

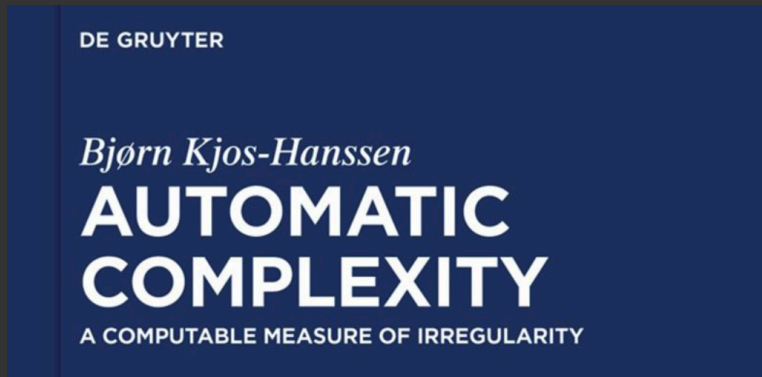
>>> Conditional automatic complexity and its metrics

Name: Bjørn Kjos-Hanssen (University of Hawai'i at Mānoa)

Date: December 15, 2023, 10:02am-10:16am

COCOON 2023

>>> Book coming in 2024



>>> Abstract

Li, Chen, Li, Ma, and Vitányi (2004) introduced a similarity metric based on Kolmogorov complexity. It followed work by Shannon in the 1950s on a metric based on entropy. We define two computable similarity metrics, analogous to the Jaccard distance and Normalized Information Distance, based on conditional automatic complexity and show that they satisfy all axioms of metric spaces.

```
>>> Outline
```

1. Motivation

2. Definitions

3. History

4. Complexity of complexity

5. Bounds

6. Conditional complexity

>>> A good notion of complexity $C(x)$, $x = 01101001$

- * C should be computable (in single-exponential time?)
- * C should be well defined
(not just 'up to a constant')
- * The problem: $C(x) \leq k$ should be NP-complete: a general search problem.
- * C should be defined in a not too convoluted way and accessible to early-stage researchers
- * C should be robust: several reasonable definitions turn out to be equivalent.

This matches the intuition that the true complexity of a pattern should be hard but not impossible to discern.

Automatic complexity $A(x)$ has these properties (except that robustness is a bit of an open question, and NP-completeness is only known for partitions in place of words).

back to Outline

>>> DFA

A deterministic finite automaton (DFA) is a 5-tuple $M = (Q, \Sigma, \delta, q_0, F)$ where

- * Q is a finite set of *states*;
- * Σ is a finite *alphabet*;
- * $\delta : Q \times \Sigma \rightarrow Q$ is a *transition function*;
- * $q_0 \in Q$ is an *initial state*;
- * $F \subseteq Q$ is a set of *final states*.

>>> NFA

A nondeterministic finite automaton (NFA) is a 5-tuple $M = (Q, \Sigma, \delta, q_0, F)$ where

- * Q is a finite set of *states*;
- * Σ is a finite *alphabet*;
- * $\delta : Q \times \Sigma \rightarrow \mathcal{P}(Q)$ is a *transition function*;
- * $q_0 \in Q$ is an *initial state*;
- * $F \subseteq Q$ is a set of *final states*.

>>> Language recognized by DFA

Let ε be the empty word. We define $\delta^* : Q \times \Sigma^* \rightarrow Q$ by

$$\begin{aligned}\delta^*(q, \varepsilon) &= q \\ \delta^*(q, xa) &= \delta(\delta(q, x), a)\end{aligned}$$

for $x \in \Sigma^*$ and $a \in \Sigma$.

$$L(M) = \{x : \delta^*(q_0, x) \in F\}$$

is the language recognized by M .

>>> Language recognized by NFA

Let ε be the empty word. We define $\delta^* : Q \times \Sigma^* \rightarrow \mathcal{P}(Q)$ by

$$\begin{aligned}\delta^*(q, \varepsilon) &= \{q\} \\ \delta^*(q, xa) &= \bigcup_{r \in \delta(q, x)} \delta(r, a)\end{aligned}$$

for $x \in \Sigma^*$ and $a \in \Sigma$.

$$L(M) = \{x : \delta^*(q_0, x) \in F\}$$

is the language recognized by M .

