frac{1}{2}\log\left(2\pi e \sigma^2\right)$. It is noticed that the Boltzmann entropy for Gaussian sources does not depend on mean vectors.

A stochastic process $\{X_t, t \in \mathbf{Z}\}$ is called autoregressive of first order (denoted $AR(1)$) if its elements are given by the relation

$$X_t = \phi X_{t-1} + \xi_t, \tag{5}$$

where $\phi \in \mathbf{R}$ and $\{\xi_t, t \in \mathbf{Z}\}$ is a sequence of independent, identical distributed random variables (the white noise which generates the process). An $AR(1)$ process is stationary if $\mid \phi \mid < 1$.

Theorem 3.7 *Let* $\{X_t, t \in \mathbf{Z}\}$ *be a stationary Gaussian process. Then* $\{X_t, t \in \mathbf{R}\}$ *is a Markov chain if and only if it is an autoregressive process* $AR(1)$*, given by the relation (5), where* $\mid \phi \mid < 1$ *and* $\{\xi_t, t \in \mathbf{Z}\}$ *is a Gaussian white noise with variance* 1.

For proofs see [14, 20]. Using Corollary 3.5, and Propositions 3.6, 3.7 one can obtain the following property.

Proposition 3.8 *The value of the entropy rate of a stationary Gaussian Markov source is independent of its* $AR(1)$ *representation, and is equal to*

$$\bar{h}\left(X\right) = \frac{1}{2} \log\left(2\pi e\right).$$

4 Continuous Time Markov Sources

Let $\left(A^T, \mathcal{K}^\mathcal{T}, \mu\right)$ be a continuous time information source, with $T = \mathbf{R}$, such that its output is the stochastic process $\{X_t, t \in \mathbf{R}\}$. Defining the entropy rate of such a source is rather complicate, even for stationary Gaussian sources, where canonical representations are available.

4.1 Gaussian sources

We consider, first, the case $A = \mathbf{R}$. The statisticians' approach is based on the fact that observation is made only discretely, for example at every k units of time. Then the mathematical model of observed values is given by $X^{(k)} = \{X_{nk} \mid n \in \mathbf{Z}\}$. The observed process $X^{(k)}$ may be called the discretization process of $\{X_t, t \in \mathbf{R}\}$ with time interval k.

Proposition 4.1 *Let $\left(A^T, \mathcal{K}^{\mathcal{T}}, \mu\right)$ be a Markov stationary Gaussian information source, with $T = \mathbf{R}$ and $A = \mathbf{R}$. Then the process $X^{(k)} = \{X_{nk}, n \in \mathbf{Z}\}$ is an $AR(1)$ process, hence a discrete time Markov stationary Gaussian process. Thus, the entropy rate may be evaluated when the initial source is observed with time interval k.*

For the proof of this result we refer to [20]. It goes without saying that, in general, $\{X_t, t \in \mathbf{R}\}$ cannot be recovered from $X^{(k)}$, and we only can estimate the structure of the initial information source from the observed discretization process.

4.2 Pure jump sources

Now we consider the case of a finite alphabet $A = \{1, \ldots, s\}$ and time $T = [0, \infty)$. Let $\{X_t, t \geq 0\}$ be the outcome of the Markov source $\left(A^T, \mathcal{K}^{\mathcal{T}}, \mu\right)$ and let us assume that the transition probabilities

$$p_t\left(j \mid i\right) = \Pr\left(X_{s+t} = j \mid X_s = i\right),$$

are independent of s and continuous at every t, including $t = 0$.

Suppose the following limits exist and are finite:

$$q_{ij} = \lim_{t \to 0} \frac{p_t\left(j \mid i\right) - \delta_{ij}}{t}, i, j = 1, \ldots s, \tag{6}$$

where δ_{ij} is Kronecker's symbol. Then the process $\{X_t, t \geq 0\}$ is called a *Markov pure-jump process*, with the infinitesimal generator $Q = \parallel q_{ij} \parallel_{i,j=1,\ldots,s}$.

