The entropy of (formal) languages and dimensions of subsets in CANTOR space

Ludwig Staiger

Martin-Luther-Universität Halle-Wittenberg
Institut für Informatik

EQINOCS, Grenoble, January 23, 2013

Notation: Strings and Languages

Finite Alphabet $X = \{0, \dots, r-1\}$, cardinality $|X| = r$

Finite strings (words) $w = x_1 \cdots x_n \in X^*$, $x_i \in X$

Length $|w| = n$

Prefix $v \sqsubseteq w$ if $v = x_1 \cdots x_m$, $w = x_1 \cdots x_n$ and $m \leq n$

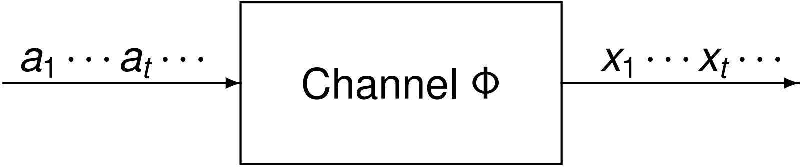
Languages $W \subseteq X^*$

Infinite strings (ω-words) $\xi = x_1 \cdots x_n \cdots \in X^\omega$

Prefixes of infinite strings $\xi[0..n] \in X^*$, $|\xi[0..n]| = n$

ω-Languages $F \subseteq X^\omega$

SHANNON's Channel Capacity



$$C(\Phi) := \lim_{t\to\infty} \frac{\log_r |\{x_1 \cdots x_t : x_1 \cdots x_t \text{ is an output of } \Phi\}|}{t}$$

Entropy of languages (CHOMSKY/MILLER '58)

The *structure function* of a language $W \subseteq X^*$:

$$s_W(n) := |\{w : w \in W \wedge |w| = n\}|$$

The *entropy* of a language $W \subseteq X^*$:

$$\mathbf{H}_W := \limsup_{n\to\infty} \frac{\log_r(1 + s_W(n))}{n}$$

Proposition

Let $W \subseteq X^*$. Then

$$\limsup_{n\to\infty} \frac{s_W(n)}{r^{\alpha\cdot n}} = \begin{cases} 0, & \text{if } \alpha > \mathbf{H}_W \text{ and} \\ \infty, & \text{if } \alpha < \mathbf{H}_W\,. \end{cases}$$

Entropy of languages: analytic functions (KUICH '70)

The *structure generating function* of a language $W \subseteq X^*$:

$$\begin{aligned} \mathfrak{s}_W : \mathbb{C} &\to \mathbb{C} \cup \{\infty\} \\ \mathfrak{s}_W(t) &:= \sum_{n \in \mathbb{N}} s_W(n) \cdot t^n \end{aligned}$$

$\operatorname{rad} W := \left(\limsup_{n\to\infty} \sqrt[n]{s_W(n)} \right)^{-1}$ is its *convergence radius*.

Proposition

1. $\mathbf{H}_W = -\log_r \operatorname{rad} W$ *if* W *is infinite,*
2. $|\mathfrak{s}_W(t)| < \infty$ *if* $|t| < \operatorname{rad} W$*, and*
3. $\mathfrak{s}_W(t) = \infty$ *for* $t > \operatorname{rad} W$ *if we consider* $\mathfrak{s}_W : \mathbb{R}_+ \to \mathbb{R}_+ \cup \{\infty\}$ *as a non-negative (monotone) function.*

Entropy of languages: properties and computability

$$\begin{array}{llcl} \text{finitely stable} & \mathbf{H}_{W\cup V} & = & \max\{\mathbf{H}_W, \mathbf{H}_V\} \\ \text{product} & \mathbf{H}_{W\cdot V} & = & \max\{\mathbf{H}_W, \mathbf{H}_V\} \text{ if } W\cdot V \neq \emptyset \end{array}$$

Theorem

1. *The entropy of regular (rational) languages is computable* [Chomsky/Miller '58]*.*
2. *The entropy of unambiguous context-free languages is computable* [Kuich '70]*.*
3. *The entropy of context-sensitive languages is uncomputable* [Kaminger '70]*.*

Entropy of languages: compression

Let $\gamma :\subseteq X^* \to X^*$ be a (partial) mapping.

