Rauzy dimension and finite-state dimension

Verónica Becher Olivier Carton Santiago Figueira

June 26, 2024

Abstract

In a paper of 1976, Rauzy studied two complexity notions, $\underline{\beta}$ and $\overline{\beta}$, for infinite sequences over a finite alphabet. The function $\underline{\beta}$ is maximum exactly in the Borel normal sequences and $\overline{\beta}$ is minimum exactly in the sequences that, when added to any Borel normal sequence, the result is also Borel normal. Although the definition of $\underline{\beta}$ and $\overline{\beta}$ do not involve finite-state automata, we establish some connections between them and the lower $\underline{\dim}$ and upper $\overline{\dim}$ finite-state dimension (or other equivalent notions like finite-state compression ratio, aligned-entropy or cumulative log-loss of finite-state predictors). We show tight lower and upper bounds on $\underline{\dim}$ and $\overline{\dim}$ as functions of $\underline{\beta}$ and $\overline{\beta}$, respectively. In particular this implies that sequences with $\overline{\dim}$ zero are exactly the ones that that, when added to any Borel normal sequence, the result is also Borel normal. We also show that the finite-state dimensions $\underline{\dim}$ and $\overline{\dim}$ are essentially subadditive. We need two technical tools that are of independent interest. One is the family of local finite-state automata, which are automata whose memory consists of the last k read symbols for some fixed integer k. We show that compressors based on local finite-state automata are as good as standard finite-state compressors. The other one is a notion of finite-state relational (non-deterministic) compressor, which can compress an input in several ways provided the input can always be recovered from any of its outputs. We show that such compressors cannot compress more than standard (deterministic) finite-state compressors.

Contents

1 Introduction and statement of results

In [8] Gérard Rauzy studies two variants of complexity functions aiming at measuring the randomness of sequences over a finite alphabet. These functions are based on the difficulty to predict a symbol based of the next symbols. For any integer $b \geqslant 2$, Rauzy defines two functions $\{0, \ldots, b-1\}^{\mathbb{N}} \to [0, (b-1)/b]$, called $\overline{\beta}$ and $\underline{\beta}$, and he proves two theorems. The first one says: the sequences that are Borel normal to base b, henceforth the b-normal sequences, are exactly the infinite sequences over the alphabet $\{0, \ldots, b-1\}$ with maximum $\underline{\beta}$ value $(b-1)/b$. The second theorem says: the infinite sequences over the alphabet $\{0, \ldots, b-1\}$ that when added to any b-normal sequence preserve b-normality are exactly those for which its $\overline{\beta}$ value is zero. Here we rely on the identification of sequences with real numbers in $[0, 1]$: the infinite sequence x over the alphabet $\{0, \ldots, b-1\}$ corresponds to the real $0.x$ in base b. It is known that the set of sequences with $\overline{\beta}$ value is zero has Hausdorff dimension zero [3]. All eventually periodic sequences and all Sturmian sequences are in this set.

A very nice characterisation, proved by Dai, Lathrop, Lutz, and Mayordomo in [6] shows b-normal sequences coincides with sequences with maximum strong finite-state dimension, another complexity function for sequences, here called lower finite-state dimension and notated $\underline{\dim}$. Thus, the set of infinite sequences with maximum $\underline{\beta}$ value coincides with the set of sequences with maximum $\underline{\dim}$. In the present work we establish connections between these two complexity functions for the case of infinite binary sequences.

We show two main results. In Theorem 1, we give tight upper and lower bounds for the lower finite-state dimension in terms of the Rauzy's functions. In particular, this theorem characterizes infinite sequences whose binary expansion has minimum or maximum $\underline{\beta}$, as those with, respectively, minimum or maximum lower finite-state dimension. Theorem 1 also establishes a similar result but for $\overline{\beta}$ instead of $\underline{\beta}$, and a variant of $\underline{\dim}$, called upper finite-state dimension, notated $\overline{\dim}$.

Our second main result is Theorem 2 and it generalizes Rauzy's second theorem for the notions of finite-state dimension. Theorem 2 essentially says that the finite-state dimensions $\underline{\dim}$ and $\overline{\dim}$ are subadditive.

It is known that the lower and upper finite-state dimensions coincide with other notions of complexity of an infinite sequence x. These notions include the so called finite-state compression ratios [6] (roughly the best amount of compression that a finite state compressor can achieve along prefixes of x in relation to the length of such prefixes), the cumulative log-loss (roughly a measure of the least number of errors that a finite-state predictor can achieve along the prefixes of x) and the aligned entropy of x (roughly the entropy of the random variable whose distribution probability corresponds to the relative frequency of fixed length non-overlapping blocks along x, for longer and longer blocks). All these complexity notions –whose formal definitions will be duly given, and will always split into the upper and lower versions coincide for each sequence x: they capture different aspects of the same phenomenon x, but, when such aspects are translated into a complexity value in $[0, 1]$, they all coincide. It is important to bring up these alternative but equivalent complexity functions because along our work we will use, depending on the aspect of x we want to explore, the one that fits better. Since for binary sequences the image of Rauzy's $\underline{\beta}$ and $\overline{\beta}$ is a number in $[0, 1/2]$, there is no hope that these functions would give yet another complexity coinciding with dimension (or compression ratios or cumulative log-loss, or aligned entropy). However, Theorem 1 aims at sharply relating them.

For the proofs of Theorem 1 and 2 we need two technical tools that are of independent interest. One is the family of local finite-state automata, which are automata whose memory consists in the last k read symbols for some fixed integer k. We show that compressors based on local finite-state automata are as good as standard finite-state compressors. The other is a notion of finite-state relational (non-deterministic) compressor, which can compress an input in several ways provided the input can always be recovered from any of its outputs. We show that such compressors cannot compress more than standard (deterministic) finite-state compressors.

We now give the definitions and the precise statements of our results. Although many of the results we present here hold for arbitrary integer bases,for simplicity we just consider a binary alphabet, that is $b=2$. Fix $\mathbb{B}=\{0,1\}$. The binary expansion of real number $x\in\mathbb{R}^+=\{x\in\mathbb{R}: x\geqslant 0\}$ is a sequence of digits $a_0,a_1,\ldots$ in $\mathbb{B}^{\mathbb{N}}$ such that

$$x=\lfloor x\rfloor+\sum_{i\geqslant 1}a_{i-1}2^{-i}.$$

For the rational numbers having two representations we assume the one ending with infinitely many 0s. As usual we identify positive reals in $[0,1]$ with its binary expansion, namely sequences in $\mathbb{B}^{\mathbb{N}}$. We write $\mathbb{B}^*$ for the set of finite sequences of symbols in $\mathbb{B}$, and its elements are called words. The positions in words and sequences are numbered starting from 0. If $x=x_0x_1x_2\cdots$ is a finite or infinite sequence then $x[i]=x_i$, $w[i:j]$ is the factor $x_i\cdots x_j$ and $w[i:j)$ is the factor $x_i\cdots x_{j-1}$. The length of a word w is $|w|$.

For each positive integer $\ell\geqslant 1$ and for each word $w\in\mathbb{B}^*$, Rauzy in [8] defines $\beta_\ell(w)$ and $\gamma_\ell(w)$ as follows,

$$\beta_\ell(w)\triangleq\min_{f:\mathbb{B}^\ell\to\mathbb{B}}\frac{1}{|w|}\sum_{i=0}^{|w|-\ell-1}(1-\delta_{w[i],f(w[i+1:i+\ell])})$$

$$\gamma_\ell(w)\triangleq\min_{f:\mathbb{B}^\ell\to\mathbb{B}}\frac{1}{|w|}\sum_{i=0}^{|w|-\ell-1}(1-\delta_{f(w[i:i+\ell)),w[i+\ell]})$$

where f ranges over all functions from $\mathbb{B}^\ell$ to $\mathbb{B}$ and $\delta_{a,b}$ is the Kronecker delta whose value is 1 if $a=b$ and 0 otherwise. Note that $1-\delta_{w[i],f(w[i+1:i+\ell])}$ (respectively $(1-\delta_{f(w[i:i+\ell)),w[i+\ell]})$ is 0 if the symbols $w[i]$ and $f(w[i+1:i+\ell])$ (respectively $w[i+\ell]$ and $f(w[i:i+\ell))$ match and it is 1 otherwise. Therefore, the summation gives the number of mismatches between the symbol at position i (respectively at position $i+\ell$) and the symbol given by the function f applied to the factor of length ℓ after (respectively before) position i. Rauzy extends these definitions [1] to $\underline{\beta_\ell},\overline{\beta_\ell},\underline{\beta},\overline{\beta}:\mathbb{B}^{\mathbb{N}}\to[0,1/2]$ and $\underline{\gamma_\ell},\overline{\gamma_\ell},\underline{\gamma},\overline{\gamma}:\mathbb{B}^{\mathbb{N}}\to[0,1/2]$,

$$\underline{\beta_\ell}(x)\triangleq\liminf_{n\to\infty}\beta_\ell(x[0:n))\quad\text{and}\quad\overline{\beta_\ell}(x)\triangleq\limsup_{n\to,\infty}\beta_\ell(x[0:n)),$$

$$\underline{\beta}(x)\triangleq\lim_{\ell\to\infty}\underline{\beta_\ell}(x)\quad\text{and}\quad\overline{\beta}(x)\triangleq\lim_{\ell\to\infty}\overline{\beta_\ell}(x),$$

$$\underline{\gamma_\ell}(x)\triangleq\liminf_{n\to\infty}\gamma_\ell(x[0:n))\quad\text{and}\quad\overline{\gamma_\ell}(x)\triangleq\limsup_{n\to\infty}\gamma_\ell(x[0:n)),$$

$$\underline{\gamma}(x)\triangleq\lim_{\ell\to\infty}\underline{\gamma_\ell}(x)\quad\text{and}\quad\overline{\gamma}(x)\triangleq\lim_{\ell\to\infty}\overline{\gamma_\ell}(x).$$

These four limits exist because $\underline{\beta_\ell}(x)$, $\overline{\beta_\ell}(x)$, $\underline{\gamma_\ell}(x)$ and $\overline{\gamma_\ell}(x)$ are decreasing as functions in ℓ while x is fixed. The following proposition, whose proof is in Section 7, shows that it is worth distinguishing between β and γ.

Proposition 1. *The two functions $\underline{\beta}$ and $\underline{\gamma}$ are different, and so are $\overline{\beta}$ and $\overline{\gamma}$.*

Let $\mathfrak{h}:[0,1]\to[0,1]$ be the classical *entropy* function $\mathfrak{h}(\alpha)=-\alpha\log\alpha-(1-\alpha)\log(1-\alpha)$ where log is the logarithm in base 2 and it is assumed the usual convention that $0\log 0=0$. The graph of the function $\mathfrak{h}$ is pictured in Figure 1.

A function $\mu:\mathbb{B}^*\to\mathbb{R}^+$ such that for every $w\in\mathbb{B}^*$ $\mu(w0)+\mu(w1)=\mu(w)$, is a called a *measure*. For a real number s and a measure μ, the function defined by $\mu^{(s)}(w)\triangleq 2^{s|w|}\mu(w)$ is known as an s-*gale*. It satisfies the equality $\mu^{(s)}(w0)+\mu^{(s)}(w1)=2^s\mu^{(s)}(w)$ for each word $w\in\mathbb{B}^*$. A sequence $x\in\mathbb{B}^{\mathbb{N}}$ is called *generic* for a measure μ if the limit frequency of w is equal to $\mu(w)$ for each finite

[1] In [8] Rauzy uses β'_ℓ and for γ_ℓ. He gives his theorem on $\overline{\beta}$ and $\underline{\beta}$ and just points [8, Remarque page 212] that $\overline{\beta}$ and $\overline{\gamma}$ are 0 on the same sequences, and $\underline{\beta}$ and $\underline{\gamma}$ are 1/2 on the same sequences.