Let us put

$$q_i = -q_{ii}, \tag{7}$$

assume $q_i > 0$, for every $i = 1, \ldots, s$, and notice that

$$\sum_{j \neq i} q_{ij} = q_i, \; i = 1, \ldots, s. \tag{8}$$

We also assume that $p_t\left(j \mid i\right) > 0$ for all $i, j \in A$ and all $t > 0$. Then, the stationary distribution of the process $\{\pi_j, j \in A\}$ exists, and satisfies the following relations:

$$\lim_{t \to \infty} p_t\left(j \mid i\right) = \pi_j > 0, \text{ for every } i \in A,$$

$$\sum_{j \in A} \pi_j = 1, \; \sum_{i \in A} \pi_i \cdot p_t\left(j \mid i\right) = \pi_j, \text{for all } t.$$

The source broadcasting has the following constructive development:

- At $t = 0$ the source broadcasts the signal i with probability π_i.
- The emission time of this first signal is a random variable T_0, with probability density

$$f_{T_0}\left(t\right) = q_i \exp\left(-q_i t\right), \; t > 0.$$

- At time $t = T_0$ the signal j $(j \neq i)$ is broadcast with probability q_{ij}/q_i.
- The random emission time of j is T_1, with probability density

$$f_{T_1}\left(t\right) = q_j \exp\left(-q_j t\right), \; t > 0.$$

- At time $t = T_0 + T_1$ the process jumps to the signal k $(k \neq j)$ with probability q_{jk}/q_j, and so on.

Let $\{Z_0, Z_1, \ldots\}$ be the successive states the system passes through. The bivariate discrete-time process $\{(Z_n, T_n), n = 0, 1, \ldots\}$ is a Markov process (called *the embedded process*) with the state space $A \times [0, \infty)$, initial probabilities

$$\Pr(Z_0 = i, T_0 > t) = \pi_i \exp(-q_i t),$$

and transition probabilities

$$\Pr(Z_{n+1} = j, T_{n+1} > t \mid Z_n = i, T_n = u) = \begin{cases} (q_{ij}/q_i) \exp(-q_i t), & j \neq i, \\ 0, & j = i. \end{cases}$$

Let us suppose that the emission of the source is observed during a fixed interval of time $[0, t]$ and let us denote by $v = ((z_0, t_0), \ldots, (z_{n-1}, t_{n-1}), z_n)$ the recorded trajectory of the embedded process. The probability density corresponding to this sample is

$$f_t(v) = \begin{cases} \pi_{z_0} \exp(-q_{z_0} t), & \text{if } v = (z_0), \\ \pi_{z_0} \prod_{j=0}^{n-1} q_{z_j z_{j+1}} \exp\left[-\left(q_{z_j} - q_{z_n}\right) t_j - q_{z_n} t\right], & \text{if } \sum_{j=0}^{n-1} t_j < t, \\ 0, & \text{otherwise.} \end{cases}$$

Let $n_t(i, j)$ be the total number of jumps from i to j during $[0, t]$ and let $r_t(i)$ be the total time during which signal i is broadcast. Then

$$f_t(v) = K \prod_{i,j \in A,\, i \neq j} (q_{ij})^{n_t(i,j)} \prod_{i \in A} \exp(-q_i \cdot r_t(i)),$$

where K is a positive constant, independent of the elements of Q.

We define the Boltzmann entropy for the observation interval $[0, t]$ by

$$h_t = -\int f_t(v) \log f_t(v) \, d\mu(v)$$

and the entropy rate of the source by

$$\bar{h} = \lim_{t \to \infty} \frac{h_t}{t}.$$

By direct calculation one can obtain the expression of h_t:

$$h_t = -\log K - \frac{t}{\rho} \sum_{i,j \in A,\, i \neq j} Q^{ii} q_{ij} \log q_{ij} + \frac{t}{\rho} \sum_{i \in A} Q^{ii} q_i,$$

where Q^{ii} be the (i, i) cofactor of the matrix Q and ρ be the product of the non-zero eigenvalues of Q.