Definition

Let $\Sigma = \{0, 1\}$ and $x \in \Sigma^*$. The automatic complexity $A(x)$ of x is the least $|Q|$ over all DFAs M with $L(M) \cap \Sigma^{|x|} = \{x\}$.

Example

$A(0^n) = 2$ for all $n \geq 1$.

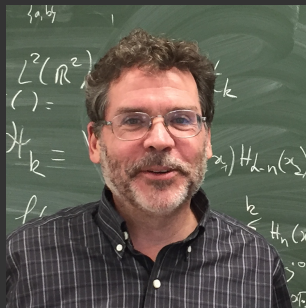


Figure: Jeff Shallit

>>> State sequences

Let M be an NFA. An *accepting state sequence* for $x = x_0 \dots x_{n-1}$ in M , where $x_i \in \Sigma$ for $0 \leq i < n$, is a sequence (q_0, \dots, q_n) where $q_i \in Q$ and

$$q_{i+1} \in \delta^*(q_i, x_i)$$

for each $0 \leq i \leq n$, and $q_n \in F$.

Let $P(x, M)$ be the set of accepting state sequences for x in M .

>>> Nondeterministic automatic complexity A_N

We say that an NFA M path-uniquely accepts x if

- * $|P(x, M)| = 1$;
- * $|P(x, M)| = 0$ for all $y \neq x$, $|y| = |x|$.

$A_N(x) = A_{Nu}(x)$ is the minimum of $|Q|$ over all NFAs M which path-uniquely accept x .

>>> Robustness question

Definition

$A_{Ne}(x)$ is the minimum of $|Q|$ over all NFAs M with $L(M) \cap \Sigma^{|x|} = \{x\}$.

- * A_{Ne} is a direct analogue of A ;
- * A_{Nu} is easier to compute in practice.

The main open robustness question for A_N :

Question

$$A_{Ne} = A_{Nu}?$$

>>> Comparative disagreement

If X is a set and (Y, \leq) a linear order then any $f: X \rightarrow Y$ induces a total preorder on X by $a \leq_f b \iff f(a) \leq f(b)$.
 f, g induce *distinct* orders if $\leq_f \neq \leq_g$.
 f, g induce *incompatible* orders if there exists a, b with $f(a) < f(b)$, $g(b) < g(a)$.

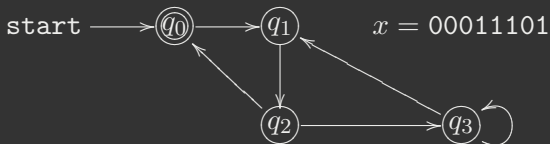
>>> Edge (transition) complexity

An edge of M is a pair (q_1, q_2) with $q_2 \in \delta(q_1, a)$ for some $a \in \Sigma$ (the label of the edge).

The set of edges of M is $E(M)$.

$E_N(x)$ is the minimal $|E(M)|$ over all NFAs M path-uniquely accepting x .

>>> E_N and A_N induce incompatible orders



	x	y
A_N	4	5
E_N	6	5



$y = 00001000$

back to Outline

>>> Pre-history

- * Turing [1936]: Universal Turing machine
- * McCulloch and Pitts [1943]: Idea of finite automata
- * Rabin and Scott [1959]: Finite automata introduced
- * Kolmogorov [1965] Kolmogorov complexity of a string
 $x = 0111001$, say, = the length of the shortest program
printing x
- * Kolmogorov [1974]: Kolmogorov's structure functions
- * Diwan [1986]: Complexity based on context-free grammars,
word chains
- * Yu et al. [1994]: State complexity questions for finite
automata

>>> The 2000s: the beginning

- * Shallit and Wang [2001], 2nd Workshop on Descriptive Complexity of Automata, Grammars and Related Structures (London, Ontario, July 27 -- 29, 2000): definitions and the first results.
- * I independently defined automatic complexity in a Discrete mathematics (Math 301) class in 2009

>>> The 2010s

- * Hyde and Kjos-Hanssen [2014, 2015]: Nondeterministic version and first results thereon: cubefree words are maximally complex
- * Kjos-Hanssen [2014, 2015]: automatic structure functions, thanks to a suggestion of Vereshchagin in Singapore 2014
- * Kjos-Hanssen [2017a]: Quantum automatic complexity: counting eigenstates vs. counting states
- * Kjos-Hanssen [2017b, 2018]: Automatic complexity of linear shift register sequences is $n/2 - O(\log^2 n)$
- * Kjos-Hanssen [2017c]: maximally complex words not a CFL.

