ON THE DISTRIBUTION OF COMPLEXITIES OF FINITE WORDS, WITH CONNECTIONS TO CONSTRUCTIVE IMMUNITY

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

MATHEMATICS

MAY 2023

By

Samuel Birns

Dissertation Committee:

Bjørn Kjos-Hanssen, Chairperson
Ruth Haas
David Ross
Rufus Willett
Malik Younsi
Michelle Seidel

ACKNOWLEDGMENTS

I am grateful firstly and most obviously to my advisor, Bjørn Kjos-Hanssen. My research interests and career goals changed drastically during my time in graduate school, and I found Bjørn's advice and guidance on both mathematical and academic matters invaluable. I could not have asked for a more encouraging, supportive, and - above all - patient advisor.

I have met many wonderful people in Hawai‘i who have supported me in one way or another. In particular, I am fortunate to have crossed paths with my former housemates: Michael, Lean, Austin, Joseph, and Tom, with whom I have shared a home away from home.

Finally, I would like to thank my mother, Olive, my sister, Joanna, and my partner, Audrey. I am and will always be humbled by their support and belief in me. They have coaxed and cajoled from me a better version of myself, leading indirectly but surely to the completion of this thesis.

ABSTRACT

This thesis presents results on complexities of finite and infinite words, where "complexity" is defined using finite state and Turing machines, respectively. Specifically, the complexity of a word w is defined to be the minimum number of states in a finite state machine needed to uniquely output w among all words of length $|w|$. A theorem counting the number of words of length n and deterministic hidden Markov model complexity q is proven, giving insight into the distribution of deterministic hidden Markov model complexity. In particular, this theorem allows for the computation of deterministic hidden Markov model complexity in polynomial time, an improvement over the exponential-time naive computation. Non-deterministic hidden Markov models of various states are shown to be realized by deterministic hidden Markov models, and the strengths of languages defined by hidden Markov model complexity are investigated. Finally, analogs of these results are expanded to infinite words and to finite-state gambler complexity.

When discussing complexities of infinite words in the next topic, Σ_1^0 and constructively Σ_1^0 dense sets are introduced. It is shown that these classes are distinct and that they occur in non-Δ_2^0 degrees, high degrees, and c.e. degrees. Then, connections from Σ_1^0 sets to a problem on the computable dense subsets of $\mathbb{Q}$ is established. Finally, results on effective notions of the complexities of finite words are presented, providing a link between the two topics.

TABLE OF CONTENTS

Acknowledgments . . . . . **iii**
Abstract . . . . . **iv**
List of Tables . . . . . **vi**
List of Figures . . . . . **vii**
1 Introduction **1**
2 The distribution of deterministic HMM complexity **4**
2.1 Preliminaries 4
2.2 Asymptotic counts of words of given complexities 6
2.3 Calculations on high-complexity words 17
2.3.1 Binary words 17
2.3.2 Alphabets of arbitrary size 22
2.4 Exact HMM complexities of given words 24
3 Further results on finite-state machine complexity **30**
3.1 Complexity of probabilistic HMMs 30
3.1.1 Transition-probabilistic HMMs 30
3.1.2 Emission-probabilistic HMMs 32
3.2 HMM complexity as language 35
3.2.1 Regularity of complexity for fixed length 35
3.2.2 Regularity of fixed complexity 38
3.3 Complexity via other finite-state machines 39
3.3.1 Finite-state gamblers 39
3.3.2 Büchi automata 41
4 Constructive immunity **43**
4.1 Introduction 43
4.2 Σ^0_1-density 44
4.2.1 Numberings 45
4.3 Prevalence of Σ^0_1-density 46
4.3.1 Closure properties and Σ^0_1-density 46
4.3.2 Cofinality in the Turing degrees of constructive Σ^0_1-density 49
4.3.3 Non-Δ^0_2 degrees 49
4.3.4 High degrees 50
4.3.5 Progressive approximations 51
5 Further results on finite words and effectiveness **52**
5.1 Density for finite complexity 52
5.2 Fibonacci words 54
5.3 Pushdown automata and HMM complexity 55
5.4 Effective witnesses of deterministic HMM complexity 57
Index **59**
Bibliography **60**

LIST OF TABLES

2.1 Eventually integral pairs of complexity 3 13
2.2 Binary words with eventually integral complexity 5 17

LIST OF FIGURES

2.1 Almost all words have high complexity 20
2.2 “Average” complexity increases logarithmically 21
2.3 $\mathcal{P}(n, k)$ for various n and k 21
2.4 $s_q(n)$ for various n and q 29
2.5 $s_q(15)$ for various beta distributions 29

3.1 Possible 2-state half-transition-probabilistic HMMs 31
3.2 Accepting DFA for L_2 37
3.3 0^n-accepting finite-state gambler 41
3.4 $(01)^n$-accepting finite-state gambler 41
3.5 Büchi automaton accepting $(01)^\omega$ 42

4.1 Co-immunity results 47

CHAPTER 1
INTRODUCTION

Given a word over some alphabet, what makes it *complex*? Most people would say that 10111111110111000001 [1] is more complex than 01010101010101010101, though both have the same probability $\left(\frac{1}{2^{20}}\right)$ of being chosen among all binary words of length 20. Notions of complexity bridge this gap; roughly, a word's complexity is defined to be the size (in bytes, states, etc.) of the smallest computer program that can output the word. A wide variety of established notions of theoretical program (or "machine") make explorations in complexity a fruitful research area.

The classical definition of complexity has many presentations, such as in [22], and relies on a Turing machine as its model of computation. More specifically, the *Kolmogorov complexity* of a finite binary string σ is

$$K(\sigma) := \min\{|\tau| : U(\tau) = \sigma\}$$

where U is a prefix-free universal Turing machine and τ ranges over the set of finite binary strings. However, a major drawback of Kolmogorov complexity and the use of Turing machines in general is that Turing machines are not *computable*, i.e. for a given σ, the question "What is $K(\sigma)$?" may not be decidable. As an alternative, complexity for finite words has been considered with *finite-state machines*, a class of theoretical machines. In [29], Shallit and Wang defined the *automatic complexity* of a word x to be the number of states of a *deterministic finite automaton* M that uniquely accepts x among all words of length $|x|$. That is,

$$A(x) := \min\left\{|M| : L(M) \cap \Sigma^{|x|} = \{x\}\right\}$$

where $L(M)$ is the language of words accepted by the deterministic finite automaton M. A third alternative was proposed in [19], where *quantum* deterministic automata were considered:

$$Q(x) := \min\{|Q| : \exists \alpha, \omega \in \mathbb{CP}^q, \delta_x \alpha = \omega\}$$

where $\delta_x \in PU(q)$, the projective unitary group of q dimensions. Bounds for $A(x)$ and $Q(x)$ were shown in [17] and [19], among others. However, deterministic automata are a very restrictive class of machines; it would be natural to ask for a definition of complexity that is somehow more descriptive than $A(x)$ and $Q(x)$, but not as complicated as $K(\sigma)$ as to be unable to be computed for specific words.

Chapter 2 aims to bridge this gap by defining complexity in terms of hidden Markov models (HMMs). Define the deterministic HMM complexity $D(x)$ to be

$$D(x) := \min\{|H| : H \text{ emits } x\}^2$$

[1] This sequence was constructed via a Python random number generator.

[2] If a deterministic HMM emits a word, then it emits that word with probability 1 by definition, so it is not necessary to repeat the uniqueness conditions in the definitions of $A(x)$ and $Q(x)$.

Chapter 2 establishes initial properties of $D(x)$ and concludes with a result on the distribution of $D(x)$: given

$$s_q(n) := |\{x : |x| = n \wedge D(x) = q\}|$$

a formula for $s_q(n)$ whenever $n \geq 2q - 1$ is derived, establishing the asymptotic behavior of $s_q(n)$. This, in turn, leads to a computational improvement from exponential to polynomial time in calculating $s_q(n)$ for $n \geq 2q - 1$. Additionally, links between deterministic and general HMMs are established when the number of states is low, showing that deterministic HMMs capture much of the power of general HMMs.

Chapter 3 extends the results of Chapter 2 to present proofs that the families of languages

$$L_{n,1} = \{x \in \{0, \ldots, n-1\}^* : D(x) < |x|\}$$
$$L_{n,2} = \{x \in \{0, 1\}^* : D(x) = |x|\}$$

are regular. Chapter 3 then defines complexity for finite words via finite-state gamblers and for infinite words via Büchi automata, and proves results analogous to those in Chapter 2.

Chapter 4, which appeared in [3], departs from the decidable machines used in Chapter 2 to consider problems related to *Martin-Löf randomness*, which has many classical expositions, such as [14]. A *real* $A \in 2^\omega$ is Martin-Löf random if there does not exist a computable collection of computably enumerable sets U_n, $n \in \omega$, such that $A \in \bigcap_{n \in \omega} U_n$; intuitively, A is Martin-Löf random if it is "unusual" relative to any possible computable "test." It is a well-known and deep theorem that A is Martin-Löf random if and only if there exists some natural number C such that for all n, $K(A \upharpoonright n) > n - C$, i.e. if almost all initial segments of A are maximally complex. Chapter 4 arose after investigating the following question: "If $S \subseteq \mathbb{Q}$ is 'random' in the sense of Martin-Löf, when does $\overline{S} = \mathbb{R}$?" where $\overline{S}$ is the closure of S in the usual topology. More precisely,

For which $A \subseteq \omega$ is $\nu(A)$ dense in the usual topology, for all injective computable numberings ν?

The answer to this question is shown to be exactly the *co-immune* sets. This leads naturally to considering the class of *construtively* Σ^0_1 *dense* sets which have nonempty intersection with each infinite computably enumerable set, and where the witness for this intersection is uniformly computable. It is shown that the concept of constructive Σ^0_1 density is distinct from non-constructive Σ^0_1 density and that Σ^0_1 dense sets are represented in the non Δ^0_2 degrees, high degrees, and c.e. degrees.

Finally, Chapter 5 provides a link between the results in Chapters 2 and 3 by exploring the role of effectiveness in complexities of finite words. Upper bounds on the time complexities for the problems of determining whether a hidden Markov model witnesses the complexity of a given word, and finding a string to append to a word to make it sufficiently complex, are presented. Chapter 3 is then revisited by proving that, although the number of states in a DFA accepting the language $L_{n,2}$ grows exponentially in n, only a linearly increasing number of states in n is required in a pushdown automata that accepts $L_{n,2}$.

CHAPTER 2
THE DISTRIBUTION OF DETERMINISTIC HMM COMPLEXITY

2.1 Preliminaries

Definition 2.1. An *alphabet* is a finite set. Given an alphabet Σ, a *finite word* over Σ is a sequence $x_0x_1\dots x_n$ for some $n \in \mathbb{N}$, where $x_i \in \Sigma$ for all $i \leq n$. Let $\mathbb{N}$ denote the natural numbers; an *infinite word* over A is a sequence $x_0x_1\dots$, where $x_i \in \Sigma$ for all $i \in \mathbb{N}$. If w is a finite word, $|w|$ is the length of w. Write Σ^n for the set of words of length n over Σ, $\Sigma^{<\omega}$ for the set of finite words over Σ, and Σ^ω for the set of infinite words over Σ.

The term "word" will be used in place of "finite word" everywhere unless otherwise specified.

The main tool used in this chapter to describe the complexity of words is the *hidden Markov Model.* The definition of a hidden Markov model will be built up through auxiliary definitions.

Definition 2.2. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, where Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω, and μ is a probability measure on Ω. Let E be a measurable space. An E-valued *random variable* is a measurable function $X : \Omega \to E$.

Definition 2.3 ([20])**.** Let Q be a finite set. A function $X : \mathbb{N} \times \Omega \to Q$ is a *stochastic process* if for every $n \in \mathbb{N}$, $X(n, \cdot) : \Omega \to Q$ is a Q-valued random variable.

Intuitively, arguments from $\mathbb{N}$ will represent time.

Fix finite sets Q and Σ. Q will represent a set of "hidden" states, and Σ will be an alphabet of "output" symbols. Let X and Y be stochastic processes $X : \mathbb{N} \times \Omega \to Q$ and $Y : \mathbb{N} \times \Omega \to \Sigma$. Write $X_n(\omega) = X(n, \omega)$ for $\omega \in \Omega$ and similarly for Y. Additionally, for $O \in \Omega$, write $\Pr(O)$ for $\mu(\{O\})$.

Definition 2.4 ([20])**.** X is a *Markov process* if

$$Pr(X_n = a \mid X_{n-1}) = Pr(X_n = a \mid X_0, \dots, X_{n-1})$$

for all $a \in Q$ and $n \in \mathbb{N}$.

Definition 2.5 ([20])**.** The pair (X, Y) is a *hidden Markov model* if X is a Markov process and

$$Pr(Y_n \in S \mid X_n = x_n) = Pr(Y_n \in S \mid X_1 = x_1, \dots, X_n = x_n)$$

for all $n \geq 1$, $x_1, \dots, x_n \in Q$, and every $S \subseteq \Sigma$.

Intuitively, Definition 2.4 states that if X is a Markov process, then for all n and a, the probability of X equalling a at time n depends only on the values of X at time $n-1$. In particular, for all n and $a_1, a_2 \in Q$, $Pr(X_n = a_1 \mid X_{n-1} = a_2)$ is constant, and for all m, $Pr(X_n = a_1 \mid X_{n-1} = a_2) = Pr(X_m = a_1 \mid X_{m-1} = a_2)$. Q was assumed to be finite by definition, so let $|Q| = n$ and $Q = q_0, q_1, \dots q_{n-1}$. Then X can be associated with the matrix

$$A = \begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0(n-1)} \\ a_{10} & \ddots & \cdots & a_{1(n-1)} \\ \vdots & \ddots & \cdots & \cdots \\ a_{(n-1)0} & a_{(n-1)1} & \cdots & a_{(n-1)(n-1)} \end{bmatrix}$$

where $a_{ij} = Pr(X_n = q_j \mid X_{n-1} = q_i)$.

Similarly, Definition 2.5 states that if (X, Y) is a hidden Markov model, the probability that Y takes on values in A at time n depends only on X_n. Let $|\Sigma| = m$ and $\Sigma = \{\sigma_0, \sigma_1, \dots, \sigma_{m-1}\}$. Y can then be associated with the matrix

$$B = \begin{bmatrix} b_{00} & b_{01} & \cdots & b_{0(m-1)} \\ b_{10} & \ddots & \cdots & b_{1(m-1)} \\ \vdots & \ddots & \cdots & \cdots \\ b_{(n-1)0} & b_{(n-1)1} & \cdots & b_{(n-1)(m-1)} \end{bmatrix}$$

where $b_{ij} = Pr(Y_n = \sigma_j \mid X_n = q_i)$. Informally, the matrix A represents the probabilities of transitioning between states of a hidden Markov model, and the matrix B represents the probabilities of emitting output symbols given the hidden Markov model's state. The terms *transition matrix* and *emission matrix* will be used to refer to A and B, respectively. Transition and emission matrices will be explicitly discussed in Section 3.1.

Through the rest of this section, hidden Markov models will be referred to by the acronym HMM. The single variable H will be used to refer to HMMs, without directly referencing the underlying stochastic processes if the discussion does not require it.

Definition 2.6. Let $w = w_0 \dots w_n$ be a word over an alphabet Σ and let $H = (X, Y)$ be a hidden Markov model. A sequence $\omega = \omega_0, \omega_1, \dots, \omega_n \in \Omega$ is an *emitting sequence for* w if $Y_i(\omega_i) = w_i$ for all $i \leq n$. Let $\mathcal{E}(w)$ be the set of emitting sequences for w. The *probability of* H *emitting* w is $\sum\limits_{\omega \in \mathcal{E}(w)} \prod\limits_{i \leq n} \Pr(Y_i = \omega_i)$.

Definition 2.7 ([21])**.** Let w be a word over an alphabet Σ. The *HMM complexity* of w, written $C_{\text{HMM}}(w)$, is the minimum number of states in a HMM H such that if $w' \neq w$ is a word with $|w'| = |w|$, then the probability of H emitting w' is strictly less than the probability of H emitting w.

This section will be primarily be concerned with upper bounds of HMM complexity. In particular, to show that the complexity of w is not greater than n for some natural number n, it suffices to show that there exists a HMM with n states that outputs w with probability 1. The following theorem produces such a bound, which indirectly proves that HMM complexity is well-defined.

Definition 2.8. The *empty word* over any alphabet is the (necessarily unique) word of length 0. Write ϵ for the empty word and define $u^0 := \epsilon$.

Definition 2.9. Let $u = u_0u_1 \dots u_n$ and $v = v_0v_1 \dots v_m$ be words over an alphabet Σ. $u \frown v := u_0u_1 \dots u_nv_0v_1 \dots v_m$ is the *concatenation* of u and v.

Definition 2.10. Let $w = x_0x_1 \dots x_{n-1}x_n \dots x_m$ be a word over an alphabet Σ with $n \leq m$. Then $w \upharpoonright n$ is the word $x_0x_1 \dots x_{n-1}$.

Definition 2.11. Let $w = x_0x_1 \dots x_{n-1}x_n \dots x_m$ be a word with $n \leq m$. Then $w[n :]$ is the word $x_nx_{n+1} \dots x_m$. In this notation, write $w[-n :]$ for $w[|w| - n :]$.

Definition 2.12. Let w be a word over an alphabet Σ and let $n \geq |w|$. $w^{\frac{n}{|w|}} := w^{\left\lfloor \frac{n}{|w|} \right\rfloor} \frown w^{n \mod |w|}$.

Example 2.13. $(010)^{\frac{5}{3}} = 01001$

Theorem 2.14 ([21])**.** *The minimum number of states of an HMM such that w occurs with probability 1 is* $\min \{|u| + |v| : \exists u, v \in \Sigma^{<\omega}, \exists p \in \mathbb{Q}, w = uv^p\}$.

Theorem 2.15 ([21])**.** *Any binary word of length n starting with 0 except $0^{n-1}1$ can be written as uv^p with $|v| > 0$ and $p > 1$ with $p \in \mathbb{Q}$.*

2.2 Asymptotic counts of words of given complexities

Definition 2.16 ([1])**.** A nonempty word w is *primitive* if there does not exist a word u and a natural number $k > 1$ such that $w = u^k$.

Definition 2.17. The Möbius function $\mathcal{M}(n)$ is defined as

$$\mathcal{M}(n) := \begin{cases} 0 & \text{if } n \text{ is divisible by a square} > 1 \\ (-1)^i & \text{if } n = p_1p_2 \dots p_i, \text{ where } p_j \text{ for } j \leq i \text{ are distinct primes} \end{cases}$$

Definition 2.18. Let Σ be an alphabet with $|\Sigma| = m$. For $n > 0$, define $\mathcal{P}(n, m)$ to be the number of primitive words of length n over Σ. That is, $\mathcal{P}(n, m) := \{w \in \Sigma^n : w \text{ is primitive}\}$.

Theorem 2.19 ([1]). *For $n > 0$, $\mathcal{P}(n,m) = \sum\limits_{d|n} \mathcal{M}(d) m^{\frac{n}{d}}$.*

The argument m will be omitted from $\mathcal{P}(n,m)$ when Σ is clear from the context.

Much of this chapter will be concerned with with complexities based on the special class of HMMs defined below.

Definition 2.20. A *deterministic HMM* is a hidden Markov model such that for every state s_1, there is exactly one (not necessarily distinct) state s_2 such that s_1 transitions to s_2 with probability 1.

Definition 2.21. For $w \in \Sigma^*$, the *deterministic HMM complexity* of w is the minimal number of states in a deterministic HMM that accepts w.

Write $D(w)$ for the deterministic HMM complexity of w.

Theorem 2.22. *For any word w, $D(w) = \min\{|u| + |v| : \exists u, v \in \Sigma^{<\omega}, \exists p \in \mathbb{Q}, w = uv^p\}$.*

Proof. Let H witness the deterministic HMM complexity of w. Since H is deterministic, H emits w with probability 1. Further, since H witnesses HMM complexity, the number of states of H is minimal. By Theorem 2.14, H must have exactly $\min\{|u| + |v| : \exists u, v \in \Sigma^{<\omega}, \exists p \in \mathbb{Q}, w = uv^p\}$ states. □

Theorem 2.23. *Deterministic HMM complexity is well-defined. In particular, for every $w \in \Sigma^*$, $D(w) \leq |w|$.*

Proof. Let $w = w_1 w_2 \dots w_k$. Consider the HMM with states $s_1, \dots, s_k$ such that for all i, $1 \leq i \leq n$, s_i outputs w_i with probability 1. □

Definition 2.24. Given a word w over Σ, the *integer HMM complexity* of w is

$$I(w) := \min\left\{|u| + |v| : \exists u, v \in \Sigma^{<\omega}, k \in \mathbb{N} \text{ such that } w = uv^k\right\}$$

The motivation for restricting to integer powers in the definition of integer HMM complexity is to be able to answer questions about the asymptotic behavior of HMM complexity. For example, $010 = (01)^{\frac{3}{2}}$ has complexity 2 under the usual definition, but considering $010 = (01)0^1$ with complexity 3 makes sense when considering 010 as a stand-in for 010^k for arbitrarily large k. This intuition is made precise by the following definition:

Definition 2.25. A pair of words (u, v) has has *eventually integral complexity* if $I(uv) = |uv|$ and for all but finitely many n, $I(uv^n) = D(uv^n)$.