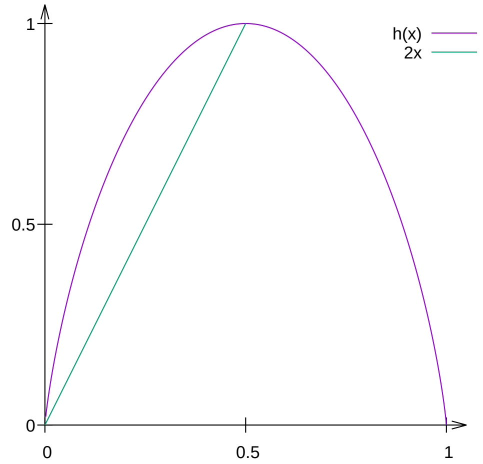

Figure 1: Graph of the function $\mathfrak{h}$

word $w \in \mathbb{B}^*$. It is a classical result that for a large class of invariant measures including measures induced by irreducible Markov chains, generic sequences for μ do exist. We use it in several places to construct examples. A measure μ is a *finite-state measure* if there is a deterministic finite-state automaton $A = \langle Q, \mathbb{B}, \delta, q_0 \rangle$ with $\delta : Q \times \mathbb{B} \to Q$ and a *betting function* $\theta : Q \to [0, 1]$ such that $\mu(w0) = (1 - \theta(\delta(q_0, w)))\mu(w)$ and $\mu(w1) = \theta(\delta(q_0, w))\mu(w)$ for each word $w \in \mathbb{B}^*$. Let $\mathcal{M}$ be the set of all finite-state measures. We say that the s-gale $\mu^{(s)}$ *succeeds* (respectively *strongly succeeds*) on the sequence x if $\limsup_{n\to\infty} \mu^{(s)}(x[0:n]) = \infty$ (respectively $\liminf_{n\to\infty} \mu^{(s)}(x[0:n]) = \infty$). The *lower finite-state dimension* $\underline{\dim}(x)$ and the *upper finite-state dimension* $\overline{\dim}(x)$ of $x \in \mathbb{B}^{\mathbb{N}}$ are respectively defined by

$$\underline{\dim}(x) \triangleq \inf\{s : \exists \mu \in \mathcal{M} \text{ such that } \mu^{(s)} \text{ succeeds on } x\}$$
$$\overline{\dim}(x) \triangleq \inf\{s : \exists \mu \in \mathcal{M} \text{ such that } \mu^{(s)} \text{ strongly succeeds on } x\}.$$

Theorem 1.

1. *For every $x \in \mathbb{B}^{\mathbb{N}}$,*

$$2\underline{\gamma}(x) \leqslant \underline{\dim}(x) \leqslant \mathfrak{h}(\underline{\gamma}(x)) \quad \textit{and} \quad 2\overline{\gamma}(x) \leqslant \overline{\dim}(x) \leqslant \mathfrak{h}(\overline{\gamma}(x)),$$
$$2\underline{\beta}(x) \leqslant \underline{\dim}(x) \leqslant \mathfrak{h}(\underline{\beta}(x)) \quad \textit{and} \quad 2\overline{\beta}(x) \leqslant \overline{\dim}(x) \leqslant \mathfrak{h}(\overline{\beta}(x)).$$

2. *For every real numbers α and ε such that $0 \leqslant \alpha \leqslant 1/2$ and $\varepsilon > 0$, there exist sequences x and x' such that $\underline{\gamma}(x) = \overline{\gamma}(x) = \underline{\beta}(x) = \overline{\beta}(x) = \underline{\gamma}(x') = \overline{\gamma}(x') = \underline{\beta}(x') = \overline{\beta}(x') = \alpha$ and $\underline{\dim}(x) = \overline{\dim}(x) \leqslant 2\alpha + \varepsilon$ and $\underline{\dim}(x') = \overline{\dim}(x') = h(\alpha)$.*

The two extreme cases $\overline{\gamma}(x) = 0$ and $\underline{\gamma}(x) = 1/2$ deserve some comments. Since $\mathfrak{h}(0) = 0$, a sequence $x \in \mathbb{B}^{\mathbb{N}}$ satisfies $\overline{\gamma}(x) = 0$ if and only if it satisfies $\overline{\dim}(x) = 0$. Since $\mathfrak{h}(1/2) = 1$, a sequence $x \in \mathbb{B}^{\mathbb{N}}$ satisfies $\underline{\gamma}(x) = 1/2$ if and only if it satisfies $\underline{\dim}(x) = 1$, that is, x is a b-normal sequence. This characterization of normality is already proved in [8].

For $x, y \in \mathbb{B}^{\mathbb{N}}$ we write $x + y$ for the binary expansion of the addition of the real numbers denoted by x and y.

Theorem 2. *For every $x, y \in \mathbb{B}^{\mathbb{N}}$,*

$$\underline{\dim}(x) - \overline{\dim}(y) \leqslant \underline{\dim}(x + y) \leqslant \underline{\dim}(x) + \overline{\dim}(y),$$
$$\overline{\dim}(x) - \overline{\dim}(y) \leqslant \overline{\dim}(x + y) \leqslant \overline{\dim}(x) + \overline{\dim}(y).$$

By Theorem 1, $\underline{\gamma}(x) = 1/2$ is equivalent to $\underline{\dim}(x) = 1$ which is, in turn, equivalent to x being b-normal. In the same way, $\overline{\gamma}(y) = 0$ is equivalent to $\overline{\dim}(y) = 0$. In the special case $\underline{\dim}(x) = 1$ and $\overline{\dim}(y) = 0$, Theorem 2 states that $\underline{\dim}(x + y) = \underline{\dim}(x) = 1$ which exactly means that adding y such that $\overline{\dim}(y) = 0$ to a b-normal number x preserves b-normality.

To prove Theorem 2 we use that addition is invertible in the sense that given $x + y$ and y, the number x is determined. We also use that addition $x + y = z$ can be computed by a synchronous transducer. Synchronous means here that the three heads reading x, y and z move synchronously. Theorem 2 holds for any operation having the same properties, such as the digit-wise addition without carry, also known as xor.

2 Preliminaries

In this section we introduce some definitions, properties and results that we need for our developments. The ones of Subsection 2.1 are from [6, Sec. 3] and [1, Sec. 7], and the ones of Subsection 2.2 are from [4, Sec. 2.1].

2.1 Measures, gales and dimension

In any finite-state measure via the deterministic finite-state automaton A and betting function θ, we associate a *weight* to each transition of A: the weight of $q \xrightarrow{0} q'$ (respectively $q \xrightarrow{1} q'$) is $1 - \theta(q)$ (respectively $\theta(q)$).

The weight of a finite run is the product of the weights of its transitions. The value of $\mu(w)$ is the weight of the unique run in A from the initial state q_0 and with label w. It is not required that $\mu(\lambda) = 1$ but since $\alpha\mu$ is also a measure for each positive real number α, it can always be assumed that $\mu(\lambda) = 1$. If μ_1 and μ_2 are two measures, then $\mu_1 + \mu_2$ is also a measure. The function mapping each word $w \in \mathbb{B}^*$ to $2^{-|w|}$ is called the *uniform measure*.

Let μ be a measure, If s is a non-positive real, each s-gale $\mu^{(s)}$ satisfies $\mu^{(s)}(w0) \leqslant \mu^{(s)}(w)$ and $\mu^{(s)}(w1) \leqslant \mu^{(s)}(w)$ for each word $w \in \mathbb{B}^*$. These relations imply that $\mu^{(s)}(w)$ is bounded by $\mu^{(s)}(\lambda)$ and that the s-gale cannot succeed on any sequence. It follows that the interesting s-gales are those for $s > 0$. For each positive real number $s > 1$, the s-gale defined by $\mu^{(s)}(w) \triangleq 2^{(s-1)|w|}$, where μ is the uniform measure, strongly succeeds on all sequences. Then, from these remarks, we have $0 \leqslant \underline{\dim}(x) \leqslant \overline{\dim}(x) \leqslant 1$. Let μ be the measure defined by $\mu(w) = 1$ if $w \in 0^*$ and $\mu(w) = 0$ otherwise. For each $s > 0$, the s-gale $\mu^{(s)}$ strongly succeeds on the sequence $0^{\mathbb{N}}$. It follows that $\underline{\dim}(0^{\mathbb{N}}) = \overline{\dim}(0^{\mathbb{N}}) = 0$.

The next proposition states that, for each finite family of measures, there is another measure which is almost as good as each one of the family. We say *almost* because we have the $2^{-\varepsilon|w|}$ factor. This proposition allows us to replace each finite family of measures (respectively predictors, compressors that we define next) by a single one which is almost as good as each one in the family.

Proposition 2 (reformulation of [6, Theorem 4.5])**.** *Let $\mu_1, \ldots, \mu_k$ be k finite-state measures. For each positive real number $\varepsilon > 0$, there exists a finite-state measure μ such that for each $w \in \mathbb{B}^*$,*

$$\mu(w) \geqslant \frac{2^{-\varepsilon|w|}}{k} \sum_{i=1}^{k} \mu_i(w).$$

For each positive integer $k \geqslant 1$, let $\mathcal{M}_k$ be the set of measures computed by a finite-state automaton with at most k states and such that the betting function θ satisfies $2^k\theta(q) \in \mathbb{N}$ for each state q of the automaton. This latter condition on θ specifies that $\theta(q)$ belongs to the finite set $\{i2^{-k} : 0 \leqslant i \leqslant 2^k\}$. It is needed to guarantee that $\mathcal{M}_k$ is finite. A measure μ is called *non-vanishing* if $\mu(w) \neq 0$ for each word $w \in \mathbb{B}^*$.

Lemma 1. *Let μ be a finite-state measure. For each positive real number $\varepsilon > 0$, there exists a non-vanishing finite-state measure μ' such that $\mu'(w) \geqslant 2^{-\varepsilon|w|}\mu(w)$ for each word $w \in \mathbb{B}^*$ and $\mu' \in \mathcal{M}_k$ for some integer $k \geqslant 1$.*

2.2 Predictors and cumulative log-loss

A *predictor* is a function π from $\mathbb{B}^* \times \mathbb{B}$ to $[0,1]$ such that $\pi(w,0) + \pi(w,1) = 1$ for each word $w \in \mathbb{B}^*$. A predictor is a *finite-state predictor* if there is a deterministic finite-state automaton A satisfying the following condition. For each word w, the values $\pi(w,0)$ and $\pi(w,1)$ are fully determined by the state reached by reading w in A from the initial state. This means that if the states reached by reading w and w' in A are equal, then $\pi(w,0) = \pi(w',0)$ and $\pi(w,1) = \pi(w',1)$. The *cumulative log-loss* $\xi_\pi(w)$ of the predictor π on a word w is defined by

$$\xi_\pi(w) \triangleq \sum_{i=0}^{|w|-1} \log \frac{1}{\pi(w[0:i), w[i])}.$$

Let Π be the set of all finite-state predictors. Let Π_k be the set of predictors computed by a finite-state automaton with at most k states and such that for each pair (w,b) in $\mathbb{B}^* \times \mathbb{B}$, the value $\pi(w,b)$ belongs to $\{i2^{-k} : 0 \leqslant i \leqslant 2^k\}$. Note that the set Π_k is finite. The *lower* and *upper* *cumulative log-loss* of a sequence $x \in \mathbb{B}^{\mathbb{N}}$ are respectively defined as follows,

$$\underline{\xi}(x) \triangleq \inf_{\pi\in\Pi} \liminf_{n\to\infty} \frac{\xi_\pi(x[0:n))}{n}$$
$$\overline{\xi}(x) \triangleq \inf_{\pi\in\Pi} \limsup_{n\to\infty} \frac{\xi_\pi(x[0:n))}{n}.$$

We show that predictors are merely a different presentation of measures and gales. More precisely, we show that there is a one-to-one correspondence between gales and predictors. Let s be a real number and $\mu^{(s)}$ be an s-gale where μ is a measure such that $\mu(\lambda) = 1$. Consider the predictor $\pi : \mathbb{B}^* \times \mathbb{B} \to [0,1]$ defined as follows for each $w \in \mathbb{B}^*$ and each $b \in \mathbb{B}$,

$$\pi(w,b) \triangleq \begin{cases} \mu(wb)/\mu(w) & \text{if } \mu(w) > 0 \\ 1/2 & \text{if } \mu(w) = 0. \end{cases}$$

Conversely, for a given predictor π, a measure μ can be defined inductively on the length of the word by

$$\mu(\lambda) \triangleq 1, \quad \mu(w0) \triangleq \pi(w,0)\mu(w) \quad \text{and} \quad \mu(w1) \triangleq \pi(w,1)\mu(w).$$

The measure μ and the predictor π are related by the following identity. For every integer $n \geqslant 1$,

$$\xi_\pi(x[0:n)) + \log \mu^{(s)}(x[0:n)) = sn.$$

This formula implies the following identities that were already pointed in [4].

Lemma 2. *For every $x \in \mathbb{B}^{\mathbb{N}}$, $\underline{\dim}(x) = \underline{\xi}(x)$ and $\overline{\dim}(x) = \overline{\xi}(x)$.*

3 Local and almost local automata

In this section we develop some tools that are needed to prove the inequalities $2\underline{\gamma}(x) \leqslant \underline{\dim}(x)$ and $2\overline{\gamma}(x) \leqslant \overline{\dim}(x)$ stated in Theorem 1.