Theorem 4.2 *Let us consider the Markov source $\left(A^T, \mathcal{K}^\mathcal{T}, \mu\right)$ with $A = \{1, \ldots, s\}$, $T = [0, \infty)$ and the infinitesimal generator Q given by the relations (6), (7), (8). Let Q^{ii} be the (i, i) cofactor of Q and ρ be the product of the non-zero eigenvalues of Q. Then the entropy rate of the source exists and is given by*

$$\bar{h} = \frac{1}{\rho} \sum_{i,j \in A} Q^{ii} q_{ij} (1 - \log q_{ij}).$$

5 Entropy and Complexity

In this section we explore some connections between program-size complexity, algorithmic probability and entropy of information sources with a binary[1] alphabet and discrete time.

[1] All results actually hold for an arbitrary finite alphabet, cf. [3].

5.1 Discrete Markov sources

Consider a discrete Markov binary information source, i.e. a finite ergodic Markov chain (see [21] with alphabet (states) $A = \{s_1, s_2, \ldots, s_m\}$ with the following property: for every $1 \leq j \leq m$, there exist two states s_{i0}, s_{i1} such that the transition probability from s_j to s_{i0} is p_i, the transition probability from s_j to s_{i1} is $1 - p_i$, and the transition probability from s_j to any s_k with $k \neq i0, i1$ is 0. We assume that each p_i is a *computable real*, that is, there is an algorithm which when presented a non-negative integer l produces the first l digits of the binary expansion of that number. Transitions from s_j to s_{i0} are labelled by 0 and transitions from s_j to s_{i1} are labelled by 1. The source generates a binary string by starting in some arbitrary fixed state, and producing the labels of transitions it takes. We denote by Pr the probability distribution of strings generated by the source. According to proposition 3.2, the entropy is defined by

$$\bar{H} = -\sum_{i=1}^{m} a_i (p_i \log p_i + (1 - p_i) \log(1 - p_i)).$$

The next result was proven in [22]:

Theorem 5.1 *Let U be a universal machine. Then,*

$$\bar{H} = \lim_{n \to \infty} \frac{1}{n} \sum_{|x|=n} H_U(x) \Pr(x).$$

To understand better the phenomenon let's consider a special case of Markov information sources, namely a Bernoulli source. To this aim consider the set of all binary strings of length n and assign a probability to each digit 0,1: $\Pr(0) = P_0, \Pr(1) = P_1$, $P_0 + P_1 = 1, 0 \leq P_0, P_1, \leq 1$. The alphabet is $A = \{0, 1\}$ and the probability of a string $x = a_1 a_2 \ldots a_n$ is $\prod_{i=1}^{n} \Pr(a_i)$. Shannon's entropy of the source becomes $\bar{H} = -P_0 \log P_0 - P_1 \log P_1$. Let $x_1, x_2, \ldots x_{2^n}$ be all strings of length n arranged in order of decreasing probability, $r \in (1/2, 1)$, and let $k(n)$ be the least integer such that $\sum_{i=1}^{k(n)} \Pr(x_i) > r$.

The intuition, expressed in [2], is that "the most likely strings have a complexity asymptotically equal to the entropy". The precise form was conjectured in [2] and proven in [18].

Theorem 5.2 *Let U be a universal machine. Then,*

$$\bar{H} = \lim_{n \to \infty} \frac{1}{nk(n)} \sum_{i=1}^{k(n)} H_U(x_i).$$

In fact a stronger result is true (note that both Theorems 5.1 and 5.2 have been stated in terms of blank end-marker complexity K; however, they can be re-phrased in terms of program-size complexity due to the observation stated in [26] that on average it doesn't matter which complexity we use as $\mid H_U(x) - K_W(x) \mid \leq o(n)$ for all strings on length n).

