Probabilistic automatic complexity of finite strings

Kenneth Gill [*]

Abstract. We introduce a new complexity measure for finite strings using probabilistic finite-state automata (PFAs), in the same spirit as existing notions employing DFAs and NFAs, and explore its properties. The PFA complexity $A_P(x)$ is the least number of states of a PFA for which x is the most likely string of its length to be accepted. The variant $A_{P,\gamma}(x)$ adds a real-valued parameter γ specifying a required lower bound on the gap in acceptance probabilities between x and other strings. We prove $A_{P,\gamma}$ is γ-computable for all γ, relate A_P to the DFA and NFA complexities, and obtain a complete classification of binary strings with $A_P = 2$. Finally, we discuss several other variations on A_P with a view to obtaining additional desirable properties.

Keywords: probabilistic automaton, finite-state automaton, automatic complexity, iterated function system, algorithmic information theory

1. Introduction

Informally, the Kolmogorov complexity of a finite string w is the size of the smallest Turing machine which outputs w given no input. As a function, it is well-known to be noncomputable, and moreover only defined up to an additive constant. These drawbacks have motivated several authors to define complexity measures based on models of computation less powerful than the Turing machine, such as context-free grammars [2, 3]. In 2001, Shallit and Wang introduced one such measure using deterministic finite-state automata (DFAs), defining $A_D(w)$ to be the number of states of the smallest

Address for correspondence: gillmathpsu@posteo.net

[*] Most of this work is part of the author's Ph.D. dissertation at Penn State University [1]. The author is grateful to his thesis advisors Jan Reimann and Linda Westrick for their invaluable help and support, to Jake Canel for several helpful conversations, and to Bjørn Kjos-Hanssen for suggesting the proof of Theorem 5.2 as well as inspiring the project as a whole. Finally, the author would like to thank the anonymous reviewers for their detailed comments which led to many improvements in the presentation. This research was supported in part by NSF grant DMS-1854107.

DFA for which w is the only string of its length to be accepted [4]. This measure is computable, well-defined, and there is a polynomial-time algorithm to recover w from a witness for $A_D(w)$. Later in 2013, Hyde defined a similar measure replacing DFAs with nondeterministic finite-state automata (NFAs) [5, 6]. A_N shares the advantages of A_D over Kolmogorov complexity while additionally being invariant under reversal and avoiding "dead states" (nonaccepting states with no out-transitions) often present among witnesses for A_D merely to satisfy the requirement of totality of the transition function. The study of A_N has been continued by Kjos-Hanssen, see e.g. [7, 8, 9, 10], as well as the recent book [11]. Topics which have been investigated include the structure of the set of "A_N-random" strings (having maximal complexity for their length), computational complexity of finding witnesses to several variations on A_N, and bounds on the NFA complexity of various interesting families of strings, such as the square-free or overlap-free words. Connections to information theory abound, including notions of information distance and algorithmic fractal dimension defined through A_N.

Inspired by the aforementioned work, we investigate what happens when deterministic or nondeterministic machines are replaced by probabilistic ones (PFAs), wherein each state transition occurs with some probability and each word w is assigned a probability of acceptance $\rho_M(w)$ by the PFA M. We view M as describing w if w uniquely maximizes ρ_M among all strings of the same length. This property can be phrased in terms of the so-called gap function, which measures how well M separates w from other strings:

Definition 1.1. The *gap function* of a PFA M reading from the alphabet Σ is the map from Σ^* to $[-1, 1]$ given by

$$\operatorname{gap}_M(w) = \min\{\, \rho_M(w) - \rho_M(z) : z \in \Sigma^{|w|} \text{ and } z \neq w \,\}. \tag{1}$$

If $\Sigma^{|w|} \setminus \{w\}$ is empty, by convention we set $\operatorname{gap}_M(w) = \rho_M(w)$. In particular, if λ is the empty string, $\operatorname{gap}_M(\lambda) = \rho_M(\lambda)$.

The gap function is positive iff w is more likely than any other string of the same length to be accepted, and we define the PFA complexity to be the least number of states needed for this to happen:

Definition 1.2. The *probabilistic automatic complexity* (PFA complexity) of w with respect to the alphabet Σ is

$$A_P(w, \Sigma) = \min\{\, k \,:\, \text{there is a } k\text{-state PFA } M \text{ reading from } \Sigma \text{ such that } \operatorname{gap}_M(w) > 0 \,\}. \tag{2}$$

We always presume w to be in Σ^*. If Σ is understood from context, we simply write $A_P(w)$. In particular, unless otherwise remarked upon, it is assumed that Σ consists exactly of the letters appearing in w.

This definition is probably the one most directly analogous to the definitions of A_D and A_N, and it turns out that A_P is also computable, as A_D and A_N are (see Theorem 5.1 below). Unfortunately, unlike A_N but like A_D, A_P is not alphabet-independent: if $w \in \Sigma_1^* \cap \Sigma_2^*$, it may be that $A_P(w, \Sigma_1) \neq A_P(w, \Sigma_2)$. (See [11, Theorem 1.28] for the statement about A_D and A_N.) For example, $A_P(a^n, \{a\}) = 1$, but $A_P(a^n, \{a, b\}) = 2$ if $b \neq a$. This is because no nonempty string can have complexity 1 over an alphabet with more than one letter: a PFA with one state assigns the same probability to every string. We have that $A_P(a^n, \Sigma) = 2$ whenever $|\Sigma| \geq 2$ by Theorem 4.1

and Corollary 4.18. The same issue does not arise for A_N since one is free in an NFA to omit any transitions involving irrelevant letters. Later we will also see that the PFA complexity of a string and its reversal can be different, like A_D and unlike A_N.

One peculiarity of A_P is that M can witness $A_P(w)$ while $\rho_M(w)$ is not very high, or not much different from $\rho_M(z)$ for other strings z of the same length. Is M really a good representation of w if it can only slightly distinguish w from other strings? What if all potential witnesses M have this property? We address this problem by defining a strengthened version of A_P which includes a real-valued parameter allowing one to specify a required lower bound on the gap between $\rho(w)$ and the next highest probability:

Definition 1.3. The *probabilistic automatic complexity of w with gap $\gamma \in [0, 1)$* is

$$A_{P,\gamma}(w, \Sigma) = \min\{\, k \ : \text{ there is a } k\text{-state PFA } M \text{ reading from } \Sigma \text{ such that } \mathrm{gap}_M(w) > \gamma \,\}.^{1} \quad (3)$$

Thus $A_{P,0}(w, \Sigma) = A_P(w, \Sigma)$. As with A_P, we omit Σ when understood from context and write $A_{P,\gamma}(w)$. For any γ, an automaton M witnessing $A_{P,\gamma}(w)$ must not only witness $A_P(w)$, but must also give w a probability at least γ higher than any other string of its length. So, by increasing the parameter γ, we increase the degree by which M recognizes w and decrease the ambiguity in determining which string of length $|w|$ is described by M. (This idea is further elaborated on after the statement of Corollary 4.20 below.)

It turns out that $A_{P,\gamma}$ is always computable from a description of γ:

Theorem 5.1. For every finite alphabet Σ and every $\gamma \in [0, 1)$, the function $w \mapsto A_{P,\gamma}(w, \Sigma)$ is γ-computable.