Proposition 2.26. *Let $u = 01$ and $v = 0$. Then (u, v) has eventually integral complexity.*

Proof. Since $I(010^n) = 3$ for all n, it suffices to show that $D(010^n) = 3$ for all but finitely many n. It will be shown that $D\left(uv^n\right) = 3$ for all $n \geq 2$. Fix $n \geq 2$ and suppose that $D\left(010^n\right) = 2$. By Theorem 2.22, $010^n = uv^p$ for $|u| + |v| = 2$ and $p \in \mathbb{Q}$. The definitions of u and v imply that either $|v| = 1$ or $|v| = 2$. Assume $|v| = 1$. Then $p \in \mathbb{N}$, and if $v = 0$ then $uv^p \upharpoonright 2 \neq 010^n \upharpoonright 2 = 01$, and if $v = 1$ then $uv^p \upharpoonright 1 \neq 010^n \upharpoonright 1 = 0$. So assume that $|v| = 2$ and let $010^n = x^p$ for some binary word x of length 2. It must be the case that $x = 01$, as $01 = (010^n) \upharpoonright 2 = x$. But $n \geq 2$ implies that $|010^n| \geq 4$ and therefore that $p \geq 2$, so it must be the case that $x(4) = 1$, a contradiction. □

Proposition 2.27. *Let $u = 010$ and let $v = 10$. Then (u, v) does not have eventually integral complexity.*

Proof. For all n, $I\left(uv^n\right) = 3$ since $uv^n = 0(10)^{n+1}$, but $D\left(uv^n\right) = 2$ by setting $u = \epsilon$, $v = 01$. □

Lemma 2.28. *For any word w, $D(w) \leq I(w)$.*

Proof. Let $w = uv^k$ for some $k \in \mathbb{N}$ with $|u| + |v|$ minimal. Then $D(w) \leq |u| + |v|$ by considering k as an element of $\mathbb{Q}$. □

Fix a natural number q. The following results are concerned with counting the size of

$$\mathcal{S}(q) := \{(u, v) \mid |uv| = q \,\text{and}\ (u, v) \text{ has eventually integral complexity}\}$$

Definition 2.29. Given a word w, a pair of words (u, v) is a *decomposition* of w if $|u| + |v| = D(w)$ and there exists $p \in \mathbb{Q}$ such that $uv^p = w$. That is, a decomposition of w witnesses the deterministic HMM complexity of w.

Definition 2.30 is defined analogously to Definition 2.29.

Definition 2.30. Given a word w a pair of words (u, v) is a *integral decomposition* of w if $|u| + |v| = I(w)$ and there exists $k \in \mathbb{N}$ such that $uv^k = w$.

Write

$$\Delta(w) := \{(u, v) : \ (u, v) \text{ is a decomposition of } w \ \}$$

and analogously,

$$\Delta_{\mathbb{N}}(w) := \{(u, v) : \ (u, v) \text{ is an integer decomposition of } w \ \}$$

Lemma 2.33, due to Shallit, provides a sufficient condition on w for which $|\Delta(w)| = 1$ and $|\Delta_{\mathbb{N}}(w)| = 1$, respectively.

Lemma 2.31 ([28] Theorem 2.3.3)**.** *Let $x, y \in \Sigma^+$. The following are equivalent:*

1. $xy = yx$.

2. *There exists $z \in Z^+$ and integers $k, l > 0$ such that $x = z^k$ and $y = z^l$.*

3. *There exist integers $i, j > 0$ such that $x^i = y^j$.*

Lemma 2.32 ([28] Theorem 2.3.6)**.** *Let x and y be nonempty words. Then $xy = yx$ if and only if there exist rational numbers $\alpha, \beta \geq 2$ such that $x^\alpha = y^\beta$.*

Lemma 2.33 ([28] Theorem 2.3.2)**.** *Let, $x, y, z \in \Sigma^+$. Then $xy = yz$ if and only if there exist $u \in \Sigma^+$, $v \in \Sigma^*$, and an integer $e \geq 0$ such that $x = uv$, $z = vu$, and $y = (uv)^e u$.*

Lemma 2.34. *Let a, b, c, and d be words over Σ with b and d primitive. Let $j, k \in \mathbb{Q}$ with $j, k \geq 2$, and assume that $ab^j = cd^k$, $|a| + |b| = |c| + |d|$, and either $a \neq c$ or $b \neq d$. Then there exist words u and v and a rational number l so that $ab^j = uv^l$ and $|a| + |b| > |u| + |v|$.*

Proof. Without loss of generality assume that $|a| \leq |c|$. Since $|a|+|b| = |c|+|d|$, $ab = cd$ and $b = c'd$ for some terminal segment c' of c. So $(c'd)^j = c'd^k$, which implies that $c'dc'd$ and $c'dd$ are both prefixes of $c'd^k$; by considering their respective lengths, $c'dd$ is a prefix of $c'dc'd$. But this implies that d is a prefix of $c'd$, so $c'd = de$ for some word e. By Lemma 2.33, there exists a nonempty word u, a word v, and a natural number p such that $c' = uv$, $d = (uv)^p u$, and $e = vu$. But then $b = c'd = uv(uv)^p u$, so $ab^j = a(uv)^l$ for some $l \in \mathbb{Q}$; further, since $u \neq \epsilon$, $|u| + |v| < |b|$. □

In simpler terms, Lemma 2.34 states that over-counting words of a given complexity can only happen when the witnessing HMMs contain no loops.

Lemma 2.35. *Let w be a word and let $(u, v) \in \Delta(w)$, with $uv^p = w$ for some $p \in \mathbb{Q}$. If $p \geq 2$, then $|\Delta(w)| = 1$.*

Proof. Let $(u_1, v_1) \in \Delta(w)$ and assume $(u, v) \neq (u_1, v_1)$. By Lemma 2.34, there exist words x and y and $q \in \mathbb{Q}$ such that $xy^q = w$ and $|x| + |y| < |u| + |v|$, contradicting the minimality of (u, v). □

Lemma 2.36. *Let w be a word and let $(u, v) \in \Delta_{\mathbb{N}}(w)$, with $uv^n = w$ for some $n \in \mathbb{N}$. If $n \geq 2$, $|\Delta_{\mathbb{N}}(w)| = 1$.*

Proof. The proof is the same as that of Lemma 2.35, considering integer decompositions and integer powers. □

$\delta(w)$ will be written to represent the unique decomposition of a word w when it is clear that it exists, and similarly $\delta_{\mathbb{N}}(w)$ will be written to represent a unique integer decomposition.

Lemma 2.37. *Let u and v be words and let $p, q \in \mathbb{Q}$ with $p, q \geq 2$. Then $\delta(uv^p) = \delta(uv^q)$.*

Proof. Note that by Lemma 2.35, $\delta(uv^p)$ and $\delta(uv^q)$ are the unique decompositions of uv^p and uv^q, respectively, so the statement of the theorem is justified. If $p = q$ then the result is trivial, so assume that $p \neq q$. Without loss of generality assume that $p > q$. Let $\Delta(uv^p) = (u_1, v_1)$ and let $\Delta(uv^q) = (u_2, v_2)$. Let $u_1 v_1^{p_1} = uv^p$ and let $u_2 v_2^{q_2} = uv^q$. Since $p_1 \geq p > 1$, $u_1 v_1^r = uv$ for some r, so $u_1 v_1^{r'} = uv^q$ for some $r' \in \mathbb{Q}$, which implies that $|u_1| + |v_1| \leq |u_2| + |v_2|$ by the definition of decompositions. Similarly, since $q_2 \geq q > 1$, $u_2 v_2^s = uv$ for some $s \in \mathbb{Q}$, so $u_2 v_2^{s'} = uv^p$ for some $s' \in \mathbb{Q}$, so $|u_2| + |v_2| \leq |u_1| + |v_1|$. So $(u_1, v_1) = (u_2, v_2)$ by Lemma 2.35. □

Lemma 2.38. *Let u and v be words and let $m, k \in \mathbb{N}$ with $m, k \geq 2$. Then $\delta_{\mathbb{N}}(uv^m) = \delta_{\mathbb{N}}(uv^k)$.*

Proof. Note that by Lemma 2.36, $\delta_{\mathbb{N}}(uv^m)$ and $\delta_{\mathbb{N}}(uv^k)$ are the unique decompositions of uv^m and uv^k, respectively, so the statement of the theorem is justified. The proof is the same as that of Lemma 2.37, considering integer powers in lieu of rational powers and appealing to Lemma 2.36 instead of Lemma 2.35. □

Lemma 2.39. *Let u and v be words. If there exists $k \in \mathbb{N}$ such that $(u, v) \in \Delta_{\mathbb{N}}(uv^k)$, then $\delta_{\mathbb{N}}(uv^n) = (u, v)$ for all $n \geq 2$.*

Proof. By Lemma 2.38, $\delta_{\mathbb{N}}(uv^n) = \delta_{\mathbb{N}}(uv^k)$. □

Lemma 2.40. *Let (u, v) have eventual integral complexity. Then, for all $n \geq 2$, $\delta_{\mathbb{N}}(uv^n) = (u, v)$.*

Proof. By the definition of eventually integral complexity, $I(uv) = |uv|$, and $|uv| = |u| + |v|$, so the integer decomposition of uv is (u, v) by Lemma 2.36. The result then holds by Lemma 2.39. □

Lemma 2.41. *Let (u, v) have eventual integral complexity. Then, for all but finitely many n, $\delta(uv^n) = \delta_{\mathbb{N}}(uv^n)$.*

Proof. Let N be large enough so that for all $n > N$, $I(uv^n) = D(uv^n)$. Let $N_1 := \max(N, 3)$ and fix $n > N_1$. By Lemma 2.40, $\delta_{\mathbb{N}}(uv^n) = (u, v)$. Let $\delta(uv^n) = (u_1, v_1)$. Since $n \geq 2$ and $|u| + |v| = I(uv^n) = D(uv^n) = |u_1| + |v_1|$, $(u, v) = (u_1, v_1)$ by Lemma 2.35. □

Lemma 2.42. *Let u and v be words with v primitive, and assume that there does not exist k such that $u = v^k$. Then for all $n \geq 2$, $\delta_{\mathbb{N}}(uv^n) = (u, v)$.*

Proof. Let $n > 1$ and let $(u_1, v_1) = \Delta_{\mathbb{N}}(uv^n)$ with $u_1 v_1^{n_1} = uv^n$ for some $n_1 \in \mathbb{N}$. Assume $|u_1| + |v_1| < |u| + |v|$. For any k, consider $k + \langle |v_1| \rangle := \{k + |v_1| \cdot m \mod |v| : m \in \mathbb{N}\}$. For some natural number k, $v(0) = v_1(k) = v(|v_1| \mod |v|)$, and in general, $v(0) = v(i)$ for all $i \in k + \langle |v_1| \rangle$. Similarly, $v(j) = v(i)$ for all $i \in (j + k) + \langle |v_1| \rangle$. Since $|u_1| + |v_1| < |u| + |v|$, $n_1 \geq n$, and since $n > 1$, for every $i < |v|$ there exists $j < |v_1|$ such that $i \in (j + k) + \langle |v_1| \rangle$. So $v = (v \upharpoonright r)^{\frac{|v|}{r}}$ where $r = \gcd(|v|, |v_1|)$, contradicting the assumption that v is primitive. □

Lemma 2.43. *Let u and v be words with v primitive. Let $n \geq 2$ and let $\Delta(uv^n) = (u_1, v_1)$. Then $|v| = |v_1|$.*

Proof. Let $u_1 v_1^p = uv^n$ for some $p \in \mathbb{Q}$. First, assuming that $|u| \leq |u_1|$, cancelling yields

$$\begin{aligned} v^n &= u_1' v_1^p \\ v' v^{n-1} &= v_1^p \\ v^{n-1} &= v_1' v_1^{p-1} \end{aligned}$$

for u_1' a suffix of u_1, v' a suffix of v, and v_1' a suffix of v_1. Then, if $|v_1| < |v|$, v is not primitive by Lemma 2.32 and Lemma 2.31. Similarly, assuming that $|u_1| \leq |u|$,

$$\begin{aligned} u' v^n &= v_1^p \\ v^n &= v_1' v_1^{p-1} \end{aligned}$$

for u' a suffix of u and v_1' a suffix of v_1. Again, if $|v_1| < |v|$, then v is not primitive by Lemma 2.32 and Lemma 2.31. □

Definition 2.44. Given words w_1 and w_2 with $|w_1| = |w_2| = n$, w_1 is a *cyclic shift* of w_2 if there exists k such that for all $i < n$, $w_1(i) = w_2\left((i + k) \mod n\right)$.

Example 2.45. 0011 is a cyclic shift of 1001, with $k = 1$.

Lemma 2.46. *Let u and v be words. For any $n \in \mathbb{N}$ with $n > 1$, let $\delta(uv^n) = (u_1, v_1)$ and let $\delta_{\mathbb{N}}(uv^n) = (u_2, v_2)$. Then v_1 is a cyclic shift of v_2.*

Proof. Without loss of generality assume v is primitive; if v is not primitive and $v = v'^k$ for some word v' and $k \in \mathbb{N}$, rewrite v^n as $(v')^{nk}$. By Lemma 2.42, $\delta_{\mathbb{N}}(uv^n) = (u, v)$, and if $\delta(uv^n) = (u_1, v_1)$, $|v_1| = |v|$ by Lemma 2.43. Since $n > 1$, there are at least two occurrences of both v_1 and v in uv^n. Further, since $|u_1| + |v_1| \leq |u| + |v|$, $|u_1| \leq |u|$, so the second occurrence of v_1 occurs during the first occurrence of v, which implies that v_1 is a cyclic shift of v. □

Definition 2.47. Given words w and w', w' is a *suffix* of w if $w' \neq \epsilon$, and there exists $k > 0$ such that $|w'| = |w| - k$ and for all $i < |w'|$, $w'(i) = w(i + k)$.

Theorem 2.48. *For any words u and v, (u, v) has eventually integral complexity if and only if v is primitive and for each suffix u' of u, $u' \neq v \upharpoonright |u'|$.*

Proof. Let (u, v) has eventually integral complexity. If v is not primitive and $v = x^m$ for some word x and natural number m, then $uv^n = ux^{nm}$ for all n, contradicting that (u, v) has eventually integral complexity. Assume that $u' = v \restriction |u'|$ for some suffix u' of u. Let $u_1 = u \restriction |u'|$ and let $v_1 = u'v \restriction |u'|$. Then

$$\begin{aligned} u_1 v_1 &= \big(u \restriction |u'|\big)\, u'v \restriction |u'| \\ &= \big((u \restriction |u'|)u'\big)\, v \restriction |u'| \\ &= uv \restriction |u'| \end{aligned}$$

so in particular, $|u_1| + |v_1| < |u| + |v|$. But for any $n > 1$,

$$\begin{aligned} uv^n &= \big((u \restriction |u'|)u'\big)\, \big((v \restriction |u'|)v[-|u'| :]\big)^n \\ &= \big((u \restriction |u'|)u'\big)\, \big((v \restriction |u'|)u'\big)^n \\ &= (u \restriction |u'|)\, \big(u'v \restriction |u'|\big)^n u' \\ &= u_1 v_1^n u' \\ &= u_1 v_1^{\frac{n|v_1|+|u'|}{|v_1|}} \end{aligned}$$

so $\delta(uv^n) \neq (u, v)$. But by Lemma 2.40, $\delta_{\mathbb{N}}(uv^n) = (u, v)$, and by Lemma 2.41, $\delta_{\mathbb{N}}(uv^n) = \Delta(uv^n)$, a contradiction.

Now assume that (u, v) does not have eventually integral complexity. Further, assume that v is primitive; the objective of the proof will be to show that there exists a suffix u' of u such that $u' = v[-|u'| :]$. Fix $n > 1$. If $u = v^k$ for some k, then v is a suffix of u, so assume that $u \neq v^k$ for all k. By Lemma 2.42, $\delta_{\mathbb{N}}(uv^n) = (u, v)$. Let $\delta(uv^n) = (u_1, v_1)$. Then

$$\begin{aligned} |u_1| + |v_1| &= D(uv^n) \\ &\leq I(uv^n) \\ &= |u| + |v| \end{aligned}$$

where the equalities follow from the definitions of decompositions and integer decompositions, respectively, and the inequality follows from Lemma 2.28. If $D(uv^n) = I(uv^n)$, then $(u, v) = (u_1, v_1)$. But for any m, $\delta(uv^m) = (u_1, v_1)$ by Lemma 2.37, and $\delta_{\mathbb{N}}(uv^m) = (u, v)$ by Lemma 2.38, so $D(uv^m) = I(uv^m)$ and (u, v) has eventually integral complexity. So $D(uv^n) < I(uv^n)$ and $|u_1| + |v_1| < |u| + |v|$. v_1 is a cyclic shift of v by Lemma 2.46, so $|v_1| = |v|$ which implies $|u_1| < |u|$. Since $\delta(uv^n) = (u_1, v_1)$, $u_1 v_1^p = uv^n$ for some $p \in \mathbb{Q}$. In particular, $u = u_1 v_1'$ for some initial segment v_1' of v_1. Let $k = |u| - |u_1|$ and let $u' = u[-k :]$. Then $v_1 \restriction k = u'$. But since $v(0) = v_1(k)$ and v_1 is a cyclic shift of v, $v_1 \restriction k = v[-k :]$.

□

u	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	0	1	00	01	10	11
v	001	010	011	100	101	110	01	10	1	0	1	0

Table 2.1: Eventually integral pairs of complexity 3

Theorem 2.49. *For all natural numbers q, over an alphabet of size m,*

$$|\mathcal{S}(q)| = \mathcal{P}(q) + \sum_{i=1}^{q-1} m^{i-1}\,(m-1)\,\mathcal{P}(q-i)$$

Proof. Consider uv with $|uv| = q$ and assume $|u| = i$. There are $\mathcal{P}(q-i)$ primitive words v. By Theorem 2.48, uv has eventually integral complexity if and only if no suffix of u is a prefix of v. So there are $m-1$ choices for the last letter of u to guarantee no suffix of u is a prefix of v. There are then m^{i-1} choices for the remaining letters of u, completing the proof. □

Example 2.50. By Theorem 2.49, there are $\mathcal{P}(3)+2\cdot\mathcal{P}(2)+2(2-1)\mathcal{P}(1) = 12$ eventually integral pairs of complexity 3 over an alphabet of size 2 where u and v are binary words. The 12 pairs are listed in Table 2.1.

For $q > 0$, let $\mathcal{C}(q) := \mathcal{P}(q) + \sum_{i=1}^{q-1} m^{i-1}\,(m-1)\,\mathcal{P}(q-i)$, so that in particular $\mathcal{C}(q) = |\mathcal{S}(q)|$. Note that $\mathcal{C}(q)$ has been studied in the literature, and in particular exists as sequence #A059412 in the On-line Encyclopedia of Integer Sequences. $\mathcal{C}(q)$ was first studied by Shallit et. al. in [13], and the link from their work to $\mathcal{C}(q)$ was provided by a comment of Michael Vielhaber.

Definition 2.51 ([13])**.** $f_k(q)$:= the number of pairwise non-isomorphic minimal DFAs with q states over a k-letter alphabet.

Theorem 2.52 ([25], [13])**.** *For all $q \in \mathbb{N}$ with $q > 0$, $\mathcal{C}(q) = f_1(q)$.*

$\mathcal{C}(q)$ does not count the number of words of length n with complexity q; such a formula should, of course, also depend on n. Assuming $n = q$ still does not give an accurate count of the words of complexity q, as $\mathcal{C}(q)$ counts *pairs* (u, v), which over-counts when concatenation is applied. However, consider the case when n increases relative to a fixed q. Words of length n and complexity q are formed from elements of $\mathcal{S}(q)$ by taking increasingly large powers of v, making over-counting due to concatenation less likely. Therefore, for $n >> q$, $\mathcal{C}(q)$ should accurately count the number of words of length n with complexity k. This intuition is made precise by the following results.

Lemma 2.53. *Let a, b, c, and d be words. Assume that $ab^j = cd^k$ for $j, k \in \mathbb{Q}$, $|a|+|b| = |c|+|d|$, and that either $a \neq c$ or $b \neq d$. If $|ab^j| \geq 2|ab| - 1$, then there exist words u and v and $l \in \mathbb{Q}$ such that $ab^j = uv^l$ and $|a|+|b| > |u|+|v|$.*

Proof. Let $n = |ab^j|$ and let $q = |ab|$. Since $|ab^j| = |a| + j|b|$ and $|ab| = |a| + |b|$, rearranging the terms in $n \geq 2q - 1$ yields that $(j-2)|b| + 1 \geq |a|$. Since $|ab^j| = |cd^k|$ and $|ab| = |cd|$, the same logic yields that $(k-2)|d| + 1 \geq |c|$. There are three possible cases:

Case 1: $j, k \geq 2$. Then the result immediately follows by Lemma 2.34.

Case 2: $j, k < 2$. Considering that $j < 2$, there are two subcases to consider: either $|b| = 0$ or $|b| > 0$. In the first subcase, $|a| \leq 1$ implies that $|ab^j| = |cd^k| = 1$ so the result is trivial. In the second subcase $j < 2$ implies that $|a| \leq 1 + (j-2)|b| < 1$, which implies that $|a| = 0$. But then $(2-j)|b| \leq 1$, so $b^j = b^2[: -1]$. Applying the same logic in this subcase to the fact that $k < 2$ yields that $d^k = d^2[: -1]$, which implies that $a = c = \epsilon$ and $b = d$, a contradiction,

Case 3: exactly one of j and k is at least two. Without loss of generality assume that $j \geq 2$ and $k < 2$. By the arguments in the previous case, either $|d| = 0$ and $|c| = 1$, which immediately leads to a contradiction, or $|c| = 0$. So $ab^j = d^k$. Let b^- denote $b \restriction |b| - 1$. By the argument in the proof of Lemma 2.34, $abab^-$ and abb are prefixes of the same word. There are two possible subcases: either $|a| > 0$ or $|a| = 0$. In the first subcase, abb is a prefix of $abab^-$, which implies that b is a prefix of ab^-. So $ab^- = bg$ for some word g, which implies that $ab^- = b^-h$ for some word h. Then there exist u, v, and l satisfying the conclusion by Lemma 2.33. On the other hand, if $|a| = 0$, then $j|b| = k|d|$. But it is also the case that $|b| = |d|$, so $2 \leq j = k$, contradicting the assumption. □

Theorem 2.54. *For n and $q \leq n$, the function $f_{n,q} : \mathcal{S}(q) \to \Sigma^n$ defined by $(u, v) \mapsto uv^{\frac{n-|u|}{|v|}}$ is injective whenever $n \geq 2q - 1$.*

Proof. Fix $(u, v) \in \mathcal{S}(q)$. First, assume that $\frac{n-|u|}{|v|} < 1$. By the definition of $\mathcal{S}(q)$, $|uv| = q$ and (u, v) has eventually integral complexity. Since (u, v) has eventually integral complexity, $I(uv) = |uv|$. So $n < |u| + |v| = |uv| = I(uv) = q$ and the result holds vacuously.