Theorem 5.3 *Let U be a universal machine. For every $\varepsilon > 0$ let*

$$\mathcal{H}^{\varepsilon}_{\backslash} = \{\S \in \mathcal{A}^* \mid |\S| = \backslash, \bar{\mathcal{H}} - \varepsilon < \frac{\mathcal{H}_{\mathcal{U}}(\S)}{\backslash} < \bar{\mathcal{H}} + \varepsilon\}.$$

Then,

$$\lim_{n \to \infty} \Pr(\mathcal{H}^{\varepsilon}_{\backslash}) = \infty.$$

5.2 Entropy of Computable Semi-Distributions

A function $P : A^* \to [0, 1]$ such that $\sum_{x \in A^*} P(x) \leq 1$ is called a *semi-distribution* over the strings. In case $\sum_x P(x) = 1$, P is a *distribution*. Any distribution P can be extended to a probability distribution μ, defined on the σ-field generated by cylinders.

A semi-distribution P is semi-computable from below (above) in case the set $\{(x, r) \mid x \in A^*,\ r \in \mathbf{Q},\ P(x) > r\}$ ($\{(x, r) \mid x \in A^*,\ r \in \mathbf{Q},\ P(x) < r\}$) is c.e. A semi-distribution P is computable if it is semi-computable from below and from above. For example, the algorithmic probability P_M

defined by (1) is a semi-distribution semi-computable from below. If $M = U$ is a universal machine, then $\Omega_U = \sum_{x\in A^*} P_U(x)$ is a c.e. and random real, a Chaitin's Omega number [4]. The function $P(x) = 2^{-2|x|-1}$ is a computable distribution. Computability is preserved via the extension μ of P; see, for example, [5, 3].

A *prefix-code (instantaneous code)* for strings is an one-one function $C : D \to A^*$, $D \subset A^*$ such that $C(D)$ is prefix-free. For example, $C : A^n \to A^*$ given by $C(x) = x$ is a prefix-code. Another example: for every surjective machine M, $C_M(x) = x^*_M$ is a prefix-code (here $x^*_M = \min\{u \mid M(u) = x\}$, where the minimum is taken according to the quasi-lexicographical ordering of strings); universal machines are surjective.

To motivate the next result we re-phrase Shannon-Fano theorem (see [19]; compare also with theorem 5.3) in terms of stationary Markov sources. Consider a stationary Markov information source with a finite alphabet $\left(A^{\mathbf{N}}, \mathcal{K}^{\mathbf{N}}, \mu\right)$ and denote by $\Pr(x)$ the probability corresponding to the distribution μ of the source. The *average length* of the prefix-code $C : A^n \to A^*$ is the number

$$L_{C,\Pr} = \sum_{x\in A^n} \Pr(x) \cdot |C(x)|.$$

Theorem 5.4 *Let $\left(A^{\mathbf{N}}, \mathcal{K}^{\mathbf{N}}, \mu\right)$ be a stationary Markov information source with a finite alphabet A. For every positive number $\varepsilon > 0$ there exists n_0 such that for every positive integer $n \geq n_0$ there exists a prefix-code $C : A^n \to A^*$ such that*

$$\bar{H} - \varepsilon < \frac{L_{C,\Pr}}{n} < \bar{H} + \varepsilon.$$

Consider now prefix-codes $C : A^* \to A^*$. The *average code-string length* of a prefix-code C with respect to a semi-distribution P is the number

$$L_{C,P} = \sum_{x\in A^*} P(x) \cdot |C(x)|.$$

The *minimal average code-string length* with respect to a semi-distribution P is

$$L_P = \inf\{L_{C,P} \mid C \text{ prefix-code}\}.$$

The *entropy* of a semi-distribution P is

$$\mathcal{H}_P = -\sum_{x\in A^*} P(x) \cdot \log P(x).$$

Shannon's classical argument [24] (see more in [13]) can be expressed for semi-distributions as follows:

Theorem 5.5 *The following inequalities hold true for every semi-distribution P:*

$$\mathcal{H}_P - 1 \leq \mathcal{H}_P + \left(\sum_x P(x)\right) \log \left(\sum_x P(x)\right) \leq L_P \leq \mathcal{H}_P + 1.$$

If P is a distribution, then $\log(\sum_x P(x)) = 0$, so we get the classical inequality $\mathcal{H}_P \geq L_P$. However, this inequality is not true for every semi-distribution. For example, take $P(x) = 2^{-2|x|-3}$, $C(x) = x_1x_1 \ldots x_nx_n01$, and note that $L_P \leq L_{C,P} = \mathcal{H}_P - \frac{1}{4}$.