Of course, if γ is a computable number, this just means that $w \mapsto A_{P,\gamma}(w)$ is computable. (Section 5 clarifies precisely what is meant by a "description" of γ.) We also establish the almost-everywhere uniform computability of $A_{P,\gamma}(w)$ as a function of both w and γ in Theorem 5.3, although its proof does not extend to the case $\gamma = 0$.

Our other main result about A_P is the following complete classification of binary strings with complexity 2, which arguably helps to vindicate A_P by showing that it does appear to capture some intuitive structure in strings.

Theorem 4.1. If $\Sigma = \{i, j\}$ and $w \in \Sigma^*$, we have $A_P(w, \Sigma) = 2$ if and only if w is of the form

$$i^n j^m, \qquad i^n j^m i, \qquad i^n (ji)^m, \qquad \text{or} \quad i^n (ji)^m j \qquad (22)$$

for some $n \geq 0$, $m \geq 1$.

One can say a bit more: a consequence of the proof of Theorem 4.1 is that whenever $w \in \Sigma^*$ with $|\Sigma| = 2$ and $\Sigma' \supset \Sigma$, then $A_P(w, \Sigma) = 2$ implies $A_P(w, \Sigma') = 2$ (Corollary 4.18). In other words,

[1] This is a slightly different definition from that originally given in the author's dissertation [1], which required $\mathrm{gap}_M(w) \geq \gamma$ rather than $>$. The author has come to feel that the present definition is more natural. For the most part, only minor amendments to the proofs of results involving $A_{P,\gamma}$ were needed as a result of this change.

the property of being a nonconstant binary string with complexity 2 is alphabet-independent, and so we can write $A_P(w) = 2$ without ambiguity. (A string w is *constant* if it is of the form a^n for some $n \geq 0$, where a is a single letter. Otherwise w is *nonconstant*.)

The class of strings with $A_P = 2$ is far larger than that with $A_N = 2$, which is exactly the set of strings of the form ij^n, $i^n j$, $(ij)^n$, or $(ij)^n i$, as classified in [5]. In fact, A_N is unbounded on strings of the form $i^n j^m$, which implies

Corollary 4.20. The quantity $A_N(w) - A_P(w)$ may be arbitrarily large among binary w.

$A_{P,\gamma}$ has a philosophically attractive feature not shared by A_P which we now describe for the sake of further motivating its study. Suppose one is given an automaton M as a "black box", that is, with no information whatsoever about its inner workings. All one can do is run it with some input string, and check whether it accepts or rejects the string. Suppose further that an experimenter wishes to test whether this automaton witnesses an upper bound for $A_P(w)$ for some string w. Then the experimenter needs not only to check whether each $z \in \Sigma^{|w|}$ is accepted, but whether or not it will be accepted with a lower bound λ on its probability of acceptance, for each λ in turn. (This would enable them to decide if there is some particular w, λ with $\rho_M(w) > \lambda$ but $\rho_M(z) < \lambda$ whenever $|z| = |w|$ and $z \neq w$. In other words, they would estimate a lower bound on $\text{gap}_M(w)$.) The experimenter can only accomplish this by running the machine repeatedly on each input w to get some sense of the expected value of $\rho_M(w)$, up to some acceptable margin of error ε.

In his original paper introducing PFAs, Rabin [12] discusses a similar endeavor in the context of establishing experimentally that w is in a given stochastic language, where a language is stochastic if it is of the form $\{ w \in \Sigma^* : \rho_M(w) > \lambda \}$ for some PFA M and $\lambda \in [0, 1]$, called the *cut-point*. As he points out, the law of large numbers implies that as long as $\rho_M(w) \neq \lambda$, there is a finite number $N = N(w, \varepsilon)$ such that running N trials, counting the number s of successes, and comparing s/N with λ will correctly determine if $\rho_M(w) > \lambda$ with probability $1 - \varepsilon$. But, as he goes on to say, finding $N(w, \varepsilon)$ would depend on knowing $\rho_M(w)$ in the first place.

Rabin's solution is to only consider cut-points λ which are isolated for M, that is, such that $|\rho_M(w) - \lambda| \geq \gamma$ for all $w \in \Sigma^*$ and some $\gamma > 0$. If one wants to run the above experiment to test if $\rho_M(w) > \lambda$ when λ is isolated, then the number of trials N needed to determine this within margin of error ε now only depends on γ and ε, regardless of M. Knowledge of $\rho_M(w)$ is not needed. Of course, this is not a solution from a practical point of view if no such cut-point is given at the outset, because now the experimenter would need to determine if λ is isolated for M and (if so) a lower bound for its degree of isolation γ. The problem of determining if a given rational cut-point is isolated for a given PFA is known to be Σ^0_2-complete [13, Theorem 1].

But—back to our black-box experiment—if one specifies γ at the outset and looks for a witness for an upper bound on $A_{P,\gamma}(w)$ rather than $A_P(w)$, the problem disappears and we still get that N depends only on γ and ε, with both of these parameters now being chosen by the experimenter. To see why, let a single trial consist of running every word of length $|w|$ through the machine M once. If $s(w, N)$ is the number of acceptances of w in N trials, then there is a function $N = N(\gamma, \varepsilon)$ such that for each $z \in \Sigma^{|w|}$, one correctly concludes with probability at least $1 - \varepsilon'$ that $\rho_M(w) - \rho_M(z) > \gamma$ given $[s(w, N) - s(z, N)]/N > \gamma$, assuming $\rho_M(w) - \rho_M(z) \neq \gamma$. Here ε' is chosen small enough that $(1 - \varepsilon')^{|\Sigma|^{|w|}} > 1 - \varepsilon$. Since acceptances and rejections of words are presumed to be independent events, it follows that after $N(\gamma, \varepsilon)$ trials, one correctly concludes with probability at least $1 - \varepsilon$ that $\mathrm{gap}_M(w) > \gamma$.

This paper establishes several basic properties of A_P and $A_{P,\gamma}$ and hopefully justifies their study as having intrinsic interest, but many avenues of investigation are left unexplored. This is in part due to both A_P and $A_{P,\gamma}$ proving somewhat combinatorially difficult to reason with, aside from a few of our results which follow quickly from straightforward matrix calculations. In particular there is nothing we can say about the asymptotic behavior of either quantity: no example is known at the time of writing which even suggests A_P can be greater than 3. Indeed, the original motivation behind proving the classification theorem (Theorem 4.1) was to show that A_P can be greater than 2, which was until then unclear. Then the most fundamental question we leave unanswered is probably

Question 1.4. Is A_P unbounded? If not, what is its maximum value? Similarly when restricted to a given finite alphabet, and similarly for $A_{P,\gamma}$.