γ is a *compression* for $W \subseteq X^*$ if $\mathrm{dom}(\gamma) \supseteq W$ and γ is one-to-one.

$\tau(w) := |\gamma(w)|/|w|$ is the *compression ratio* for $w \in W$.
The *average compression ratio* on $W \subseteq X^*$ is

$$\tau_{\mathrm{ave}}(W) := \limsup_{n \to \infty} \frac{\sum_{w \in W \cap X^n} \tau(w)}{s_W(n)} .$$

Theorem ([Hansel, Perrin, Simon '92])

For every infinite $W \subseteq X^*$ and every compression $\gamma :\subseteq X^* \to X^*$ of W we have

$$\mathbf{H}_W \leq \tau_{\mathrm{ave}}(W) .$$

Entropy of languages: star languages W^*

Definition (Kleene star)

$$W^* := \{w_1 \cdots w_\ell : \ell \geq 0 \wedge w_i \in W \text{ for } 1 \leq i \leq \ell\}$$

$$\mathfrak{s}_{W^*}(t) \leq \frac{1}{1 - \mathfrak{s}_W(t)} \text{ for } 0 \leq t < \infty$$

with equality if W is a code.

Theorem ([*St'88*])

Let $W \subseteq X^*$ be an infinite language. Then for every $\varepsilon > 0$ there is a finite subset $U \subseteq W$ such that

$$\mathbf{H}_{W^*} - \mathbf{H}_{U^*} < \varepsilon.$$

Entropy of languages: regular languages I

Lemma

If $W \subseteq X^*$ is a regular language accepted by a k-state automaton. Then

$$s_W(n) \le s_{\mathbf{infix}(W)}(n) \le (k+1)^2 \cdot \sum_{i=0}^{2k} s_W(n+i).$$

Corollary

If $W \subseteq X^*$ is a regular language then

$$\mathbf{H}_W = \mathbf{H}_{\mathbf{pref}(W)} = \mathbf{H}_{\mathbf{infix}(W)}.$$

Lemma

If $\emptyset \ne W \subseteq X^*$ is regular and a finite union of codes then

$$\mathbf{H}_W < \mathbf{H}_{W^*}.$$

Entropy of languages: regular languages II

Theorem ([Merzenich and *St.* '94])

Let W be regular and $s_W(n) \leq c \cdot r^{\mathbf{H}_W \cdot n}$ for some $c > 0$ and all $n \in \mathbb{N}$ and $\mathbf{H}_W = \mathbf{H}_{W \cap w \cdot X^*}$ for all $w \in \mathbf{pref}(W)$.
If $V \subseteq W$ is a regular language such that $V \cap w \cdot X^* \subset W \cap w \cdot X^*$ for all $w \in \mathbf{pref}(W)$ then $\mathbf{H}_V < \mathbf{H}_W$.

Lemma ([*St'85*])

Let $\emptyset \neq W \subseteq X^*$ be regular. Then there are constants $c_1, c_2 > 0$ such that

$$c_1 \cdot r^{\mathbf{H}_{W^*} \cdot n} \leq_{\mathrm{i.o.}} s_{W^*}(n) \leq c_2 \cdot r^{\mathbf{H}_{W^*} \cdot n}.$$

Corollary (Folklore: forbidden subwords)

If $\emptyset \neq W \subseteq X^*$ is regular and irreducible and $u \in \mathbf{infix}(W)$ then $\mathbf{H}_{W \setminus X^* u X^*} < \mathbf{H}_W$.