Now assume that $\frac{n-|u|}{|v|} \geq 1$ and let $(u_1, v_1) \in \mathcal{S}(q)$ with $(u_1, v_1) \neq (u, v)$. If $u_1v_1 \neq uv$ then $u_1v_1^k \neq uv^k$ for all natural numbers $k \geq 1$, so assume $u_1v_1 = uv$. It must then be the case that $|v_1| \neq |v|$; otherwise, $v_1 = v$ and then $u_1 = u$. Without loss of generality assume that $|v_1| > |v|$. Let $(u', v') \in \mathcal{S}(q)$ be arbitrary such that $\frac{n-|u'|}{|v'|} \geq 1$, and let $\frac{n-|u'|}{|v'|} = \frac{m'}{v'} + 1$ for some natural number m'. Then,

$$\begin{aligned} m' + |v'| &= n - |u'| \\ m' &= n - q \end{aligned}$$

so in particular, m' is a constant that does not depend on the choice of (u', v'). Let m denote this constant. Then,

$$
\begin{aligned}
u_1 v_1^{\frac{m}{|v_1|}+1} &= uv^{\frac{m}{|v|}+1} \iff && (2.1)\\
u_1 v_1 v_1 \restriction m &= uvv \restriction m \iff && (2.2)\\
v_1 \restriction m &= v \restriction m && (2.3)
\end{aligned}
$$

Additionally, $u_1 v_1^p \neq uv^q$; otherwise, by Lemma 2.53, there exists words u_2, v_2 and $l \in \mathbb{Q}$ such that $u_2 v_2^l = uv^q$ and $|u_2| + |v_2| < |u| + |v|$. This implies that $D(uv^{q'}) \leq |u_2| + |v_2| < |u| + |v|$ for all $q' \geq q$, but $I(uv^n) = |u| + |v|$ for all n by Lemma 2.41, contradicting the assumption that $(u, v) \in \mathcal{S}(q)$. So assume that $m \leq |v_1|$. Since $n \geq 2q - 1$,

$$
\begin{aligned}
|u_1 v_1^{\frac{m}{|v_1|}+1}| &= |u_1| + |v_1| + m\\
&> 2q - 2\\
&= 2\left(|u_1| + |v_1|\right) - 2
\end{aligned}
$$

so $|u_1| + |v_1| - 2 < m \leq |v_1|$, which implies that $|u_1| < 2$. Substituting u and v for u_1 and v_1 in the calculation above also shows that $|u| < 2$. Since $|v_1| > |v|$, $|u_1| < |u|$, so $|u_1| = 0$ and $|u| = 1$. By Equation (2.3), $v_1 \restriction m = v \restriction m$. Since $n \geq 2q - 1$, $m \geq q - 1$, so $v_1 \restriction q - 1 = v \restriction q - 1$, which is equal to v since $|v| = q - 1$. But since $v_1 = u_1 v_1 = uv$, $v = v_1[1:]$. So for all i with $0 \leq i < q$, $v_1(i) = v(i) = v_1(i + 1)$. In other words, $v_1 = x^q$ for some letter x, contradicting the assumption that $(u_1, v_1) \in \mathcal{S}(q)$. □

In Theorem 2.55 and Theorem 2.56, $f_{n,q}(u, v)$ will be written for $f_{n,q}((u, v))$ where $(u, v) \in \mathcal{S}(q)$.

Theorem 2.55. *Let $f_{n,q}$ be as in Theorem 2.54. For every word w of length n and HMM complexity q, there exists $(u, v) \in \mathcal{S}(q)$ such that $f_{n,q}(u, v) = w$.*

Proof. Let $w \in \Sigma^n$, and let $w = uv^p$ for $|u| + |v| = q$ and $p \in \mathbb{Q}$. Since $|w| = n$, $p = \frac{n-|u|}{|v|}$. So if $(u, v) \in \mathcal{S}(q)$, then $f_{n,q}((u, v)) = w$; it therefore suffices to show that (u, v) is an eventually integral pair. Let k be any natural number with $k \geq p$. Assume towards a contradiction that there exists words u_1 and v_1 and a natural number k_1 such that $uv^k = u_1 v_1^{k_1}$ and $|u_1| + |v_1| < |u| + |v|$. There are three possible cases:

Case 1: $|u| = |u_1|$. Then $k|v| = k_1|v_1|$ which implies that $k \mid k_1$. Let $k \cdot m = k_1$ for some m and let $k - r = p$ for $r \in \mathbb{Q}$. Substituting yields

$$
\begin{aligned}
uv^p &= uv^{k-r}\\
&= uv_1^{\frac{k_1}{m}-r}
\end{aligned}
$$

contradicting the assumption that the HMM complexity of w is q.

Case 2: $|u| < |u_1|$. Then $v^k = u'v_1^{k_1}$, where u' is a terminal segment of u_1. After cancelling u', $(v')^s = v_1^{k_1}$, where v' is a terminal segment of v and $s = \frac{k|v|-|u'|}{|v'|}$. So $v' = v_1^{\frac{k_1}{s}}$ and $uv^p = u_1 v'^{\frac{p}{k_1 s}}$. Additionally, $|v'| = |v_1| - |u_1| + |u|$, so

$$\begin{aligned} |u_1| + |v'| &= |u| + |v_1| \\ &< |u| + |v| \end{aligned}$$

again contradicting that assumption that the HMM complexity of w is q.

Case 3: $|u_1| < |u|$. The proof of this subcase will proceed analogously to case 2. By cancelling u_1, $u'v^k = v_1^{k_1}$ where u' is a terminal segment of u. So then $v^k = v'^s$, where v' is a terminal segment of v_1 and $s = \frac{k_1|v_1|-|u'|}{|v'|}$. So $v' = \frac{k}{s}$ and $uv^p = uv'^{\frac{p}{ks}}$. Additionally, $|v'| = |u_1| + |v_1| - |u|$, so

$$\begin{aligned} |u| + |v'| &= |u_1| + |v_1| \\ &< |u| + |v| \end{aligned}$$

again contradicting the assumption that the HMM complexity of w is q. □

Theorem 2.56. *Let n and q be natural numbers with $n \geq 2q - 1$. Let $(u, v) \in \mathcal{S}(q)$ and let $w = f_{n,q}(u, v)$. Then $D(w) = q$.*

Proof. Let $p = \frac{n-|u|}{|v|}$, so that $uv^p = w$. Since $|u| + |v| = q$,

$$\begin{aligned} n - |u| &\geq 2(|u| + |v|) - |u| - 1 \\ &= 2|v| + |u| - 1 \end{aligned}$$

and hence

$$\begin{aligned} p &= 2 + \frac{|u| - 1}{|v|} \\ &> 1 \end{aligned}$$

Then, if $D(w) < q$ and $\delta(w) = (u_1, v_1)$ for $(u_1, v_1) \neq (u, v)$, $u_1 v_1^q = w$ for some $q \geq p$, contradicting Lemma 2.37. □

Taken together, Theorem 2.54 and Theorem 2.56 show that eventually integral pairs of complexity q correspond exactly to words of length n with HMM complexity q whenever $n \geq 2q - 1$. This idea will be referenced again in Theorem 2.86 of Section 2.4.

Let $s_q(n)$ be the number of words of length n with complexity q.

u	v	$\frac{n-\lvert u\rvert}{\lvert v\rvert}$	$uv^{\frac{n-\lvert u\rvert}{\lvert v\rvert}}$
ϵ	001	5/3	00100
ϵ	010	5/3	01001
ϵ	011	5/3	01101
ϵ	100	5/3	10010
ϵ	101	5/3	10110
ϵ	110	5/3	11011
0	01	2	00101
1	10	2	11010
00	1	3	00111
01	0	3	01000
10	1	3	10111
11	0	3	11000

Table 2.2: Binary words with eventually integral complexity 5

Question 2.57. For n and q with $n < 2q - 1$, what is $\mathcal{C}(q) - s_q(n)$?

Of course, Question 2.57 can be answered by providing an explicit formula for this difference, and there is no reason to believe that deriving such a formula is out of reach. However, answering Question 2.57 may not be necessarily to determine complexity distributions of arbitrary lengths; solving the weaker Question 2.58 and the related Question 2.59 may be sufficient.

Question 2.58. For a fixed n, for which q with $q \leq n < 2q - 1$ is $\mathcal{C}(q) - s_q(n)$ maximized? What is the upper bound of $\mathcal{C}(q) - s_q(n)$ in terms of k?

Question 2.59. For each n, let $q_n := \text{argmax}_q \mathcal{C}(q) - s_q(n)$. What is the asymptotic behavior of $\mathcal{C}(q_n) - s_{q_n}(n)$?

Example 2.60. Table 2.2 shows the injectivity of $f_{5,3}$ with $\Sigma = \{0, 1\}$: Indeed, these are exactly the binary words of length 5 with HMM complexity 3.

Example 2.61. The bound in Theorem 2.54 is sharp: if $n = 4$ with q and Σ as before, $(001)^{\frac{4}{3}} = 0(01)^{\frac{3}{2}}$.

2.3 Calculations on high-complexity words

2.3.1 Binary words

So far, the results presented in this chapter have been concerned with counting words with arbitrary complexity by finding sufficient conditions on word length for which integer complexity coincides with HMM complexity. In this section, the number of words of relatively *high* HMM complexity for a *fixed* length will be established. The key observation for doing so is that if a word of length n

ends in a subword of length k that occurs elsewhere in the word, the HMM complexity of the word is at most $n-k$, as the word can be outputted by "reusing" states for the last k letters.

In this section, all words will be assumed to be over the alphabet $\{0,1\}$.

Definition 2.62. For a natural number $k \geq 2$, the *k-bonacci numbers* are a sequence $\left\{F_n^k\right\}_{n=0}^{\infty}$ defined recursively by

$$F_n^k = \begin{cases} 0 & n < k-1 \\ 1 & n = k-1 \\ \sum_{i=1}^k F_{n-i}^k & n \geq k \end{cases}$$

Note that when $k = 2$, the k-bonacci numbers are the classic Fibonacci sequence. The k-bonacci numbers are also known as the *k-step Fibonacci numbers.*

Definition 2.63. Let w and w' be words. w' is a *subword* of w if there exists n such that for all $i < |w'|$, $w'_i = w_{i+n}$.

Note that if $|w| < |w'|$, then w' cannot be a subword of w.

Lemma 2.64. *For all k and $i < k$, $F_{k+i}^k = 2^i$.*

Proof. Proof by strong induction on i. $F_k^k = 1$ immediately. Assuming that $F_{k+j}^k = 2^j$ for all $j \leq i$ and $i+1 < k$,

$$\begin{aligned} F_{k+i+1}^k &= \sum_{j=1}^{k} F_{i+1-j}^k \\ &= \sum_{j=-1}^{i} F_{k+j}^k \\ &= 1 + \sum_{j=0}^{i} 2^j = 2^{i+1} \end{aligned}$$

completing the proof. The assumption that $i+1 < k$ is used to rewrite the indices in the second equality, since $F_{i+1-j}^k = 0$ for all $j > i-k+2$. □

Definition 2.65. Let n and k be natural numbers and let w be a word of length k. Define

$$\mathcal{S}_{n,w} := \{w' \in \{0,1\}^n \mid w \text{ is not a subword of } w'\}$$

and let $\mathcal{C}_{n,w} := |\mathcal{S}_{n,w}|$.

Definition 2.66. A word w is a *monoword* if $w_i = w_j$ for all $i, j < |w|$.

The following result is a generalization of arguments in [4].

Theorem 2.67. *For all natural numbers n and k and words w of length k, $F^k_{n+k} \leq \mathcal{C}_{n,w}$, with equality if and only if either $n < k$ or w is a monoword.*

Proof. Write $w = w_0 \dots w_{k-1}$ for $w_i \in \{0,1\}$. If $n < k$, then w is not a subword of any word of length n, so $\mathcal{C}_n = 2^n = F^k_{n+k}$ by Lemma 2.64. Assume $n \geq k$ and let $w' \in \mathcal{S}_n$. There are k distinct possible cases:

$$\begin{aligned} w'_{n-1} &\neq w_{k-1}, \text{ or} \\ w'_{n-1} = w_{k-1} \text{ and } w'_{n-2} &\neq w_{k-2}, \text{ or} \\ w'_{n-1} = w_{k-1} \text{ and } w'_{n-2} = w_{k-2} \text{ and } w'_{n-3} &\neq w_{k-3}, \text{ or} \\ &\dots \\ w'_{n-i} = w_{k-i} \text{ for all } 1 \leq i < k \text{ and } w'_{n-k} &\neq w_0 \end{aligned}$$

But by definition,

$$\begin{aligned} &\mathcal{S}_{n-1,w} \subseteq \left\{ w \restriction_{n-1} \mid w \in \{0,1\}^n \wedge \left(w'_{n-1} \neq w_{k-1}\right)\right\} \\ &\mathcal{S}_{n-2,w} \subseteq \left\{ w \restriction_{n-2} \mid w \in \{0,1\}^n \wedge \left(w'_{n-1} = w_{k-1} \text{ and } w'_{n-2} \neq w_{k-2}\right)\right\} \\ &\qquad \dots \\ &\mathcal{S}_{n-k,w} \subseteq \left\{ w \restriction_{n-k} \mid w \in \{0,1\}^n \wedge \left(w'_{n-i} = w_{k-i} \text{ for all } 1 \leq i < k \text{ and } w'_{n-k} \neq w_0\right)\right\} \end{aligned}$$

so $\mathcal{C}_{n,w} \geq \sum_{i=1}^k \mathcal{C}_{n-i,w}$.

Now assume w is a monoword. It suffices to show that all of the set inequalities above are all equalities; in this case, $\mathcal{C}_{n,w} = \sum_{i=1}^k \mathcal{C}_{n-i,w}$ and the result follows from the recursive definition of F^k_{n+k}. Fix i with $1 \leq i \leq k$ and let

$$w' \in \left\{ w \restriction_{n-i} \mid w \in \{0,1\}^n \wedge \left(w'_{n-j} = w_{k-j} \text{ for all } 1 \leq j < i \text{ and } w'_{n-i} \neq w_{k-i}\right)\right\}$$

Since $w'_{n-i} \neq w_{k-i}$ and w is a monoword, w'_{n-i} cannot be a letter in a subword containing w, and $w' \in \mathcal{S}_{n-i,w}$.

Conversely, assume w is not a monoword. Let i be greatest such that $w_i \neq w_{k-1}$ and $i \neq k-1$. Then

$$w \in \left\{ w \restriction_{n-i} \mid w \in \{0,1\}^n \wedge \left(w'_{n-j} = w_{k-j} \text{ for all } 1 \leq j < i \text{ and } w'_{n-i} \neq w_{k-i}\right)\right\}$$

so

$$\mathcal{S}_{n-i,w} \subsetneq \left\{ w \restriction_{n-i} \mid w \in \{0,1\}^n \wedge \left(w'_{n-j} = w_{k-j} \text{ for all } 1 \leq j < i \text{ and } w'_{n-i} \neq w_{k-i}\right)\right\}$$

and $\mathcal{C}_{n,w} > \sum_{i=1}^k \mathcal{C}_{n-i,w}$. □

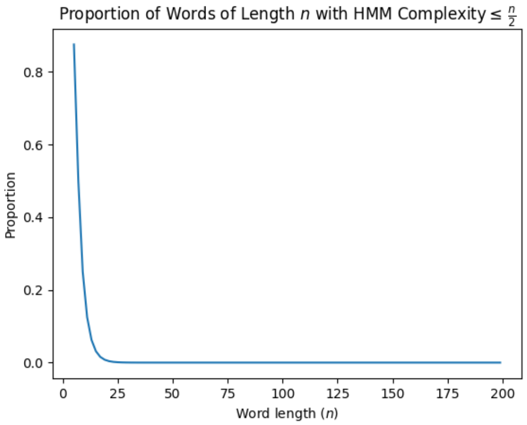

Figure 2.1: Almost all words have high complexity

Theorem 2.68. *Let n and k be natural numbers with $k \leq \lfloor \frac{n}{2} \rfloor$. There are at most $2^k \left(2^{n-k} - F_n^k\right)$ words w of length n such that $w' := w_{n-k} \dots w_{n-1}$ is a subword of $w \restriction_{n-k}$.*

Proof. Let w be a word of length n and fix $w' := w_{n-k} \dots w_{n-1}$. There are at least F_n^k words of length $n-k$ that do not contain w' by Theorem 2.67, so there are at *most* $2^{n-k} - F_n^k$ words of length $n-k$ that *do* contain w'. There are 2^k possibilities for w', which completes the proof. □

Theorem 2.69. *For all natural numbers n and k with $k \leq \lfloor \frac{n}{2} \rfloor$, there are at most $2^k \left(2^{n-k} - F_n^k\right)$ words w of length n such that $D(w) \leq n-k$.*

Proof. If the last k letters of w are contained in $w \restriction_{n-k}$, the HMM complexity of w is at most $n-k$, so the result immediately follows from Theorem 2.68. □

For natural numbers k and n with $k \leq \lfloor \frac{n}{2} \rfloor$, let $\mathcal{P}(n,k)$ be the proportion of words w of length n with $D(w) = n-k$.

Corollary 2.70. $\mathcal{P}(n,k) \leq 1 - \frac{F_n^k}{2^{n-k}}$.

Proof. Immediate by dividing the quantity in Theorem 2.69 by 2^n. □

Example 2.71. As shown in Figure 2.1, the proportion of words of length n with complexity at most $\frac{n}{2}$ approaches zero as $n \to \infty$. This is intuitive, as "almost all" words should be complex.

Example 2.72. Define

$$f(n) := \max\left\{k \mid \mathcal{P}(n,k) > \tfrac{1}{2}\right\}$$

Then $f(n) \to \log n$ as $n \to \infty$, as in Figure 2.2. This is shown for specific values of n and k in Figure 2.3.

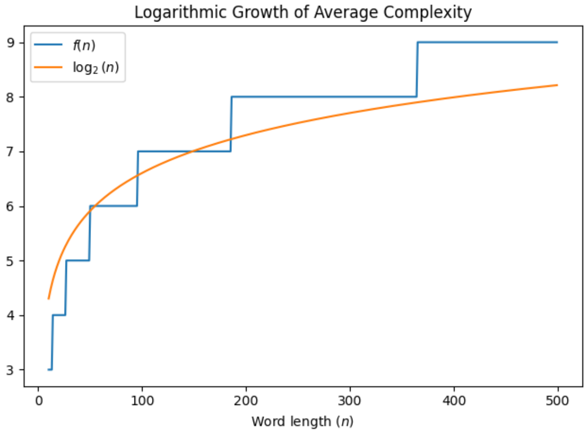

Figure 2.2: “Average” complexity increases logarithmically

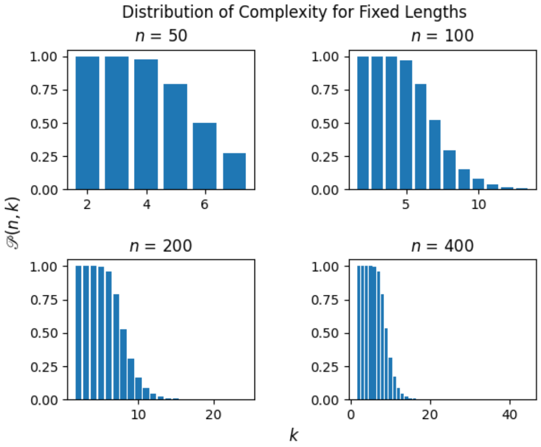

Figure 2.3: $\mathcal{P}(n,k)$ for various n and k

2.3.2 Alphabets of arbitrary size

The results of the last section will be extended to alphabets of arbitrary size.

Definition 2.73. The *base-m k*-bonacci numbers $\widehat{F}_n^{m,k}$ are defined recursively by

$$\widehat{F}_n^{m,k} := \begin{cases} m^n & n < k \\ m^n - 1 & n = k \\ (m-1) \cdot \sum_{i=1}^{k} F_{n-i}^{m,k} & n > k \end{cases}.$$

Definition 2.74. For a natural number m and alphabet $\mathcal{A}$ of size m, define

$$\mathcal{S}_{n,w} := \{w' \in \{0,1\}^n \mid w \text{ is not a subword of } w'\}$$

and let $\mathcal{C}_{n,w} := |\mathcal{S}_{n,w}|$.

This can be understood as overwriting the previous definitions of $\mathcal{S}_{n,w}$ and $\mathcal{C}_{n,w}$ in Definition 2.65. The implicit dependence on alphabet size here will be understood.