Remark 1.5. Probabilistic finite automata were independently introduced in 1963 by Michael Rabin and J. W. Carlyle [12, 14]. Carlyle's stochastic sequential machines are transducers with both input and output behavior, while Rabin's PFAs—which are sometimes also called stochastic acceptors—can only accept an input string with some probability. The present work focuses only on PFAs as defined by Rabin, although Carlyle-style machines have found wide applicability in machine learning and pattern recognition; see [15] for a modern survey. A notion of transducer complexity of finite strings has also been studied [16], but the approach taken there is most like that of the Kolmogorov complexity rather than A_D. We leave the probabilistic analogue for future work. There is, however, an idea related to A_P which has been studied for transducers in the machine learning literature. Given a probabilistic finite-state transducer T, x is called the *most probable string* or *consensus string* of T if it is generated by T with maximal probability among all strings, not just among those with the same length [15]. One might ask if this notion should be adapted to PFAs, defining the complexity of x instead as the smallest size (in some sense) of a PFA accepting x with unique highest probability among all strings. But we will see in Proposition 4.10 that a single PFA can simultaneously witness $A_P(x)$ for every one of an infinite family of strings x of similar structure. This ability arguably lends A_P a descriptive advantage over a notion resulting from viewing a PFA as only describing its most probable string.

Remark 1.6. This paper is a rewritten and expanded version of the second chapter of the author's dissertation [1], and Proposition 3.1, Corollary 3.2, Theorem 4.2, Proposition 4.11, Corollary 4.19, and Theorem 5.3 were already present in the latter work, as well as the content of Section 4.1 and any lemmata used in the proof of Theorem 4.2. Propositions 3.3 and 3.4 also appeared there in a less general form. All other results are new to the present article.

The structure of the rest of the paper is as follows. After collecting some formal definitions in the next section, we state a few preliminary results in Section 3, including Proposition 3.1 relating A_P and $A_{P,\gamma}$ to A_D and A_N. Here we also discuss a few examples of the calculation of A_P and $A_{P,\gamma}$. Section 4, which takes up over half of the paper, consists entirely of the proof of Theorem 4.1. This proof exploits a correspondence between PFAs and iterated function systems detailed in Section 4.1. Section 5 is devoted to proving the computability of A_P and $A_{P,\gamma}$, which involves techniques from computable analysis as well as an application of a classical result in model theory. Finally, in Section 6 we discuss several further variations on A_P with an eye to mitigating its potential flaws as a complexity measure.

2. Preliminaries

Our notation is mostly standard. Let Σ^* be the set of finite strings over the finite alphabet Σ. Write either xy or $x^\frown y$ for the concatenation of the strings x and y.

Definition 2.1. A *probabilistic finite-state automaton* (PFA) is an abstract machine specified by a tuple $M = (Q, \Sigma, P, \vec{\pi}, \vec{\eta})$, where

- $Q = \{ q_1, \dots, q_n \}$ is the set of states;
- Σ is a finite alphabet;
- P is a set of $n \times n$ right-stochastic matrices $\{ P_a : a \in \Sigma \}$ describing the transition probabilities. For each $a \in \Sigma$, $(P_a)_{ij}$ is the probability of going from q_i to q_j when letter a is read;
- $\vec{\pi}$ is a row vector of length n giving a probability distribution on initial states, so $\vec{\pi}_j$ is the probability of the machine starting in state q_j; and
- $\vec{\eta}$ is a column vector of length n determining the set of accepting states, with $\vec{\eta}_i$ being 1 if q_i is accepting and 0 otherwise.

M is said to be *over* Σ if it reads from the alphabet Σ. When Σ is not important, we will omit its mention, and likewise we usually identify Q with the set $\{1, \dots, n\}$ for some n. Thus we may specify a PFA by giving only $\vec{\pi}$, $\vec{\eta}$, and the matrices P_a. If all entries of $\vec{\pi}$ and each P_a are rational numbers, then we refer to M as *rational*. A PFA can also be represented as a digraph, with edges labeled by transition probabilities. For example, Figure 1 depicts the PFA over the alphabet $\{0, 1\}$ with

$$P_0 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad P_1 = \begin{pmatrix} .5 & 0 & .5 \\ 0 & 1 & 0 \\ .5 & .5 & 0 \end{pmatrix}, \quad \vec{\pi} = (1, 0, 0), \quad \vec{\eta} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{4}$$

Definition 2.2. If M is a PFA and $x = x_1 x_2 \cdots x_\ell$ is a string, let

$$P_M(x) = P_{x_1} P_{x_2} \cdots P_{x_\ell}. \tag{5}$$

Then the *acceptance probability of* x with respect to M is

$$\rho_M(x) = \vec{\pi} P_M(x) \vec{\eta}. \tag{6}$$

If M is understood from context we may simply write $\rho(x)$, and similarly $\operatorname{gap}(x)$.

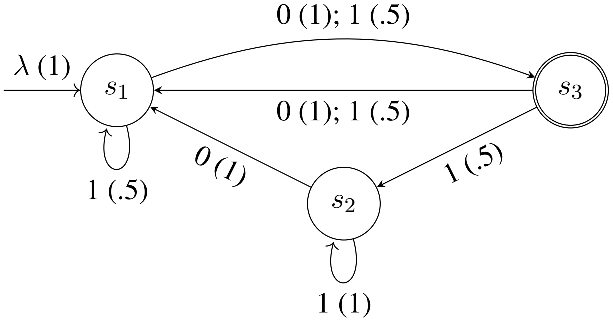


Figure 1: An example of a PFA. Numbers in parentheses are transition probabilities, so that the PFA starts in state s_1 with probability 1. Here s_3 is the unique accepting state.

One can view a DFA as the special case of a PFA in which $\vec{\pi}$ is a coordinate vector and all P_as are zero-one matrices. An NFA is then a slight relaxation of a DFA where $\vec{\pi}$ may be any zero-one vector and each P_a may be any zero-one matrix. Of course, DFAs and NFAs are usually represented as digraphs, but it is convenient for us to think of them via their transition matrices since we will manipulate them directly alongside PFAs. The precise definitions of the DFA and NFA complexities are as follows:

Definition 2.3. (Shallit and Wang [4])
The *deterministic automatic complexity* of a finite string x is

$$\begin{aligned} A_D(x) = \min\{k : \text{there is a } k\text{-state DFA accepting } x \\ \text{uniquely among strings of length } |x|\}. \end{aligned} \tag{7}$$

In other words, thinking of a witnessing DFA M as a PFA, this says $\text{gap}_M(x) = 1$, or equivalently $\text{gap}_M(x) > 0$ since the gap function takes only the values 0 and 1 when M is deterministic. It follows immediately that $A_P(x) \le A_D(x)$ for all x.

Definition 2.4. (Hyde [5])
The *nondeterministic automatic complexity* of x is

$$\begin{aligned} A_N(x) = \min\{k : \text{there is a } k\text{-state NFA accepting } x \\ \text{and with a unique accepting path of length } |x|\}. \end{aligned} \tag{8}$$

Every DFA witnessing $A_D(x)$ is an NFA that accepts x with a unique accepting path of length $|x|$ (by virtue of its determinism), so $A_N(x) \le A_D(x)$ for all x.

For the reader's convenience, we repeat here the definitions of the gap function, A_P, and $A_{P,\gamma}$ from the introduction.

Definition 2.5. The *gap function* of the PFA M over Σ is the map from Σ^* to $[-1, 1]$ given by

$$\text{gap}_M(w) = \min\{\, \rho_M(w) - \rho_M(z) : |z| = |w| \text{ and } z \ne w \,\}. \tag{9}$$

By convention we take $\mathrm{gap}_M(w) = \rho_M(w)$ if the minimum is over the empty set. Then the *probabilistic automatic complexity* of w *over* Σ *with gap* $\gamma \in [0,1)$ is

$$A_{P,\gamma}(w,\Sigma) = \min\{\, k \;:\; \text{there is a } k\text{-state PFA } M \text{ over } \Sigma \text{ such that } \mathrm{gap}_M(w) > \gamma \,\}. \tag{10}$$

The *probabilistic automatic complexity* of w is then

$$A_P(w,\Sigma) = A_{P,0}(w,\Sigma). \tag{11}$$

If Σ is understood from context then we simply write $A_P(w)$ and $A_{P,\gamma}(w)$.