X^ω as CANTOR space

Metric: $\rho(\eta,\xi) := \inf\{r^{-|w|} : w \in \mathbf{pref}(\eta) \cap \mathbf{pref}(\xi)\}$

Balls in (X^ω,ρ): $w \cdot X^\omega = \{\eta : w \in \mathbf{pref}(\eta)\}$

$= \mathbb{B}_\varepsilon(\xi) = \{\eta : \rho(\xi,\eta) < \varepsilon\}$

when $w \in \mathbf{pref}(\xi)$ and $|w| = \lfloor -\log_r \varepsilon \rfloor + 1$

Diameter: $\operatorname{diam} w \cdot X^\omega = r^{-|w|}$

Open sets: $W \cdot X^\omega = \bigcup_{w \in W} w \cdot X^\omega$

Closure: $\mathrm{cl}_\rho(F) = \{\xi : \mathbf{pref}(\xi) \subseteq \mathbf{pref}(F)\}$

Theorem

(X^ω,ρ) *is a compact metric space.*

Dimensions in CANTOR space

1. entropy dimension, *or* box-counting dimension, *or* MINKOWSKI dimension *etc.* [Tricot '81]; upper and lower case
2. HAUSDORFF dimension
3. Packing dimension, *or* modified upper box-counting dimension

Dimensions measure to some extent the density of subsets in CANTOR space [Falconer '90].

Box-counting dimension

Idea: $s_{\mathbf{pref}(F)}(n)$ is the minimum number of balls of diameter r^{-n} to cover the set $F \subseteq X^\omega$.

Definition

lower box-counting dimension

$$\underline{\dim}_B(F) := \liminf_{n\to\infty} \log_r(s_{\mathbf{pref}(F)}(n)+1)/n$$

upper box-counting dimension

$$\overline{\dim}_B(F) := \limsup_{n\to\infty} \log_r(s_{\mathbf{pref}(F)}(n)+1)/n = \mathbf{H}_{\mathbf{pref}(F)}$$

Properties

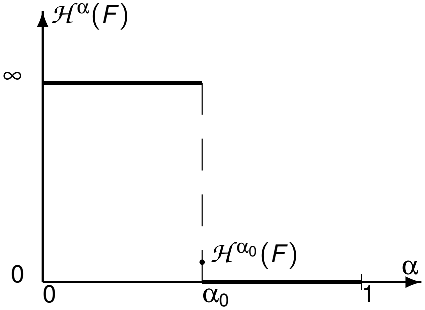
monotone	$E \subseteq F$	$\rightarrow$	$\underline{\dim}_B(E) \leq \underline{\dim}_B(F)$
finitely stable	$\overline{\dim}_B(E \cup F)$	$=$	$\max\{\overline{\dim}_B(E), \overline{\dim}_B(F)\}$
shift invariant	$\dim_B(w \cdot F)$	$=$	$\dim_B(F)$ (dim$_B$ for both)
closure	$\dim_B(F)$	$=$	$\dim_B(\mathrm{cl}_\rho(F))$

HAUSDORFF Measure

For $\mathcal{H}^{\alpha}(F;W) := \sum_{v\in W}(\mathrm{diam}\, v\cdot X^{\omega})^{\alpha} = \sum_{v\in W} r^{-\alpha\cdot|v|}$ the function

$$\mathcal{H}^{\alpha}(F) := \lim_{n\to\infty} \inf\left\{\mathcal{H}^{\alpha}(F;W) : W\cdot X^{\omega}\supseteq F \wedge \inf\{|v| : v\in W\}\geq n\right\}$$

is a metric outer measure on X^{ω}.