Theorem 2.75. *Let m, k, and n be natural numbers, and let w be a word of length k over an alphabet $\mathcal{A}$ of size m. Then $\widehat{F}_n^{m,k} \leq C_{n,w}$, with equality if and only if either $n < k$ or w is a monoword.*

Proof. The idea behind this proof is analogous to Theorem 2.68. If $n < k$, then w is not a subword of any word of length n; there are m^n such words. If $n = k$ and w' is a word of length n, w is a subword of w' if and only if $w = w'$; there are $m^n - 1$ words of length n not equal to w. Let $n > k$ and let w' be a subword of length n such that w' is not a subword of w. Write $w = w_0 \dots w_{k-1}$ for $w_i \in \{0,1\}$. As before, there are k distinct possible cases:

$$\begin{aligned} w'_{n-1} &\neq w_{k-1}, \text{ or} \\ w'_{n-1} = w_{k-1} \text{ and } w'_{n-2} &\neq w_{k-2}, \text{ or} \\ w'_{n-1} = w_{k-1} \text{ and } w'_{n-2} = w_{k-2} \text{ and } w'_{n-3} &\neq w_{k-3}, \text{ or} \\ &\dots \\ w'_{n-i} = w_{k-i} \text{ for all } 1 \leq i < k \text{ and } w'_{n-k} &\neq w_0 \end{aligned}$$

Also as before,

$$\begin{aligned} &\mathcal{S}_{n-1,w} \subseteq \left\{ w \restriction_{n-1} \mid w \in \{0,1\}^n \wedge \left(w'_{n-1} \neq w_{k-1} \right) \right\} \\ &\mathcal{S}_{n-2,w} \subseteq \left\{ w \restriction_{n-2} \mid w \in \{0,1\}^n \wedge \left(w'_{n-1} = w_{k-1} \text{ and } w'_{n-2} \neq w_{k-2} \right) \right\} \\ &\quad \dots \\ &\mathcal{S}_{n-k,w} \subseteq \left\{ w \restriction_{n-k} \mid w \in \{0,1\}^n \wedge \left(w'_{n-i} = w_{k-i} \text{ for all } 1 \leq i < k \text{ and } w'_{n-k} \neq w_0 \right) \right\} \end{aligned}$$

so

$$\begin{aligned}
&\mathcal{C}_{n-1} \leq (m-1) \cdot \left| \left\{ w \in \mathcal{A}^n \mid w'_{n-1} \neq w_{k-1} \right\} \right| \\
&\mathcal{C}_{n-2} \leq (m-1) \cdot \left| \left\{ w \in \mathcal{A}^n \mid w'_{n-1} = w_{k-1} \text{ and } w'_{n-2} \neq w_{k-2} \right\} \right| \\
&\quad \cdots \\
&\mathcal{C}_{n-k} \leq (m-1) \cdot \left| \left\{ w \in \mathcal{A}^n \mid w'_{n-i} = w_{k-i} \text{ for all } 1 \leq i < k \text{ and } w'_{n-k} \neq w_0 \right\} \right|
\end{aligned}$$

where the $m-1$ is the number of letters in $\mathcal{A}$ not equal to w'_{n-i} for all i. So $\mathcal{C}_n \geq (m-1) \cdot \sum_{i=1}^{k} C_{n-i}$.

The same argument as in Theorem 2.69 shows that equality holds if and only if w is a monoword.

□

The notation $\widehat{F}_n^{m,k}$ was used to distinguish between the previously established F_n^k. It may be expected that $\widehat{F}_n^{2,k} - F_n^k$, but the sequences actually differ up to indices since $\widehat{F}_n^{m,k}$ is never equal to zero. This relationship will be established more closely in a later example Example 2.77.

Theorem 2.76. *For all natural numbers m, n, and k with $k \leq \lfloor \frac{n}{2} \rfloor$, there are at most*

$$m^k \left(m^{n-k} - \widehat{F}_{n-k}^{m,k} \right)$$

words w of length n over an alphabet of size m with $D(w) \leq n-k$.

Proof. If the last k letters of w are contained in $w \upharpoonright_{n-k}$, the HMM complexity of w is at most $n-k$, so the result immediately follows from Theorem 2.75. □

Example 2.77. For all natural numbers n and k, $\widehat{F}_n^{2,k} = F_{n+k}^k$.

Proof. Fix k. By Lemma 2.64 and the definition of $\widehat{F}$, $\widehat{F}_n^{2,k} = F_{n+k}^k = 2^n$ for all $n < k$. The proof immediately follows by the recursive definitions of F and $\widehat{F}$. □

For natural numbers m, k, and n with $k \leq \lfloor \frac{n}{2} \rfloor$, let $\widehat{\mathcal{P}(n,k)}$ be the proportion of words w of length n over an alphabet of size m with $D(w) = n-k$.

Corollary 2.78. $\widehat{\mathcal{P}(n,k)} \leq 1 - \frac{\widehat{F}_{n-k}^{m,k}}{m^{n-k}}$

Proof. Immediate by dividing the quantity in Theorem 2.76 by m^n. □

Remark 2.79. Let $w = w_1 \dots w_n$ be a word and let $q < n$. For $0 \leq i < n-q$, let A_i be the event that $w_i \dots w_{i+q} = w_{n-q} \dots w_n$ Then

$$\begin{aligned}
\Pr\left(w[-q:] \preceq w \upharpoonright q \right) &= \Pr\left(\cup_{i=0}^{n} A_i \right) \\
&\leq \sum_{i=0}^{n-q-1} \Pr\left(A_i \right) \\
&= (n-q) 2^{-(q+1)}
\end{aligned}$$

which goes to zero if q is larger than $\log_2(n)$. Indeed, let $\epsilon > 0$ and let $c = (\log_2(n))^{1+\epsilon}$. Then

$$\frac{n-q}{2^{q+1}} < \frac{n}{2n^{1+\epsilon}}$$

The distribution of $D(w)$ for random w with $|w| = n$ is then similar to the *Gumbel distribution*, the distribution of $n - Z(w)$, where $Z(w)$ is the length of the longest subsequence of zeroes in w, as discussed in [4]. If $0 \leq q < n$, the number of binary words that contain 0^{n-q} is

$$\sum_{i=0}^{q} \binom{q}{i}(i+1) = 2^{q-1}(q+2)$$

where each term in the left summand represents the possible arrangements of i ones in the subword that does not contain 0^{n-q}. In particular, the Gumbel distribution does not depend on n. In the next section it will be shown that $D(w)$ also does not depend on n when n is sufficiently large.

2.4 Exact HMM complexities of given words

Although we consider different machines as in [2], notation will be borrowed to extend results in Section 2.2 with a partial answer to Question 2.57.

Definition 2.80 ([2])**.** For an alphabet Σ and natural numbers n and q, define

$$s_q(n) := |\,\{w \in \Sigma^n : D(w) = q\}\,|$$

Although $s_q(n)$ implicitly depends on the size of the underlying alphabet Σ, this fact will not need to be addressed and so Σ is left out of the notation for brevity.

Definition 2.81. For $x \in \mathbb{Z}$, $x \vee 0 := \max(x, 0)$.

Lemma 2.82 (*The inclusion-exclusion principle*)**.** *Let $A_1, \ldots A_n$ be finite sets. Then*

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\varnothing \neq J \subseteq \{1,\ldots,n\}} (-1)^{|J|+1} \left|\bigcap_{j \in J} A_j\right|$$

Lemma 2.83. *Let r and m be natural numbers. Let u be a word over an alphabet of size m with $|u| = s < r$. Let*

$$s' := \begin{cases} \min(|u'|) & \exists l \in \mathbb{Q},\, 1 < l \wedge (u')^l = u \\ s & \textit{else} \end{cases}$$

$$P := \left\{ p : p \mid r \wedge s' \leq \frac{r}{p},\ p \textit{ is prime} \right\}$$

For each $\varnothing \neq P' \subseteq P$, define

$$p_{P'} := \begin{cases} \prod_{p \in P'} p & P' \neq \varnothing \\ 1 & P' = \varnothing \end{cases}$$

$$s_{P'} := |P'|$$

The number of primitive words w of length r such that $w \upharpoonright s = u$ is equal to

$$\sum_{P' \subseteq P} (-1)^{s_{P'}+1} \cdot m^{\left(\frac{r}{p_{P'}} - s\right) \vee 0}$$

Proof. The proof will proceed instead by counting the *non*-primitive words of length r that start with u. Let

$$A_d = \left\{ w \in \Sigma^r : w \upharpoonright_s = u \wedge w \text{ is a } \tfrac{r}{d}\text{-th power} \right\}$$

The goal is to calculate $\left|\bigcup_{d|r} A_d\right|$ by applying Lemma 2.82. If $d \mid r$, then $A_d \subseteq A_{d'}$ for all $d' \mid d$. Therefore,

$$\left|\bigcup_{d|r} A_d\right| = \left|\bigcup_{p : p|r \wedge p\,\text{prime}} A_p\right|$$

However, most such A_p will be empty. A_p is nonempty if and only if there exists a word $w \in \Sigma^r$ such that both the first $\frac{r}{p}$ letters of w and the first s bits of w are fixed. If u is primitive and is not equal to $(u')^l$ for any $l \in \mathbb{Q}$ with $l > 1$, then it must be the case that $\frac{r}{p} \geq s$; otherwise, $\frac{r}{p} \geq |u'|$, where u' is the shortest word such that $(u')^l = u$ for $l \in \mathbb{Q}$. So $\left|\bigcup_{d|r} A_d\right| = \bigcup_{p \in P} A_p$. Now, let $\varnothing \neq P' \subseteq P$; calculating $\left|\bigcap_{p \in P'} A_p\right|$ finishes the proof. $\bigcap_{p \in P'} A_p = A_{\mathrm{lcm}(p \in P')}$, and since P' contains distinct primes, $\mathrm{lcm}(p \in P') = \prod_{p \in P'}$. Let $p_{P'} = \prod_{p \in P'}$. For $w \in A_{p_{P'}}$, $\left(\frac{r}{p} - s'\right) \vee 0$ letters of w are not fixed, which implies that $|A_{p_{P'}}| = m^{\left(\frac{r}{p_{P'}} - s\right) \vee 0}$. So the number of words of length r starting with u is equal to

$$\begin{aligned} m^{r-s} - \sum_{\varnothing \neq P' \subseteq P} (-1)^{s_{P'}+1} \cdot m^{\left(\frac{r}{p_{P'}} - s\right) \vee 0} &= \\ m^{r-s} - \left(\sum_{\varnothing \neq P' \subseteq P} (-1)^{s_{P'}+1} \cdot m^{\left(\frac{r}{p_{P'}} - s\right) \vee 0} + (-1)^{s_{\varnothing}+1} \cdot m^{\left(\frac{r}{p_{\varnothing}} - s\right) \vee 0} \right) &= \\ \sum_{P' \subseteq P} (-1)^{s_{P'}+1} \cdot m^{\left(\frac{r}{p_{P'}} - s\right) \vee 0} & \end{aligned}$$

completing the proof.

□

Lemma 2.84. *Let r, m and l be natural numbers. Let u be a word over an alphabet of size m with $|u| = s < \lfloor \frac{r}{2} \rfloor$ and $l \leq s$. Let P, s', and $p_{P'}$ be as in Lemma 2.83. Define*

$$\begin{aligned} \mathrm{IdxFi} &= \{i : i < s\} \cup \{i : l \leq i < l + s\} \\ \mathrm{IdxFr} &= \{i < r : i \notin \mathrm{IdxFi}\} \end{aligned}$$

the set of fixed and free indices in u. For each $p_{P'}$, define

$$\begin{aligned} M_{P'} &= \left\{ i < s : \exists j \in \mathrm{IdxFi}, i \equiv j \mod \frac{r}{p_{P'}} \right\} \\ T_{P'} &= \left\{ i < \frac{r}{p_{P'}} : \forall j \in \mathrm{IdxFr}, i \not\equiv j \mod \frac{r}{p_{P'}} \right\} \end{aligned}$$

and let

$$\mathcal{F}_P = \left\{ \varnothing \neq P' \subseteq P : |M_{P'}| \leq 1 \right\} \cup \left\{ \varnothing \neq P' \subseteq P : \forall j, k \in M_{P'}, u[j] = u[k] \right\}$$

The number of primitive words w of length r such that $w \restriction s = u$ and $w[l : l+s] = u$ is equal to

$$m^{r-s} - \sum_{P' \in \mathcal{F}_P} (-1)^{s_{P'}+1} \cdot m^{|T_{P'}|}$$

Proof. The proof is analogous to that of Lemma 2.83. As in that proof, Lemma 2.82 will be used to count the non-primitive words satisfying the conditions. Define A_d as in Lemma 2.83. As before, if p is a prime divisor of r, A_p may be empty. Consider IdxFi, the *fixed indices* of any possible word in A_p. A word in A_p is a $\frac{r}{p_{P'}}$-th power, and therefore any index in IdxFi modulo $\frac{r}{p_{P'}}$ is also fixed. If $u[j] \neq u[k]$ for any two indices j and k in IdxFi modulo $\frac{r}{p_{P'}}$, then A_p is empty. Hence $\left|\bigcup_{d|r} A_d\right| = \left|\bigcup_{P' \in \mathcal{F}_P} A_{p_{P'}}\right|$. Let $P' \in \mathcal{F}_P$; it suffices to calculate $|A_{p_{P'}}|$. The possible free letters of a word in $A_{p_{P'}}$ are the ones less than $\frac{r}{p_{P'}}$ such that they are not equal to any letters in IdxFr, the *free indices*, modulo $\frac{r}{p'_P}$. This is exactly the definition of $T_{P'}$, finishing the proof.

□

Lemma 2.85. *Let r and s be natural numbers with $s < \lfloor \frac{r}{2} \rfloor$. There are*

$$\mathcal{P}(r) - \sum_{i=1}^{s} \sum_{u \in \Sigma^i} F_2(r, u, r - i)$$

primitive words w of length r such that for all $k \leq s$, $w \restriction k = w[r - k :]$.

Proof. For any $i \leq s$ and for any word u of length i, there are $F_2(r, u, r-i)$ primitive words of length r such that $u = w \restriction k = w[r-k:]$ by Lemma 2.84. The result then follows by summing over all $u \in \Sigma^i$ and all $i \leq s$. □

Results thus far have enabled calculation of $s_q(n)$ for the regimes $n \leq q$ and $n \geq 2q-1$. Specifically,

Theorem 2.86. *Let m, n, and q be natural numbers. For natural numbers n and q and an alphabet Σ with $|\Sigma| = m$,*

$$s_q(n) = \begin{cases} 0 & n < q \\ m(m-1)^{n-1} & n = q \\ \mathcal{C}(q) & n \geq 2q-1 \end{cases}$$

Proof. If $n < q$, the result is immediate since $D(w) \leq |w|$ for any word w by Theorem 2.23. If $n = q$, then a word w has complexity q if and only if w_{n-1} does not occur in $w \restriction_{n-1}$. Fixing one of m possibilities for the last letter, there are $(m-1)^{q-1}$ choices for the remaining $n-1$ letters.

Recall the map $f_{n,q}$ from the set of eventually integral pairs of complexity q, $\mathcal{S}(q)$, to words of length n defined by $f_{n,q}(u,v) = uv^{\frac{n-|u|}{|v|}}$. By Theorem 2.55,

$$\{w \in \Sigma^n : D(w) = q\} \subseteq \{f_{n,q}(u,v) : (u,v) \in \mathcal{S}(q)\}$$

and by Theorem 2.56, if $n \geq 2q-1$ then

$$\{f_{n,q}(u,v) : (u,v) \in \mathcal{S}(q)\} \subseteq \{w \in \Sigma^n : D(w) = q\}$$

So whenever $n \geq 2q-1$,

$$\begin{aligned} s_q(n) &= |\{w \in \Sigma^n : D(w) = q\}| \\ &= |\{f_{n,q}(u,v) : (u,v) \in \mathcal{S}(q)\}| \\ &= |\mathcal{S}(q)| \\ &= \mathcal{C}(q) \end{aligned}$$

where the first equality follows by the definition of $s_q(n)$, the third equality follows since $f_{n,q}$ is injective for $n \geq 2q-1$ by Theorem 2.54, and the last equality follows by Theorem 2.49. □

Remark 2.87. Since $\mathcal{C}(n)$ can be computed in polynomial time in n, $s_q(n)$ can also be computed in polynomial time in n. This is an improvement over naively computing $s_q(n)$ for each word of length n, a computation that would run in exponential time in n.

Lemma 2.83 and Lemma 2.84 were formulated with the goal of extending Theorem 2.86 to calculate $s_q(n)$ for $q < n < 2q-1$. Recall that if (u,v) is an eventually integral pair of complexity

q, $uv^{1+\frac{n-q}{|v|}}$ is a word of length n of complexity q; further, all words of length n and complexity q are formed from eventually integral pairs with complexity q. Let (u_1, v_1) and (u_2, v_2) have eventually integral complexity with $|u_1| + |v_1| = |u_2| + |v_2| = q$, and assume that $(u_1, v_1) \neq (u_2, v_2)$ and $u_1v_1 = u_2v_2$. Without loss of generality assume that $|v_1| > |v_2|$. If $u_1v_1^{1+\frac{n-q}{|v_1|}} = u_2v_2^{1+\frac{n-q}{|v_2|}}$, then $v_1 \upharpoonright_{n-q} = v_2 \upharpoonright_{n-q}$. But since $u_1v_1 = u_2v_2$, this implies that $v_1 \upharpoonright_{n-q} = v_1[l : l + n - q]$, where $l = |v_1| - |v_2|$. If $l \geq n - q$, then two distinct subwords of v_1 are equal; otherwise, an initial segment of v_1 of length $l + 1$ is fixed. Naively, iterating over all possible lengths $|v_1|$ and $|v_2|$, all possible initial segments of v_1 of length $n - q$, calculating the number of primitive words of length $|v_1|$ such that there exists words of length $|v_2|$ with $v_1 \upharpoonright_{n-q} = v_2 \upharpoonright_{n-q}$, and deriving the number of eventually integral pairs (u_1, v_1) of the corresponding v_1 should be the difference between $\mathcal{C}(q)$ and $s_q(n)$. That is, the following algorithm should return $\mathcal{C}(q) - s_q(n)$:

```
1. Initialize a variable $Tot$.
2. For each $i$ with $n - q < i \leq q$,
    (a) For each $j$ with $1 \leq j < i$,
    (b) If $i - j < n - q$,
            $Tot += m^{i-1}(m-1) \sum_{u \in \Sigma^{n-q}} F_1(i, u)$
        Else,
            $Tot += m^{i-1}(m-1) \sum_{u \in \Sigma^{n-q}} F_2(i, u, i - j)$
3. Return $Tot$
```

There are two problems with this algorithm. First, fixing v_1 and simply counting the number of v_2 such that $v_1 \upharpoonright_{n-q} = v_2 \upharpoonright_{n-q}$ ignores the fact that if (u_1, v_1) being an eventually integral pair with $u_1v_1 = u_2v_2$ does not imply that (u_2, v_2) is an eventually integral pair. For example, letting $n = 6$, $q = 5$, $i = 4$, and $j = 1$, $(1, 0100)$ is an eventually integral pair correctly associated with 101001, which has length 6 and complexity 5, but $(10100, 0)$ is not an eventually integral pair. Second, primitive words of length q do not necessarily have complexity q. For example, if $n = 5$ and $q = 4$, $(\epsilon, 0010)$ would be counted by the algorithm as a possible eventually integral pair but 00100 has HMM complexity 3, via $(001)^{\frac{5}{3}}$.

Figure 2.4 shows $s_q(n)$ for various n and q.

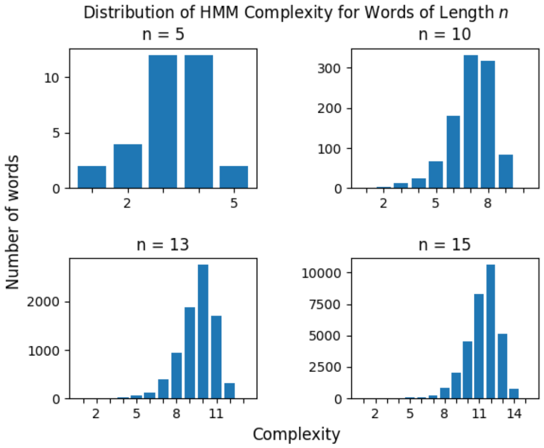

Figure 2.4: $s_q(n)$ for various n and q

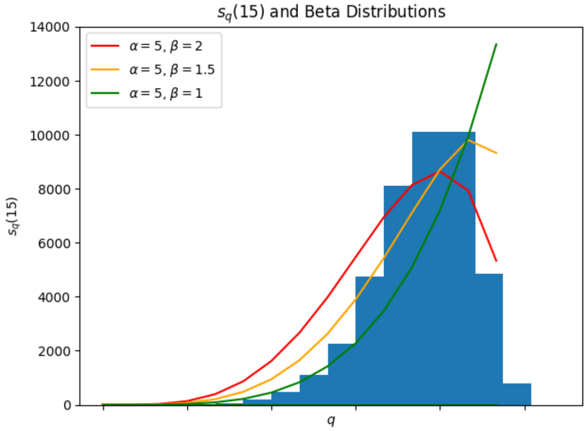

Figure 2.5: $s_q(15)$ for various beta distributions

Definition 2.88. The *beta function* $B(z_1, z_2)$ is defined by

$$B(z_1, z_2) := \int_0^1 t^{z_1-1}(1-t)^{z_2-1}dt$$

Definition 2.89. Given constants α and β, the probability density function of the *beta distribution* is defined for each x with $0 \leq x \leq 1$ by

$$f(x, \alpha, \beta) := \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where B is the beta function. Figure 2.5 graphs various beta distributions against the distribution of HMM complexity among words of length 15; preliminary results suggest that the two are similar.

CHAPTER 3
FURTHER RESULTS ON FINITE-STATE MACHINE COMPLEXITY

3.1 Complexity of probabilistic HMMs

So far, only deterministic HMMs and their associated definitions of complexity have been considered. This section aims to expand the results in this chapter to probabilistic HMMs.

3.1.1 Transition-probabilistic HMMs

Here, a specific class of probabilistic HMMs will be considered. Recall from Section 2.1 that if $\{a_{ij}\}$ is the transition matrix of a HMM, a_{ij} is the probability that the HMM transitions from state i to state j.

Definition 3.1. A *half-transition-probabilistic HMM* is a HMM H such that if $A = \{a_{ij}\}$ is the transition matrix of H, then $a_{ij} \in \left\{0, \frac{1}{2}, 1\right\}$ for all i and j.