3. First results on A_P

In this section we establish a few basic properties of A_P and $A_{P,\gamma}$, beginning by relating them to A_D and A_N:

Proposition 3.1. (i) For any x, $A_P(x) \le A_N(x) + 1$.

(ii) $A_P(x) \le A_{P,\gamma}(x) \le A_D(x)$ for all x and $\gamma \in [0,1)$. For every x, there is a $\gamma' > 0$ such that $A_{P,\gamma}(x) = A_P(x)$ for all $\gamma \in [0,\gamma')$.

Proof:

(i) Let $M = (Q,\Sigma,P,\vec{\pi},\vec{\eta})$ be an NFA witnessing $A_N(x)$. Uniqueness of M's accepting path for x means in particular that $\vec{\pi}$ is a coordinate vector. Then define a PFA $M' = (Q',\Sigma,P',\vec{\pi}',\vec{\eta}')$ as follows. Let $Q' = Q \cup \{q\}$, where q is a new state not occurring in Q, to be listed after all other states. Let $\vec{\pi}' = [\vec{\pi}|0]$ and $\vec{\eta}' = [\vec{\eta}|0]$, where $[\vec{a}|\vec{b}]$ denotes the concatenation of the vectors $\vec{a}$ and $\vec{b}$. Write X^i for the ith row of any matrix X ($i \ge 1$). For each $a \in \Sigma$, let P'_a be built as follows from P_a: if P^i_a has at least one nonzero entry, let $(P'_a)^i = [P^i_a|0]/(\sum P^i_a)$. Otherwise, let $(P'_a)^i = [P^i_a|1] = (0,\ldots,0,1)$. Finally, if $|Q| = k$, then append a new row $(P'_a)^{k+1} = (0,\ldots,0,1)$ (this corresponds to the new state q).

Then M' still has a unique accepting path of length $|x|$; in particular, x is the only string of length $|x|$ with $\rho_{M'}(x)$ positive. Therefore M' witnesses an upper bound for $A_P(x)$.

(ii) If $\gamma < \gamma'$ then $A_{P,\gamma}(x) \le A_{P,\gamma'}(x)$, and if M is a DFA witnessing $A_D(x)$ then $\mathrm{gap}_M(x) = 1$. This gives the first statement. For the second statement, one can for example pick $\gamma' = \mathrm{gap}_M(x)/2$ for any witness M for $A_P(x)$.

□

Corollary 3.2. For all x, $A_P(x) \le \lfloor |x|/2 \rfloor + 2$.

Proof:
Hyde showed in [5, Theorem 3.1] that $A_N(x) \le \lfloor |x|/2 \rfloor + 1$, so the bound immediately follows from the proposition. □

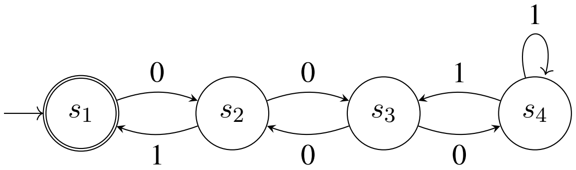


Figure 2: An NFA witnessing that $A_N(0001101) = 4$.

The procedure described in the first part of Proposition 3.1 demonstrates that if $A_N(x)$ is witnessed by an NFA such that every state has at least one out-transition for every letter, then $A_P(x) \leq A_N(x)$ (because there are no rows of all zeros in the transition matrices, and the "dead state" q need not be added).

As an example of this construction, according to Bjørn Kjos-Hanssen's website,[2] $A_N(0001101) = 4$ via the NFA depicted in Figure 2. Here s_1 is both the initial and accepting state. In matrix form, this NFA can be represented as

$$P_0 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad P_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad \vec{\pi} = (1,0,0,0), \quad \vec{\eta} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{12}$$

To transform this into a PFA, we need to add a fifth state due to the rows of zeros, and from the construction we get

$$P_0' = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, P_1' = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \vec{\pi}' = (1,0,0,0,0), \vec{\eta}' = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{13}$$

However, this is hardly optimal as a witness for A_P, since actually $A_P(0001101) = 3$ via

$$P_0 = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}, \quad P_1 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \vec{\pi} = (1,0,0), \quad \vec{\eta} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{14}$$

These and all other numerical examples mentioned below were verified with the SageMath mathematical software, using a code library developed by the author in order to more easily compute with PFAs and weighted automata in general [17]. Using this library, one could for example verify that (14) witnesses an upper bound for $A_P(0001101)$ as follows, entering the code into the Sage command line while `WeightedAutomaton.py` is in the working directory:

[2] `https://math.hawaii.edu/wordpress/bjoern/complexity-of-0001101/`

```
sage.all.load("WeightedAutomaton.py")
A = WeightedAutomaton({’0’: [[0,1/2,1/2],
                             [0,1/2,1/2],
                             [0,1,0]],
                       ’1’: [[0,0,1],
                             [1,0,0],
                             [0,1,0]]},
                      [1,0,0],[1,0,0])
A.is_highest(’0001101’)
```

The output `True` signifies that `A` assigns 0001101 the unique highest probability among strings of length 7. If one so desires, one can run `A(’0001101’)` or `A.prob(’0001101’)` to find that this probability is $13/16$, and running `A.gap(’0001101’)` computes $\mathrm{gap}_{\mathtt{A}}(0001101)$ to be $1/16$. The above code is also present in the cited GitHub repository for the `WeightedAutomaton` library along with similar code verifying the claimed properties of all other explicit examples given in this paper. It may be found in the file `paperexamples.ipynb`.

No string x is presently known for which $A_P(x)$ is equal to the maximum possible value $A_N(x)+1$. Direct computations using the abovementioned code library have shown that all binary strings x of length 10 or less have $A_P(x) \leq 3$, whereas many such strings have $A_N(x) = 4$, 5, or 6. Theorem 4.3 implies that every x with $A_N(x) = 2$ also has $A_P(x) = 2$. Remembering that "$A_P(x)$" without specifying an alphabet is supposed to denote $A_P(x, \Sigma)$ where Σ is exactly the set of letters used in x, we have that $A_P(x) = 1$ if and only if $x = a^n$ for some $n \geq 0$, and this is also true for A_N.

So far we have not mentioned any examples involving $A_{P,\gamma}$. Experimentally, it appears that one has to make the value of γ quite low in order to get small values of $A_{P,\gamma}$, for all but very short strings. This makes intuitive sense in view of the proof of Theorem 4.2, and more generally the phenomenon of stability of a contractive iterated function system: all orbits converge to the attractor, and correspondingly acceptance probabilities will tend to cluster together for longer strings, at least for a generic automaton. As an example, if $x = 0110$, then a simple witness for $A_P(x) = 2$ is the PFA M given by

$$P_0 = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}, \quad P_1 = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}, \quad \vec{\pi} = (1,0), \quad \vec{\eta} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{15}$$

It follows from Proposition 4.10 that M witnesses that $A_P(01^m0) = 2$ for all m, and one can calculate that $\mathrm{gap}_M(0110) = 1/16$, $\mathrm{gap}_M(01^30) = 1/32$, $\mathrm{gap}_M(01^40) = 1/64$, and so on.