HAUSDORFF dimension

Definition (HAUSDORFF dimension)

$$\dim_H F = \sup\{\alpha : \alpha = 0 \vee \mathcal{H}^\alpha(F) = \infty\} = \inf\{\alpha : \mathcal{H}^\alpha(F) = 0\}$$

Properties

monotone	$E \subseteq F$	$\to$	$\dim_H(E) \leq \dim_H(F)$
countably stable	$\dim_H \bigcup_{i\in\mathbb{N}} F_i$	$=$	$\sup_{i\in\mathbb{N}} \dim_H(F_i)$
shift invariant	$\dim_H(w \cdot F)$	$=$	$\dim_H(F)$

HAUSDORFF dimension: a combinatorial property

Definition (Uniformly bounded growth of subtrees)

A subset $F \subseteq X^\omega$ is said to have *uniformly bounded growth* provided for every $n \in \mathbb{N}$ and all $\varepsilon > 0$ the condition

$$\lim_{|w|\to\infty} \frac{s_{\mathbf{pref}(F)\cap w\cdot X^*}(n+|w|)}{s_{\mathbf{pref}(F)}(n)\cdot r^{\varepsilon\cdot|w|}} = 0$$

holds true.

Theorem (*St.* '89)

Let $F \subseteq X^\omega$. If F has uniformly bounded growth then

$$\dim_H \mathrm{cl}_\rho(F) = \overline{\dim}_B F = \mathbf{H}_{\mathbf{pref}(F)}\,.$$

Packing dimension

Definition (Packing *or* modified upper box-counting dimension)

$$\dim_P F := \inf\left\{\sup_{i\in\mathbb{N}} \overline{\dim}_B F_i : \bigcup_{i\in\mathbb{N}} F_i \supseteq F\right\}$$

Properties

monotone	$E \subseteq F$	$\rightarrow$	$\dim_P(E) \leq \dim_P(F)$
countably stable	$\dim_P \bigcup_{i\in\mathbb{N}} F_i$	$=$	$\sup_{i\in\mathbb{N}} \dim_P(F_i)$
shift invariant	$\dim_P(w \cdot F)$	$=$	$\dim_P(F)$

Proposition (Relations)

$\dim_H F \leq \dim_P F \leq \overline{\dim}_B F$ *and* $\dim_H F \leq \underline{\dim}_B F \leq \overline{\dim}_B F$

Entropy characterisations

Let $W \subseteq X^*$ and define the limits

$$\begin{aligned} \textit{i.o.-limit} \quad \overrightarrow{W} &:= \{\xi : \xi \in X^\omega \wedge |\mathbf{pref}(\xi) \cap W| = \infty\} \text{ and} \\ \textit{a.e.-limit} \quad W^{\uparrow} &:= \{\xi : \xi \in X^\omega \wedge |\mathbf{pref}(\xi) \setminus W| < \infty\} \end{aligned}$$

Proposition ([Rogers '70, *St.* '93, Hitchcock '05])

Let $F \subseteq X^\omega$. Then

$$\begin{aligned} \dim_H F &:= \inf\{\mathbf{H}_W : W \subseteq X^* \wedge F \subseteq \overrightarrow{W}\} \textit{ and} \\ \dim_P F &:= \inf\{\mathbf{H}_W : W \subseteq X^* \wedge F \subseteq W^{\uparrow}\}. \end{aligned}$$

Entropy characterisations: effectivisation

Definition (Σ_2-definable ω-languages)

$F \subseteq X^\omega$ is Σ_2-*definable* $:\Leftrightarrow$
there is a computable set $M_F \subseteq \mathbb{N} \times X^*$ such that

$$\xi \in F \quad \longleftrightarrow \quad \exists i \in \mathbb{N} : \forall w \in \mathbf{pref}(\xi) : (i, w) \in M_F \ .$$

Corollary

$F \subseteq X^\omega$ is Σ_2-definable if and only if there is a computable $W \subseteq X^*$ such that $F = W^\uparrow$.