>>> The 2020s

- * Kjos-Hanssen [2021a]: Infinite Fibonacci words have intermediate complexity rate
- * Jordon and Moser [2021]: normal sequences with ``non-maximal'' automatic complexity. Open problem though: are there infinite words x with $A(x \upharpoonright n)/n \rightarrow 1$?
- * Incompressibility theorem (answering question from 2001; Kjos-Hanssen [2021b])

back to Outline

>>> Complexity of complexity

- * Problem: given x and k , determine whether $A(x) \leq k$.
- * Problem: given x and k , determine whether $A_N(x) \leq k$.

These problems are in NP. It is not known whether they are NP-complete.

What's known:

- * Related problems are not in CFL or coCFL, and not in SAC^0 (recognized by semi-unbounded fan-in constant depth circuits) or $coSAC^0$.
- * A related problem of automatic complexity of equivalence relations is NP-complete.

back to Outline

```
>>> Automatic complexity rate
```

Deterministic:

$$\frac{A(x)}{|x|}$$

Nondeterministic:

$$\frac{A_N(x)}{|x|}$$

>>> Upper bounds

Theorem ([Shallit and Wang, 2001, Theorem 6])

The automatic complexity rate of almost all binary words is upper bounded by $3/4 + \epsilon$ for each $\epsilon > 0$.

Theorem ([Kjos-Hanssen, 2019, Theorem 26])

The automatic complexity rate of almost all b -ary words is upper bounded by $\frac{1}{2} + \frac{1}{2b} + \epsilon$ for each $\epsilon > 0$.

As $b \rightarrow \infty$ the rate bound is $1/2$ which is sharp.

>>> Upper bounds

Words with fixed alphabet size cannot get too close to the $A(x)$ upper bound of $|x| + 1$; there is a logarithmic gap:

Theorem ([Shallit and Wang, 2001, Theorem 5])

Let x be a k -ary word.

If $n \geq k^t + t$ then $A(x) \leq n + 2 - t$.

The proof involves looking for repeated subwords.

Theorem (K., in prep.)

$$A(x) \leq |x| + 3 - \frac{1}{2}\sqrt{|x|}.$$

The proof involves looking at runs from $0^*, 1^*, (01)^*, (10)^*$.

Question

Is $A(x) \leq |x|/2 + o(|x|)$?

>>> Upper bound of A_N

Theorem (Hyde (2013), see [Hyde and Kjos-Hanssen, 2015, Theorem 3])

Let $x \in \Sigma^n$. Then $A_N(x) \leq \frac{n}{2} + 1$.