However, it is not necessarily the case that gaps strictly decrease for longer strings. The PFA

$$P_0 = \begin{pmatrix} 0 & 1 & 0 \\ 2/3 & 0 & 1/3 \\ 1/3 & 0 & 2/3 \end{pmatrix}, \quad P_1 = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 2/3 & 1/3 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \vec{\pi} = (0,0,1), \quad \vec{\eta} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \tag{16}$$

witnesses that $A_P(1^n010^3) \leq 3$ for all $n \leq 15$, and each of these strings is given probability $\frac{91}{243}$ with gap $\frac{1}{243}$ (the same almost certainly holds for all n, though we have not yet shown this). This is unsurprising since $\vec{\pi}$ is the (left) Perron-Frobenius eigenvector of P_1. Of course, such a relation is easily destroyed by perturbing the entries of $\vec{\pi}$ and P_1.

It does not seem very easy to simultaneously make γ large and $A_{P,\gamma}$ small even for short strings. One can get a gap of about 0.5609 for $x = 0110$ by using the 3-state PFA

$$P_0 = \begin{pmatrix} 0 & 0 & 1 \\ 0.22151 & 0.77485 & 0.00364 \\ 0.9995 & 0 & 0.0005 \end{pmatrix}, P_1 = \begin{pmatrix} 0 & 0.5622 & 0.4378 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \vec{\pi} = (1,0,0), \vec{\eta} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{17}$$

But among 2-state PFAs, the highest gap known for x at the time of writing is approximately 0.1775, via

$$P_0 = \begin{pmatrix} 0.16748 & 0.83252 \\ 0.98999 & 0.01001 \end{pmatrix}, \quad P_1 = \begin{pmatrix} 0.66116 & 0.33884 \\ 0 & 1 \end{pmatrix}, \quad \vec{\pi} = (1,0), \quad \vec{\eta} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{18}$$

These automata were found by numerically optimizing the results of brute-force searching. In the second case, the search was over roughly 850,000 2-state PFAs and turned up only one giving x a gap greater than 1/6 (it was about 0.1719). Among the same set of PFAs, the largest gap found for 01^30 was approximately 0.1178.

The next result should be compared with the facts that $A_D(xyz) \geq A_D(y)$ and $A_N(xyz) \geq A_N(y)$ for any strings x, y, z. (See [10, Lemma 12] and [5, Theorem 2.4]. The statement for A_N can be derived from $A_N(xy) \geq A_N(x)$ and the invariance of A_N under string reversal.)

Proposition 3.3. For all strings x, y and all $\gamma \in [0,1)$, we have $A_{P,\gamma}(xy) \geq A_{P,\gamma}(y)$.

Proof:
Given γ, let $M = (Q, \Sigma, P, \vec{\pi}, \vec{\eta})$ witness $A_{P,\gamma}(xy)$. Let $M' = (Q, \Sigma, P, \vec{\pi}', \vec{\eta})$ be a PFA with the same configuration as M except for its initial state distribution, which will be $\vec{\pi}' = \vec{\pi} P_M(x)$. Since $P_M(x)$ is a stochastic matrix, $\vec{\pi}'$ is still a probability vector. By definition, $P_M(xw) = P_M(x)P_M(w)$ for all strings w, so we have $\rho_{M'}(w) = \rho_M(xw)$ and consequently

$$\mathrm{gap}_{M'}(w) = \min\{\, \rho_M(xw) - \rho_M(xz) \;:\; z \in \Sigma^{|w|} \setminus \{w\} \,\} \geq \mathrm{gap}_M(xw) \tag{19}$$

for all w. In particular, $\mathrm{gap}_{M'}(y) \geq \mathrm{gap}_M(xy) > \gamma$, so M' witnesses an upper bound for $A_{P,\gamma}(y)$.

□

One property of A_N which motivated its introduction, as mentioned above, is that $A_N(x) = A_N(\overleftarrow{x})$, where $\overleftarrow{x}$ is the reversal of x:

$$\overleftarrow{x} = x_n \cdots x_2 x_1 \quad \text{if} \quad x = x_1 x_2 \cdots x_n. \tag{20}$$

By Theorem 4.1, the class of strings x with $A_P(x) = 2$ is not closed under reversal, so A_P does not share this property.

In Section 6 we will take up the question of how one might recover the property by modifying A_P. Equality of $A_P(x)$ and $A_P(\overleftarrow{x})$ is possible in at least some cases:

Proposition 3.4. If $A_P(x)$ is witnessed by a PFA $M = (Q, \Sigma, P, \vec{\pi}, \vec{\eta})$ such that each $P_a \in P$ is doubly stochastic, and such that all nonzero entries of $\vec{\pi}$ are equal, then $A_P(\overleftarrow{x}) \leq A_P(x)$. If M witnesses $A_{P,\gamma}(x)$ and one can additionally take $\vec{\pi}$ and $\vec{\eta}$ to have the same number of nonzero entries, then $A_{P,\gamma}(\overleftarrow{x}) \leq A_{P,\gamma}(x)$.

Proof:
The idea is more or less the content of Exercise A.2.8 of Chapter 3 of [18]. Define the PFA $M' = (Q, \Sigma, P', \vec{\pi}', \vec{\eta}')$ by $P'_a = P_a^T$ for each $a \in \Sigma$ and $\vec{\pi}' = \vec{\eta}^T/s$, where $s = \sum \vec{\eta}$. If each entry of $\vec{\pi}$ is either 0 or $1/n$ (for some $n \geq 1$), then let $\vec{\eta}' = n\vec{\pi}^T$.

Intuitively, M' represents the automaton obtained by operating M in reverse. We have $P_{M'}(\overleftarrow{x}) = P_M(x)^T$, so

$$\begin{aligned} \rho_{M'}(\overleftarrow{x}) = (\vec{\eta}^T/s)P_{M'}(\overleftarrow{x})(n\vec{\pi}^T) &= ns^{-1}\left(\vec{\pi}P_M(x)\vec{\eta}\right)^T \\ &= ns^{-1}\left(\rho_M(x)\right)^T = ns^{-1}\rho_M(x). \end{aligned} \tag{21}$$

The same calculation shows $\rho_{M'}(\overleftarrow{y}) = ns^{-1}\rho_M(y)$ for all y, so if $\rho_M(x) > \rho_M(y)$, then $\rho_{M'}(\overleftarrow{x}) > \rho_{M'}(\overleftarrow{y})$. Therefore M' witnesses $A_P(\overleftarrow{x}) \leq A_P(x)$ since M' and M have the same number of states.