Theorem (*St.* '98)

If $F \subseteq X^\omega$ is a Σ_2-definable set then

$$\dim_H F = \inf \{ \mathbf{H}_W : W \subseteq X^* \wedge W \text{ is computable } \wedge F \subseteq \overrightarrow{W} \} \, .$$

ω-power languages and regular ω-languages

ω-power languages: $W^\omega := \{w_1 \cdots w_i \cdots : w_i \in W \wedge |w_i| > 0\}$

Proposition

$$\dim_H W^\omega = \mathbf{H}_{W^*} \quad \textit{and} \quad \dim_P \mathrm{cl}_\rho(W^\omega) = \overline{\dim}_B W^\omega$$

Regular ω-languages: $F = \bigcup_{i=1}^n V_i \cdot W_i^\omega$ where all languages V_i, W_i are regular.

Proposition

1. *If $W \subseteq X^*$ is regular then* $\dim_H W^\omega = \overline{\dim}_B W^\omega$
2. *If $F \subseteq X^\omega$ is regular then* $\dim_H F = \dim_P F$.

Regular ω-languages: density

Proposition

Let $\emptyset \neq F \subseteq X^{\omega}$ be regular, $\alpha = \dim_H F$. Then $\mathcal{H}^{\alpha}(F) > 0$.

Theorem (Measure-category theorem [*St.* '98])

Let $\emptyset \neq F \subseteq X^{\omega}$ be regular, $0 < \alpha = \dim_H F$, $\mathcal{H}^{\alpha}(F) < \infty$ and $\dim_H(F \cap w \cdot X^{\omega}) = \dim_H F$ whenever $F \cap w \cdot X^{\omega} \neq \emptyset$.
Then for every regular $E \subseteq F$ the following conditions are equivalent:

1. *E is of first* Baire *category in F,*
2. *$\mathcal{H}^{\alpha}(E) = 0$, and*
3. *$\dim_H E < \dim_H F$.*

Applications: tail exchange property

Theorem (Semenov '84, Perrin and Schupp '86)

Let $E \subseteq {}^{-\omega}X^{\omega}$ be an automaton definable set of bi-infinite words, and a let $\mathbf{x}, \mathbf{y} \in {}^{-\omega}X^{\omega}$ having the same set of factors which in addition occur to both sides infinitely often.
Then $\mathbf{x} \in E$ implies $\mathbf{y} \in E$.

Theorem ([*St.* '98, *St.* '12])

Let $F \subseteq X^{\omega}$ be regular, ξ, η have recurrent tails and $\mathbf{infix}_{\infty}(\xi) = \mathbf{infix}_{\infty}(\eta)$ be a regular language.
If $\xi \in F$ then there are $w \in \mathbf{pref}(\xi)$ and an $m \in \mathbb{N}$ such that $w \cdot \eta[m..\infty] \in F$.

Further applications: KOLMOGOROV complexity

β-entropy: definition

Recall

$$\sum_{w\in W}\left(\frac{1}{r}\right)^{s\cdot|w|}=\begin{cases}\infty, & \text{if } s<\mathbf{H}_W \text{ and}\\ <\infty, & \text{if } s>\mathbf{H}_W.\end{cases}$$

Let $\beta:(X^*,\cdot)\to((0,\infty),\cdot)$ be a morphism (valuation,distribution).

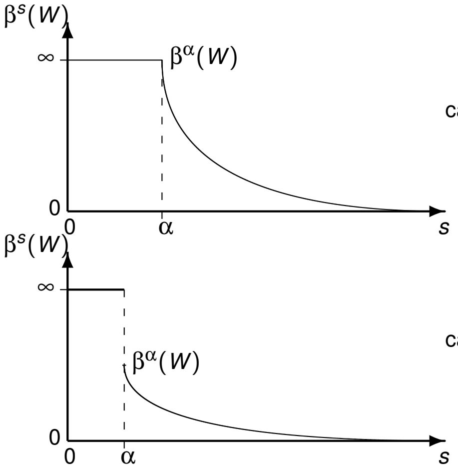
Definition (β-entropy)

Let $W\subseteq X^*$. The unique point $\alpha\in(0,\infty)\cup\{\infty\}$ for which

$$\beta^s(W):=\sum_{w\in W}\beta(w)^{s\cdot|w|}=\begin{cases}\infty, & \text{if } s<\alpha \text{ and}\\ <\infty, & \text{if } s>\alpha\end{cases}$$

holds is referred to as the β-*entropy* $\mathbf{H}_W^{\beta}$ of W.