A finite-state automaton is called k-*local* if its memory consists in the last k read symbols. A finite-state automaton is called *local* if it is k-local for some integer k. The graph of a local finite-state automaton is also known as a *de Bruijn graph*. Formally, let $k \geqslant 0$ be a non-negative integer. The state set of the k-local finite-state automaton is the set $\mathbb{B}^k$ of words of length k and its transition set is $\{au \xrightarrow{b} ub : a,b \in \mathbb{B}, u \in \mathbb{B}^{k-1}\}$. Said differently, for each word $w = aub$ of length $k+1$, there is a transition from its prefix au of length k to its suffix ub of length k and labeled by its last symbol b. The 2- and 3-local automata are pictured in Figure 2. The state reached after reading a long enough word w is the suffix of length k of w whatever the starting

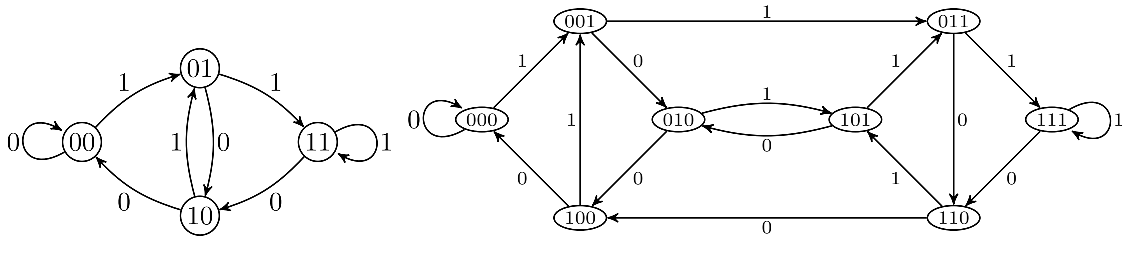

Figure 2: 2- and 3-local automata

state is. This means that the starting state is not relevant for the long term behaviour of such an automaton.

A finite-state automaton is (k, m)-*almost local* if its memory consists in the last k read symbols read so far modulo m. A finite-state automaton is called *almost local* if it is (k, m)-almost local for some non-negative integers k and m. Formally, let $k, m \geqslant 0$ be two non-negative integers. The state set of the (k, m)-almost local automaton is the set $\mathbb{B}^k \times \{0, \ldots, m-1\}$ and its transition set is the set $\{(au, i) \xrightarrow{b} (ub, j) : a, b \in \mathbb{B}, u \in \mathbb{B}^{k-1}, j \equiv i + 1 \bmod m\}$. Thus, the (k, m)-almost local automaton is defined by the product of the de Bruijn graph of order n and a simple cycle of length k.

A finite-state measure (resp. finite-state predictor) is said to be (k, m)-almost local if it can be defined by a (k, m)-almost local finite-state automaton. The same holds for the k-local case.

Let σ be the *shift function* which maps each either finite or infinite sequence $b_0b_1b_2 \cdots \in \mathbb{B}^{\geqslant 1} \cup \mathbb{B}^{\mathbb{N}}$ to the sequence $b_1b_2b_3 \cdots$ obtained by removing the first symbol. For each integer n, let us denote by σ^n the composition $\sigma \circ \cdots \circ \sigma$ n times. Notice that the domain of σ^n is $\mathbb{B}^{\geqslant n} \cup \mathbb{B}^{\mathbb{N}}$. By a slight abuse of notation, we write $\mu\sigma^n$ for the composition $\mu \circ \sigma^n$. Note that $\mu\sigma^n$ is not a measure because it is not defined for words in $\mathbb{B}^{<n}$. However, if $\mu\sigma^n$ is extended to short words by setting $\mu\sigma^n(w) = 2^{n-|w|}$ for $w \in \mathbb{B}^{\leqslant n}$, then $\mu\sigma^n$ becomes a measure which coincides with the uniform measure for short words. In the sequel, we always assume that $\mu\sigma^n$ has been extended to short words by setting $\mu\sigma^n(w) = 2^{n-|w|}$ for $w \in \mathbb{B}^{\leqslant n}$. Note that if μ is non-vanishing, then $\mu\sigma$ is also non-vanishing.

The following lemma states that the ratio $\mu\sigma^{m+n}(w)/\mu\sigma^n(w)$ for $|w| \geqslant k+m+n$ only depends on n and on the prefix of length $k + m + n$ of w.

Lemma 3. *Let μ be a non-vanishing (k, m)-almost local measure. For each non-negative integer n and each word $u \in \mathbb{B}^{k+m+n}$, there is a positive constant $K_{n,u}$ such that for each word $v \in \mathbb{B}^*$ $\mu\sigma^{m+n}(uv) = K_{n,u}\mu\sigma^n(uv)$.*

Proof. Since $|u| = k + m + n$, the two runs on $\sigma^n(uv)$ and $\sigma^{m+n}(uv)$ start coinciding after reading u. □

Let μ be a non-vanishing (k, m)-almost local measure for some integers $k, m \geqslant 0$. Let ε be a positive real number. Applying Proposition 2 to the family $\{\mu\sigma^i : 0 \leqslant i < m\}$ and the number ε, we get a measure μ'. It can be verified that the construction used in the proof of Proposition 2 – which is a reformulation of [6, Theorem 4.5] – provides a measure μ' which is (k', m')-almost local for some integers $k', m' \geqslant 0$.

By construction, this measure μ' satisfies for each word $w \in \mathbb{B}^*$,

$$\mu'(w) \geqslant \frac{2^{-\varepsilon|w|}}{m} \sum_{i=0}^{m-1} \mu\sigma^i(w).$$

Note that the measure μ' is non-vanishing because each measure $\mu\sigma^i$ is non vanishing. Along this section when we write μ and μ' we are referring to these measures. The following lemma shows that the ratio $\mu'\sigma^n(w)/2^{-\varepsilon|w|}\mu(w)$ for $|w| \geqslant k' + m' + n$ is lower bounded by a constant which only depends on n and on the prefix u of length $k' + m' + n$ of w.

Lemma 4. *For each non-negative integer n and each word $u \in \mathbb{B}^{k'+m'+n}$, there is a positive constant $K_{n,u}$ such that for each word $v \in \mathbb{B}^*$ $\mu'\sigma^n(uv) \geqslant K_{n,u}2^{-\varepsilon|uv|}\mu(uv)$.*

Proof. By Lemma 3 applied to μ', it suffices to prove the statement when $0 \leqslant n < m'$. Let n be such that $0 \leqslant n < m'$. The result is obvious for $n = 0$. Therefore we suppose that $1 \leqslant n \leqslant m' - 1$. Let us write $n = km + r$ where $0 < r \leqslant m$. By construction of the measure μ', this measure satisfies $\mu'(w) \geqslant 2^{-\varepsilon|w|}\mu\sigma^{m-r}(w)/m$. Applying this latter equality to $\sigma^{r+km}(w)$, we get $\mu'\sigma^n(w) \geqslant 2^{-\varepsilon|w|}\mu\sigma^{(k+1)m}(w)/m$. The result follows then by Lemma 3 applied to μ. □

From the (k', m')-almost local measure μ', we construct a k'-local measure μ'' as follows. Let θ' be the betting function from $\mathbb{B}^{k'} \times \{0, \dots, m'-1\}$ to $[0, 1]$ defining the measure μ'. Let us define the betting function θ'' from $\mathbb{B}^{k'}$ to $[0, 1]$ by

$$\theta''(u) \triangleq \alpha_u \left(\prod_{i=0}^{m'-1} \theta'(u, i) \right)^{\frac{1}{m'}}$$

where the positive real number α_u is such that

$$\alpha_u \left[\left(\prod_{i=0}^{m'-1} \theta'(u, i) \right)^{\frac{1}{m'}} + \left(\prod_{i=0}^{m'-1} (1 - \theta'(u, i)) \right)^{\frac{1}{m'}} \right] = 1.$$

Apart from the normalizing coefficient α_u, $\theta''(u)$ is defined as the geometric mean of the values $\theta'(u, i)$ for $i = 0, \dots, m' - 1$. By symmetry, $1 - \theta''(u)$ must also be the geometric mean of the values $1 - \theta'(u, i)$ for $i = 0, \dots, m' - 1$. The normalizing coefficient α_u is chosen such that this is true. We claim that $\alpha_u \geqslant 1$. By the Jensen inequality applied to the log function which is concave, the geometric mean is not greater than the arithmetic mean, that is,

$$\left(\prod_{i=0}^{m'-1} \theta'(u, i) \right)^{\frac{1}{m'}} \leqslant \frac{1}{m'} \sum_{i=0}^{m'-1} \theta'(u, i)$$

$$\left(\prod_{i=0}^{m'-1} (1 - \theta'(u, i)) \right)^{\frac{1}{m'}} \leqslant \frac{1}{m'} \sum_{i=0}^{m'-1} (1 - \theta'(u, i)).$$

Summing these two inequalities we obtain

$$\left(\prod_{i=0}^{m'-1} \theta'(u, i) \right)^{\frac{1}{m'}} + \left(\prod_{i=0}^{m'-1} (1 - \theta'(u, i)) \right)^{\frac{1}{m'}} \leqslant 1$$

which implies that $\alpha_u \geqslant 1$.

The following proposition concludes that the ratio $\mu''(w)/2^{-\varepsilon|w|}\mu(w)$ for large enough w is lower bounded by a constant.

Proposition 3. *There exists a positive constant K such that for each long enough word $w \in \mathbb{B}^*$, $\mu''(w) \geqslant K2^{-\varepsilon|w|}\mu(w)$.*

Proof. Since each coefficient α_u satisfies $\alpha_u \geqslant 1$, there exists a constant K' such that

$$\mu''(w) \geqslant K' \left(\prod_{i=0}^{m'-1} \mu'\sigma^i(w) \right)^{\frac{1}{m'}}.$$

Combining this inequality with the statement of Lemma 4, we get the result. □

The next corollary follows directly from the previous proposition and the one-to-one correspondence between s-gales and predictors. It will allow us to use a local predictor in the proof of Proposition 5.1.

Corollary 1. *For each almost local predictor π and each positive real ε, there exists a local predictor π' such that for each long enough word $x \in \mathbb{B}^*$, $\xi_{\pi'}(w) \leqslant \xi_\pi(w) + \varepsilon|w|$.*

4 Relational, functional and deterministic compressors

To prove Theorem 2, we need to work with compressors with possibly non-functional behaviour. We call them relational compressors. We also introduce, as special case of them, the standard functional (non-deterministic) and deterministic versions.

A *transducer* with input and output alphabet $\mathbb{B}$ is a tuple $\langle Q, \Delta, I, F\rangle$, where Q is a finite set of states, $\Delta \subset Q \times \mathbb{B}^* \times \mathbb{B}^* \times Q$ is a finite set of transitions, $I \subseteq Q$ is the non-empty subset of initial states, and $F \subseteq Q$ is the subset of final states. A transition (p, u, v, q) is written $p \xrightarrow{u|v} q$.

Relational compressors. A *relational compressor* is a transducer $\langle Q, \Delta, I, F\rangle$ that could possibly yield several accepting runs for the same input sequence x. We introduce notation to deal with such runs. A *finite run* η from a state p to a state q is a sequence of consecutive transitions $p \xrightarrow{u_0|v_0} q_1 \xrightarrow{u_1|u_1} q_2 \cdots q_{n-1} \xrightarrow{u_{n-1}|v_{n-1}} q$. Its *label* $\Lambda(\eta)$ is the pair (u, v) where its *input label* $\Lambda_0(\eta)$ and *output label* $\Lambda_1(\eta)$ are respectively $u \triangleq u_0 \cdots u_{n-1}$ and $v \triangleq v_0 \cdots v_{n-1}$.

Similarly, an *infinite run* η is an infinite sequence $q_0 \xrightarrow{u_0|v_0} q_1 \xrightarrow{u_1|v_1} q_2 \xrightarrow{u_2|v_2} q_3 \cdots$ of consecutive transitions. Its *label* $\Lambda(\eta)$ is the pair (x, y) where its *input label* $\Lambda_0(\eta)$ and *output label* $\Lambda_1(\eta)$ are respectively $x \triangleq u_0u_1u_2 \cdots$ and $y \triangleq v_0v_1v_2 \cdots$. Note that both the input and output labels can be finite as labels of transitions might be empty.