Complexity for half-transition-probabilistic HMMs will be defined analogously to HMM complexity.

Definition 3.2. Let H be a HMM with transition matrix $A = \{a_{ij}\}$ and let $w = w_0 w_1 \dots w_{n-1}$ be a word. Assume H has n states and fix an enumeration $q_0, q_1, \dots q_{n-1}$ of the states of H. Let $X = x_0 x_1 \dots x_{n-1}$ be a sequence of states such that $Y(x_i) = w_i$ for all $i < n$, and let $f : \{0, \dots n\} \to \{0, \dots, n\}$ be such that $q_{f(i)} = x_i$ for all $i < n$. The *probability of H emitting w* is

$$p_H(w) = \sum_X \prod_{i=0}^{n} a_{f(i)f(i+1)}$$

where the sum is taken over all possible X.

Definition 3.3. Let w be a word over an alphabet Σ. The *half-transition-probabilistic HMM complexity* of w, written $C_{htpHMM}(w)$, is the minimum number of states in a half-transition-probabilistic HMM H such that if $w' \neq w$ is a word with $|w'| = |w|$, then $p_H(w') < p_H(w)$.

Note that by this definition, for all n, the set of deterministic HMMs is a proper subset of the set of half-transition-probabilistic HMMs. This leads directly to the following observation.

Lemma 3.4. *For any alphabet Σ and word $w \in \Sigma^+$, $C_{htpHMM}(w) \leq D(w)$.*

Proof. Let H be the deterministic HMM witnessing $D(w)$. Then H is also a half-transition-probabilistic HMM, so $C_{htpHMM}(w) \leq n$, where n is the number of states in H. □

Half-transition-probabilistic HMM complexity is equal to deterministic complexity for binary words of complexity two. More precisely,

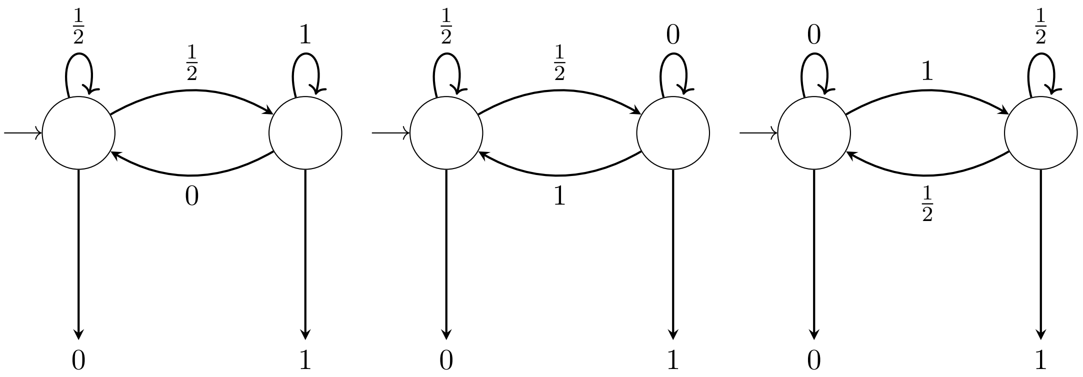

Figure 3.1: Possible 2-state half-transition-probabilistic HMMs

Theorem 3.5. *Let $\Sigma = \{0, 1\}$. Then $\{w \in \Sigma^+ \mid C_{htpHMM}(w) = 2\} = \{w \in \Sigma^+ \mid D(w) = 2\}$.*

Proof. If $n = 1$, then $\{w \in \Sigma^+ \mid C_{htpHMM}(w) = 2\} = \{w \in \Sigma^+ \mid D(w) = 2\} = \varnothing$, so assume $n > 1$. The result will be shown by arguing that for all possible half-transition-probabilistic HMMs H with two states, for every word w such that H witnesses $C_{htpHMM}(w) = 2$, $D(w) = 2$. Without loss of generality, assume that exactly one state of H emits 0 and exactly one state of H emits 1; otherwise, H can only emit either 0^+ or 1^+, neither of which have complexity 2. By assuming that the start state of such a H emits 0, it suffices to prove the result for words starting with 0. It can also be assumed that there is at least one instance of $\frac{1}{2}$ in H; otherwise, H only witnesses deterministic complexity. Finally, let A be the transition matrix of H. If $a_{00} = 1$, then H will never transition out of the start state, making H equivalent to a one-state HMM. The possible transition matrices for H are then

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

These correspond to the half-transition-probabilistic HMMs in Figure 3.1.

Let H_1, H_2, and H_3 be the half-transition-probabilistic HMMs in Figure 3.1, respectively.

Claim 1: For all $n > 1$ and $w \in \Sigma^n$, if $w \neq 01^{n-1}$ then $p_{H_1}(w) < p_{H_1}(01^{n-1})$.

Proof of claim 1: Let A_1 be the transition matrix of H_1. Since $a_{00} = a_{01} = \frac{1}{2}$, $p_{H_1}(w) < \frac{1}{2}$ for any word w. But $p_{H_1}(01^{n-1}) = \frac{1}{2}$ by direct computation.

Claim 2: For all $n > 1$ and $w \in \Sigma^n$, if $w \neq (01)^{\frac{n}{2}}$ then $p_{H_2}(w) < p_{H_2}((01)^{\frac{n}{2}})$.

Proof of claim 2: First, note that for all $n \geq 1$ and all words w of length $2n + 1$, $p_{H_2}(w) \leq \frac{1}{2^n}$. This can be proven via induction. For $n = 1$, a word of length 3 is outputted by two state transitions. If q_0 and q_1 are the states of H_2 emitting 0 and 1, respectively, the possible state transitions to emit a word of length 3 are q_0q_0, q_0q_1, q_1q_0, q_1q_1. These correspond to the calculations $a_0 0 * a_0 0$, $a_{00} * a_{01}$, $a_{01} * a_{10}$, $a_{01} * a_{11}$, which evaluate to $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{2}$, 0, respectively. The same logic and calculations on state transitions apply to the inductive step. The proof of claim 2 now follows from observing that $p_{H_2}((01)^{\frac{n}{2}}) = \frac{1}{2^n}$ for all $n > 1$.

Claim 3: For all $n > 1$ and $w \in \Sigma^n$, if $w \neq (01)^{\frac{n}{2}}$ then $p_{H_3}(w) < p_{H_2}((01)^{\frac{n}{2}})$.

Proof of claim 3: The proof of this claim is analogous to that of claim 2. First, for all $n \geq 1$ and all words w of length $2n$, $p_{H_2}(w) \leq \frac{1}{2^n}$. This can be proven via induction. Let q_0 and q_1 be the states of H_3 emitting 0 and 1, respectively. The case $n = 1$ is immediate as q_0 transitions to q_1 with probability 1. Assume true for n and consider the next two state changes required to emit a word of length $2(n+1)$. Regardless of which state H_3 is in at time n, the next two possible state transitions are q_0q_0, q_0q_1, q_1q_0, q_1q_1, corresponding to probabilities of 0, 0 $\frac{1}{2}$, $\frac{1}{2}$, respectively, completing the induction. The proof of claim 3 now follows from observing that $p_{H_3}((01)^{\frac{n}{2}}) = \frac{1}{2^n}$ for all $n \geq 1$.

Together, the three claims show that for all $w \in \Sigma^n$ for $n > 1$ and for all $0 \leq i \leq 2$, either H_i is not a witness for the half-transition-probabilistic HMM complexity of w or $w \in \left\{01^{n-1}, (01)^{\frac{n}{2}}\right\}$. But $D(01^{n-1}) = D((01)^{\frac{n}{2}}) = 2$ for all $n > 1$, finishing the proof. □

3.1.2 Emission-probabilistic HMMs

By allowing the emission matrix, and not the transmission matrix, of a HMM to be probabilistic, a different notion of HMM complexity can be derived. Recall from Section 2.1 that if $\{b_{ij}\}$ is the emission matrix of a HMM, b_{ij} is the probability that the HMM emits symbol j if it is in state i.

Definition 3.6. A *half-emission-probabilistic HMM* is a HMM H such that if B is the emission matrix of H, then $b_{ij} \in \left\{0, \frac{1}{2}, 1\right\}$ for all i and j.

Definition 3.7. Let w be a word over an alphabet Σ. The *half-emission-probabilistic HMM complexity* of w, written $C_{hepHMM}(w)$, is the minimum number of states in a half-emission-probabilistic HMM H such that if $w' \neq w$ is a word with $|w'| = |w|$, then $p_H(w) < p_H(w')$.

The notions of half-emission-probabilistic HMMs and half-transmission-probabilistic HMMs do not coincide.

Theorem 3.8. *There exists a half-transmission-probabilistic HMM H_1 such that for any half-emission-probabilistic HMM H_2, there is a word w such that $p_{H_1}(w) \neq p_{H_2}(w)$.*

Proof. Let H_1 be the half-transmission-probabilistic HMM with transmission matrix $\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$ and assume there exists a half-emission-probabilistic HMM H_2 such that for any word w, $p_{H_1}(w) = p_{H_2}(w)$. First, note that for any $w \in \Sigma^+$,

$$\begin{cases} p_{H_1}(w) = \frac{1}{2^{n+1}} & \exists n, k \text{ such that } w = 0^n 1^k \\ 0 & \text{else} \end{cases}$$

In particular, $p_{H_1}(0) = p_{H_1}(1) = \frac{1}{2}$, which implies that the starting state of H_2 must emit 0 and 1 each with probability $\frac{1}{2}$. If H_2 only contains this state, then $p_{H_2}(00) = \frac{1}{4} \neq p_{H_1}(00)$, so H_2 contains at least two states. If H_2 emits 0 and 1 each with probability $\frac{1}{2}$, then again $p_{H_2}(00) = \frac{1}{4}$, but if H_2 emits 1 with probability 1, then $p_{H_2}(001) = 0$, a contradiction. □

The strict requirement that the transmission (resp. emission) probabilities of half-transmission-probabilistic (resp. half-emission-probabilistic) HMMs only take values in $\left\{0, \frac{1}{2}, 1\right\}$ will be dropped when considering the dual question of whether every emission-probabilistic HMM can be simulated by a transmission-probabilistic HMM.

Definition 3.9. A *transmission-probabilistic* (resp. *emission-probabilistic*) HMM is a HMM whose transmission (resp. emission) matrix takes values in $\mathbb{R} \cap [0, 1]$.

Definition 3.10. A *delayed HMM* is a HMM that does not emit a symbol until after it makes its first transition.

Transition-probabilistic delayed HMMs will be used as a substitute for "ordinary" transition-probabilistic HMMs because a transition-probabilistic HMM always emits the same symbol from its start state with probability 1, so it is trivial to show that there exist emission-probabilistic HMMs that cannot be simulated by any transition-probabilistic HMM.

Theorem 3.11. *For any emission-probabilistic two-state HMM H_1, there exists a transition-probabilistic delayed HMM H_2 such that for any word w, $p_{H_1}(w) = p_{H_2}(w)$.*

Proof. Let H_1 have emission matrix $\begin{bmatrix} x & 1-x \\ 1-y & y \end{bmatrix}$. There are three possible transition matrices for H_1:

Case 1: the first row of the transition matrix of H_1 is $[1, 0]$. Consider the transition-probabilistic delayed HMM with transition matrix $\begin{bmatrix} x & 1-x \\ x & 1-x \end{bmatrix}$. Prove that $p_{H_1}(w) = p_{H_2}(w)$ for all words w by structural induction. $p_{H_1}(0) = p_{H_2}(0) = x$ and $p_{H_1}(1) = p_{H_2}(1) = 1 - x$ by definition. Now assume that $p_{H_1}(w) = p_{H_2}(w)$ for a word w. Then

$$p_{H_1}(w0) = p_{H_1} \cdot x = p_{H_2} \cdot x = p_{H_2}(w0)$$

and

$$p_{H_1}(w1) = p_{H_1} \cdot (1 - x) = p_{H_2} \cdot (1 - x) = p_{H_2}(w1)$$

completing the induction.

Case 2: the transmission matrix of H_1 is $\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$. Consider the transition-probabilistic delayed HMM with transition matrix

$$\begin{bmatrix} 0 & x & 1-x \\ 0 & 1-y & y \\ 0 & 1-y & y \end{bmatrix}$$

and emission matrix

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Prove that $p_{H_1}(w) = p_{H_2}(w)$ for all words w by structural induction. $p_{H_1}(0) = p_{H_2}(0) = x$ and $p_{H_1}(1) = p_{H_2}(1) = 1 - x$ by definition. Now assume that $p_{H_1}(w) = p_{H_2}(w)$ for a word w. Then

$$p_{H_1}(w0) = p_{H_1} \cdot (1 - y) = p_{H_2} \cdot (1 - y) = p_{H_2}(w0)$$

and

$$p_{H_1}(w1) = p_{H_1} \cdot y = p_{H_2} \cdot y = p_{H_2}(w1)$$

completing the induction.

Case 3: the transmission matrix of H_1 is $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Consider the transition-probabilistic delayed HMM with transition matrix

$$\begin{bmatrix} 0 & x & 0 & 0 & 1-x \\ 0 & 0 & 1-y & y & 0 \\ 0 & x & 0 & 0 & 1-x \\ 0 & x & 0 & 0 & 1-x \\ 0 & 0 & 1-y & y & 0 \end{bmatrix}$$

and emission matrix

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

Prove that $p_{H_1}(w) = p_{H_2}(w)$ for all words w by structural induction. $p_{H_1}(0) = p_{H_2}(0) = x$ and $p_{H_1}(1) = p_{H_2}(1) = 1 - x$ by definition. Now assume that $p_{H_1}(w) = p_{H_2}(w)$ for a word w. Let s_{11} be the state of H_1 that corresponds to the first row of the emission matrix of H_1, and let s_{12} be the state that corresponds to the second. Similarly, for $1 \leq i \leq 5$ define s_{2i} to be the state of H_2 corresponding to the ith row of H_2. There are two subcases: either H_1 is in s_{11} after emitting w or H_2 is in s_{12}. First, assume that H_1 is in s_{11}. Since H_1 transitions between s_1 and s_2 with probability 1, if H_1 is in s_{11} then $|w|$ must be odd. This implies that H_2 must be in either state s_{22} or s_{25}. To emit $w0$, H_1 transitions to s_{12} and

$$p_{H_1}(w0) = p_{H_1} \cdot (1 - y) = p_{H_2} \cdot (1 - y) = p_{H_2}(w0)$$

and

$$p_{H_1}(w1) = p_{H_1} \cdot y = p_{H_2} \cdot y = p_{H_2}(w1)$$

as required. Now assume that H_1 is in s_{12} after emitting w. By following the same logic, $|w|$ is even, so

$$p_{H_1}(w0) = p_{H_1} \cdot x = p_{H_2} \cdot x = p_{H_2}(w0)$$

and

$$p_{H_1}(w1) = p_{H_1} \cdot (1-x) = p_{H_2} \cdot (1-x) = p_{H_2}(w1)$$

as required. □

3.2 HMM complexity as language

3.2.1 Regularity of complexity for fixed length

Definition 3.12. Let Σ be an alphabet. A *language over* Σ is a set $L \subseteq \Sigma^*$.

In this section, deterministic HMM complexity will be used to define languages over various alphabets, and results on the properties of these languages will be derived.

Definition 3.13. Let Σ be an alphabet. A *regular expression over* Σ is a well-formed string over $\Sigma \cup \{\epsilon, \varnothing, (,), +, *\}$.

Here, $+$ represents a set union. Given a language Σ, note that any regular expression over Σ defines a subset of Σ^*.

Example 3.14. Let $\Sigma = \{0, 1\}$ and let φ be the regular expression $0^* + 1^*$. Then φ defines the set $\{\epsilon, 0, 1, 00, 11, 000, 111, \dots\}$.

Given an alphabet Σ and a regular expression φ, let $\widehat{\varphi} \subseteq \Sigma^*$ be the set defined by φ.

Definition 3.15. A language L is *regular* if there exists a regular expression φ such that $\widehat{\varphi} = L$.

Fix $n \in \mathbb{N}$ with $n > 1$. [1] Let $\Sigma_n = \{0, \dots, n-1\}$ and write Σ_n^* for $(\Sigma_n)^*$. Let L_n be the language $\{w \in \Sigma_n^* \mid D(w) < |w|\}$.

Theorem 3.16. *L_n is regular.*

The proof will be done by strong induction on n. The inductive hypothesis will be used by appealing to two lemmas and a well-known result.

Lemma 3.17. *Let φ be a permutation on Σ_n, which induces a morphism $\varphi^* : \Sigma_n^* \to \Sigma_n^*$. Then for all $w \in \Sigma_n^*$, $w \in L_n$ if and only if $\varphi^*(w) \in L_n$.*

[1] The results here also hold for $n = 0$ (by defining $\Sigma_0 := \varnothing$) and $n = 1$, but $n = 2$ is the first "interesting" case. In particular, using $n = 2$ as a base case demonstrates the intuition behind the proof most clearly.

Proof. If M is a HMM witnessing $w \in L$, relabel the state changes of M according to φ. □

Lemma 3.18. $L_n \subseteq L_{n+1}$.

Proof. If M is a HMM witnessing $w \in L_n$, to create a deterministic HMM M' on Σ_{n+1} witnessing $w \in L_{n+1}$, create a state s such that s transitions to itself with probability 1, and a transition from each existing state to s whenever the letter n appears. □

Lemma 3.19 ([28] Theorem 3.9.5)**.** *Let L be any language and let Σ be any alphabet. For $x, y \in \Sigma^*$, define $x \sim_L y$ if for all $z \in \Sigma^*$, $xz \in L \leftrightarrow yz \in L$. Then $\sim_L$ is an equivalence relation on Σ^* and L is regular if and only if $\sim_L$ partitions Σ^* into finitely many equivalent classes.*

And now back to,

Proof of the theorem. The goal of the proof will be to show that for all n, $\sim_{L_n}$ partitions Σ_n^* into finitely many equivalence classes. The result will then follow by Lemma 3.19

Let $n = 2$. Then $[\epsilon]$, $[0]$, $[1]$, $[00]$, $[01]$, $[11]$, and $[010]$ are the only equivalence classes mod $\sim_{L_2}$. To see this, fix $w \in \Sigma_2^*$ and consider the following cases.

- Case 1: $w \in L_2$. There are three subcases:

  1a. $0 \in w$ and $1 \in w$. Then $w \in [010]$.

  1b. Either $0 \in w$ or $1 \in w$, but not both. In the former case, $w \in [00]$, and in the latter, $w \in [11]$.

  1c. $w = \epsilon$. Then $w \in [\epsilon]$ trivially.

- Case 2: $w \notin L$. Similarly to the first case, there are two subcases:

  2a. $0 \in w$ and $1 \in w$. Then $w \in [01]$.

  2b. Either $0 \in w$ or $1 \in w$, but not both. In the former case, $w \in [0]$, and in the latter, $w \in [1]$.

Now assume true for all $k \leq n$ and fix $w \in \Sigma_{n+1}^*$. The breakdown of cases is analogous to the proof of the base case.

- Case 1: $w \in L_{n+1}$. There are two subcases:

  1a. $i \in w$ for all $i \in \Sigma_{n+1}$. Then $w \in [01 \ldots n0]$.

  1b. $w \in (\Sigma')^*$ for some $\Sigma' \subsetneq \Sigma$. Let $|\Sigma'| = k$ for $k < n+1$ and let $\varphi : \Sigma' \to \Sigma_k$ be a bijection. By the inductive hypothesis, L_k is regular, so $\varphi(w)$ is in one of at most finitely many equivalence classes induced by $\sim_{L_k}$. Then if $\varphi(w) \in [v]$ for some $v \in \Sigma_k^*$, $w \in \left[\varphi^{-1}(v)\right]$ by the lemmas and there are at most finitely many such $\left[\varphi^{-1}(v)\right]$.

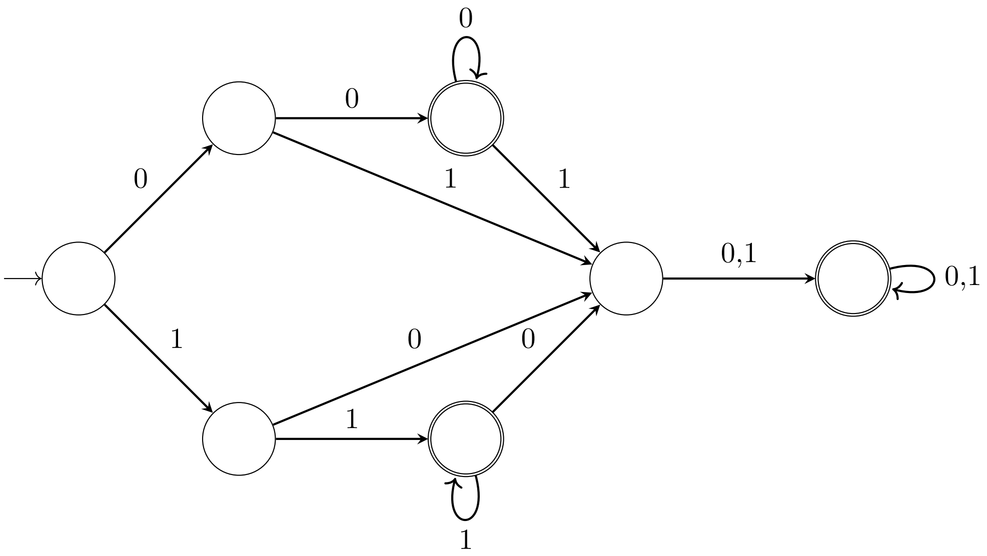

Figure 3.2: Accepting DFA for L_2

- Case 2: $w \notin L_{n+1}$.

  2a. $i \in w$ for all $i \in \Sigma_{n+1}$. Then $w \in [01 \ldots n]$.

  1b. $w \in (\Sigma')^*$ for some $\Sigma' \subsetneq \Sigma$. Then apply the same reasoning as in case 1b.