β-entropy: typical plots



case $\beta^\alpha(W) = \infty$

case $\beta^\alpha(W) < \infty$

β-entropy: regular languages

Theorem ([Mauldin/Williams '88, Bandt '89])

If β is a computable mapping and W is regular then $\mathbf{H}^{\beta}_{W}$ is computable.

Lemma

If $W \subseteq X^*$ is a regular language then $\mathbf{H}^{\beta}_{W} = \mathbf{H}^{\beta}_{\mathbf{pref}(W)} = \mathbf{H}^{\beta}_{\mathbf{infix}(W)}$.

Lemma ([Fernau/*St.* '01])

If $\emptyset \neq W \subseteq X^*$ is regular and a finite union of codes and $\mathbf{H}^{\beta}_{W} < \infty$ then

$$\mathbf{H}^{\beta}_{W} < \mathbf{H}^{\beta}_{W^*}.$$

β-entropy: star languages

Definition (c-essential domain for β)

$$V_{\beta,c} := \{w : w \in X^* \wedge \beta(w) \leq c^{|w|}\}$$

Remark. If $\max\{\beta(x) : x \in X\} \leq c < 1$ then $V_{\beta,c} = X^*$.

Theorem ([Fernau/*St.* '01])

Let $W \subseteq V_{\beta,c}$ for some $c < 1$. Then, for every $\varepsilon > 0$, there is a finite subset $U \subseteq W$ such that

$$\mathbf{H}^{\beta}_{W^*} - \mathbf{H}^{\beta}_{U^*} < \varepsilon \,.$$

β-metric in CANTOR-space

Definition (β-metric)

$$\rho_\beta(\xi,\eta) = \begin{cases} 0 & \text{, if } \xi = \eta \text{ , and} \\ \min\{\beta(w) : w \in \mathbf{pref}(\xi) \cap \mathbf{pref}(\eta)\} & \text{, otherwise.} \end{cases}$$

Fact

1. *If* $\beta(x) < 1$ *for all* $x \in X$ *then* (X^ω, ρ_β) *is (topologically) homeomorphic to the usual* CANTOR*-space* (X^ω, ρ).
2. *If* $\beta(x) \geq 1$ *for some* $x \in X$ *then the space* (X^ω, ρ_β) *contains isolated points.*

HAUSDORFF-dimension in (X^ω, ρ_β)

$$\mathcal{H}_\beta^\alpha(F) := \lim_{\varepsilon\to 0} \inf\left\{ \sum_{w\in W} r^{-\alpha\cdot|w|} : F \subseteq W\cdot X^\omega \wedge \forall w(w\in W \to \beta(w)\le\varepsilon)\right\}$$

$$\dim_H^\beta F = \sup\{\alpha : \alpha = 0 \vee \mathcal{H}_\beta^\alpha(F) = \infty\} = \inf\{\alpha : \mathcal{H}_\beta^\alpha(F) = 0\}$$

Lemma (Countable stability)

$$\dim_H^{(\beta)} \bigcup_{i\in\mathbb{N}} F_i = \sup_{i\in\mathbb{N}} \dim_H^{(\beta)} F_i$$

Theorem (Entropy characterisation via the *i.o.*-limit)

Let $F \subseteq \overrightarrow{V_{\beta,c}}$ *for some* $c < 1$. *Then*

$$\dim_H^{(\beta)} F = \inf\left\{ H_W^\beta : W \subseteq V_{\beta,c} \wedge F \subseteq \overrightarrow{W}\right\}.$$

ω-power languages and regular ω-languages

Lemma

If $c \in (0,1)$ *and* $W \subseteq V_{\beta,c}$*, then* $\dim_H^{(\beta)} W^\omega = \mathbf{H}_{W^*}^{\beta}$.