An infinite run $q_0 \xrightarrow{x|y} \infty$ is *accepting* if 1) both x and y are infinite, 2) q_0 is initial and 3) q_i is final for infinitely many indices i. This is the classical Büchi acceptance condition. A pair (x, y) is *accepted* if it is the label $\Lambda(\eta)$ of some accepting run η. If η is either a finite or an infinite run, $\eta[m : n)$ is the finite run $q_m \xrightarrow{u_m|v_m} q_{m+1} \xrightarrow{u_{m+1}|u_{m+1}} q_{m+2} \cdots q_{n-1} \xrightarrow{u_{n-1}|v_{n-1}} q_n$ of length $n - m$.

Although a relational compressor can compress an input in different ways, it has to be always possible to recover the input from any of its outputs. So an additional requirement holds for any relational compressor: if (x, y) and (x', y) are accepted then $x = x'$. A state is *useful* if it occurs in some accepting run. Since the realized relation remains unchanged when useless states are removed, we always assume in the sequel that the set Q only contains useful states.

Let η be an accepting run $q_0 \xrightarrow{u_0|v_0} q_1 \xrightarrow{u_1|v_1} q_2 \xrightarrow{u_2|v_2} q_3 \cdots$ in a relational compressor C. According to the notation introduced above, $\Lambda_0(\eta[0 : n))$ and $\Lambda_1(\eta[0 : n))$ are respectively $u_0 \cdots u_{n-1}$ and $v_0 \cdots v_{n-1}$. The *compression ratios* $\underline{\rho}_C(\eta)$ and $\bar{\rho}_C(\eta)$ of C along the run η are respectively defined by

$$\underline{\rho}_C(\eta) \triangleq \liminf_{n\to\infty} \frac{|\Lambda_1(\eta[0 : n))|}{|\Lambda_0(\eta[0 : n))|} \quad \text{and} \quad \bar{\rho}_C(\eta) \triangleq \limsup_{n\to\infty} \frac{|\Lambda_1(\eta[0 : n))|}{|\Lambda_0(\eta[0 : n))|}.$$

Let $\mathcal{C}$ be the class of all relational compressors. The *lower compression ratio* $\underline{\rho} : \mathbb{B}^{\mathbb{N}} \to [0, 1]$ and the *upper compression ratio* $\bar{\rho}(x) : \mathbb{B}^{\mathbb{N}} \to [0, 1]$ are respectively defined by

$$\underline{\rho}(x) \triangleq \inf_{C\in\mathcal{C}} \inf_{\Lambda_0(\eta)=x} \underline{\rho}_C(\eta) \quad \text{and} \quad \bar{\rho}(x) \triangleq \inf_{C\in\mathcal{C}} \inf_{\Lambda_0(\eta)=x} \bar{\rho}_C(\eta), \tag{1}$$

where η ranges over all accepting runs in C such that $\Lambda_0(\eta) = x$.

Functional compressors. A *functional compressor* C is a relational compressor such that the relation it realizes is actually a function: if (x, y) and (x, y') are accepted by C then $y = y'$. Hence functional compressors compute on-to-one functions. In this case we write $C(x) = y$. Notice that C may be non deterministic, having possibly more than one accepting run to compute $C(x)$.

If C is a functional compressor then $\underline{\rho}_C$ and $\bar{\rho}_C$ are defined in the same way as for the relational compressors and $\underline{\rho}^{\mathrm{f}}$, $\bar{\rho}^{\mathrm{f}}$ are defined as in (1) except that $\mathcal{C}$ is the class of all functional compressors.

Deterministic compressors A *deterministic compressor* is a relational compressor $\langle Q, \Delta, I, F\rangle$ where Δ is in fact a function $\delta : Q \times \mathbb{B}^* \times \mathbb{B}^* \to Q$ and I is a singleton. If C is a deterministic compressor then $\underline{\rho}_C$ and $\bar{\rho}_C$ are defined in the same way as for the relational compressors and $\underline{\rho}^{\mathrm{d}}$, $\bar{\rho}^{\mathrm{d}}$ are defined as in (1) except that $\mathcal{C}$ is the class of all deterministic compressors. Of course, any deterministic compressor is functional and any functional compressor is relational. The equivalence between $\underline{\rho}^{\mathrm{d}} = \underline{\rho}^{\mathrm{f}}$ and $\bar{\rho}^{\mathrm{d}} = \bar{\rho}^{\mathrm{f}}$ follows from [2].

A deterministic compressor is said to be (k, m)-almost local if it can be computed by a (k, m)-almost local finite-state automaton.

4.1 Aligned entropy equals compression ratio of relational compressors

For w and u two words in $\mathbb{B}^*$, the number $|w|_u$ of *occurrences* of u in w is and the number $\|w\|_u$ of *aligned occurrences* of u in w are respectively defined by

$$\begin{aligned} |w|_u &\triangleq \#\{i : w[i : i + |u|) = u\}, \\ \|w\|_u &\triangleq \#\{i : w[i : i + |u|) = u \text{ and } i = 0 \bmod |u|\}. \end{aligned}$$

For instance $|0000|_{00} = 3$ while $\|0000\|_{00} = 2$. The notation $\|w\|_u$ is usually used when the length $|w|$ is a multiple of the length $|u|$ but this is not required by the definition.

Now we consider the notion of *aligned entropy* of a given word, also known as aligned block entropy which is defined by the frequency of non-overlapping blocks in the given word. For each positive integer $\ell \geqslant 1$, the normalized *aligned ℓ-entropy* $ah_\ell(w)$ of a word $w \in \mathbb{B}^*$ is defined as follows:

$$ah_\ell(w) \triangleq -\frac{1}{\ell} \sum_{u \in \mathbb{B}^\ell} f_u \log f_u \quad \text{where} \quad f_u = \frac{\|w\|_u}{\lfloor |w|/\ell \rfloor}$$

with the usual convention $0 \log 0 = 0$. By *normalized*, we mean that the summation is divided by ℓ so that the inequalities $0 \leqslant ah_\ell(w) \leqslant 1$ hold. The *aligned entropies* $\underline{ah}(x)$ and $\overline{ah}(x)$ of a sequence x are respectively defined as follows.

$$\begin{aligned} \underline{ah}(x) &\triangleq \liminf_{\ell\to\infty} \underline{ah}_\ell(x) \quad \text{where} \quad \underline{ah}_\ell(x) \triangleq \liminf_{n\to\infty} ah_\ell(x[0 : n)) \\ \overline{ah}(x) &\triangleq \liminf_{\ell\to\infty} \overline{ah}_\ell(x) \quad \text{where} \quad \overline{ah}_\ell(x) \triangleq \limsup_{n\to\infty} ah_\ell(x[0 : n)) \end{aligned}$$

Although we shall not use it in our proofs, we mention the notion of *non-aligned entropy*, also known as non-aligned block entropy.

$$h_\ell(w) \triangleq -\frac{1}{\ell} \sum_{u \in \mathbb{B}^\ell} f_u \log f_u \quad \text{where} \quad f_u = \frac{|w|_u}{|w| - |u| + 1}$$

with the usual convention $0 \log 0 = 0$. The *non-aligned entropies* $\underline{h}(x)$ and $\overline{h}(x)$ of a sequence x are respectively defined as follows.

$$\begin{aligned} \underline{h}(x) &\triangleq \liminf_{\ell\to\infty} \underline{h}_\ell(x) \quad \text{where} \quad \underline{h}_\ell(x) \triangleq \liminf_{n\to\infty} h_\ell(x[0 : n)) \\ \overline{h}(x) &\triangleq \liminf_{\ell\to\infty} \overline{h}_\ell(x) \quad \text{where} \quad \overline{h}_\ell(x) \triangleq \limsup_{n\to\infty} h_\ell(x[0 : n)) \end{aligned}$$

These functions should not be confused with the classical notion of entropy $\mathfrak{h}$ appearing in Theorem 1. Recently, Kozachinski and Shen show in [7, Theorem 1] that for every every $x \in \mathbb{B}^{\mathbb{N}}$, $\underline{ah}(x) = \underline{h}(x)$ and $\overline{ah}(x) = \overline{h}(x)$.

A proof that for every $x \in \mathbb{B}^{\mathbb{N}}$, $\bar{\rho}^{\mathrm{d}}(x) = \overline{ah}(x)$ appears in [9]. The inequality $\overline{ah}(x) \leqslant \bar{\rho}^{\mathrm{d}}(x)$. first appeared in [10]. Moving from $\bar{\rho}^{\mathrm{d}}(x) = \overline{ah}(x)$ to $\underline{\rho}^{\mathrm{d}}(x) = \underline{ah}(x)$ is not difficult.

The next theorem is our main result in this section and proves that the lower and upper compression ratio obtained with relational compressors coincides with lower and upper aligned entropy.

Theorem 3. *For every $x \in \mathbb{B}^{\mathbb{N}}$, $\underline{ah}(x) = \underline{\rho}(x)$ and $\overline{ah}(x) = \bar{\rho}(x)$.*

In [9] it is shown the same result but restricted to deterministic compressors, namely, it is shown that $\overline{ah} = \bar{\rho}^{\mathrm{d}}$, and from that equality is not difficult to derive $\underline{ah} = \underline{\rho}^{\mathrm{d}}$. Altogether we have that $\bar{\rho} = \bar{\rho}^{\mathrm{f}} = \bar{\rho}^{\mathrm{d}}$ and $\underline{\rho} = \underline{\rho}^{\mathrm{f}} = \underline{\rho}^{\mathrm{d}}$.

The proof of Theorem 3 is divided in Lemmas 6, 7 and 8 . We start with a generalization of Kraft's inequality. It is the only place that we use that if (x, y) and (x', y) are accepted, then necessarily $x = x'$.

For a relational compressor C, let us denote by $L_C(w)$ the minimum number of symbols written by a finite run in C with input label w, that is,

$$L_C(w) \triangleq \min\{|v| : \text{there exists a run } p \xrightarrow{w|v} q \text{ in } C\}.$$

Lemma 5. *Let C be a relational compressor with $\#Q$ states and let r_C be the maximum number of symbols written by a single transition of C. For each integer $\ell \geqslant 0$,*

$$\sum_{w \in \mathbb{B}^\ell} 2^{-L_C(w)} \leqslant (\#Q)^2(1 + \ell r_C).$$

Proof. We claim that for each integer k the cardinality of the set $\{w \in \mathbb{B}^\ell : L_C(w) = k\}$ is at most $(\#Q)^2 2^k$. Let p and q be two states of C and v be a word. Let us prove that the set $\{w \in \mathbb{B}^\ell : p \xrightarrow{w|v} q\}$ has cardinality at most 1. Suppose that there are 2 distinct words w_1, w_2 in this set. Since the transducer is trim, there is a finite run $i \xrightarrow{u|v'} p$ from an initial state i and a final run $q \xrightarrow{x|y} \infty$. It follows that both pairs $(uw_1x, v'vy)$ and $(uw_2x, v'vy)$ are accepted and this is a contradiction. If a word w belongs to $\{w \in \mathbb{B}^\ell : L_C(w) = k\}$, then there are two states p, q and a word v of length k such that w belongs to $\{w \in \mathbb{B}^\ell : p \xrightarrow{w|v} q\}$. This proves the upper bound for the cardinality of $\{w \in \mathbb{B}^\ell : L_C(w) = k\}$ since there are $(\#Q)^2$ possible choices for p and q and 2^k possible choices for v:

$$\begin{aligned}\sum_{w \in \mathbb{B}^\ell} 2^{-L_C(w)} &= \sum_{k=0}^{\ell r_C} \#\{w \in \mathbb{B}^\ell : L_C(w) = k\} 2^{-k} \\ &\leqslant \sum_{k=0}^{\ell r_C} (\#Q)^2 = (\#Q)^2(1 + \ell r_C).\end{aligned}$$

□

Lemma 6 (adapted from [9])**.** *For every $x \in \mathbb{B}^{\mathbb{N}}$, $\underline{ah}(x) \leqslant \underline{\rho}(x)$ and $\overline{ah}(x) \leqslant \bar{\rho}(x)$.*