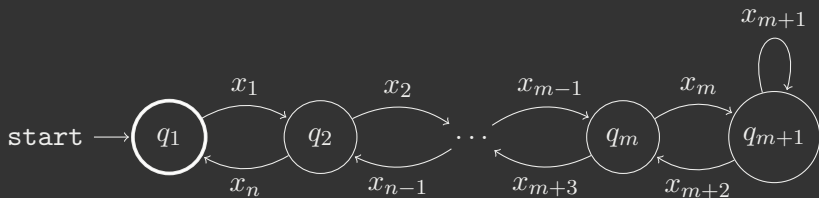


Figure: An NFA uniquely accepting $x = x_1x_2\dots x_n$, $n = 2m + 1$

>>> Shifted versions

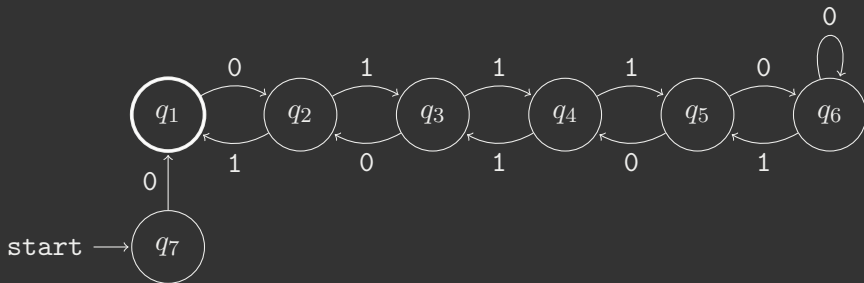
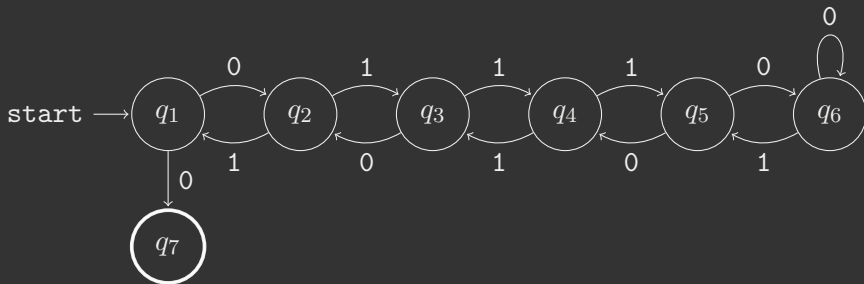


Figure: NFAs uniquely accepting $x = 011100101010$ and

$y = 001110010101$.

>>> Deficiency

Definition

The A_N -deficiency of x is $\lfloor n/2 \rfloor + 1 - A_N(x)$.

Theorem (Shallit and Wang 2001)

If $x \in \Sigma^n$ is square-free then $A(x) \geq \frac{n+1}{2}$.

Therefore square-free words have A_N -deficiency 0. So Hyde's theorem is sharp.

>>> Lower bounds

Theorem ([Shallit and Wang, 2001, Theorem 8])

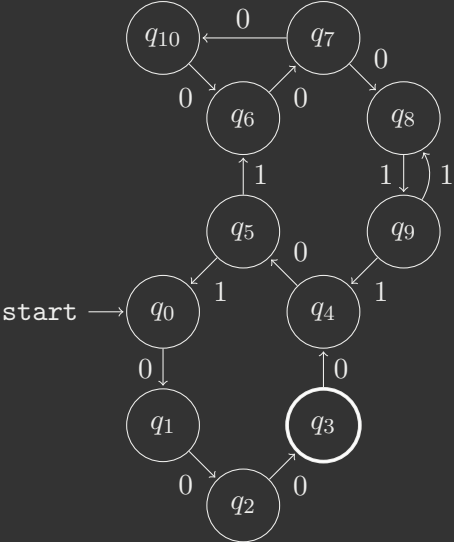
$A(x) \geq n/13$ for almost all binary x .

They stated that Holger Petersen has informed them that the result can be strengthened to $A(x) \geq n/7$.

Theorem ([Kjos-Hanssen, 2021b, Theorem 18])

$A(x) \geq n/(2 + \epsilon)$ for almost all x , for any $\epsilon > 0$.

Beros et al. [2019]: the digraphs representing the witnessing automata are planar, in fact they are **trees of cycles**. The cycles are added in Kleene--Brouwer order. Example: $0^5 1 0^5 1^6 0 1 0^3$:



back to Outline

>>> Tracks

Definition (product of words)

Let Γ and Δ be alphabets. Let $n \in \mathbb{N}$, $x \in \Gamma^n$ and $y \in \Delta^n$. When no confusion with binomial coefficients is likely, we let $\binom{a}{b} = (a, b) \in \Sigma \times \Delta$. The *product* of x and y , $x \times y \in (\Gamma \times \Delta)^n$, is defined to be the word

$$\binom{x_0}{y_0} \binom{x_1}{y_1} \cdots \binom{x_{n-1}}{y_{n-1}},$$

which we may also denote as $\binom{x}{y}$.