Lemma

If $c \in (0,1)$ *and* $W \subseteq V_{\beta,c}$ *is a regular language, then*

$$\dim_H^{(\beta)} W^\omega = \dim_H^{(\beta)} \mathrm{cl}_\beta(W^\omega).$$

References

- C. Bandt.
  Self-similar sets 3. Constructions with sofic systems.
  Monatsh. Math., 108:89–102, 1989.
- N. Chomsky and G.A. Miller,
  Finite-state languages,
  Inform. Control 1:91–112, 1958.
- K.J. Falconer,
  Fractal Geometry.
  Wiley, 1990.
- H. Fernau and L. Staiger,
  Iterated function systems and control languages,
  Inform. and Comput., 168:125–143, 2001.

References

G. Hansel, D. Perrin, and I. Simon.
Entropy and compression,
in: *STACS'92*, A. Finkel and M. Jantzen (Eds.),
Lect. Notes in Comput. Sci. 577, Springer-Verlag, Berlin, 515–530 1992.

J.M. Hitchcock.
Correspondence principles for effective dimension.
Theory Comput. Systems 38:559–571, 2005.

F. P. Kaminger.
The noncomputability of the channel capacity of context-sensitive languages.
Inform. Control, 17:175–182, 1970.

W. Kuich.
On the entropy of context-free languages.
Inform. Control, 16:173–200, 1970.

References

- R. D. Mauldin and S. C. Williams.
  Hausdorff dimension in graph directed constructions.
  Trans. AMS, 309(2):811–829, 1988.
- W. Merzenich and L. Staiger.
  Fractals, dimension, and formal languages.
  RAIRO Inf. théor. Appl., 28(3–4):361–386, 1994.
- D. Perrin and P.E. Schupp.
  Automata on the integers, recurrence distinguishability, and the equivalence and decidability of monadic theories,
  in: Proc. 1st Symp. Logic in Computer Science, IEEE Press, 301–304, 1986.
- C. A. Rogers,
  Hausdorff Measures.
  Cambridge University Press, 1970.

References

A. Semenov.
Decidability of monadic theories,
in: Math. Found. of Comput. Sci. (M. Chytil and V. Koubek eds.) *Lect. Notes in Comput. Sci.* 176, Springer-Verlag, Berlin, 162–175, 1984.

L. Staiger.
The entropy of finite-state ω-languages.
Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform., 14(5):383–392, 1985.

L. Staiger.
Ein Satz über die Entropie von Untermonoiden.
Theor. Comput. Sci., 61:279–282, 1988.

L. Staiger.
Combinatorial properties of the Hausdorff dimension.
J. Statist. Plann. Inference, 23:95–100, 1989.

References

L. Staiger.
Kolmogorov complexity and Hausdorff dimension.
Inform. and Comput., 103:159–194, 1993.

L. Staiger,
A tight upper bound on Kolmogorov complexity and uniformly optimal prediction,
Theory Comput. Systems 31:215–229, 1998.

L. Staiger.
Rich ω-words and monadic second-order arithmetic.
in: *CSL'97, Selected Papers*,
Lecture Notes in Comput. Sci. 1414, Springer-Verlag, Berlin, 478–490, 1998.

References

L. Staiger.
The Kolmogorov complexity of infinite words,
Theor. Comput. Sci. 383:187–199, 2007.

L. Staiger.
Asymptotic subword complexity,
in: *Languages Alive*,
Lecture Notes in Comput. Sci. 7300, Springer-Verlag, Heidelberg, 236–245, 2012.

C. Tricot.
Douze définitions de la densité logarithmique.
CR de l' Académie des Sciences (Paris), série I, 293:549–552, 1981.