Proof. We follow the proof that $\underline{ah}(x) \leqslant \underline{\rho}^{\mathrm{f}}(x)$ and $\overline{ah}(x) \leqslant \bar{\rho}^{\mathrm{f}}(x)$ in [9] but in the more general context of relational compressors. We give here the proof for $\overline{ah}(x) \leqslant \bar{\rho}(x)$. The proof for $\underline{ah}(x) \leqslant \underline{\rho}(x)$ requires to take a limit inferior at the end rather than a limit superior. Let C be a relational compressor with $k = \#Q$ states. Without loss of generality, it can be assumed that the length of the input label of each transition is exactly one (see [2, Theorem 4.3]). If the length of an input label is greater than 1, the transition can be split into several transitions adding new fresh states. Let $L_C(w)$ be the length of the shortest output that C produces when reading w, where the shortest is taken over all possible finite runs with w as input label. Let η be an infinite run of C such that $\Lambda_0(\eta) = x$ and $\Lambda_1(\eta)$ is infinite. We claim that $\underline{ah}(x) \leqslant \underline{\rho}_C(\eta)$ and $\overline{ah}(x) \leqslant \bar{\rho}_C(\eta)$. Let ℓ be a fixed integer. Then

$$\sum_{w\in\mathbb{B}^\ell} \|\Lambda_0(\eta[0:n))\|_w L_C(w) \leqslant |\Lambda_1(\eta[0:n))|.$$

Then, setting $K(n) \triangleq \lfloor |\Lambda_0(\eta[0:n))|/\ell \rfloor$ and dividing by $\ell K(n)$ gives

$$\frac{1}{\ell}\sum_{w\in\mathbb{B}^\ell} \frac{\|\Lambda_0(\eta[0:n))\|_w}{K(n)} L_C(w) \leqslant \frac{|\Lambda_1(\eta[0:n))|}{\ell K(n)}.$$

$$0 \leqslant \frac{|\Lambda_1(\eta[0:n))|}{\ell K(n)} + \frac{1}{\ell}\sum_{w\in\mathbb{B}^\ell} \frac{\|\Lambda_0(\eta[0:n))\|_w}{K(n)} \log(2^{-L_C(w)})$$

and adding $ah_\ell(\Lambda_0(\eta[0:n)))$ on both sides,

$$ah_\ell(\Lambda_0(\eta[0:n))) \leqslant \frac{|\Lambda_1(\eta[0:n))|}{\ell K(n)} + \frac{1}{\ell}\sum_{w\in\mathbb{B}^\ell} \frac{\|\Lambda_0(\eta[0:n))\|_w}{K(n)} \log\left(\frac{K(n)2^{-L_C(w)}}{\|\Lambda_0(\eta[0:n))\|_w}\right).$$

By Jensen inequality applied to the log function,

$$ah_\ell(\Lambda_0(\eta[0:n))) \leqslant \frac{|\Lambda_1(\eta[0:n))|}{\ell K(n)} + \frac{1}{\ell}\log\left(\sum_{w\in\mathbb{B}^\ell} 2^{-L_C(w)}\right).$$

By the generalized Kraft's inequality of Lemma 5,

$$ah_\ell(\Lambda_0(\eta[0:n))) \leqslant \frac{|\Lambda_1(\eta[0:n))|}{\ell K(n)} + \frac{1}{\ell}\log((\#Q)^2(1+\ell r_C)).$$

Taking the limit superior when $n \to \infty$ and using $\lim_{n\to\infty} |\Lambda_0(\eta[0:n))|/\ell K(n) = 1$ gives

$$\overline{ah}_\ell(x) \leqslant \bar{\rho}_C(\eta) + \frac{1}{\ell}\log((\#Q)^2(1+\ell r_C))$$

and then taking the limit inferior when $\ell \to \infty$ and the infimum over all runs η in all relational compressors C yields $\overline{ah}(x) \leqslant \bar{\rho}(x)$. ☐

Since $\bar{\rho} = \bar{\rho}^{\mathrm{f}} = \bar{\rho}^{\mathrm{d}}$ and $\underline{\rho} = \underline{\rho}^{\mathrm{f}} = \underline{\rho}^{\mathrm{d}}$, Lemmas 7 and 8 are implied by $\underline{\rho}^{\mathrm{d}}(x) \leqslant \underline{ah}(x)$ and $\bar{\rho}^{\mathrm{d}}(x) \leqslant \overline{ah}(x)$ shown in [9]. The inequality $\underline{\rho}(x) \leqslant \underline{ah}(x)$ follows directly from the following lemma. This result is known, but here we need that the constructed compressor is almost local.

Lemma 7. *Let $x \in \mathbb{B}^{\mathbb{N}}$. For each positive real number $\varepsilon > 0$, there exists an almost local compressor C such that $\underline{\rho}_C(x) \leqslant \underline{ah}(x) + \varepsilon$.*

Proof. By definition of $\underline{ah}(x)$, there exists an integer $\ell \geqslant 1$ and an increasing sequence $(n_k)_{k\geqslant 0}$ of integers such that $ah_\ell(x[0:\ell n_k)) < \underline{ah}(x) + \varepsilon$ for each integer $k \geqslant 0$. By replacing the sequence $(n_k)_{k\geqslant 0}$ by one of its sub-sequences, it can be assumed that the limit $\lim_{k\to\infty} \|x[0:\ell n_k)\|_w/n_k$ exists for every word w of length ℓ. Let f_w be this limit for w of length ℓ. Note that the numbers f_w satisfy the relations $\sum_{w\in\mathbb{B}^\ell} f_w = 1$ and $-\frac{1}{\ell}\sum_{w\in\mathbb{B}^\ell} f_w \log f_w \leqslant \underline{ah}_\ell(x) + \varepsilon$. Let $m \geqslant 1$ be an integer to be fixed later. For each word u of length ℓm, let f_u be defined by $f_u \triangleq f_{w_1} \cdots f_{w_m}$ where u is factorized $u = w_1 \cdots w_m$ with $|w_i| = \ell$ for each $1 \leqslant i \leqslant m$. It is easily checked that

$\sum_{u \in \mathbb{B}^{\ell m}} f_u = 1$. This equality implies that there exists a family $(v_u)_{u \in \mathbb{B}^{\ell m}}$ of words such that the set $\{v_u : u \in \mathbb{B}^{\ell m}\}$ is prefix-free and that $|v_u| \leqslant \lceil -\log f_u \rceil$ holds for each word $u \in \mathbb{B}^{\ell m}$. Let us consider the deterministic compressor C that replaces each word u of length ℓm by the word v_u. Such a compressor is of course almost local. Since $\{v_u : u \in \mathbb{B}^{\ell m}\}$ is prefix-free, this encoding is clearly one-to-one. It remains to show that C compresses as expected. Without loss of generality, it can be assumed that each integer n_k is a multiple of m.

$$\begin{aligned}
|C(x[0:\ell n_k))| &\leqslant \sum_{u \in \mathbb{B}^{\ell m}} \|x[0:\ell n_k)\|_u |v_u| \\
&\leqslant \sum_{u \in \mathbb{B}^{\ell m}} \|x[0:\ell n_k)\|_u \lceil -\log f_u \rceil \\
&\leqslant \frac{n_k}{m} - \sum_{u \in \mathbb{B}^{\ell m}} \|x[0:\ell n_k)\|_u \log f_u \\
&\leqslant \frac{n_k}{m} - \sum_{w \in \mathbb{B}^{\ell}} \|x[0:\ell n_k)\|_w \log f_w .
\end{aligned}$$

Dividing by ℓn_k gives

$$\frac{|C(x[0:\ell n_k))|}{\ell n_k} \leqslant \frac{1}{\ell m} - \frac{1}{\ell} \sum_{w \in \mathbb{B}^{\ell}} \frac{\|x[0:\ell n_k)\|_w}{n_k} \log f_w .$$

Choosing m large enough such that $1/\ell m \leqslant \varepsilon$ and letting k go to infinity gives

$$\underline{\rho}(x) \leqslant \varepsilon - \frac{1}{\ell} \sum_{w \in \mathbb{B}^{\ell}} f_w \log f_w \leqslant \underline{ah}(x) + 2\varepsilon .$$

Since the inequality holds for every $\varepsilon > 0$, the proof is complete. □

The inequality $\bar{\rho}(x) \leqslant \overline{ah}(x)$ follows directly from the following lemma.

Lemma 8. *Let $x \in \mathbb{B}^{\mathbb{N}}$. For each positive real number $\varepsilon > 0$, there exists an almost local compressor C such that $\bar{\rho}_C(x) \leqslant \overline{ah}(x) + \varepsilon$.*

Proof. By definition of $\overline{ah}(x)$, there exists an integer $\ell \geqslant 1$ and an integer $N \geqslant 0$ such that for all integers $n \geqslant N$, the inequality $ah_\ell(x[0:n)) < \overline{ah}(x) + \varepsilon$ holds. For each word w of length ℓ and each integer $n \geqslant N$, let $f_{w,n}$ be the ratio $\|x[0:n)\|_w / \lfloor n/\ell \rfloor$. Note that the numbers $f_{w,n}$ satisfy the equality $\sum_{w \in \mathbb{B}^{\ell}} f_{w,n} = 1$ and that the aligned entropy $ah_\ell(x[0:n))$ is equal to $-\frac{1}{\ell} \sum_{w \in \mathbb{B}^{\ell}} f_{w,n} \log f_{w,n}$. Let $m \geqslant 1$ be an integer to be fixed later. For each word u of length ℓm, and each integer $n \geqslant N$, let $f_{u,n}$ be defined by $f_{u,n} \triangleq f_{w_1,n} \cdots f_{w_m,n}$ where u is factorized $u = w_1 \cdots w_m$ with $|w_i| = \ell$ for each $1 \leqslant i \leqslant m$. It is easily checked that $\sum_{u \in \mathbb{B}^{\ell m}} f_{u,n} = 1$. This equality implies that there exists a family $(v_{u,n})_{u \in \mathbb{B}^{\ell m}}$ of words such that the set $\{v_{u,n} : u \in \mathbb{B}^{\ell m}\}$ is prefix-free and that $|v_{u,n}| \leqslant \lceil -\log f_{u,n} \rceil$ holds for each word $u \in \mathbb{B}^{\ell m}$. Consider the deterministic compressor C_n that replaces each word u of length ℓm by the word $v_{u,n}$. Since $\{v_{u,n} : u \in \mathbb{B}^{\ell m}\}$ is prefix-free, this encoding is clearly one-to-one. Note that the size of C_n depends only on ℓm and not on n. It remains to show that C_n compresses as expected,

$$\begin{aligned}
|C_n(x[0:n))| &\leqslant \sum_{u\in\mathbb{B}^{\ell m}} \|x[0:n)\|_u |v_{u,n}| \\
&\leqslant \sum_{u\in\mathbb{B}^{\ell m}} \|x[0:n)\|_u \lceil -\log f_{u,n}\rceil \\
&\leqslant \frac{n}{\ell m} - \sum_{u\in\mathbb{B}^{\ell m}} \|x[0:n)\|_u \log f_{u,n} \\
&\leqslant \frac{n}{\ell m} - \sum_{w\in\mathbb{B}^{\ell}} \|x[0:n)\|_w \log f_{w,n}.
\end{aligned}$$

Dividing by n gives

$$\frac{|C_n(x[0:n))|}{n} \leqslant \frac{1}{\ell m} - \frac{1}{\ell}\sum_{w\in\mathbb{B}^{\ell}} f_{w,n}\log f_{w,n}.$$

Choosing m large enough such that $1/\ell m \leqslant \varepsilon$ and letting k go to infinity gives

$$\bar{\rho}(x) \leqslant \limsup_{n\to\infty} \min_{C\in\mathcal{C}_k} \frac{|C(x[0:n))|}{n} \leqslant \varepsilon - \frac{1}{\ell}\sum_{w\in\mathbb{B}^{\ell}} f_w \log f_w \leqslant \overline{ah}(x) + 2\varepsilon.$$

where k is the number of states of all C_n which depends on ℓm. Since the relation holds for each $\varepsilon > 0$, the proof is complete. Note that changing ε requires to change ℓ and m and thus k. ☐

Observe that Lemma 6 is about compression ratio for the class of all relational compressors, while Lemmas 7 and 8 are about compression ratio for the class of all almost local compressors.

5 Proof of Theorem 1

The proof of Theorem 1 is divided into three sections. Section 5.1 is devoted to the proof of the inequalities $2\underline{\gamma}(x) \leqslant \underline{\dim}(x) \leqslant \mathfrak{h}(\underline{\gamma}(x))$ and $2\overline{\gamma}(x) \leqslant \overline{\dim}(x) \leqslant \mathfrak{h}(\overline{\gamma}(x))$. Section 5.2 is devoted to the remaining inequalities $2\underline{\beta}(x) \leqslant \underline{\dim}(x) \leqslant \mathfrak{h}(\underline{\beta}(x))$ and $2\overline{\beta}(x) \leqslant \overline{\dim}(x) \leqslant \mathfrak{h}(\overline{\beta}(x))$. Altogether this shows item 1 of Theorem 1. Finally, Section 5.3 is devoted to showing item 2 of Theorem 1.