Definition (projections of a word)

Let Γ and Δ be alphabets. Let $n \in \mathbb{N}$, $x \in \Gamma^n$ and $y \in \Delta^n$. The projections π_1 and π_2 are defined by $\pi_1(x \times y) = x$, $\pi_2(x \times y) = y$.

The product $x \times y$ (Shallit [2023]) is also known as the track $x\#y$ (Kjos-Hanssen [2024]).

A rainbow word is defined inductively: the empty word is one, and if x, y are rainbow words then so is xy provided x and y have no common symbols.

Theorem

For all words x, y , we have $\max\{A_N(x), A_N(y)\} \leq A_N(x \times y)$.
There exist words x, y with $\max\{A^-(x), A^-(y)\} \not\leq A^-(x \times y)$.

Proof.

Let x be a word of some length n with $A^-(x) > n/2 + 1$. An example can be found among the maximum-length sequences for linear feedback shift registers as observed in Kjos-Hanssen [2018]. Let y be a rainbow word of the same length. Whenever y is a rainbow word, so is $x \times y$. Therefore $A^-(x \times y) \leq n/2 + 1 < A^-(x)$ by a general upper bound theorem due to Hyde (COCOON 2014).

Definition

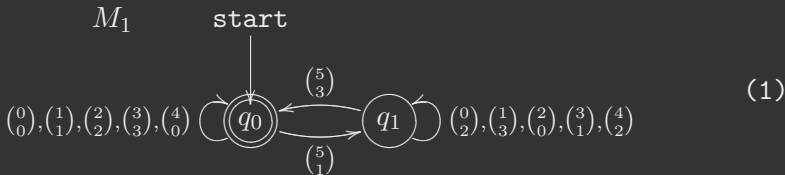
Let Γ and Δ be alphabets. Let $n \in \mathbb{N}$ and $x \in \Gamma^n, y \in \Delta^n$. The *conditional (nondeterministic) automatic complexity of x given y* , $A_N(x \mid y)$, is the minimum number of states of an NFA over $\Gamma \times \Delta$ such that Item i and Item ii hold.

- (i) Let m be the number of accepting walks of length $n = |x| = |y|$ for which the word w read on the walk satisfies $\pi_1(w) = y$. Then $m = 1$.
- (ii) M accepts $y \times x$.

Note that M may accept many words of length n , but only one of the form $y \times (?)$.

>>> Example

Let $y = (012345)^k$ for some large k , and let $x = (0123)^l$ where $4l = 6k$, so that $|x| = |y|$. We have $A_N(x \times y) = \text{lcm}(4, 6) = 12$, and $A_N(y) = 6$. The fact that $A_N(x \mid y) = 2$ is witnessed by the NFA in (1).



Theorem

$$A_N(x \times y) \leq A_N(x \mid y) \cdot A_N(y).$$

We show $A_N(x) \leq A_N(y \times x) \leq A_N((y \times x) \times y) \leq A_N(x \mid y) \cdot A_N(y)$.
Actually $A_N(y \times x) = A_N((y \times x) \times y)$ always holds since we can use an invertible morphism $(b, a) \leftrightarrow ((b, a), b)$. Construction for the proof: Let NFAs

$$M_1 = (Q_1, \Gamma \times \Delta, \delta_1, q_{0,1}, F_1), \quad M_2 = (Q_2, \Gamma, \delta_2, q_{0,2}, F_2)$$

be given. The product is

$$M_1 \times_1 M_2 = (Q_1 \times Q_2, \Delta, \delta, (q_{0,1}, q_{0,2}), F_1 \times F_2)$$

where $(r, r') \in \delta((q, q'), a)$ if $r \in \delta_1(q, (b, a))$ and $r' \in \delta_2(q', b)$ for some b .