□

From the Myhill-Nerode theorem, equivalence classes of $\sim_L$ correspond to states of a DFA accepting L. The correctness of the base case of the theorem can be verified by explicitly constructing an accepting DFA for L_2, shown in Figure 3.2.

To construct a witnessing DFA for L_n, $n > 2$, consider the n subsets $\Sigma'_1, \ldots, \Sigma'_n$ of Σ_n corresponding to each choice of $n-1$ elements of Σ_n. Let M_0 be the DFA accepting L_{n-1}. For each Σ'_i, let M_i be the DFA defined by relabeling the arrows of M_0 according to a bijection $\Sigma_{n-1} \to \Sigma'_i$. Then "glue" together M_i, $0 \leq i \leq n$. For example, if $n = 3$, M_0 would be the DFA shown above, M_1 would be a DFA on $\{0, 2\}$ and M_2 would be a DFA on $\{1, 2\}$.

The benefit of the proof of Theorem 3.16 is that it immediately shows how to construct a witnessing DFA. However, Theorem 3.16 has a simpler proof which will be presented here for completeness.

Lemma 3.20. *If L is a regular language, then L^c is also regular.*

Proof. Lemma 3.19 implies that L is regular if and only if L is accepted by a DFA. Let D be a DFA that accepts L, and let D' be the DFA formed by turning all the accept states into non-accepting states, and all non-accepting states into accept states. Then D' recognizes L^c. □

Theorem 3.21. *L_n is regular.*

Proof. By Lemma 3.20, it suffices to show that $(L_n)^c = \{w : D(w) = |w|\}$ is regular. $D(w) = |w|$ if and only if $w[-1] \notin w \upharpoonright_{|w|-1}$. For $0 \leq i < n$, let φ_i be the regular expression

$$(0 + 1 + \cdots + (i-1) + (i+1) + \cdots + (n-2) + (n-1))^* i$$

That is, φ_i represents the set of words w such that the last letter of w is i, and i does not appear in $w \upharpoonright_{|w|-1}$. Let

$$\varphi = \varphi_0 + \varphi_1 + \cdots + \varphi_{n-1}$$

Then $\widehat{\varphi} = (L_n)^c$ by definition. □

3.2.2 Regularity of fixed complexity

The results here can be considered as dual to the results in Section 3.2.1. Specifically, while Section 3.2.1 dealt with the regularity of the language of binary words of arbitrary complexity and fixed length, this section will discuss the regularity of the language of binary words of fixed complexities and arbitrary lengths. Specifically, if $\Sigma = \{0, 1\}$, the language $L_n := \{w \in \Sigma^* \mid D(w) = n\}$ is regular for all n.

Lemma 3.22. *Let u and v be words with $u \preceq v$. Then $D(u) \leq D(v)$.*

Proof. If H is a HMM that witnesses the deterministic HMM complexity of v, then H also emits u. □

Lemma 3.23. *Let n be a natural number and let w be a word such that $|w| > 2n$ and $D(w) < n$. Then, for any word z, $D(wz) \neq n$.*

Proof. Let $w = uv^p$ for words u and v and for $p \in \mathbb{Q}$. Assume that $|u| + |v|$ is minimal. $|u| + |v| < n$, so $p > 1$. Let z be any word. There are two cases to consider:

Case 1: Either $p \in \mathbb{N}$ and $z = v^q$ for some $q \in \mathbb{Q}$, or $p = k + \frac{r}{s}$, $z \upharpoonright_{s-r} = v[s-r:]$, and $z[s-r : |v|] = v \upharpoonright_{s-r}$, $z = \left(z_{|v|}\right)^q$ for some $k, r, s \in \mathbb{N}$ and $q \in \mathbb{Q}$. In either of these subcases, $v^p z = v^{p'}$ for some $p' \in \mathbb{Q}$, so $D(wz) = D(w) < n$.

Case 2: Neither of the subcases in case 1 holds. Assume that $uv^p z = u_1 v_1^{p_1}$ for $p_1 \in \mathbb{Q}$ and words u_1 and v_1 with $|u_1| + |v_1| < |u| + |v|$. Then $uv^p = u_1 v_1^{p_2}$ for some $p_2 \in \mathbb{Q}$, contradicting the fact that $|u| + |v|$ is minimal. But this subcase implies that $v^p z \neq a^q$ for any word a and $q \in \mathbb{Q}$. So $D(uv^p z) \geq |uv^p| > n$. □

Lemma 3.24. *Fix a natural number n. Let $L_n := \{w \in \Sigma^* \mid D(w) = n\}$ and let $\sim_{L_n}$ be the relation defined in Lemma 3.19. Let (u, v) be an eventually integral pair with $|u| + |v| = n$. Let $p_1, p_2 \in \mathbb{Q}$ with $p_1, p_2 > n - |u|$ and let $p_1 = \frac{r_1}{|v|}$, $p_2 = \frac{r_2}{|v|}$. Then $uv^{p_1} \sim_{L_n} uv^{p_2}$ if and only if $r_1 = r_2$.*

Proof. Let $r_1 = r_2$. Then for all words z, $uv^{p_1}z \in L_n$ if and only if $z \restriction_{r_1}= v[|v| - r_1 :]$, $z[|v| - r_1 : |v|] = v \restriction_{|v|-r}$, and $z = \left(z_{|v|}\right)^q$ for some $q \in \mathbb{Q}$ if and only if $uv^{p_2}v \in L_n$.

Let $r_1 \neq r_2$. Let $z = v[|v| - r_1 :]$. Then $uv^{p_1}z \in L_n$ but $uv^{p_2}z \notin L_n$, so $uv^{p_1} \not\sim_{L_n} uv^{p_2}$. □

Lemma 3.25. *Let n, L_n, and $\sim_{L_n}$ be as in Lemma 3.24. Let $(u_1, v_1) \neq (u_2, v_2)$ be two eventually integral pairs with $|u_1| + |v_1| = |u_2| + |v_2| = n$. Then, for all $p_1 \in \mathbb{Q}$ with $p_1 > n - |v_1|$ and $p_2 \in \mathbb{Q}$ with $p_2 > n - |v_2|$, $u_1v_1^{p_1} \not\sim u_2v_2^{p_2}$.*

Proof. Let $p_1 = \frac{r_1}{|v_1|}$. Let $z = v[|v| - r_1 :]$. Then $u_1v_1^{p_1}z \in L_n$ but $u_2v_2^{p_2}z \notin L_n$. □

Lemma 3.26. *Let n, L_n, and $\sim_{L_n}$ be as in Lemma 3.24. Let w_1 and w_2 be two words with $|w_1|, |w_2| > n$. Then $w_1 \sim_{L_n} w_2$.*

Proof. Let z be any word. If $D(w_1) > n$, then $D(w_1z) > n$ by Lemma 3.22, and if $D(w_1) < n$, $D(w_1z) \neq n$ by Lemma 3.23, and similarly for w_2. So $w_1 \sim_{L_n} w_2$ trivially. □

Lemma 3.27. *Let n, L_n, and $\sim_{L_n}$ be as in Lemma 3.24. Then $\sim_{L_n}$ partitions $\{w \in \Sigma^+ : |w| > n\}$ into finitely many equivalent classes.*

Proof. By Lemma 3.26, all words of length $> n$ with complexity not equal to n are in a single equivalence class. By Lemma 3.24 and Lemma 3.25, all words of length $> n$ with complexity equal to n are in one of the equivalence classes corresponding to eventually integral pairs. □

Theorem 3.28. *L_n is regular for all n.*

Proof. By Lemma 3.27, $\sim_{L_n}$ partitions $\Sigma^{n+} := \{w \in \Sigma^+ : |w| > n\}$ into finitely many equivalent classes. But since $\Sigma^* \setminus \Sigma^{n+}$ is finite, $\sim_{L_n}$ partitions Σ^* into finitely many equivalence classes. The result follows immediately by Lemma 3.19. □

3.3 Complexity via other finite-state machines

3.3.1 Finite-state gamblers

Here we investigate complexity via *finite-state gamblers*, as discussed in [12]. Informally, a finite-state gambler is a hidden Markov model on a binary alphabet that, instead of outputting elements from a finite set, outputs a sequence of rational numbers, representing its capital at some time t. This capital comes from a list of *betting functions* that output the capital at time $t + 1$, given the capital and state at time t. More precisely,

Definition 3.29 ([12])**.** A *k-account finite-state gambler* is a tuple

$$G = (Q, \delta, \overrightarrow{\beta}, q_0, \overrightarrow{c_0})$$

where

- Q is a nonempty, finite set of states
- $\delta : Q \times \{0,1\} \to Q$ is the transition function
- $\overrightarrow{\beta} : Q \to (\mathbb{Q} \cap [0,1])^k$, where exponentiation denotes Cartesian product, is the betting function
- $q_0 \in Q$ is the initial state
- $\overrightarrow{c_0}$ is the initial capital vector.

Here, it will only be necessary to consider k-account finite-state gamblers for $k = 1$; the term "finite-state gambler" will be understood to mean "1-account finite-state gambler." β and c_0 will be written for $\overrightarrow{\beta}$ and $\overrightarrow{c_0}$, respectively.

Definition 3.30 ([12])**.** Given a finite-state gambler G, the *martingale* of G is a function

$$d_G : \{0,1\}^* \to [0,\infty)$$

defined recursively by

$$\begin{aligned} d_G(\epsilon) &= c_0 \\ d_G(wb) &= 2d_G(w)\left[(1-b)(1-\beta(\delta(w)) + b\beta(\delta(w))\right] \end{aligned}$$

A significant difference between finite-state gamblers and hidden Markov models is that finite-state gamblers have continuous output. This can be bridged by focusing on a narrower class of finite-state gamblers defined below.

Definition 3.31. An *all-or-nothing finite-state gambler* is a finite-state gambler with betting function $\beta : Q \to \{0,1\}$.

The following definition and results demonstrate the link between finite-state gamblers and complexity.

Definition 3.32. Given a word w, define the *gambling complexity* of w by

$$\Gamma(w) := \min\{|Q| \mid d_G(w) > 0\}$$

where Q are the states of an all-or-nothing finite-state gambler G.

In the remainder of this section, all finite-state gamblers will be assumed to be all-or-nothing finite-state gamblers with $c_0 = 1$. (The choice of c_0 is arbitrary; any positive value will suffice.)

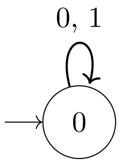

Figure 3.3: 0^n-accepting finite-state gambler

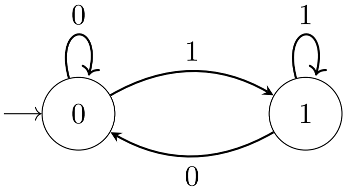

Figure 3.4: $(01)^n$-accepting finite-state gambler

Example 3.33. For every n, $\Gamma(0^n) = 1$.

Proof. Consider the finite-state gambler Figure 3.3 G defined by with $\beta(0) = 1$. That is, G always bets all of its capital that the next bit is 0. Immediately $d_G(0^n) = n$. □

Example 3.34. For every n, $\Gamma((01)^n) = 2$.

Proof. Consider the finite-state gambler Figure 3.4 G defined by with $\beta(0) = 1$ and $\beta(1) = 0$. That is, when in state 0, G bets all of its capital that the next bit is 1, and when in state 1, G bets all of its capital that the next bit is 0. Immediately, $d_G((01)^n) = 2n$, so $\Gamma((01)^n) \leq 2$.

To see that $d_G((01)^n)$ cannot be 1, note that any finite-state gambler with 1 state either always bets all of its capital that the next bit is 0, or always bets all of its capital that the next bit is 0. In either case, $d_G((01)^n) = 0$. □

Note that a finite-state gambler can be viewed as a deterministic hidden Markov model by considering its betting function as an emission function while keeping the same underlying states and transition function, and vice versa to view a deterministic hidden Markov model as a finite-state gambler. This immediately implies

Theorem 3.35. *For all w, $\Gamma(w) = D(w)$.*

3.3.2 Büchi automata

Definition 3.36. A *Büchi automaton* [7] is an automaton whose inputs are infinite words. An input is accepted if it passes through an accept state infinitely many times.

Example 3.37. The Büchi automaton Figure 3.5 accepts $(01)^\omega$ but not 0^ω.

Definition 3.38. For X in Σ^ω, the Büchi complexity of X is the minimum number of states in a Büchi automaton B such that X is the only word in A^ω that B accepts.

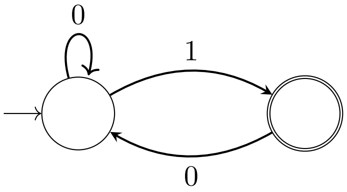

Figure 3.5: Büchi automaton accepting $(01)^\omega$

Theorem 3.39. *For every n, there is a word with Büchi complexity n.*

Proof. Let w be a primitive words of length n. Then w^ω has Büchi complexity n. □

Theorem 3.40. *There is a word with infinite Büchi complexity.*

Proof. Take any word that does not end in the infinite power of a single word, such as the Thue-Morse word [1]. □

The above theorem is also a consequence of the following result.

Theorem 3.41. *There are only countably many words with finite Büchi complexity.*

Proof. A word with finite Büchi complexity is by definition composed of a finite-length initial segment and a finite-length ω-base, and there are countably many such combinations. □

Theorem 3.42. *For each n, the average number of words with Büchi complexity $\leq n$ is $\frac{\sum_{i=1}^{n} \mathcal{C}(i)}{2^n}$.*

Proof. By Theorem 2.49, these are the number of words with eventually integral complexity, and by Theorem 2.54, Büchi complexity supersedes HMM complexity for all sufficiently long words. □

CHAPTER 4
CONSTRUCTIVE IMMUNITY

4.1 Introduction

Effectively immune sets, introduced by Smullyan in 1964 [30], are well-known in computability as one of the incarnations of diagonal non-computability, first made famous by Arslanov's completeness criterion. Let ω denote the natural numbers. A set $A \subseteq \omega$ is *effectively immune* if there is a computable function h such that $|W_e| \leq h(e)$ whenever $W_e \subseteq A$, where $\{W_e\}_{e\in\omega}$ is a standard enumeration of the computably enumerable (c.e.) sets.

There is a more obvious effectivization of immunity (the lack of infinite computable subsets), however: *constructive immunity*, introduced by Xiang Li [23] who actually (and inconveniently) called it "effective immunity".

Definition 4.1. A set A is *constructively immune* if there exists a partial recursive ψ such that for all x, if W_x is infinite then $\psi(x)\downarrow$ and $\psi(x) \in W_x \setminus A$.

The Turing degrees of constructively immune sets and the related Σ^0_1-dense sets have not been considered before in the literature, except that Xiang Li implicitly showed that they include all c.e. degrees. We prove in Section 4.3 that the Turing degrees of Σ^0_1-dense sets include all non-Δ^0_2 degrees, all high degrees, and all c.e. degrees. We do not know whether they include *all* Turing degrees.

The history of the study of constructive immunity seems to be easily summarized. After Xiang Li's 1983 paper, Odifreddi's 1989 textbook [24] included Li's results as exercises, and Calude's 1994 monograph [8] showed that the set $RAND^C_t = \{x : C(x) \geq |x| - t\}$ is constructively immune, where C is Kolmogorov complexity. Schafer 1997 [27] further developed an example involving minimal indices, and Brattka 2002 [5] gave one example in a more general setting than Cantor space. Finally in 2008 Ferbus-Zanda and Grigorieff proved an equivalence with constructive Σ^0_1-density.

Definition 4.2 (Ferbus-Zanda and Grigorieff [16])**.** A set $A \subseteq \omega$ is Σ^0_1-*dense* if for every infinite c.e. set C, there exists an infinite c.e. set D such that $D \subseteq C$ and $D \subseteq A$.

If there is a computable function $f : \omega \to \omega$ such that for each W_e, $W_{f(e)} \subseteq A \cap W_e$, and $W_{f(e)}$ is infinite if W_e is infinite, then A is *constructively* Σ^0_1-*dense*.

We should note that while the various flavors of immune sets are always infinite by definition, Ferbus-Zanda and Grigorieff do not require Σ^0_1-dense sets to be co-infinite.

The Σ^0_1-dense sets form a natural Π^0_4 class in 2^ω that coincides with the simple sets on Δ^0_2 but is prevalent (in fact exists in every Turing degree) outside of Δ^0_2 by Theorem 4.25 below.

4.2 Σ^0_1-density

To show that there exists a set that is Σ^0_1-dense, but not constructively so, we use Mathias forcing. A detailed treatment of the computability theory of Mathias forcing can be found in [9].

Definition 4.3. A *Mathias condition* is a pair (d, E) where $d, E \subseteq \omega$, d is a finite set, E is an infinite computable set, and $\max(d) < \min(E)$. A condition (d_2, E_2) *extends* a condition (d_1, E_1) if

- $d_1 = d_2 \cap (\max d_1 + 1)$, i.e., d_1 is an initial segment of d_2,
- E_2 is a subset of E_1, and
- d_2 is contained in $d_1 \cup E_1$.

A set A is *Mathias generic* if it is generic for Mathias forcing.

Theorem 4.4. *If A is Mathias generic, then*

1. $\omega \setminus A$ is Σ^0_1-dense.

2. $\omega \setminus A$ is not constructively Σ^0_1-dense.

Proof. 1. Let W_e be an infinite c.e. set. Let (d, E) be a Mathias condition.

Case (i): $E \cap W_e$ is finite. Then for any Mathias generic A extending the condition (d, E), $\omega \setminus A$ contains an infinite subset of W_e, in fact a set of the form $W_e \setminus F$ where F is finite.

Case (ii): $E \cap W_e$ is infinite. Then $E \cap W_e$ is c.e., hence has an infinite computable subset D. Write $D = D_1 \cup D_2$ where D_1, D_2 are disjoint infinite c.e. sets. The condition (d, D_1) extends (d, E) and forces a Mathias generic A extending it to be such that $\omega \setminus A$ has an infinite subset in common with W_e, namely D_2.

We have shown that for each infinite c.e. set W_e, each Mathias condition has an extension forcing the statement that a Matias generic A satisfies

$$\omega \setminus A \text{ has an infinite c.e. subset in common with } W_e. \tag{*}$$

Thus by standard forcing theory it follows that each Mathias generic satisfies $(*)$.
2. Let f be a computable function. It suffices to show that for each Mathias generic A, there exists an i such that W_i is infinite and $W_{f(i)}$ is either finite, or not a subset of W_i, or not a subset of $\overline{A}$. For this, as in (1) above it suffices to show that for each condition (d, D) there exists a condition (d', E') extending (d, E) and an i such that W_i is infinite and $W_{f(i)}$ is either finite, or not a subset of W_i, or not a subset of $\overline{A}$ for any A extending (d', D').

Let (d, E) be a Mathias condition and write $D = W_i$. If $W_{f(i)}$ is finite or not a subset of W_i then we are done. Otherwise there exists a condition (d', E') extending (d, E) such that $E' \cap W_{f(i)}$ is nonempty. This can be done by a finite extension (making only finitely many changes to the condition). □

Theorem 4.5 ([16, Proposition 3.3])**.** *A set $Z \subseteq \omega$ is constructively immune if and only if it is infinite and $\omega \setminus Z$ is constructively Σ^0_1-dense.*

Since Ferbus-Zanda and Grigorieff's paper has not gone through peer review, we provide the proof.

Proof. $\Leftarrow$: Let the function g witness that $\omega \backslash Z$ is constructively Σ^0_1-dense. Define a partial recursive function φ by stipulating that $\varphi(i)$ is the first number in the enumeration of $W_{g(i)}$, if any.

$\Rightarrow$: Define a partial recursive function $\mu(i,n)$ by

- $\mu(i,0) = \varphi(i)$;
- $\mu(i,n+1) = \varphi(i_n)$, where i_n is such that $W_{i_n} = W_i \setminus \{\mu(i,m) : m \leq n\}$.

Let g be total recursive so that $W_{g(i)} = \{\mu(i,m) : m \in \omega\}$. If W_i is infinite then all $\mu(i,m)$'s are defined and distinct and belong to $W_i \cap Z$. Thus, $W_{g(i)}$ is an infinite subset of $W_i \cap Z$. □

Recall that a c.e. set is *simple* if it is co-immune.

Theorem 4.6 (Xiang Li [23])**.** *Let A be a set and let $\{\phi_x\}_{x\in\omega}$ be a standard enumeration of the partial computable functions.*

1. *If A is constructively immune then A is immune and $\overline{A}$ is not immune.*
2. *If A is simple then $\overline{A}$ is constructively immune.*
3. *$\{x : (\forall y)(\phi_x = \phi_y \to x \leq y)\}$ is constructively immune.*

4.2.1 Numberings

A *numbering* of a countable set $\mathcal{A}$ is an onto function $\nu : \omega \to \mathcal{A}$. The theory of numberings has a long history [15]. Numberings of the set of rational numbers $\mathbb{Q}$ provide an application area for Σ^0_1-density. Rosenstein [26, Section 16.2: Looking at $\mathbb{Q}$ effectively] discusses computable dense subsets of $\mathbb{Q}$. Here we are mainly concerned with noncomputable sets.

Proposition 4.7. *Let $A \subseteq \omega$. The following are equivalent:*

1. *$\nu(A)$ is dense for every injective computable numbering ν of $\mathbb{Q}$;*
2. *A is co-immune.*

Proof. (1) $\implies$ (2): We prove the contrapositive. Suppose $\overline{A}$ contains an infinite c.e. set W_e. Consider a computable numbering ν that maps W_e onto $[0,1] \cap \mathbb{Q}$. Then $\nu(A)$ is disjoint from $[0,1]$ and hence not dense.