The proofs in Section 5.1 establish correspondences between the Rauzy dimension $\underline{\gamma}$ and $\overline{\gamma}$ and predictors. The proof of the inequality $\overline{\dim}(x) \leqslant \mathfrak{h}(\overline{\gamma}(x))$ is rather standard. The proof of the inequality $2\gamma(x) \leqslant \overline{\dim}(x)$ is more difficult, as it requires switching from a finite-state predictor to a sliding window. The first step is to obtain an almost local deterministic compressor using the classical proof that $\underline{\rho}(x) \leqslant \underline{ah}(x)$. Then, we convert this almost local deterministic compressor into an almost local measure using the proof in [6] that $\underline{\dim}(x) \leqslant \underline{\rho}(x)$. From this almost local measure we construct a local measure using the tools developed in Section 3. As measures and predictors are two presentations of the same object, we obtain a local predictor to be used with sliding window.

5.1 Proof for Rauzy $\underline{\gamma}$ and $\overline{\gamma}$

Proposition 4. *For every $x \in \mathbb{B}^{\mathbb{N}}$, $2\underline{\gamma}(x) \leqslant \underline{\dim}(x)$ and $2\overline{\gamma}(x) \leqslant \overline{\dim}(x)$.*

Proof. By Lemma 2, $\underline{\dim}(x) = \underline{\xi}(x)$ and $\overline{\dim}(x) = \overline{\xi}(x)$, so we will show $2\underline{\gamma}(x) \leqslant \underline{\xi}(x)$ and $2\overline{\gamma}(x) \leqslant \overline{\xi}(x)$. Let $\varepsilon > 0$ be a positive real number. There exists then a predictor π such that $\liminf_{n\to\infty} \xi_\pi(x[0:n))/n \leqslant \underline{\dim}(x) + \varepsilon$. By Corollary 1, it can be assumed that this predictor π is k-local for some integer $k \geqslant 1$. We claim that there exists a function $f : \mathbb{B}^k \to \mathbb{B}$ such that

$2\gamma_f(x[0:n]) \leqslant \xi_\pi(x[0:n])$ holds for each large enough integer n where γ_f is defined for each word $w \in \mathbb{B}^*$ by

$$\gamma_f(w) \triangleq \sum_{i=0}^{|w|-k-1} (1 - \delta_{f(w[i:i+k)),w[i+k]}).$$

For each word $w \in \mathbb{B}^k$ and $a \in \mathbb{B}$, let α_{wa} be the number $|x[0:n)|_{wa}/n$. Let f be the function defined by $f(w) = a$ for each $w \in \mathbb{B}^k$ and $a \in \mathbb{B}$ if $\alpha_{w\bar{a}} \leqslant \alpha_{wa}$ with the usual notation $\bar{0} = 1$ and $\bar{1} = 0$. In the case $\alpha_{w0} = \alpha_{w1}$, we choose arbitrary $f(w) = 1$. If f is defined in this way, then $\gamma_f(x[0:n))/n$ is equal to

$$\frac{\gamma_f(x[0:n))}{n} = \sum_{w \in \mathbb{B}^k} \min(\alpha_{w0}, \alpha_{w1})$$

Suppose that the local predictor π is defined by the betting function θ from $\mathbb{B}^k$ to $[0,1]$. The *log-loss rate* of π on $x[0:n)$ is then given by

$$\frac{\xi_\pi(x[0:n))}{n} = -\sum_{w \in \mathbb{B}^k} \big(\alpha_{w0} \log(1-\theta(w)) + \alpha_{w1} \log \theta(w)\big).$$

It is a classical result that the real number $p \in [0,1]$ being fixed, the expression $-p \log q - (1-p)\log(1-q)$ is minimal when $q = p$. It follows that

$$\begin{aligned}\frac{\xi_\pi(x[0:n))}{n} &\geqslant -\sum_{w \in \mathbb{B}^k} \left(\alpha_{w0} \log \frac{\alpha_{w0}}{\alpha_{w0}+\alpha_{w1}} + \alpha_{w1} \log \frac{\alpha_{w1}}{\alpha_{w0}+\alpha_{w1}}\right) \\ &= \sum_{w \in \mathbb{B}^k} (\alpha_{w0}+\alpha_{w1}) \mathfrak{h}\left(\frac{\alpha_{w0}}{\alpha_{w0}+\alpha_{w1}}\right).\end{aligned}$$

The fact that the function $\mathfrak{h}$ is concave and $\mathfrak{h}(0) = 0$ and $\mathfrak{h}(1/2) = 1$ imply that $2p \leqslant \mathfrak{h}(p)$ for each $p \in [0,1/2]$. Applying this inequality to $p = 2\min(\alpha_{w0}, \alpha_{w1})/(\alpha_{w0}+\alpha_{w1})$ which belongs to $[0,1/2]$, we get

$$\frac{2\min(\alpha_{w0}, \alpha_{w1})}{\alpha_{w0}+\alpha_{w1}} \leqslant \mathfrak{h}\left(\frac{\alpha_{w0}}{\alpha_{w0}+\alpha_{w1}}\right).$$

Combining this latter inequality with the one obtained above, we get the inequality $2\gamma_k(x[1:n)) \leqslant \xi_\pi(x[0:n))/n$. Taking the limit inferior or superior when n goes to infinity completes the proof of the two inequalities. □

We need to prove the following.

Proposition 5. *For every $x \in \mathbb{B}^{\mathbb{N}}$, $\underline{\dim}(x) \leqslant \mathfrak{h}(\underline{\gamma}(x))$ and $\overline{\dim}(x) \leqslant \mathfrak{h}(\overline{\gamma}(x))$.*

Observe that the statement of this Proposition 5 is quite tight. Let $0 \leqslant \alpha < \alpha' \leqslant 1/2$ be two different real numbers. There exists two sequences x_0 and x_1 such that

$$\underline{\gamma}(x_0) = \overline{\gamma}(x_0) = \underline{\gamma}(x_1) = \overline{\gamma}(x_1) = (\alpha + \alpha')/2$$

but

$$\underline{\dim}(x_0) = \overline{\dim}(x_0) = \mathfrak{h}((\alpha+\alpha')/2)$$

and

$$\underline{\dim}(x_1) = \overline{\dim}(x_1) = (\mathfrak{h}(\alpha) + \mathfrak{h}(\alpha'))/2.$$

Since the function $\mathfrak{h}$ is concave $(\mathfrak{h}(\alpha) + \mathfrak{h}(\alpha'))/2 < \mathfrak{h}((\alpha+\alpha')/2)$.

Let x_0 be a generic sequence for the Bernoulli measure μ_0 given $\mu_0(0) = (\alpha+\alpha')/2$ and $\mu_0(1) = 1 - (\alpha+\alpha')/2$. Let x_1 be a sequence such that its subsequences of symbols at even and odd positions are respectively generic sequences for the Bernoulli measures μ and μ' given by $\mu(0) = \alpha$, $\mu(1) = 1 - \alpha$, $\mu'(0) = \alpha'$ and $\mu'(1) = 1 - \alpha'$.

Before giving the proof of Proposition 5 we give the following lemma.

Lemma 9. *For every pair of real numbers α and α' such that $0 \leqslant \alpha \leqslant \alpha' \leqslant 1/2$,*

$$-\alpha \log \alpha' - (1-\alpha)\log(1-\alpha') \leqslant \mathfrak{h}(\alpha) + 2(\alpha' - \alpha).$$

Proof. Since the log function is increasing $-\log \alpha' \leqslant -\log \alpha$ and thus $-\alpha \log \alpha' \leqslant -\alpha \log \alpha$. By the mean value theorem

$$-\log(1-\alpha') + \log(1-\alpha) = (\alpha' - \alpha) g'(\nu)$$

where $\alpha \leqslant \nu \leqslant \alpha'$ and g' is the derivative of the function $g(x) = -\log(1-x)$. Since $g'(x) = 1/(1-x)$, $g'(\nu)$ is bounded by 2 for each $\nu \in [0, 1/2]$. It follows that

$$0 \leqslant -\log(1-\alpha') + \log(1-\alpha) \leqslant 2(\alpha' - \alpha)$$

and

$$-(1-\alpha)\log(1-\alpha') \leqslant -(1-\alpha)\log(1-\alpha) + 2(\alpha' - \alpha).$$

Combining this later inequality with $-\alpha \log \alpha' \leqslant -\alpha \log \alpha$ completes the proof. □

Now we are ready for the delayed proof.

Proof of Proposition 5. Lemma 2 asserts that for every $x \in \mathbb{B}^{\mathbb{N}}$, $\underline{\dim}(x) = \underline{\xi}(x)$ and $\overline{\dim}(x) = \overline{\xi}(x)$. We start proving $\underline{\xi}(x) \leqslant \mathfrak{h}(\underline{\gamma}(x))$. It is sufficient to show that, for each real number α such that $0 < \alpha < 1/2$, if $\underline{\gamma}(x) \leqslant \alpha$, then $\underline{\xi}(x) \leqslant \mathfrak{h}(\alpha)$. Let us suppose that $\underline{\gamma}(x) \leqslant \alpha < 1/2$ and let $\varepsilon > 0$ be a positive real number such that $\alpha + \varepsilon < 1/2$. There exists then a positive integer ℓ such that $\underline{\gamma_\ell}(x) < \alpha + \varepsilon$. There exists an increasing sequence of integers $(n_k)_{k \geqslant 0}$ such that $\lim_{k\to\infty} \gamma_\ell(x[0:n_k)) = \alpha' < \alpha + \varepsilon$. Since there are finitely many functions from $\mathbb{B}^\ell$ to $\mathbb{B}$, it can be assumed that the value $\gamma_\ell(x[0:n_k))$ is always achieved with the same function f from $\mathbb{B}^\ell$ to $\mathbb{B}$. Let us use this function f to define the predictor π for each pair $(w, b) \in \mathbb{B}^* \times \mathbb{B}$ by

$$\pi(w,b) = \begin{cases} 1/2 & \text{if } |w| < \ell, \\ 1-\alpha' & \text{if the suffix } v \text{ of length } \ell \text{ of } w \text{ satisfies } f(v) = b, \\ \alpha' & \text{if the suffix } v \text{ of length } \ell \text{ of } w \text{ satisfies } f(v) \neq b. \end{cases}$$

The log-loss rate of the predictor π is given by

$$\xi_\pi(w) = \ell - |w|\gamma_\ell(w)\log \alpha' - |w|(1-\gamma_\ell(w))\log(1-\alpha')$$

for each word w satisfying $|w| \geqslant \ell$. If follows from $\lim_{k\to\infty} \gamma_\ell(x[0:n_k)) = \alpha'$ that

$$\lim_{k\to\infty} \xi_\pi(x[0:n_k])/n_k = \mathfrak{h}(\alpha')$$

and thus $\underline{\xi}(x) \leqslant \mathfrak{h}(\alpha') \leqslant \mathfrak{h}(\alpha + \varepsilon)$ because h is increasing on the interval $[0, 1/2]$. Since this inequality holds for each positive real number ε, we get $\underline{\xi}(x) \leqslant \mathfrak{h}(\alpha)$ by continuity of the function h.