If $\vec{\pi}$ and $\vec{\eta}$ have the same number of nonzero entries, then $n = s$, and so $\rho_{M'}(\overleftarrow{y}) = \rho_M(y)$ for all $y \in \Sigma^*$. Hence $\mathrm{gap}_{M'}(\overleftarrow{x}) = \mathrm{gap}_M(x)$ and the second statement follows. □

As a corollary of this fact together with Theorem 4.1 below, for most binary strings x such that $A_P(x) = 2$, the latter cannot be witnessed by a PFA as in the proposition. $A_P(\overleftarrow{x}) = A_P(x)$ holds whenever both quantities can be witnessed by such a PFA. An example is given by (16), which witnesses $A_P(1^n010^3) = 3$ and can be turned into a witness for $A_P(0^3101^n) \leq 3$ for all $n \leq 15$ (at least) using the procedure in the proof of Proposition 3.4. Since none of the latter family of strings with $n \geq 2$ can have complexity 2 by Theorem 4.1, it follows that $A_P(0^3101^n) = 3$ for $2 \leq n \leq 15$ (at least).

4. Classification of binary strings with $A_P = 2$

This section is devoted to proving the following theorem, restated from the introduction for the reader's convenience:

Theorem 4.1. If $\Sigma = \{i, j\}$ and $w \in \Sigma^*$, we have $A_P(w, \Sigma) = 2$ if and only if w is of the form

$$i^n j^m, \qquad i^n j^m i, \qquad i^n (ji)^m, \qquad \text{or} \quad i^n (ji)^m j \tag{22}$$

for some $n \geq 0$, $m \geq 1$.

This set of strings is significantly larger than the set of binary strings with NFA complexity 2. As classified in [5], the strings with $A_N(w) = 2$ consist exactly of

$$i^m j, \qquad ij^m, \qquad (ij)^m, \qquad \text{and} \quad (ij)^m i \tag{23}$$

for all m. All that can be generally said about $A_N(i^n j^m)$, for instance, is that it is no more than $\min\{n, m\} + 1$ [5, Example 4.1]. The proof of Corollary 4.20 shows that it is unbounded (see Section 4.4).

The proof of this theorem will occupy a substantial portion of the rest of the paper, and we split it into two halves, the forward and reverse directions:

Theorem 4.2. For $w \in \{i, j\}^*$, if $A_P(w) = 2$, then w is of the form

$$i^n j^m, \qquad i^n j^m i, \qquad i^n (ji)^m, \qquad \text{or} \quad i^n j(ij)^m \qquad \text{for some } n \geq 0, m \geq 1. \tag{24}$$

Theorem 4.3. The values of $A_P(i^n j^m)$, $A_P(i^n j^m i)$, $A_P(i^n (ji)^m)$, and $A_P(i^n j(ij)^m)$ are equal to 2 for all $n \geq 0$ and $m \geq 1$.

Before proceeding to the very long proof, we take a moment to outline the plan of attack. Every PFA M corresponds to an iterated function system (IFS) $F = \{ f_a : a \in \Sigma \}$, along with a starting vector $\vec{x}_0$, in such a way that ρ_M is completely recovered through the orbits of $\vec{x}_0$ under F. (This is detailed in Section 4.1.) The correspondence also works in reverse, and so we can use an IFS as a proxy for a PFA. In particular, the question "for each n, which string of length n receives the unique highest probability according to M?" is transmuted into the question "for each n, which sequence of compositions of F of length n results in the unique highest value when starting from $\vec{x}_0$?"

When M has two states and reads from the alphabet $\{0, 1\}$, we get a one-dimensional IFS consisting of affine maps $f_0, f_1 \colon [0, 1] \to [0, 1]$ together with a starting value $x_0 \in [0, 1]$. Theorem 4.2 is proved in Section 4.2 by starting with an arbitrary such IFS and undertaking an exhaustive combinatorial analysis of its dynamics in order to classify the strings of each length which receive maximal probabilities. All of these strings are then seen to lie in one of the families in (22). (There is a vast literature on the dynamics of IFSs, but most work is concerned with their asymptotics and properties of their attractors, so is not directly helpful in this endeavor.) With a little extra work, in Proposition 4.10 we also get a characterization of the set of strings whose probabilities are *uniquely* maximal with respect to any given IFS. The structure of this set depends in an essential way on the slopes of the lines f_0 and f_1. For a generic IFS in which both slopes are positive, the maximal-probability strings are of the form $i^n j^m$ for some fixed n and all $m \geq 0$. When both slopes are negative, instead we get maximal-probability strings of the form $i^n (ji)^m$ and $i^n j(ij)^m$ for a fixed n and all m. And when one slope is positive and the other is negative, we can get either $i^n j^m$ for a fixed n and all m, or $i^n j^m i$ for a fixed n and all m. Section 4.2 gives a more detailed summary of the techniques used in the proof and describes the overall breakdown into subcases.

In Section 4.3, we show that for every string w listed in (22) there are affine maps f_0, f_1 on $[0, 1]$ and an $x_0 \in [0, 1]$ such that the IFS (f_0, f_1, x_0) corresponds to a PFA which witnesses that $A_P(w) \leq 2$. This is done in a rather indirect way which exploits the "fixed n and all m" pattern referenced above—that is, roughly speaking, a generic two-state PFA witnesses the complexity of a family $\mathcal{S}$ of strings consisting of a fixed constant prefix followed by any number of repetitions of a fixed string of length 1 or 2, with in some cases a single extra letter tacked onto the end. Hence it is enough to find, for each n and each possible repeated pattern s, a PFA such that the members of $\mathcal{S}$ have prefixes of length n and pattern s. (In other words, one needs only to worry about the "fixed n" since "all m" is automatically satisfied.) To do so, we split the strings in (22) into seven subsets based on s, with each set corresponding to a certain subcase of the proof of Theorem 4.2 in which elements of that set must receive maximal probabilities according to a given IFS. For each of these sets and each n, we prove that there is an IFS meeting exactly the conditions of the corresponding subcase and whose family of maximal-probability strings has a prefix of length n. A detailed summary of how exactly this is accomplished is given at the start of Section 4.3.

4.1. The iterated function system approach

An *iterated function system* (IFS) on a compact metric space X is a dynamical system consisting of a finite set of continuous maps $f_1, \ldots, f_n$ on X, viewed as inducing a semigroup action on X under composition. If X is $\mathbb{R}^n$ or a compact subset of it, and the maps f_i are affine maps, then the IFS is called *affine*. It is well-known that the attractors of contractive IFSs are fractals, and the use of affine IFSs for efficient representation of fractal images has been studied [19, 20, 21].

Our interest in IFSs is, for present purposes, limited to the fact that one may obtain an IFS through the acceptance probability function of a PFA, and in doing so shed light on the family of strings whose complexity the PFA witnesses. Connections between IFSs and PFAs are already known: Culik and Dube [22, 20] in effect use PFAs as one method of generating fractal images, as an alternative to directly employing IFSs. They also introduce probabilistic affine automata, a generalization of PFAs in which each input letter corresponds to an affine map to be applied with some probability. (See [23] for a more recent study of this idea.)

Kocić and Simoncelli in [24] demonstrated a correspondence between IFSs given by a set of stochastic matrices and affine IFSs on lower-dimensional simplices. We present this correspondence in a more elementary formulation adapted to PFAs, showing that the PFA's acceptance probability function descends to the IFS in a natural fashion. Let $M = (S, \Sigma, P, \vec{\pi}, \vec{\eta})$ be a PFA with k states. If there are 0 or k accepting states, then ρ_M is identically 0 or 1, respectively, so assume without loss of generality that there are between 1 and $k-1$ accepting states. By permuting the states of M (and hence the rows and columns of $\vec{\pi}$, $\vec{\eta}$, and each P_σ), we may assume that the kth state is not accepting.