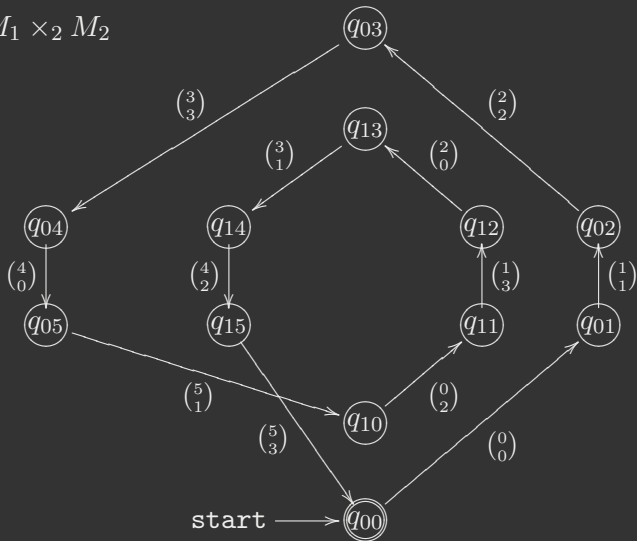
(We can also form $M_1 \times_2 M_2$ where $(r, r') \in \delta((q, q'), (b, a))$ if $r \in \delta_1(q, (b, a))$ and $r' \in \delta_2(q', b)$.)

The product $M_1 \times_2 M_2$ can be viewed as the natural NFA for

$$\varphi((L(M_1) \times L(M_2)) \cap \Sigma^*)$$

where Σ is a ``diagonal'' subalphabet of $\Gamma \times \Delta$, and φ is the projection morphism $((b, a), c) \rightarrow (b, a)$. Then $M_1 \times_1 M_2$ is obtained by also applying the morphism $(b, a) \rightarrow a$.

$$M_1 \times_2 M_2$$



(2)

>>> Metrics

Using the conditional automatic complexity we obtain metrics where Vitányi et al. only obtained approximate metrics for Kolmogorov complexity.

Definition

Let x, y be words of length $n \in \mathbb{N}$. We define

$$J_{\max}(x, y) = \frac{\log \max\{A(x \mid y), A(y \mid x)\}}{\log \max\{A(x), A(y)\}}$$

>>> References I

- A. A. Beros, B. Kjos-Hanssen, and D. K. Yogi. Planar digraphs for automatic complexity. In *Theory and applications of models of computation*, volume 11436 of *Lecture Notes in Comput. Sci.*, pages 59--73. Springer, Cham, 2019. doi: 10.1007/978-3-030-14812-6_5. URL https://doi.org/10.1007/978-3-030-14812-6_5.
- A. A. Diwan. A new combinatorial complexity measure for languages. *Tata Institute, Bombay, India*, 1986.
- K. K. Hyde and B. Kjos-Hanssen. Nondeterministic automatic complexity of almost square-free and strongly cube-free words. In *Computing and combinatorics*, volume 8591 of *Lecture Notes in Comput. Sci.*, pages 61--70. Springer, Cham, 2014. doi: 10.1007/978-3-319-08783-2_6. URL https://doi.org/10.1007/978-3-319-08783-2_6.

>>> References II

- K. K. Hyde and B. Kjos-Hanssen. Nondeterministic automatic complexity of overlap-free and almost square-free words. *Electron. J. Combin.*, 22(3):Paper 3.22, 18, 2015. doi: 10.37236/4851. URL <https://doi.org/10.37236/4851>.
- L. Jordon and P. Moser. Normal sequences with non maximal automatic complexity. *CoRR*, abs/2107.05979, 2021. URL <https://arxiv.org/abs/2107.05979>.
- B. Kjos-Hanssen. Kolmogorov structure functions for automatic complexity in computational statistics. In *Combinatorial optimization and applications*, volume 8881 of *Lecture Notes in Comput. Sci.*, pages 652--665. Springer, Cham, 2014. doi: 10.1007/978-3-319-12691-3_49. URL https://doi.org/10.1007/978-3-319-12691-3_49.