Under which conditions given a semi-distribution P can we find a (universal) machine M such that $H_M(x)$ is equal, up to an additive constant, to $-\log P(x)$? In what follows we will assume that $P(x) > 0$, for every x. The main technical result was obtained in [6].

Theorem 5.6 *Assume that P is a semi-distribution and there exist a c.e. set $S \subset A^* \times \mathbf{N}$ and a constant $c \geq 0$ such that the following two conditions are satisfied for every $x \in A^*$:*

(i) $\sum_{(x,n)\in S} 2^{-n} \leq P(x)$,

(ii) *if* $P(x) > 2^{-n}$, *then* $(x, m) \in S$, *for some* $m \leq n + c$.

Then, there exists a machine M *(depending upon* S*) such that for all* x,

$$-\log P(x) \leq H_M(x) \leq (1 + c) - \log P(x).$$

Specializing P in theorem 5.6 we deduce that minimal programs are almost optimal for P.

Proposition 5.7 *Assume that* P *is a semi-distribution semi-computable from below. Then, there exists a machine* M *(depending upon* P*) such that for all* x,

$$-\log P(x) \leq H_M(x) \leq 2 - \log P(x). \tag{9}$$

Consequently, minimal programs for M *are almost optimal: the code* C_M *satisfies the inequalities:*

$$0 \leq L_{C_M,P} - \mathcal{H}_P \leq 2.$$

Minimal programs of universal machines are almost optimal for every semi-computable semi-distribution P.

Theorem 5.8 *Let* P *be a semi-distribution semi-computable from below, and* U *a universal machine. Then, there exists a constant* c_P *(depending upon* P*) such that*

$$0 \leq L_{C_U,P} - \mathcal{H}_P \leq 1 + c_P.$$

Theorem 5.8 generalizes a result in [12] proven for computable distributions; see also [23]. The result is important only for semi-distributions for which the entropy is *infinite*. For example, the entropy of the semi-distribution $P(x) = \frac{2^{-|x|}}{(|x|+2)\log(|x|+2)}$ is infinite.

5.3 Algorithmic coding theorem

A deep relation between entropy and program-size complexity appears in the algorithmic coding theorem of Chaitin and Gács (see [7, 8, 17, 3, 10]):

Theorem 5.9 *There exists a constant* $c \geq 0$ *such that for all strings* x, $|H_U(x) + \log P_U(x)| \leq 1 + c$ *(equivalently,* $H_U(x) = -\log P_U(x) + O(1)$*).*

The uncertainty given by the unknown, additive, computer-dependent, constant appearing in theorem 5.9 is a serious issue of concern for a physical theory, so various attempts have been made to eliminate it (see, for example, [23]). In [6] one characterizes all machines satisfying theorem 5.9 and one constructs a class of (universal) machines for which the inequality is satisfied with constant $c = 0$, that is, $H_U(x) = -\log P_U(x)$.

Proposition 5.10 *Let* M *be a machine and* $c \geq 0$. *The following statements are equivalent:*

(a) *for all* x, $H_M(x) \leq (1 + c) - \log P_M(x)$,

(b) *for all non-negative* n, *if* $P_M(x) > 2^{-n}$, *then* $H_M(x) \leq n + c$.