Recall that if $|w| = n$, then $\rho_M(w) = \vec{\pi} P_M(w) \vec{\eta}$, where $P_M(w) = \prod_{i=1}^n P_{w_{n-i}}$. This just means that $\rho_M(w)$ is a sum of up to $k-1$ elements of the row vector $\vec{\pi} P_M(w)$. We can think of each multiplication by a P_σ as updating the state distribution $\vec{\pi}$, and of $\vec{\pi}$ itself as representing the state distribution $\vec{\pi}(\lambda)$ after reading the empty string λ. Then let

$$\vec{\pi}(w) = \left(p_1(w), p_2(w), \ldots, p_{k-1}(w), 1 - \sum_{i<k} p_i(w) \right) = \vec{\pi}(\lambda) P_M(w) \tag{25}$$

be the state distribution after reading a string w. Now, the last component of $\vec{\pi}(w^\frown\sigma)$ only depends on its first $k-1$ components together with the first $k-1$ columns of P_σ. Since the kth state of M is not accepting, $\rho_M(w^\frown\sigma)$ thus depends only on the first $k-1$ components of $\vec{\pi}(w)$, and if we only care about recovering ρ_M then we can drop the kth component from $\vec{\pi}(w)$ without losing any information.

So, let $\vec{a}_i$ be the ith row of P_σ truncated to its first $k-1$ entries, let $\vec{y}(w) = \vec{\pi}(w) \upharpoonright (k-1)$, and let $\vec{1}_{m,n}$ be the $m \times n$ matrix of all 1s. Also let U be P_σ with its last row and column deleted (so the rows of U are the vectors $\vec{a}_i$ for $i < k$). Then for any $\sigma \in \Sigma$,

$$\vec{y}(w^\frown\sigma) = \vec{y}(w) U + \left(1 - \sum \vec{y}(w)\right) \vec{a}_k = \vec{a}_k + \vec{y}(w) \left(U - \vec{1}_{k-1,1} \vec{a}_k\right). \tag{26}$$

$\vec{y}(w)$ is an element of the $(k-1)$-dimensional unit simplex S_{k-1}, so we identify $w \mapsto \vec{\pi}(w^\frown\sigma)$ with the map $f_\sigma \colon S_{k-1} \to S_{k-1}$ that sends $\vec{x}$ to $\vec{a}_k + \vec{x}B$, where $B = U - \vec{1}_{k-1,1}\vec{a}_k$. Note that the entries of B may be negative. Multiplication by P_σ thus corresponds to composition by f_σ. If we give the IFS consisting of the functions f_σ the starting vector $\vec{x}_0 = (p_1(\lambda), p_2(\lambda), \ldots, p_{k-1}(\lambda))$, then we have

$$\rho_M(w) = \sum \big\{ \, (f_{w_n} \circ f_{w_{n-1}} \circ \cdots \circ f_{w_0}(\vec{x}_0))_i : \text{the } i\text{th state of } M \text{ is accepting} \, \big\}, \tag{27}$$

where $\vec{v}_i$ here denotes the ith component of the vector $\vec{v}$ and where $w = w_0 w_1 \cdots w_n$. Hence for any k-state PFA M there is an affine IFS on S_{k-1} whose iterations exactly recover the function ρ_M in the above fashion.

In the other direction, suppose we are given a finite set of affine maps $f_\sigma \colon \vec{x} \mapsto \vec{a} + \vec{x}B$ on S_{k-1}, where $\vec{a}$ and B depend on σ, along with a starting vector $\vec{x}_0 = (p_1, \ldots, p_{k-1})$. We build a PFA M as follows. Let $\tilde{A}$ and $\tilde{B}$ be the $k \times k$ matrices given by $\tilde{A} = \vec{1}_{k,1}\left(\vec{a} \;\middle|\; 1 - \sum \vec{a}\right)$ and

$$\tilde{B} = \begin{pmatrix} \vec{1}_{k-1,1} \\ 0 \end{pmatrix} \left(B \;\middle|\; -B\vec{1}_{k-1,1}\right) = \left(\begin{array}{c|c} \begin{matrix} \\ B \\ \\ \end{matrix} & \begin{matrix} -\sum_{i<k} B_{1,i} \\ \vdots \\ -\sum_{i<k} B_{k-1,i} \end{matrix} \\ \hline \vec{0} & 0 \end{array}\right). \tag{28}$$

Then let

$$P_\sigma = \tilde{A} + \tilde{B} \quad \text{and} \quad \vec{\pi} = \vec{\pi}(\lambda) = \left(\vec{x}_0 \;\middle|\; 1 - \sum \vec{x}_0\right) \in \mathbb{R}^k. \tag{29}$$

Also define $\vec{\pi}(w)$ for any w as before. Then P_σ is stochastic: first, each row clearly sums to 1 as the row sums of $\tilde{A}$ and $\tilde{B}$ are all 1 and 0, respectively. If $\vec{e}_i$ is the ith standard basis vector in $\mathbb{R}^{k-1}$, then $f_\sigma(\vec{e}_i)$ is the sum of $\vec{a}$ and the ith row of B, i.e., the upper left $(k-1) \times (k-1)$ submatrix of P_σ. From $\vec{a} = f_\sigma(\vec{0}) \in S_{k-1}$ and $f_\sigma(\vec{e}_i) \in S_{k-1}$ it follows that each entry of $\tilde{A} + \tilde{B}$ is in $[0,1]$.

One can check that $\vec{\pi}(\lambda)P_\sigma \restriction (k-1) = \vec{a} + \vec{x}_0 B = f_\sigma(\vec{x}_0)$. Inductively we have that $\vec{\pi}(w^\frown\sigma) \restriction (k-1) = f_\sigma(\vec{x})$ if $\vec{x} = \vec{\pi}(\lambda)P_M(w)$. Now, the data we have so far does not uniquely specify a PFA $M = (\{1, \ldots, k\}, \Sigma, \{P_\sigma\}, \vec{\pi}, \vec{\eta})$, because nothing about the vector of accepting states $\vec{\eta}$ is implied by the IFS we started with except that the kth state should not be accepting. Thus the same IFS can be made to correspond to any PFA M having the $\vec{\pi}$ and matrices P_σ given above, and such that the last state is not accepting. The equation (27) holds for any such M and w, which completes the correspondence.