To show that $\overline{\xi}(x) \leqslant \mathfrak{h}(\overline{\gamma}(x))$ it is sufficient to prove that, for each real number α such that $0 < \alpha < 1/2$, if $\overline{\gamma}(x) \leqslant \alpha$, we have $\overline{\xi}(x) \leqslant \mathfrak{h}(\alpha)$. Let us suppose that $\overline{\gamma}(x) \leqslant \alpha < 1/2$ and let $\varepsilon > 0$ be a positive real number such that $\alpha + 2\varepsilon < 1/2$. There exists then a positive integer ℓ such that $\overline{\gamma_\ell}(x) < \alpha + \varepsilon$. This means that for each integer n large enough, there exists a function $f_n : \mathbb{B}^\ell \to \mathbb{B}$ such that

$$\alpha_n \triangleq \frac{1}{n} \sum_{i=0}^{n-\ell-1} (1 - \delta_{f_n(x[i:i+\ell)), x[i+\ell]})$$

satisfies $\alpha_n \leqslant \alpha + \varepsilon$

Let m be an integer such that $2^{-m} < \varepsilon$. For each integer large enough, let α'_n be defined by $\alpha'_n = (k+1)2^{-m}$ where the integer k is such that $k2^{-m} \leqslant \alpha_n < (k+1)2^{-m}$. The real number α'_n

belongs to the finite set $\{i2^{-m} : 1 \leqslant i \leqslant 2^{m-1}\}$ and satisfies $\alpha_n < \alpha'_n \leqslant \alpha_n + \varepsilon$. Let π_n be the predictor defined for each pair $(w, b) \in \mathbb{B}^* \times \mathbb{B}$ by

$$\pi_n(w,b) \triangleq \begin{cases} 1/2 & \text{if } |w| < \ell, \\ 1-\alpha'_n & \text{if the suffix } v \text{ of length } \ell \text{ of } w \text{ satisfies } f_n(v) = b, \\ \alpha'_n & \text{if the suffix } v \text{ of length } \ell \text{ of } w \text{ satisfies } f_n(v) \neq b. \end{cases}$$

Let us point out that each predictor π_n belongs to the finite class Π_k where $k = \max(2^{\ell+1}, m)$. The log-loss rate of the predictor π_n over the prefix $x[0 : n)$, for n large enough, is given by

$$\xi_{\pi_n}(x[0:n)) = \ell - n\alpha_n \log \alpha'_n - n(1-\alpha_n)\log(1-\alpha'_n).$$

By Lemma 9, $\xi_{\pi_n}(x[0 : n))/n$ is bounded by $\ell/n + \mathfrak{h}(\alpha_n) + 2\varepsilon$ which is, in turn, bounded by $\mathfrak{h}(\alpha + 2\varepsilon) + 2\varepsilon$ for n large enough. Since π_n range over a finite set Π_k for some fixed k, this proves $\underline{\xi}(x) \leqslant \mathfrak{h}(\alpha)$. ☐

5.2 Proof for Rauzy $\underline{\beta}$ and $\overline{\beta}$

For $w \in \mathbb{B}^{\mathbb{N}}$, let $\widetilde{w}$ the reverse of w, that is if $|w| = n$, $\widetilde{w} = w[n-1]w[n-2]\cdots w[0]$. With some notation abuse, we define

$$\underline{\dim}((\widetilde{x[0:n)})_{n\geqslant 1}) \triangleq \inf\{s : \exists \mu \in \mathcal{M} \quad \text{such that } \limsup_{n\to\infty} \mu^{(s)}(\widetilde{x[0:n)}) = \infty\}$$
$$\overline{\dim}((\widetilde{x[0:n)})_{n\geqslant 1}) \triangleq \inf\{s : \exists \mu \in \mathcal{M} \quad \text{such that } \liminf_{n\to\infty} \mu^{(s)}(\widetilde{x[0:n)}) = \infty\}.$$

In view of the identities $\underline{\rho}^{\mathsf{d}} = \underline{\rho}^{\mathsf{f}} = \underline{\rho}$ and $\bar{\rho}^{\mathsf{d}} = \bar{\rho}^{\mathsf{f}} = \bar{\rho}$ we consider the family $\mathcal{C}$ of all functional compressors and define

$$\underline{\rho}((\widetilde{x[0:n)})_{n\geqslant 1}) \triangleq \inf_{C\in\mathcal{C}} \liminf_{n\to\infty} \frac{|C(\widetilde{x[0:n)})|}{n}$$
$$\bar{\rho}((\widetilde{x[0:n)})_{n\geqslant 1}) \triangleq \inf_{C\in\mathcal{C}} \limsup_{n\to\infty} \frac{|C(\widetilde{x([0:n)})|}{n}.$$

In the proof of the next proposition we use that for every $x \in \mathbb{B}^{\mathbb{N}}$, $\underline{\dim}(x) = \underline{\rho}(x)$, which is proved in [6, Theorem 7.2]; and that $\overline{\dim}(x) = \bar{\rho}(x)$, which is proved in [1, Theorem 6.18].

Proposition 6. *For all $x \in \mathbb{B}^{\mathbb{N}}$, $\underline{\dim}(x) = \underline{\dim}((\widetilde{x[0:n)})_{n\geqslant 1})$ and $\overline{\dim}(x) = \overline{\dim}((\widetilde{x[0:n)})_{n\geqslant 1})$.*

Proof. We argue for $\underline{\dim}$, since the case of $\overline{\dim}$ is alike. Since $\underline{\dim}(x) = \underline{\rho}(x)$ it suffices to prove that $\underline{\rho}(x) = \underline{\rho}((\widetilde{x[0:n)})_{n\geqslant 1})$. For each deterministic compressor $C = (Q, \delta, q_0, F)$ consider $\widetilde{C} = (Q, \widetilde{\delta}, F, \{q_0\})$ that results from reversing the transitions of C and exchanging initial and final states. Precisely, the transition relation $\widetilde{\delta} \subseteq Q \times \mathbb{B} \times \mathbb{B} \times Q$ is such that if $q \xrightarrow{u|v} p$ is in C then $p \xrightarrow{u|v} q$ is in $\widetilde{C}$. Since C is a deterministic compressor it computes a one-to-one function.

Then, $\widetilde{C}$ is possibly non-deterministic but it is a functional compressor because it computes a one-to-one function. Then, for every deterministic compressor C

$$\underline{\rho}_{\widetilde{C}}(x) = \liminf_{n\to\infty} \frac{|C(x[0:n)|}{n} = \liminf_{n\to\infty} \frac{|\widetilde{C}(\widetilde{x[0:n)})|}{n}.$$

We conclude $\underline{\rho}(x) = \underline{\rho}((\widetilde{x[0:n)})_{n\geqslant 1})$. ☐

Define

$$\underline{\gamma_\ell}((\widetilde{x[0:n)})_{n\geqslant 1}) \triangleq \liminf_{n\to\infty} \gamma_\ell(\widetilde{x[0:n)}) \quad \text{and} \quad \overline{\gamma_\ell}((\widetilde{x[0:n)})_{n\geqslant 1}) \triangleq \limsup_{n\to\infty} \gamma_\ell(\widetilde{x[0:n)})$$

$$\underline{\gamma}((\widetilde{x[0:n)})_{n\geqslant 1}) \triangleq \lim_{\ell\to\infty} \underline{\gamma_\ell}((\widetilde{x[0:n)})_{n\geqslant 1}) \quad \text{and} \quad \overline{\gamma}((\widetilde{x[0:n)})_{n\geqslant 1}) \triangleq \lim_{\ell\to\infty} \overline{\gamma_\ell}((\widetilde{x[0:n)})_{n\geqslant 1}).$$

Proposition 7. *For all $x \in \mathbb{B}^{\mathbb{N}}$, $\underline{\beta}(x) = \underline{\gamma}((\widetilde{x[0:n)})_{n\geqslant 1})$ and $\overline{\beta}(x) = \overline{\gamma}((\widetilde{x[0:n)})_{n\geqslant 1})$.*

Proof. We first show that for every $w \in \mathbb{B}^*$, for any $\ell \leqslant |w|$, $\underline{\gamma_\ell}(w) = \underline{\beta_\ell}(\widetilde{w})$. Let $n = |w| = |\widetilde{w}|$. Since for $i = 0, 1, \ldots, n-1$, $w[n-1-i] = \widetilde{w}[i]$, the definitions $\underline{\gamma_\ell}$ and $\underline{\beta_\ell}$ ensure that for $i = 0, 1, \ldots, n-\ell-1$,

$$w[n-1-i] = f(w[n-1-i-\ell : n-2-i]) \qquad \text{if and only if} \qquad \widetilde{w}[i] = f(\widetilde{w}[i+1 : i+\ell]).$$

Thus,

$$\frac{1}{n}\sum_{i=0}^{n-\ell-1}(1-\delta_{f(w[n-1-i-\ell,n-2],w[n-1-i]}) = \frac{1}{n}\sum_{i=0}^{n-\ell-1}(1-\delta_{w[i,]f(\widetilde{w}[i+1,i+\ell],)}).$$

Equivalently, we can write

$$\frac{1}{n}\sum_{i=0}^{n-\ell-1}(1-\delta_{f(w[i,i+\ell-1],w[i+\ell]}) = \frac{1}{n}\sum_{i=0}(1-\delta_{w[i,]f(\widetilde{w}[i+1,i+\ell],)}).$$

We conclude that $\underline{\gamma_\ell}(w) = \underline{\beta_\ell}(\widetilde{w})$. Using that for every $w \in \mathbb{B}^*$, $\underline{\beta}(w) = \underline{\gamma}(\widetilde{w})$ and the definition of $\underline{\beta}$ we obtain the wanted result,

$$\begin{aligned}
\underline{\beta}(x) &= \lim_{\ell\to\infty} \liminf_{n\to\infty} \underline{\beta_\ell}(x[0:n)) \\
&= \lim_{\ell\to\infty} \liminf_{n\to\infty} \underline{\gamma_\ell}(\widetilde{x[0:n)}) \\
&= \lim_{\ell\to\infty} \underline{\gamma_\ell}((\widetilde{x[0:n)}_{n\geq 1}) \\
&= \underline{\gamma}((\widetilde{x[0:n)}_{n\geq 1}).
\end{aligned}$$

The case of $\overline{\beta}$ is similar. ☐

Propositions 6 and 7 imply

$$\begin{aligned}
2\underline{\beta}(x) = 2\underline{\gamma}((\widetilde{x[0:n)})_{n\geqslant 1}) &\leqslant \underline{\dim}((\widetilde{x[0:n)})_{n\geqslant 1}) = \underline{\dim}(x) &&\leqslant h(\underline{\gamma}(x)) = h(\underline{\beta}((\widetilde{x[0:n)})_{n\geqslant 1})) \\
2\overline{\beta}(x) = 2\overline{\gamma}((\widetilde{x[0:n)})_{n\geqslant 1}) &\leqslant \overline{\dim}((\widetilde{x[0:n)})_{n\geqslant 1})) = \overline{\dim}(x) &&\leqslant h(\overline{\beta}(x)) = h(\overline{\gamma}((\widetilde{x[0:n)})_{n\geqslant 1})).
\end{aligned}$$

5.3 The bounds are sharp

The following proposition explains why the inequalities of item 1 of Theorem 1 are sharp and proves item 2.

Proposition 8. *For every real numbers α and ε such that $0 \leqslant \alpha \leqslant 1/2$ and $\varepsilon > 0$, there exist sequences x and x' such that $\underline{\gamma}(x) = \overline{\gamma}(x) = \underline{\beta}(x) = \overline{\beta}(x) = \underline{\gamma}(x') = \overline{\gamma}(x') = \underline{\beta}(x') = \overline{\beta}(x') = \alpha$ and $\underline{\dim}(x) = \overline{\dim}(x) \leqslant 2\alpha + \varepsilon$ and $\underline{\dim}(x') = \overline{\dim}(x') = h(\alpha)$.*

Proof. We start with the definition of the sequence x'. Let μ be the Bernoulli measure given by $\mu(0) = \alpha$ and $\mu(1) = 1-\alpha$ and let x' be a generic sequence for μ. It is straightforward to verify all equalities involving x'.

The construction of the sequence x is a bit more involved. Let p and q be two integers and α' be a real number such that $\alpha = p/2q + \alpha'/q$. Choosing p and q large enough can guarantee that

$h(\alpha')/q \leqslant 1/q < \varepsilon$. Let y be a normal sequence over $\mathbb{B}$ and let z be a generic sequence for the Bernoulli measure μ' given by $\mu'(0) = \alpha'$ and $\mu'(1) = 1 - \alpha'$. The sequence $x = (x_n)_{n \geqslant 0}$ is then defined by

$$x_n = \begin{cases} y_n & \text{if } 0 \leqslant n \bmod q \leqslant p-1, \\ z_n & \text{if } n \bmod q = p, \\ 0 & \text{if } p+1 \leqslant n \bmod q \leqslant q-1. \end{cases}$$

It can be verified that $\underline{\gamma}(x) = \overline{\gamma}(x) = \underline{\beta}(x) = \overline{\beta}(x) = \alpha$ and that $\underline{\dim}(x) = \overline{\dim}(x) = p/q + h(\alpha')/q \leqslant 2\alpha + \varepsilon$. □

This completes the proof of Theorem 1.

6 Proof of Theorem 2

Thanks to the following lemma, the proof of Theorem 2 requires to prove just two inequalities.

Lemma 10.