(2) $\implies$ (1): We again prove the contrapositive. Assume that $\nu(A)$ is not dense for a certain computable ν. Let $\{x_n : n \in \omega\}$ be a converging infinite sequence of rationals disjoint from $\nu(A)$. Then $\{\nu^{-1}(x_n) : n \in \omega\}$ is an infinite c.e. subset of $\overline{A}$. □

Definition 4.8. A subset A of $\mathbb{Q}$ is *co-nowhere dense* if for each interval $[a, b] \subseteq \mathbb{Q}$, $[a', b'] \subseteq A$ for some $[a', b'] \subseteq [a, b]$.

Proposition 4.9. *A set is co-nowhere dense under every numbering if and only if it is co-finite.*

Proof. Only the forward direction needs to be proven; the other direction is immediate. Let A be a co-infinite set, and define ν by letting ν map $\omega \setminus A$ onto $[0, 1]$. Then A is not co-nowhere dense. □

Proposition 4.10. *A is infinite and non-immune if and only if there exists a computable numbering with respect to which A is co-nowhere dense.*

Proof. Let A be infinite and not immune. Thus, there is an infinite $W_e \subseteq A$ for some e. Let ν be a computable numbering that maps W_e onto $\mathbb{Q} \setminus \omega$. Then A is co-nowhere dense under ν.

Conversely, let A be co-nowhere dense under some computable numbering ν. Then $\nu^{-1}([a, b])$ is an infinite c.e. subset of A for some suitable a, b. □

A set $D \subseteq \mathbb{Q}$ is *effectively dense* if there is a computable function $f(a, b)$ giving an element of $D \cap (a, b)$ for $a < b \in \mathbb{Q}$.

Proposition 4.11. *A set A is constructively Σ^0_1-dense if and only if it is effectively dense for all computable numberings.*

Proof. By Theorem 4.5, A is constructively Σ^0_1-dense if and only if it is infinite and $\omega \setminus A$ is constructively immune. Constructive immunity of $\omega \setminus A$ implies effective density of A since the witnessing function for constructive immunity can be be used to witness effective density. For the converse we exploit the assumption that we get to choose a suitable ν. □

Let A and B be sets, with B computable. We say that A is *co-immune within B* if there is no infinite computable subset of $A^c \cap B$. The following diagram includes some claims not proved in the paper, whose proof (or disproof) may be considered enjoyable exercises. The quantifiers $\exists\nu$, $\forall\nu$ range over computable numberings of $\mathbb{Q}$.

4.3 Prevalence of Σ^0_1-density

In this section we investigate the existence of Σ^0_1-density in the Turing degrees at large.

4.3.1 Closure properties and Σ^0_1-density

Proposition 4.12. *1. The intersection of two Σ^0_1-dense sets is Σ^0_1-dense.*

2. The intersection of two constructively Σ^0_1-dense sets is constructively Σ^0_1-dense.

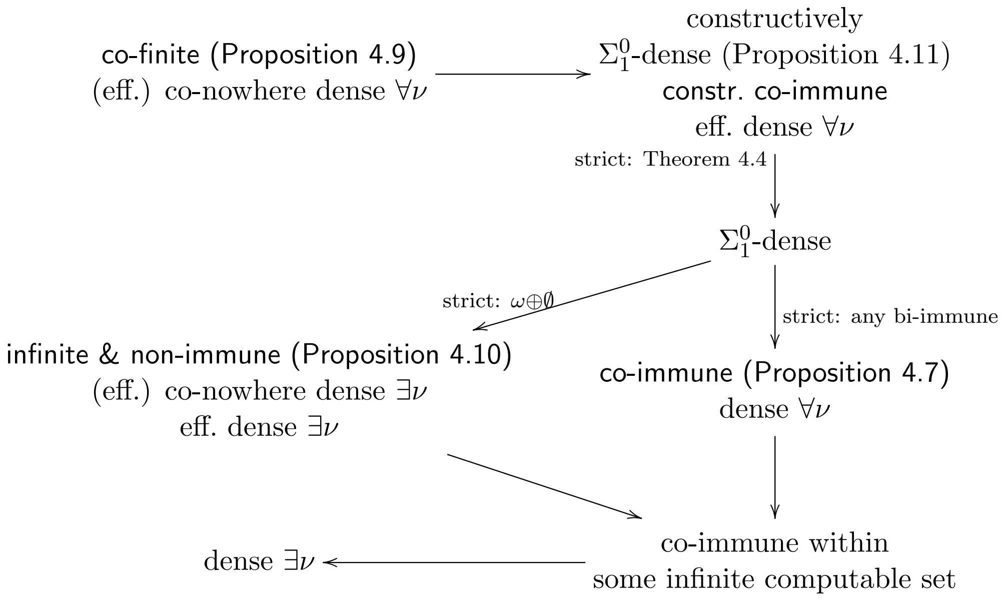
co-finite (Proposition 4.9)
(eff.) co-nowhere dense $\forall\nu$
constructively
Σ^0_1-dense (Proposition 4.11)
constr. co-immune
eff. dense $\forall\nu$
strict: Theorem 4.4
Σ^0_1-dense
strict: $\omega\oplus\emptyset$
strict: any bi-immune
infinite & non-immune (Proposition 4.10)
(eff.) co-nowhere dense $\exists\nu$
eff. dense $\exists\nu$
co-immune (Proposition 4.7)
dense $\forall\nu$
co-immune within
some infinite computable set
dense $\exists\nu$

Figure 4.1: Co-immunity results

Proof. Let A and B be Σ^0_1-dense sets. Let W_e be an infinite c.e. set. Since A is Σ^0_1-dense, there exists an infinite c.e. set $W_d \subseteq A \cap W_e$. Since B is Σ^0_1-dense, there exists an infinite c.e. set $W_a \subseteq B \cap W_d$. Then $W_a \subseteq (A \cap B) \cap W_e$, as desired. This proves (1). To prove (2), let f and g witness the effective Σ^0_1-density of A and B, respectively. Given W_e, we have $W_{f(e)} \subseteq A \cap W_e$ and then

$$W_{g(f(e))} \subseteq B \cap W_{f(e)} \subseteq A \cap B \cap W_e.$$

In other words, $g \circ f$ witnesses the effective Σ^0_1-density of $A \cap B$. □

Corollary 4.13. Bi-Σ^0_1-dense sets do not exist.

Proof. If A and A^c are both Σ^0_1-dense then by Proposition 4.12, $A \cap A^c$ is Σ^0_1-dense, which is a contradiction. □

For sets A and B, $A \subseteq^* B$ means that $A \setminus B$ is a finite set.

Proposition 4.14. *1. If A is Σ^0_1-dense and $A \subseteq^* B$, then B is Σ^0_1-dense.*

2. If A is constructively Σ^0_1-dense and $A \subseteq^* B$, then B is constructively Σ^0_1-dense.

Proof. Let W_e be an infinite c.e. set. Since A is Σ^0_1-dense, there exists an infinite c.e. set W_d such that $W_d \subseteq A \cap W_e$. Let $W_c = W_d \setminus (A \setminus B)$. Since $A \setminus B$ is finite, W_c is an infinite c.e. set. Since $W_d \subseteq A$, we have $W_c = W_d \cap (B \cup A^c) = W_d \cap B$. Then, since $W_d \subseteq W_e$, we have

$W_c \subseteq B \cap W_e$, and we conclude that B is Σ^0_1-dense. This proves (1). To prove (2), if f witnesses that A is constructively Σ^0_1-dense then a function g with $W_{g(e)} = W_{f(e)} \setminus (A \setminus B)$ witnesses that B is constructively Σ^0_1-dense. □

Proposition 4.15. *Let B be a co-finite set. Then B is constructively Σ^0_1-dense.*

Proof. The set ω is constructively Σ^0_1-dense as witnessed by the identity function $f(e) = e$. Thus by Item 2 of Proposition 4.14, B is as well. □

As usual we write $A \oplus B = \{2x \mid x \in A\} \cup \{2x + 1 \mid x \in B\}$.

Proposition 4.16. *1. If X_0 and X_1 are Σ^0_1-dense sets then so is $X_0 \oplus X_1$.*

2. If X_0 and X_1 are constructively Σ^0_1-dense sets then so is $X_0 \oplus X_1$.

Proof. Let $W_e = W_{c_0} \oplus W_{c_1}$ be an infinite c.e. set. For $i = 0, 1$, since X_i is Σ^0_1-dense there exists $W_{d_i} \subseteq X_i \cap W_{c_i}$ such that W_{d_i} is infinite if W_{c_i} is infinite. Then $W_{d_0} \oplus W_{d_1}$ is an infinite c.e. subset of $(X_0 \oplus X_1) \cap W_e$.

This proves (1). To prove (2), if d_i are now functions witnessing the effective Σ^0_1-density of X_i then $W_{d_i(c_i)} \subseteq X_i \cap W_{c_i}$, and $W_{d_0(c_0)} \oplus W_{d_1(c_1)}$ is an infinite c.e. subset of $(X_0 \oplus X_1) \cap W_e$. Thus a function g satisfying

$$W_{g(e)} = W_{d_0(c_0)} \oplus W_{d_1(c_1)},$$

where $W_e = W_{c_0} \oplus W_{c_1}$, witnesses the effective Σ^0_1-density of $X_0 \oplus X_1$. □

Theorem 4.17. *There is no Σ^0_1-dense set A such that all Σ^0_1-dense sets B satisfy $A \subseteq^* B$.*

Proof. Suppose there is such a set A. Let W_d be an infinite computable subset of A. Let G be a Mathias generic with $G \cap W^c_d = \emptyset$, i.e., $G \subseteq W_d$. Then $B := G^c$ is Σ^0_1-dense by Theorem 4.4. Thus $A \cap G^c$ is also Σ^0_1-dense by Proposition 4.12. And $G \subseteq W_d \subseteq A$ and by assumption $A \subseteq^* G^c$ so we get $G \subseteq^* G^c$, a contradiction. □

These results show that the Σ^0_1-dense sets under $\subseteq^*$ form a non-principal filter whose Turing degrees form a join semi-lattice.

Theorem 4.18. *Let A be a c.e. set. The following are equivalent:*

1. *A is co-infinite and constructively Σ^0_1-dense.*

2. *A is co-infinite and Σ^0_1-dense.*

3. *A is co-immune.*

Proof. $1 \implies 2 \implies 3$ is immediate from the definitions, and $3 \implies 1$ is immediate from Theorem 4.5 and Theorem 4.6. □

Theorem 4.19. *Every c.e. Turing degree contains a constructively Σ^0_1-dense set.*

Proof. Let $\mathbf{a}$ be a c.e. degree. If $\mathbf{a} > \mathbf{0}$ then $\mathbf{a}$ contains a simple set A, see, e.g., [31], so Theorem 4.18 finishes this case. The degree $\mathbf{0}$ contains all the co-finite sets, which are constructively Σ^0_1-dense by Proposition 4.15. □

4.3.2 Cofinality in the Turing degrees of constructive Σ^0_1-density

Definition 4.20. For $k \geq 0$, let I_k be intervals of length $k+2$ such that $\min(I_0) = 0$ and $\max(I_k) + 1 = \min(I_{k+1})$.

Let V_e be a subset of W_e defined by the following condition. We let x enter V_e at a stage where x enters W_e if doing so is permitted by the rule: Let k be such that $x \in I_k$. Then for all $j \leq k$, $|V_j \cap I_k| \leq 1$; and for all $j > k$, $V_j \cap I_k = \emptyset$.

Lemma 4.21. *There exists a c.e., co-infinite, constructively Σ^0_1-dense, and effectively co-immune set.*

Proof. Let $A = \bigcup_{e\in\omega} V_e$. V_e is c.e. by construction, and if W_e is infinite, V_e is also infinite. So $V_e = W_{f(e)}$ is the set witnessing that A is constructively Σ^0_1-dense.

Moreover A is coinfinite since $|A \cap I_k| \leq k+1 < k+2 = |I_k|$ gives $I_k \not\subseteq A$ for each k and

$$|\omega \setminus A| = \left|\left(\bigcup_{k\in\omega} I_k\right) \setminus A\right| = \left|\bigcup_{k\in\omega} (I_k \setminus A)\right| = \sum_{k\in\omega} |I_k \setminus A| \geq \sum_{k\in\omega} 1 = \infty.$$

The set A is effectively co-immune because if W_e is disjoint from A then since as soon as a number in I_k for $k \geq e$ enters W_e then that number is put into A, $W_e \subseteq \bigcup_{k<e} I_k$ so $|W_e| \leq \sum_{k<e}(k+2) = \sum_{k\leq e+1} k = \frac{(e+1)(e+2)}{2}$. □

Theorem 4.22. *For each set R there exists a constructively Σ^0_1-dense, effectively co-immune set S with $R \leq_T S$.*

Proof. Let R be any set, which we may assume is co-infinite. Let A be as in the proof of Lemma 4.21. Let $S \supseteq A$ be defined by

$$S = A \cup \bigcup_{k\in R} I_k.$$

Since $A \subseteq S$ and S is co-infinite, S is constructively Σ^0_1-dense and effectively co-immune. Since $k \in R \iff I_k \subseteq S$, we have $R \leq_T S$. □

4.3.3 Non-Δ^0_2 degrees

Lemma 4.23. *Suppose that $T \subseteq 2^{<\omega}$ is a tree with only one infinite path. Then for each length n there exists a length $m > n$ such that exactly one string of length n has an extension of length m in T.*

Proof. Suppose not, i.e., there is a length n such that for all $m > n$ there are at least two strings σ_m, τ_m of length n with extensions of length m in T. By the pigeonhole principle there is a pair (σ, τ) that is a choice of (σ_m, τ_m) for infinitely many m. Then by compactness both σ and τ must be extendible to infinite paths of T. □

Lemma 4.24. *Suppose that $T \subseteq 2^{<\omega}$ is a tree with only one infinite path A, and that T is a c.e. set of strings. Then A is Δ^0_2.*

Proof. By Lemma 4.23, for each length n there exists a length $m > n$ such that exactly one string of length n has an extension of length m in T. Using $0'$ as an oracle we can find that m and define $A \upharpoonright n$ by looking for such a string. In fact, $T \leq_T 0'$ and so its unique path $A \leq_T 0'$ as well. □

Theorem 4.25. *Given $A \in 2^\omega$, let $\widehat{A} := \{\sigma \in 2^{<\omega} \mid \sigma \prec A\}$ be the set of finite prefixes of A. If A is not Δ^0_2 then $\widehat{A}$ is co-Σ^0_1-dense.*

Proof. Let A^* be the complement of $\widehat{A}$. Let $W_e \subseteq 2^{<\omega}$ be an infinite c.e. set of strings. Let T be the set of all prefixes of elements of W_e. Then T is an infinite tree, hence by compactness it has at least one infinite path. That is, there is at least one real B such that all its prefixes are in T.

Case 1: The only such real is $B = A$. Then by Lemma 4.24, A is Δ^0_2.

Case 2: There is a $B \neq A$ such that all its prefixes are in T. Let σ be a prefix of B that is not a prefix of A. Let $W_d = [\sigma] \cap W_e$. Since all prefixes of B are prefixes of elements of W_e, there are infinitely many extensions of σ that are prefixes of elements of W_e. Consequently W_d is infinite. Thus, W_d is our desired infinite subset of $A^* \cap W_e$. □

4.3.4 High degrees

Definition 4.26. A set A is *co-r-cohesive* if its complement is r-cohesive. This means that for each computable (recursive) set W_d, either $W_d \subseteq^* A$ or $W_d^c \subseteq^* A$.

Definition 4.27 (Odifreddi [24, Exercise III.4.8], Jockusch and Stephan [18])**.** A set A is strongly hyperhyperimmune (s.h.h.i.) if for each computable $f : \omega \to \omega$ for which the sets $W_{f(e)}$ are disjoint, there is an e with $W_{f(e)} \subseteq \omega \setminus A$.

A set A is strongly hyperimmune (s.h.i.) if for each computable $f : \omega \to \omega$ for which the sets $W_{f(e)}$ are disjoint and computable, with $\bigcup_{e\in\omega} W_{f(e)}$ also computable, there is an e with $W_{f(e)} \subseteq \omega \setminus A$.

Proposition 4.28. *Every s.h.i. set is co-Σ^0_1-dense.*

Proof. Let A be s.h.i. Let W_e be an infinite c.e. set. Let W_d be an infinite computable subset of W_e. Effectively decompose W_d into infinitely many disjoint infinite computable sets,

$$W_d = \bigcup_{i\in\omega} W_{g(d,i)}.$$

For instance, if $W_d = \{a_0 < a_1 < \dots\}$ then we may let $W_{g(e,i)} = \{a_n : n = 2^i(2k+1), i \geq 0, k \geq 0\}$. Since A is s.h.i., there exists some i_e such that $W_{g(d,i_e)} \subseteq A^c$. The sets $W_{g(d,i_e)}$ witness that A^c is Σ^0_1-dense. □

Clearly r-cohesive implies s.h.i., and s.h.h.i. implies s.h.i. It was shown by Jockusch and Stephan [18, Corollary 2.4] that the cohesive degrees coincide with the r–cohesive degrees and (Corollary 3.10) that the s.h.i. and s.h.h.i. degrees coincide.

Proposition 4.29. *Every high degree contains a Σ^0_1-dense set.*

Proof. Let $\mathbf{h}$ be a Turing degree. If $\mathbf{h} \not\leq \mathbf{0}'$, then $\mathbf{h}$ contains a Σ^0_1-dense set by Theorem Theorem 4.25.

If $\mathbf{h} \leq \mathbf{0}'$ and $\mathbf{h}$ is high then since the strongly hyperhyperimmune and cohesive degrees coincide, and are exactly the high degrees [11], $\mathbf{h}$ contains a strongly hyperimmune set. Hence by Theorem Proposition 4.28, $\mathbf{h}$ contains a Σ^0_1-dense set. □

4.3.5 Progressive approximations

Definition 4.30. Let A be a Δ^0_2 set. A computable approximation $\{\sigma_t\}_{t\in\omega}$ of A, where each σ_t is a finite string and $\lim_{t\to\infty} \sigma_t = A$, is *progressive* if for each t,

- if $|\sigma_t| \leq |\sigma_{t-1}|$ then $\sigma_t \upharpoonright (|\sigma_t| - 1) = \sigma_{t-1} \upharpoonright (|\sigma_t| - 1)$ (the last bit of σ_t is the only difference with σ_{t-1});
- if $|\sigma_t| > |\sigma_{t-1}|$ then $\sigma_{t-1} \prec \sigma_t$; and
- if $\sigma_t \not\prec \sigma_s$ for some $s > t$ then $\sigma_t \not\prec \sigma_{s'}$ for all $s' \geq s$ (once an approximation looks wrong, it never looks right again).

If A has a progressive approximation then we say that A is *progressively approximable.*

Note that a progressively approximable set must be h-c.e. where $h(n) = 2^n$.

Theorem 4.31. *Let A be a progressively approximable and noncomputable set. Let $\{\sigma_t\}_{t\in\omega}$ be a progressive approximation of A. Then $\{t : \sigma_t \prec A\}$ is constructively immune.*

Proof. Let W_e be an infinite c.e. set and let T be an infinite computable subset of W_e. Since A is noncomputable, we do not have $T \subseteq \{t : \sigma_t \prec A\}$. Since the approximation $\{\sigma_t\}_{t\in\omega}$ is progressive, once we observe a t for which $\sigma_t \not\prec \sigma_s$, for some $s > t$, then we know that $\sigma_t \not\prec A$. Then we define $\varphi(e) = t$, and φ witnesses that $\{t : \sigma_t \prec A\}$ is constructively immune. □

A direction for future work may be to find new Turing degrees of progressively approximable sets.

CHAPTER 5
FURTHER RESULTS ON FINITE WORDS AND EFFECTIVENESS

5.1 Density for finite complexity

Connections between the notions of density described in Chapter 4 and the notions of complexity investigated in Chapters 2 and 3 are presented here.

Definition 5.1. $D \subseteq 2^{<\omega}$ is *dense* if for all $x \in 2^{<\omega}$, there exists $y \in D$ such that $x \preceq y$.

The "regular" notion of density defined above can be viewed as a computable analogue of Σ^0_1 density. Specifically, both notions of density make precise the notion of a class having a "typical" property. This section will connect the results of Chapter 2 to the density arguments presented in this chapter by demonstrating the densities and non-densities of various sets defined by deterministic HMM complexity. Recall that if $w \in 2^{<\omega}$, $D(w)$ is the deterministic HMM complexity of w, as defined in Definition 2.21.

Theorem 5.2. *For any fixed n, the set $\{w : D(w) = n\}$ is not dense.*

Proof. Let x be a word such that $D(x) = n + 1$. By Lemma 3.22, for any word y, if $x \preceq y$ then $D(y) \geq n + 1$, so $y \notin \{w : D(w) = n\}$. □

The next results are concerned with the density of high-complexity words. Although the set of words of a fixed complexity is not dense, as shown in Theorem 5.2, and the set of words of a given complexity fixed relative to the length of the word will be shown to not be dense in Theorem 5.4, Theorem 5.7 shows that taking a union over these sets results in a dense set.

Definition 5.3. For every $c \in \omega$, let $\mathcal{H}_c := \{w : D(w) = |w| - c\}$.

The choice of $\mathcal{H}$ in the notation of $\mathcal{H}_c$ was motivated by the following intuition: for a fixed c and words w with $|w|$ growing arbitrarily large, $\mathcal{H}_c$ can be thought of as the set of words of high complexity, up to some constant c.

Theorem 5.4. *For every fixed c, $\mathcal{H}_c$ is not dense.*

Proof. Let x be a word such that for every word u of length $c + 1$, u is a subword of x. Then, for any word y with $x \preceq y$, $D(y) \leq |y| - (c + 1)$, as the last $c + 1$ bits of y appear in y. □

Definition 5.5. Let $\mathcal{H}_{\frac{1}{2}} := \left\{w \in 2^{<\omega} : D(w) \leq \frac{|w|}{2}\right\}$.

In the following lemma and theorem, recall the definition of a *decomposition* as defined in Definition 2.29.