For any machine M satisfying one of the equivalent conditions in proposition 5.10, theorem 5.9 holds:

$$|H_M(x) + \log P_M(x)| \leq 1 + c, \tag{10}$$

and in fact, a machine M satisfies (10) if and only if (b) is satisfied. *Every universal machine* U *satisfies condition* (b), *but not all machines satisfy this condition.*

References

[1] I. V. Basawa, B. L. S. Prakasa-Rao. *Statistical Inference for Stochastic Processes*, Academic Press, New York, 1980.

[2] W. A. Beyer, L. Stein, S. M. Ulam. The notion of complexity, Los Alamos Report LA-4822, US Dept. of Commerce, Springfield, 1971.

[3] C. Calude. *Information and Randomness. An Algorithmic Perspective*, Springer Verlag, Berlin, 1994.

[4] C. S. Calude, G. J. Chaitin. Randomness everywhere, *Nature*, 400 22 July (1999), 319–320.

[5] C. Calude, I. Chiţescu. Probabilities on the space of sequences, Technical Report No 103, Computer Science Department, University of Auckland, New Zealand, 1994.

[6] C. S. Calude, H. Ishihara, T. Yamaguchi. Minimal programs are almost optimal, *Int. J. Found. Comput. Sci.*, in press.

[7] G. J. Chaitin. *Information, Randomness and Incompleteness, Papers on Algorithmic Information Theory*, World Scientific, Singapore, 1987 (2nd ed., 1990).

[8] G. J. Chaitin. *Algorithmic Information Theory*, Cambridge University Press, Cambridge, 1987 (third printing 1990).

[9] G. J. Chaitin. *The Limits of Mathematics*, Springer-Verlag, Singapore, 1997.

[10] G. J. Chaitin. *The Unknowable*, Springer-Verlag, Singapore, 1999.

[11] G. J. Chaitin. *Exploring Randomness*, Springer-Verlag, London, 2000.

[12] T. M. Cover, P. Gács, R. M. Gray. Kolmogorov's contributions to information theory and algorithmic complexity, *The Annals of Probability,* 17 (1989), 840–865.

[13] T. M. Cover, J. A. Thomas. *Elements of Information Theory*, John Wiley, New York, 1991.

[14] J. L. Doob. *Stochastic Processes*, John Wiley, New York, 1953.

[15] M. Dumitrescu. Some informational properties of Markov pure-jump processes, *Casopis pro Pěstováni Matematiky*, 113, 4 (1988), 429-434.

[16] M. Dumitrescu. An invariant of a class of Markov absorbing information sources, *An. Univ. Bucureşti, Matematica*, XLIV (1995), 21-24.

[17] P. Gács. On the symmetry of algorithmic information, *Soviet Math. Dokl.,* 15 (1974) 1477-1480; correction, Ibidem 15 (1974) 1480.

[18] S. Galatolo. A proof of the Beyer–Stein–Ulam relation between complexity and entropy, *Discrete Mathematics* 223 (2000), 367-372.

[19] S. Guiaşu. *Information Theory and Applications*, McGraw-Hill, New York, 1977.

[20] S. Ihara. *Information Theory for Continuous Systems*, World Scientific, Singapore, 1993.

[21] J. G. Kemeny, J. L. Snell. *Finite Markov Chains*, D. van Nostrand, Princeton, 1960.

[22] J. F. Lynch. A relation between complexity and entropy, *Ulam Quarterly* 3 (1995), 7-14.

[23] R. Schack. Algorithmic information and simplicity in statistical physics, *Int. J. Theor. Physics,* 36 (1997), 209–226.

[24] C. E. Shannon. A mathematical theory of communication, *Bell System Technical Journal*, 27 (1948), 379-423, 623-656.

[25] R. I. Soare. *Recursively Enumerable Sets and Degrees*, Springer-Verlag, Berlin, 1987.

[26] L. Staiger. The Kolmogorov complexity of real numbers, in: G. Ciobanu and Gh. Păun (eds.). *Proc. Fundamentals of Computation Theory*, Lecture Notes in Comput. Sci. No. 1684, Springer–Verlag, Berlin, 1999, 536-546.