1. *if* $\underline{\dim}(x+y) \leqslant \underline{\dim}(x) + \overline{\dim}(y)$ *for every* $x, y \in \mathbb{B}^{\mathbb{N}}$ *then* $\underline{\dim}(x) - \overline{\dim}(y) \leqslant \underline{\dim}(x+y)$ *for every* $x, y \in \mathbb{B}^{\mathbb{N}}$*;*
2. *if* $\overline{\dim}(x+y) \leqslant \overline{\dim}(x) + \overline{\dim}(y)$ *for every* $x, y \in \mathbb{B}^{\mathbb{N}}$ *then* $\overline{\dim}(x) - \overline{\dim}(y) \leqslant \overline{\dim}(x+y)$ *for every* $x, y \in \mathbb{B}^{\mathbb{N}}$*.*

Proof. We show item 1; item 2 is analogous. We apply the hypothesis with $x' = x+y$ and $y' = -y$ to get $\underline{\dim}(x) \leqslant \underline{\dim}(x+y) + \overline{\dim}(-y) = \underline{\dim}(x+y) + \overline{\dim}(y)$. The last equality comes from $\overline{\dim}(-y) = \overline{\dim}(y)$. □

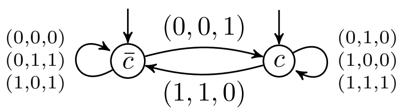

Figure 3: Transducer T_+ for addition

Proof of Theorem 2. By Lemma 10, it is sufficient to prove the two $\underline{\dim}(x+y) \leqslant \underline{\dim}(x) + \overline{\dim}(y)$ and $\overline{\dim}(x+y) \leqslant \overline{\dim}(x) + \overline{\dim}(y)$. We start with the first of the two inequalities.

Let $z = x + y$ be the sequence obtained by adding x and y. Let $\varepsilon > 0$ be a positive real number. We claim that there exists a relational compressor C_z such that $\underline{\rho}_{C_z}(z) < \underline{\rho}(x) + \bar{\rho}(y) + \varepsilon$. This relational compressor C_z is obtained by combining several transducers. Let C_x be a relational compressor such that $\underline{\rho}_{C_x}(x) < \underline{\rho}(x) + \varepsilon/4$ and let C_y be a relational compressor such that $\bar{\rho}_{C_y}(y) < \bar{\rho}(y) + \varepsilon/4$. Let T_+ be a transducer realizing addition: this transducer has three tapes and accepts triples (x, y, z) such that $z = x + y$. Each transition of T_+ is labeled by a triple (a, b, c) in $\mathbb{B}^3$. The transducer T_+ is pictured in Figure 3.

The transducer C_z is built as follows. It has one input tape where z is read and one output tape. The input z is fed to the transducer T_+ which non-deterministically produces two outputs x' and y' such that $z = x' + y'$. The output x' is not written on a tape but directly fed as input to the relational compressor C_x. Similarly, the output y' is not written on a tape but directly fed as input to the relational compressor C_y. The outputs of the relational compressors C_x and C_y are then combined on the single output tape of C_z as follows. Let m be an integer to be fixed later. The outputs of C_x and C_y are written by blocks of length m with an additional bit being 0 if the block comes from C_x and being 1 the block comes from C_y. This means that for m bits produced by either C_x or C_y, $m+1$ bits are written to the output tape of C_z. This additional bits guaranties that $C_x(x')$ and $C_y(y')$ and therefore x' and y' can be recovered from the content of the single output tape of C_z. The transducer C_z contains therefore two buffers of m bits where the outputs of C_x and C_y are temporarily stored. When one of the two buffers becomes full, its content is written to the output tape of C_z with the additional bit.

Among the runs of the relational compressor C_z, there is one run η where $x' = x$ and $y' = y$. Let us compute the compression ratio along this run η. By reading n binary symbols of its input z, the transducer T_+ produces n binary symbols of x which are fed to C_x and n binary symbols of y which are fed to C_y. For infinitely many integer n, the length of the output of C_x when reading n symbols is at most $(\underline{\rho}(x) + \varepsilon/4)n$. For all integers n large enough, the length of the output of C_y when reading n symbols is at most $(\bar{\rho}(y) + \varepsilon/4)n$. Therefore, the total output of C_z when reading n symbols is at most $(\underline{\rho}(x) + \bar{\rho}(y) + \varepsilon/2)(m+1)n/m$ for infinitely many integers n. The ratio $(m+1)/m$ comes from the additional bit for each block of length m. The integer m is chosen large enough such that $(\underline{\rho}(x) + \bar{\rho}(y) + \varepsilon/2)(m+1)/m < \underline{\rho}(x) + \bar{\rho}(y) + \varepsilon$. This proves that $\underline{\rho}_{C_z}(\eta) \leqslant \underline{\rho}(x) + \bar{\rho}(y) + \varepsilon$ and thus $\underline{\rho}(z) \leqslant \underline{\rho}(x) + \bar{\rho}(y) + \varepsilon$. This completes the proof of the first inequality.

The proof of the second inequality $\overline{\dim}(x+y) \leqslant \overline{\dim}(x) + \overline{\dim}(y)$ is almost the same but the transducer C_x must be chosen such that $\bar{\rho}_C C_x(x) < \bar{\rho}(x) + \varepsilon/4$. $\square$

7 Proof of Proposition 1

We show that the two functions $\underline{\beta}$ and $\underline{\gamma}$ are different, as are $\underline{\beta}$ and $\underline{\gamma}$. Precisely, we provide a sequence $x \in \mathbb{B}^{\mathbb{N}}$ such that $\underline{\beta}(x) = \overline{\beta}(x) \neq \underline{\gamma}(x) = \overline{\gamma}(x)$. In the rest of this section, we write $\beta(x)$ and $\gamma(x)$ for $\underline{\beta}(x) = \overline{\beta}(x)$ and $\underline{\gamma}(x) = \overline{\gamma}(x)$ respectively. This sequence x is obtained as a generic sequence for a Markov chain which is given below. The strategy to show that $\beta(x)$ and $\gamma(x)$ are indeed different is as follows. Proposition 9 below allows us to give the lower bound 11/24 for $\gamma(x)$. Lemma 11 allows us to compute β_ℓ and γ_ℓ from the frequencies in x of words of length $\ell+1$. From the frequencies of words of length 7 in x, we get that $\beta_6(x) = 9503/20736$ and thus $\beta(x) < 9503/20736 < 11/24 \leqslant \gamma(x)$.

We first describe the Markov chain. Its state space is $Q = \{0,1,2,3\}$ and its probability matrix is the following stochastic matrix,

$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \end{pmatrix}$$

Each transition with a non-zero probability is labelled with a symbol $b \in \mathbb{B}$ in such a way that different transitions leaving the same state have different labels. Therefore the Markov chain can be viewed as a deterministic finite-state automaton with probabilistic transitions. This automaton is pictured in Figure 4. A transition of the form $p \xrightarrow{b:\alpha} q$ means that the probability of the transition from p to q is $\alpha \in [0,1]$ and that its label is $b \in \mathbb{B}$,

The stationary distribution π of the Markov chain is the vector $\pi = [5/24, 1/4, 7/24, 1/4]$.

Proposition 9. *Let x be a generic sequence of a Markov chain with stationary distribution π. Let $\theta_{i,b}$ be the probability of having symbol b after state i. Then*

$$\gamma(x) \geqslant \sum_{i \in Q} \pi_i \min(\theta_{i,0}, \theta_{i,1}). \tag{2}$$

Our conjecture is that the inequality stated in Proposition 9 is actually an equality.

The proof of the proposition is based on the notion of the *snake Markov Chain* [5, Problem 2.2.4] that we now introduce. Let a Markov chain with state space Q and transition matrix P. Let $k \geqslant 1$ be a positive integer. The *snake* Markov chain of order k is the Markov chain whose state space is the set $\{q_1, \ldots, q_k \in Q^k : P_{q_1,q_2} \cdots P_{q_{k-1},q_k} \neq 0\}$ of sequences of k states in the original Markov

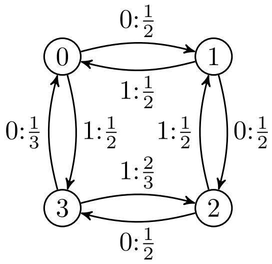

Figure 4: The Markov chain

chain. Each such sequence of k states can be viewed as a path made of $k-1$ consecutive transitions. There is a transition from state $q_1, \dots, q_k$ to state $q'_1, \dots, q'_k$ if $q'_i = q_{i+1}$ for each $1 \leqslant i \leqslant k-1$ and its probability is $P_{q'_{k-1}, q'_k} = P_{q_k, q'_k}$. If the original Markov chain is irreducible, so is the snake Markov chain. Furthermore, if $\pi = (\pi_q)_{q \in Q}$ is the stationary distribution of the original Markov chain, the stationary distribution of the snake Markov chain is given by $\pi_{q_1,\dots,q_k} = \pi_{q_1} P_{q_1,q_2} \cdots P_{q_{k-1},q_k}$. Note that the snake Markov chain of order 1 is identical to the starting Markov chain.

Proof of Proposition 9. Let us fix a positive integer $\ell \geqslant 1$ and a function $f : \mathbb{B}^\ell \to \mathbb{B}$. Let us consider the snake Markov chain of order $\ell + 1$. Each state of this latter Markov chain is a sequence of ℓ consecutive transitions forming a path of length ℓ. The label of this state is a word of length ℓ obtained by concatenating the labels of its ℓ transitions. The function f maps this word to a predicted symbol. It is important to notice that the probabilities of transitions in the snake Markov chain are essentially the same as in the original Markov chain. Since the sequence x is generic for the Markov chain, it is also generic for the snake Markov chain. Therefore for each state of the snake Markov chain, choosing the symbol with greatest probability instead of the symbol given by the function f decreases the error rate of prediction. This proves the inequality. □

Lemma 11. *Let ℓ be a positive integer and let x be sequence such that the frequency in x of each word $u \in \mathbb{B}^{\ell+1}$ does exist and is equal to α_u. Then*

$$\beta_\ell(x) = \sum_{w \in \mathbb{B}^\ell} \min(\alpha_{0w}, \alpha_{1w}) \qquad \textit{and} \qquad \gamma_\ell(x) = \sum_{w \in \mathbb{B}^\ell} \min(\alpha_{w0}, \alpha_{w1}).$$

Proof. Note that the frequency of each word w of length ℓ in x is $\alpha_{0w} + \alpha_{1w} = \alpha_{w0} + \alpha_{w1}$ and that $\alpha_{0w}/(\alpha_{0w} + \alpha_{1w})$ (respectively $\alpha_{1w}/(\alpha_{0w} + \alpha_{1w})$) is the probability of having a 0 (respectively a 1) before w. It follows that β_ℓ can be written

$$\beta_\ell(x) = \sum_{w \in \mathbb{B}^\ell} (\alpha_{0w} + \alpha_{1w}) \min\left(\frac{\alpha_{0w}}{\alpha_{0w} + \alpha_{1w}}, \frac{\alpha_{1w}}{\alpha_{0w} + \alpha_{1w}}\right).$$

and this concludes the proof. □

Acknowledgements

The authors are members of IRP SINFIN, Université Paris Cité/CNRS-Universidad de Buenos Aires-CONICET. This project has been developed with Argentine grants PICT-2021-I-A-00838, CONICET PIP 11220210100220CO, UBACyT 20020220100065BA, UBACyT 20020190100021BA. The second author is supported by ANR SymDynAr (ANR-23-CE40-0024-01).

References

[1] K. B. Athreya, J. M. Hitchcock, J. H. Lutz, and E. Mayordomo. Effective strong dimension in algorithmic information and computational complexity. *SIAM J. Comput.*, 37(3):671–705, 2007.

[2] V. Becher, O. Carton, and P.A. Heiber. Normality and automata. *Journal of Computer and System Sciences*, 81(8):1592–1613, 2015.

[3] M. Bernay. La dimension de Hausdorff de l'ensemble des nombres r-déterministes. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, Série A*, 280:539–541, 1975.

[4] C. Bourke, J. Hitchcock, and N. Vinodchandran. Entropy rates and finite-state dimension. *Theoretical Computer Science*, 349:392–406, 2005.

[5] P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 2008.

[6] J. Dai, J. Lathrop, J. Lutz, and E. Mayordomo. Finite-state dimension. *Theoretical Computer Science*, 310:1–33, 2004.

[7] A. Kozachinskiy and A. Shen. Automatic Kolmogorov complexity, normality, and finite-state dimension revisited. *Journal of Computer and System Sciences*, 118:75–107, 2021.

[8] G. Rauzy. Nombres normaux et processus déterministes. *Acta Arith.*, 29:211–225, 1976.

[9] D. Sheinwald. On the Ziv-Lempel proof and related topics. *Proceedings of the IEEE*, 82(6):866–871, 1994.

[10] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.