Lemma 5.6. *Let w be a word such that $D(w) < \frac{|w|}{2}$ and let (u,v) be a decomposition of w, with $w = uv^p$. Then $p > 1$.*

Proof. p will be minimized when $|v|$ is maximized, so without loss of generality assume that $u = \epsilon$. Then

$$|v| = D(w) < \tfrac{|w|}{2} = \tfrac{|v^p|}{2}$$

so $2|v| < |v^p|$ which immediately implies that $p > 1$. □

Theorem 5.7. *$\mathcal{H}_{\frac{1}{2}}$ is dense.*

Proof. Let $x \in 2^{<\omega}$. If $x \in \mathcal{H}_{\frac{1}{2}}$ then the proof is finished, so assume not. Let $x = uv^p$ for $p \in \mathbb{Q}$ and a decomposition (u,v) of x. By Lemma 5.6, $p > 1$. Let $n \in \omega$ be large enough so that $n \geq p$ and $(n+1)|v| > |u|$, and let v' be a word such that $|v'| = v$ and v' is not a cyclic shift of v. Consider $y = uv^nv'$ and let (w,z) be a decomposition of y with $y = wz^q$ for $q \in \mathbb{Q}$. If $q \leq 1$, then $y \in \mathcal{H}_{\frac{1}{2}}$ by applying the contrapositive of Lemma 5.6, so assume $q > 1$. Cancelling v' from both sides yields $uv^n = wz^{q'}$ for some $q' \in \mathbb{Q}$. If $q' > 1$, then by Lemma 2.34, there exists words u' and v' and $p' \in \mathbb{Q}$ such that $uv^p = u'v'^{p'}$, contradicting the assumption that (u,v) is a decomposition; so $q' \leq 1$. Since v' does not occur in v^n, there exists some k such that $z \restriction k = u[-k:]$. In particular, v^nv' is a subword of z, so

$$D(y) = |w| + |z| > (n+1)|v| > \tfrac{(n+1)|v|}{2} + \tfrac{|u|}{2} = \tfrac{|y|}{2}$$

completing the proof. □

Remark 5.8. Note that the bound $D(w) \leq \frac{|w|}{2}$ in Theorem 5.7 is equivalent to the bound in the hypothesis of Theorem 2.86. Theorem 5.7 uses this assumption via Lemma 5.6, while Theorem 2.86 uses this assumption indirectly in the proof of Theorem 2.54. Since the bound in Theorem 2.86 is sharp, it is conjectured that the bound in Theorem 5.7 is also sharp.

Definition 5.9. For every n and k let $\mathcal{L}_{n,k} = \{w : nD(w) + k \leq |w|\}$.

In contrast to $\mathcal{H}_c$, $\mathcal{L}_{n,k}$ can be thought of as a class of words with low complexity.

Remark 5.10. When $n = 2$ and $k = 1$, $\mathcal{L}_{n,k}$ is exactly the set considered in Theorem 2.86.

Theorem 5.11. *For any fixed n and k, the set $\mathcal{L}_{n,k}$ is dense.*

Proof. Fix $x \in 2^{<\omega}$ and let m be such that $n + \frac{k}{|x|} \leq m$. Then

$$nD(x^m) + k \leq n|x| + k \leq m|x| = |x^m|$$

so $x^m \in \mathcal{L}_{n,k}$.

□

5.2 Fibonacci words

Definition 5.12. Define the *Fibonacci words* $\mathcal{F}_n$ recursively by

$$\mathcal{F}_0 = \epsilon,\ \mathcal{F}_1 = 1,\ \mathcal{F}_2 = 0,\ \mathcal{F}_n = \mathcal{F}_{n-1}\mathcal{F}_{n-2} \text{ for all } n \geq 3$$

Note that for all $n \geq 3$, $\mathcal{F}_n \preceq \mathcal{F}_{n+1}$. This ensures that the following definition is well-defined.

Definition 5.13. The *infinite Fibonacci word* $\mathcal{F}$ is the word formed by taking the limit of the (finite) Fibonacci words. That is, $\mathcal{F} = \lim_{n\to\infty} \mathcal{F}_n$.

The main result of this section will be a derivation of the deterministic HMM complexity of Fibonacci words. Let F_n denote the classic Fibonacci sequence, modified so that $F_0 = 1$ and $F_1 = 2$.

Theorem 5.14 ([6])**.** *For all $n \in \omega$, there exists a unique sequence $\{r_i\}$ such that $n = \sum_{i=0}^{\infty} r_i F_i$, $r_i \in \{0, 1\}$, and for all i, at least one of r_i and r_{i+1} is zero.*

Definition 5.15. Given $n \in \omega$, the *Zeckendorf representation of n* n^* is the unique sequence $\{r_i\}$ guaranteed by Theorem 5.14.

Definition 5.16. $\mathcal{F}^n$ is the infinite word formed by removing the first n bits of $\mathcal{F}$. That is, $\mathcal{F}^n := \mathcal{F}[n :]$.

Lemma 5.17 ([10] Theorem 3.2)**.** [1] *Let n_1 and n_2 be natural numbers with $n_1 \neq n_2$. Let $n_1^* = r_0 r_1 \dots$ and let $n_2^* = s_0 s_1 \dots$. If it exists, let p be such that $r_i = s_i$ for all $i \leq p$ and $r_{p+1} \neq s_{p+1}$. Then, the length of the longest common prefix of $\mathcal{F}^{n_1}$ and $\mathcal{F}^{n_2}$ is $\sum_{i=0}^{p} (1 - r_i) F_i$.*

Lemma 5.18. *For all $n \geq 1$, $(F_n - 1)^* = \begin{cases} (10)^{\frac{n+1}{2}} 0^\omega & n \text{ odd} \\ (01)^{\frac{n}{2}} 0^\omega & n \text{ even} \end{cases}$*

Proof. It is a classic result that $\sum_{i=0}^{n} F_{2i-1} = F_{2n} - 1$ and that $\sum_{i=0}^{n} F_{2i} = F_{2n+1} - 1$; in particular, this can be seen by observing the telescoping series $F_{2i-1} = F_{2i} - F_{2(i-1)}$. The proof then follows immediately from the definition of Zeckendorf representations.

□

Lemma 5.19. *For all $n \geq 2$, $\mathcal{F}_{n-2}$ is a subword of $\mathcal{F}_n$.*

Proof. $\mathcal{F}_{n+1} = \mathcal{F}_n \mathcal{F}_{n-1} = (\mathcal{F}_{n-1}\mathcal{F}_{n-2})\mathcal{F}_{n-1}$, so the result follows by induction. □

[1] In their paper, Chuan and Ho use 1-based indices, i.e. consider w_1 to be the first letter of the word $w_1 w_2 \dots$. The relevant indices in their result are adjusted so as to not conflict with the precedence of 0-based indices established in this thesis.

Theorem 5.20. *For all n, $D\left(\mathcal{F}_n\right) = F_{n-1}$.*

Proof. $D\left(\mathcal{F}_n\right) \leq F_{n-1}$ by Lemma 5.19, so it suffices to show that $\mathcal{F}_n[-(|\mathcal{F}_{n-1}|+1):]$ is not a subword of $\mathcal{F}_n[:|\mathcal{F}_{n-1}|-1]$. Let $u = \mathcal{F}_n[-(|\mathcal{F}_{n-1}|+1):]$ and let v be a subword of $\mathcal{F}_n[:|\mathcal{F}_{n-1}|-1]$. The goal of the proof will be to use Lemma 5.17 to show that the length of the longest common prefix of u and v is strictly less than the length of u. By Lemma 5.18, the nonzero initial segment of the Zeckendorf representation of $\mathcal{F}_n - 1$ is either $(01)^{\frac{n+1}{2}}$ if n is odd or $(10)^{\frac{n}{2}}$ if n is even. Assume n is even; the case where n is odd is proved analogously. The Zeckendorf representation of words that agree with $(10)^{\frac{n+1}{2}}$ in the greatest number of indices is $(10)^{\frac{n-1}{2}}$; assume without loss of generality that that this is the Zeckendorf representation of v. Then, the longest common prefix of v and $\mathcal{F} \restriction_{|\mathcal{F}_{n-1}+1|}$ has length $\displaystyle\sum_{i=0}^{\frac{n-1}{2}} F_{\frac{n+1}{2}} = F_{n-1} < |u|$. □

5.3 Pushdown automata and HMM complexity

Definition 5.21. A *pushdown automaton* is 7-tuple $M = (Q, \Sigma, \Gamma, \delta, q_0, Z, F)$ where

- Q is a finite set (the *states*)
- Σ is a finite set (the *input alphabet*)
- Γ is a finite set (the *stack alphabet*)
- $\delta : Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma \to Q \times \Gamma^*$ is a function (the *transition function*)
- $q_0 \in Q$ is the *start state*
- $Z \in \Gamma$ is the *initial stack symbol*
- $F \subseteq Q$ is the set of *accepting states*

Intuitively, a pushdown automaton is a DFA with a fixed, finite amount of memory (the *stack*). In contrast to a DFA, whose transition function only depends on the current state and the last symbol seen, a pushdown automaton transitions based on the current state, last symbol seen, and earlier symbols that were recorded on the stack.

Recall that Section 3.2.2 established the regularity of the language family

$$L_n = \{w : D(w) = n\}$$

The main result of this section will be the explicit construction of a family of pushdown automata that recognizes L_n for each n. The existence of such a family of pushdown automata is guaranteed, as pushdown automata can recognize a strictly larger class of languages than DFAs. However, while the DFAs implicitly constructed in Section 3.2.2 to recognize L_n grow exponentially in n, it will be shown that it is sufficient to increase the number of states and stack size linearly to create a family of pushdown automata that recognize L_n.

A pushdown automaton accepts a language analogously to how a DFA accepts a language. However, the added complexity of a stack requires an auxiliary definition to formally define what it means for a pushdown automaton to accept a language.

Definition 5.22. Let P be a pushdown automaton with states Q, input alphabet Σ, and stack Γ. A *configuration of* P is an element of $Q \times \Sigma^* \times \Gamma^*$.

Let $(q, aw, X\alpha)$ and $(p, w, \beta\alpha)$ be configurations of P, where $q \in Q$, $a \in \Sigma \cup \{\epsilon\}$, $w \in \Sigma^*$, $X \in \Gamma$, $\alpha, \beta \in \Gamma^*$. Write

$$(q, aw, X\alpha) \vdash (p, w, \beta\alpha)$$

if there exists a transition $(p, \beta) \in \delta(q, a, X)$. Let $\vdash^*$ be the reflexive and transitive closure of $\vdash^*$.

Definition 5.23. Let P be a pushdown automaton and let $L \subseteq \Sigma^*$. P *accepts* L if

$$L = \{x \in \Sigma^* : \exists q \in F, \alpha \in \Gamma^*, (q_0, x, Z) \vdash^* (q, \epsilon, \alpha)\}$$

where Z is the initial stack symbol of P.

This section will focus on a special subclass of pushdown automata.

Definition 5.24. A *2-stack pushdown automaton* is a pushdown automaton such that Γ is replaced by $\Gamma \times \Gamma$ for some stack alphabet Γ.

Replacing Γ with $\Gamma \times \Gamma$ allows a 2-stack pushdown automaton to record data and write to two separate stacks. Note that a 2-stack pushdown automaton with unbounded stack size is equivalent to a Turing machine. For the rest of this section, all pushdown automata will be assumed to be 2-stack pushdown automata.

Theorem 5.25. *Let P be a pushdown automaton. Let Γ be its stack and assume that $|\Gamma| = n$, i.e. that the stack will never grow beyond height n. Fix a state $q \in Q \setminus F$ and assume that for all $w \in \Sigma^*$ and $\alpha \in \Gamma^*$,*

$$\{(p, u, \beta) \in Q \times \Sigma^* \times \Gamma^* : (q, w, \alpha) \vdash^* (p, u, \beta)\} = \{(q, w, \alpha)\}$$

Consider the following algorithm:

1. *Push the first n letters seen onto the first stack.*

2. *Pop the first stack onto the second stack.*

3. *For the next n letters, let $x \in \Sigma$ be the letter seen. Pop a letter off the second stack; call this letter y. If $x = y$, push x onto the first stack.*

4. *Pop the first stack onto the second stack.*

5. *Determine whether the word on the second stack is primitive. If it is not, or if the second stack is empty, transition to state q. Otherwise, continue.*

6. *Let k be the length of the word on the second stack. While there are more letters to be seen:*

   (a) *If the second stack is not empty, let x be the next letter seen. Pop a letter off the second stack and call this letter y. If $x = y$, push x onto the first stack. Otherwise, transition to state q.*

   (b) *Otherwise, pop the first stack onto the second stack.*

Then P recognizes L_n.

Proof. $w \in L_n$ if and only if $w = uv^p$ for some words u and v with $|u|+|v| = n$ and $|u|+|v|$ minimal. On input of a word $w = uv^p$, P writes uv to the stack, then searches the next n letters to see if w continues with another copy of v, for any possible v. If not, then w does not have complexity n and P correctly rejects w. If v is not primitive, then w can be written as uv'^p for $v' \preceq v$ and $v' \neq v$, so $D(w) < n$ and P correctly rejects w. Having established a candidate v, P then checks to see if the rest of w is composed of copies of v, accepting w if it is and rejecting if it is not. □

5.4 Effective witnesses of deterministic HMM complexity

The results in this chapter and before have been focused on the properties of HMM complexity and formulas to count the number of words of various classes defined via HMM complexity. This section will focus on the *effectiveness* of those results and of new ones.

Theorem 5.26. *Let w be a word with $D(w) > n$ for some n. Then, for any HMM H, it can be verified that H does not witness that $D(w) \leq n$ in $O(n^2)$ time.*

Proof. Input w on H. For each $i < |w|$, let $u = w \upharpoonright i$. Then, continue running $w[i:]$ on H to verify that there does not exist $v \preceq w[i:]$ such that $|u| + |v| \leq n$ and $uv^p = w$, for some $p \in \mathbb{Q}$. □

Theorem 5.27 below is analogous to Theorem 5.26: Theorem 5.26 views complexity in terms of states of a witnessing HMM, while Theorem 5.27 is concerned with the structure of the word itself.

Theorem 5.27. *Let w be a word $|w| = n$. A decomposition (u, v) of w can be found in $O(n^2)$ time.*

Proof. As in the proof of Theorem 5.26, loop through w to find potential starting indices i of v; then loop through $w[i:]$ see if $(w \upharpoonright i, w[i:])$ is a decomposition. □

Theorem 5.26 and Theorem 5.27 are worst-case bounds that cover all possible cases of HMM complexity. Many results so far have indirectly proven sharper bounds, which will be made explicit below.

Recall that $\mathcal{F}_n$ are the Fibonacci words defined in Definition 5.12 and that $\mathcal{F}$ is the infinite Fibonacci word defined in Definition 5.13.

Theorem 5.28. *For all $n \geq 3$, a decomposition of $\mathcal{F}_n$ can be found in $O(1)$ time.*

Proof. Theorem 5.20 implies that a decomposition of $\mathcal{F}_n$ is $(\epsilon, \mathcal{F}_{n-1})$, and the index of $\mathcal{F}_{n-1}$ in $\mathcal{F}_n$ can be found via the closed-form formula of the Fibonacci numbers. □

Theorem 5.29. *Given a word w, a word x of length $O(|u|)$ can be found in $O(|w|^3)$ time such that $D(wx) > \frac{|wx|}{2}$.*

Proof. This result is an effectivzation of the proof of Theorem 5.7. A decomposition (u, v) of w with $uv^p = w$ can be found in $O(|w|^2)$ time by Theorem 5.27. Let v' be such that v' is not a cyclic shift of v; such a word can be found in $O(|w|)$ time. Let n be large enough so that $n \geq p$ and $(n+1)|v| > |u|$, which can be found in $O(1)$ time. Then $D(wv^n v') > \frac{|wv^n v'|}{2}$ by Theorem 5.27. □

INDEX

$D(w)$, 7

$I(w)$, 7

$\Delta(w)$, 8

$\Delta_{\mathbb{N}}(w)$, 8

$\mathcal{C}(q)$, 13

$\mathcal{P}(n)$, 6

$\mathcal{S}(q)$, 8

$s_q(n)$, 16

$w \restriction n$, 6

$w[n:]$, 6

BIBLIOGRAPHY

[1] Jean-Paul Allouche and Jeffrey Shallit. *Automatic sequences: theory, applications, generalizations*. Cambridge university press, 2003.

[2] Achilles A. Beros, Bjørn Kjos-Hanssen, and Daylan Kaui Yogi. Planar digraphs for automatic complexity. In *International Conference on Theory and Applications of Models of Computation*, pages 59–73. Springer, 2019.

[3] Samuel D Birns and Bjørn Kjos-Hanssen. On the degrees of constructively immune sets. In *Connecting with Computability: 17th Conference on Computability in Europe, CiE 2021, Virtual Event, Ghent, July 5–9, 2021, Proceedings 17*, pages 50–59. Springer, 2021.

[4] David W. Boyd. Losing runs in Bernoulli trials. *Unpublished manuscript*, 1972.

[5] Vasco Brattka. Random numbers and an incomplete immune recursive set. In *Automata, languages and programming*, volume 2380 of *Lecture Notes in Comput. Sci.*, pages 950–961. Springer, Berlin, 2002.

[6] J.L. Brown Jr. Zeckendorf's theorem and some applications. *The Fibonacci Quarterly*, (3):163–168, 1964.

[7] J. Richard Büchi. On a decision method in restricted second order arithmetic. In *The collected works of J. Richard Büchi*, pages 425–435. Springer, 1990.

[8] Cristian Calude. *Information and randomness*. Monographs in Theoretical Computer Science. An EATCS Series. Springer-Verlag, Berlin, 1994. An algorithmic perspective, With forewords by Gregory J. Chaitin and Arto Salomaa.

[9] Peter A. Cholak, Damir D. Dzhafarov, Jeffry L. Hirst, and Theodore A. Slaman. Generics for computable Mathias forcing. *Ann. Pure Appl. Logic*, 165(9):1418–1428, 2014.

[10] Wai-Fong Chuan and Hui-Ling Ho. Locating factors of the infinite Fibonacci word. *Theoretical computer science*, 349(3):429–442, 2005.

[11] S. B. Cooper. Jump equivalence of the Δ^0_2 hyperhyperimmune sets. *J. Symbolic Logic*, 37:598–600, 1972.

[12] Jack J. Dai, James I. Lathrop, Jack H. Lutz, and Elvira Mayordomo. Finite-state dimension. *Theoretical Computer Science*, 310(1-3):1–33, 2004.

[13] Michael Domaratzki, Derek Kisman, and Jeffrey Shallit. On the number of distinct languages accepted by finite automata with n states. *Journal of Automata, Languages and Combinatorics*, 7(4):469–486, 2002.

[14] Rodney G. Downey and Denis R. Hirschfeldt. *Algorithmic randomness and complexity.* Springer Science & Business Media, 2010.

[15] Yuri L. Ershov. Theory of numberings. In *Handbook of computability theory*, volume 140 of *Stud. Logic Found. Math.*, pages 473–503. North-Holland, Amsterdam, 1999.

[16] Marie Ferbus-Zanda and Serge Grigorieff. Refinment of the "up to a constant" ordering using contructive co-immunity and alike. Application to the Min/Max hierarchy of Kolmogorov complexities. *arXiv preprint*, abs/0801.0350, 2008.

[17] Kayleigh K. Hyde and Bjørn Kjos-Hanssen. Nondeterministic automatic complexity of almost square-free and strongly cube-free words. In *Computing and Combinatorics: 20th International Conference, COCOON 2014, Atlanta, GA, USA, August 4-6, 2014. Proceedings 20*, pages 61–70. Springer, 2014.

[18] Carl Jockusch and Frank Stephan. A cohesive set which is not high. *Math. Logic Quart.*, 39(4):515–530, 1993.

[19] Bjørn Kjos-Hanssen. Superposition as memory: unlocking quantum automatic complexity. In *Unconventional Computation and Natural Computation: 16th International Conference, UCNC 2017, Fayetteville, AR, USA, June 5-9, 2017, Proceedings*, pages 160–169. Springer, 2017.

[20] Bjørn Kjos-Hanssen. *Automatic complexity: a computable measure of irregularity.* De Gruyter, 2023. To appear.

[21] Bjørn Kjos-Hanssen and Samuel D. Birns. *Statistics for Calculus Students.* Open Educational Resources, Outreach College, University of Hawai'i at Mānoa, 2019.

[22] Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

[23] Xiang Li. Effective immune sets, program index sets and effectively simple sets—generalizations and applications of the recursion theorem. In *Southeast Asian conference on logic (Singapore, 1981)*, volume 111 of *Stud. Logic Found. Math.*, pages 97–106. North-Holland, Amsterdam, 1983.

[24] Piergiorgio Odifreddi. *Classical recursion theory*, volume 125 of *Studies in Logic and the Foundations of Mathematics.* North-Holland Publishing Co., Amsterdam, 1989. The theory of functions and sets of natural numbers, With a foreword by G. E. Sacks.

[25] OEIS Foundation Inc. The On-Line Encyclopedia of Integer Sequences, 2023. Published electronically at `http://oeis.org`.

[26] Joseph G. Rosenstein. *Linear orderings*, volume 98 of *Pure and Applied Mathematics*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1982.

[27] Marcus Schaefer. A guided tour of minimal indices and shortest descriptions. *Arch. Math. Logic*, 37(8):521–548, 1998.

[28] Jeffrey Shallit. *A second course in formal languages and automata theory*. Cambridge University Press, 2008.

[29] Jeffrey Shallit and Ming-Wei Wang. Automatic complexity of strings. *Journal of Automata, Languages and Combinatorics*, 6(4):537–554, 2001.

[30] Raymond M. Smullyan. Effectively simple sets. *Proc. Amer. Math. Soc.*, 15:893–895, 1964.

[31] R. Soare. *Recursively Enumerable Sets and Degrees*. Springer, 1987.