>>> References III

- B. Kjos-Hanssen. Kolmogorov structure functions for automatic complexity. *Theoret. Comput. Sci.*, 607(part 3):435--445, 2015. ISSN 0304-3975. doi: 10.1016/j.tcs.2015.05.052. URL <https://doi.org/10.1016/j.tcs.2015.05.052>.
- B. Kjos-Hanssen. Superposition as memory: unlocking quantum automatic complexity. In *Unconventional computation and natural computation*, volume 10240 of *Lecture Notes in Comput. Sci.*, pages 160--169. Springer, Cham, 2017a. doi: 10.1007/978-3-319-58187-3_12. URL https://doi.org/10.1007/978-3-319-58187-3_12.
- B. Kjos-Hanssen. Shift registers fool finite automata. In *Logic, language, information, and computation*, volume 10388 of *Lecture Notes in Comput. Sci.*, pages 170--181. Springer, Berlin, 2017b. doi: 10.1007/978-3-662-55386-2_12. URL https://doi.org/10.1007/978-3-662-55386-2_12.

>>> References IV

- B. Kjos-Hanssen. On the complexity of automatic complexity. *Theory Comput. Syst.*, 61(4):1427--1439, 2017c. ISSN 1432-4350. doi: 10.1007/s00224-017-9795-4. URL <https://doi.org/10.1007/s00224-017-9795-4>.
- B. Kjos-Hanssen. Automatic complexity of shift register sequences. *Discrete Math.*, 341(9):2409--2417, 2018. ISSN 0012-365X. doi: 10.1016/j.disc.2018.05.015. URL <https://doi.org/10.1016/j.disc.2018.05.015>.
- B. Kjos-Hanssen. Few paths, fewer words: model selection with automatic structure functions. *Exp. Math.*, 28(1):121--127, 2019. ISSN 1058-6458. doi: 10.1080/10586458.2017.1368048. URL <https://doi.org/10.1080/10586458.2017.1368048>.

>>> References V

- B. Kjos-Hanssen. Automatic complexity of Fibonacci and tribonacci words. *Discrete Appl. Math.*, 289:446--454, 2021a. ISSN 0166-218X. doi: 10.1016/j.dam.2020.10.014. URL <https://doi.org/10.1016/j.dam.2020.10.014>.
- B. Kjos-Hanssen. An incompressibility theorem for automatic complexity. *Forum of Mathematics, Sigma*, 9:e62, 2021b. doi: 10.1017/fms.2021.58.
- B. Kjos-Hanssen. *Automatic complexity: a computable measure of irregularity*, volume 12 of *De Gruyter Series in Logic and its Applications*. De Gruyter, Berlin, 2024.
- A. Kolmogorov. Meetings of the moscow mathematical society. *Uspekhi Mat. Nauk*, 29(4(178)):153--160 (Kolmogorov on p. 155), 1974.

>>> References VI

- A. N. Kolmogorov. Three approaches to the definition of the concept ``quantity of information''. *Problemy Peredači Informacii*, 1(vyp. 1):3--11, 1965. ISSN 0555-2923.
- W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5: 115--133, 1943. ISSN 0007-4985. doi: 10.1007/bf02478259. URL <https://doi.org/10.1007/bf02478259>.
- M. O. Rabin and D. Scott. Finite automata and their decision problems. *IBM J. Res. Develop.*, 3:114--125, 1959. ISSN 0018-8646. doi: 10.1147/rd.32.0114. URL <https://doi.org/10.1147/rd.32.0114>.

>>> References VII

- J. Shallit. *The logical approach to automatic sequences---exploring combinatorics on words with Walnut*, volume 482 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2023. ISBN 978-1-108-74524-6.
- J. Shallit and M.-W. Wang. Automatic complexity of strings. volume 6, pages 537--554. 2001. 2nd Workshop on Descriptive Complexity of Automata, Grammars and Related Structures (London, ON, 2000).
- A. M. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proc. London Math. Soc.* (2), 42 (3):230--265, 1936. ISSN 0024-6115. doi: 10.1112/plms/s2-42.1.230. URL <https://doi.org/10.1112/plms/s2-42.1.230>.

>>> References VIII

S. Yu, Q. Zhuang, and K. Salomaa. The state complexities of some basic operations on regular languages. *Theoretical Computer Science*, 125(2):315--328, 1994. ISSN 0304-3975. doi: [https://doi.org/10.1016/0304-3975\(92\)00011-F](https://doi.org/10.1016/0304-3975(92)00011-F). URL <https://www.sciencedirect.com/science/article/pii/030439759200011F>.