Since we will only apply this correspondence to two-state automata in the present work, we separately outline this case for clarity. Given a two-state PFA M, write $\vec{\pi}$ as $(p, 1-p)$. Assume by permuting the states that $\vec{\eta} = (1, 0)^T$. For each $\sigma \in \Sigma$, write

$$P_\sigma = \begin{pmatrix} a_\sigma + b_\sigma & 1 - a_\sigma - b_\sigma \\ a_\sigma & 1 - a_\sigma \end{pmatrix}, \tag{30}$$

where b_σ is allowed to be negative. Then for each $w \in \Sigma^*$, we have

$$\rho(w^\frown\sigma) = \begin{pmatrix} \rho(w) & 1 - \rho(w) \end{pmatrix} \begin{pmatrix} a_\sigma + b_\sigma & 1 - a_\sigma - b_\sigma \\ a_\sigma & 1 - a_\sigma \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = a_\sigma + b_\sigma \rho(w). \tag{31}$$

We can thus associate with P_σ the "incremental probability function"

$$f_\sigma(x) = a_\sigma + b_\sigma x \tag{32}$$

mapping $[0,1]$ into itself. Viewing p as $\rho(\lambda)$, we obtain the IFS $(f_\sigma)_{\sigma\in\Sigma}$ with starting value $x_0 = p$ such that for any word $w = w_1 w_2 \dots w_n$,

$$\rho(w) = f_{w_n} \circ f_{w_{n-1}} \circ \cdots \circ f_{w_1}(x_0). \tag{33}$$

In the other direction, starting from an IFS given by affine maps f_σ on $[0,1]$ together with x_0, setting $\vec{\pi} = (x_0, 1 - x_0)$ and defining the matrices P_σ as in (30) produces a PFA whose acceptance probability function satisfies (33).

The set of w such that an upper bound for $A_P(w)$ is witnessed by M is exactly the set of w describing a sequence of compositions strictly maximizing the value along the orbit of x_0 under this IFS. This idea will be exploited heavily throughout the following section.

4.2. Proof of Theorem 4.2

We will establish in this section that any two-state PFA over a binary alphabet can only witness the complexity of strings in one of the forms given in the theorem, i.e.,

$$i^n j^m, \qquad i^n j^m i, \qquad i^n (ji)^m, \qquad \text{or} \quad i^n j (ij)^m, \tag{34}$$

if it witnesses anything at all. Permuting the underlying alphabet does not change the complexity of a string, as it corresponds merely to permuting the maps f_a of the IFS, or equivalently the transition matrices P_a of the original PFA. Therefore, any statement in this section about a string should be understood to apply equally well to its bit-flip (i.e., the result of permuting 0 and 1), by switching the roles of f_0 and f_1.

Assume we are given a two-state PFA represented by the IFS

$$f_0(x) = a + bx \qquad \text{and} \qquad f_1(x) = c + dx \tag{35}$$

with starting value $x_0 \in [0,1]$, where f_0 and f_1 map $[0,1]$ into itself. In particular, the latter means that $a, c \in [0,1]$, $b \in [-a, 1-a]$, and $d \in [-c, 1-c]$. We use the word "orbit" to mean any forward orbit of x_0 under the semigroup action generated by f_0 and f_1, that is, the orbit of x_0 under some particular sequence of compositions of f_0 and f_1. We always omit parentheses when composing functions, so that e.g. $f_0^2 x = f_0(f_0(x))$. By convention we define f_i^0 to be the identity map. If an input to a function is not specified in some (in)equality, then it should be taken to mean the (in)equality holds for all inputs, e.g., $f_0 > f_1$ means $f_0 x > f_1 x$ for all x (in $[0,1]$ or in some smaller domain specified in context). For brevity, we describe a probability as *n-maximal* (*n-minimal*) if it is maximal (minimal) among the probabilities of strings of length n. We also refer to an n-maximal (n-minimal) probability as simply an n-maximum (n-minimum). If n is clear from context, we may call such a probability maximal (minimal) or a maximum (minimum). In general, we speak of maxima and minima as though they were unique, but this is not necessary for the argument to work; see item (4) below. We say the IFS witnesses a string w if it witnesses that $A_P(w) \le 2$, i.e., $\rho(w)$ is uniquely maximal. The following elementary observations will be useful throughout:

Fact 4.4. (1) For $i \in \{0,1\}$, if f_i does not coincide with the line $y = x$, then f_i has a unique fixed point in $[0,1]$, towards which it contracts at exponential speed with rate equal to the absolute value of its slope. The fact that this point occurs in $[0,1]$ is a consequence of the assumption that f_0 and f_1 map $[0,1]$ into itself, as they must then intersect $y = x$ there. If one of the maps is $y = x$, or more generally if the two maps commute, then only constant strings can have maximal probability because then the only determining factor of $\rho(w)$ is the number of 0s and 1s in w. For this reason, outside of Lemma 4.5 below, we will always assume that f_0 and f_1 do not commute.

We will use r_0 and r_1 to denote the fixed points of f_0 and f_1, respectively. By abuse of notation, r_0 and r_1 refer either to the x-coordinates of these points or to the actual points in $[0,1]^2$. It will be clear which is meant from the context. We have $r_0 = a/(1-b)$ and $r_1 = c/(1-d)$.

(2) If f_i has positive slope, then it maps $[0, r_i)$ into itself and $(r_i, 1]$ into itself. If it has negative slope, it maps $[0, r_i)$ into $(r_i, 1]$ and vice versa. (If its slope is 0, of course, it sends every point to r_i.)

(3) If a probability x is n-maximal, then x is either the image of an $(n-1)$-maximum under a map of positive slope, or the image of an $(n-1)$-minimum under a map of negative slope. This is simply because when f_i has positive slope, $x < y$ if and only if $f_i x < f_i y$, and if f_i has negative slope then $x < y$ iff $f_i x > f_i y$. Hence we need only consider the maximum and minimum probabilities of each length in order to determine the maximal-probability strings. (Note that this need not be true for PFAs with more than two states. However, it does hold for 2-state PFAs if more maps are involved, i.e., if the size of the alphabet is increased.)

(4) Suppose f_0 and f_1 intersect at the single point $(i_x, i_y) \in [0,1]^2$, and that the maximum or minimum probability of some length turns out to equal i_x. (We always assume the maps do not coincide, since no strings can be witnessed if they do.) Then no further probabilities in the same orbit can be unique, since $f_0 i_x = f_1 i_x$. In this case, no further strings are witnessed if their probabilities are in the same orbit as i_x. We assume for simplicity that this does not happen in the arguments that follow. This does not lose any generality, because if i_x happens to be attained as the maximal or minimal probability in some orbit, then nothing changes about the behavior of the IFS except for the lack of uniqueness of the subsequent maxima and minima. (See also Proposition 4.10 below.)

Fact 4.4(3) is really the key to the whole proof of Theorem 4.2. The overall argument will need to be split into several major cases, but within each case, we have a kind of inductive structure wherein for each n, the $(n+1)$-maximum and $(n+1)$-minimum strings are obtained by appending a single letter to the n-maximum and/or n-minimum. Which letter to append to which string is determined using a set of inequalities amounting to a description of the relative effects on $\rho(w)$ of appending different strings of lengths 1, 2, and 3. For example, "$f_1 f_0 < f_0 f_1$" should be interpreted as meaning that appending 10 to a string always results in a higher probability than appending 01 to the same string. These inequalities, among other things, are collected in the following technical lemma:

Lemma 4.5. Let $f_0 = a + bx$ and $f_1 = c + dx$ be maps from $[0, 1]$ into itself. Assume that $a \leq c$, that $b > d$, and that the maps intersect at the unique point (i_x, i_y), not necessarily in $[0, 1]^2$. Observe that if we did have $(i_x, i_y) \in [0, 1]^2$, then $a \leq c$ would already imply $b > d$.

(a) If neither map is the identity, then either both maps fix i_x, both maps strictly decrease i_x, or both maps strictly increase i_x.

(b) Both maps fix i_x, i.e., $i_y = i_x$, iff $r_0 = r_1 = i_x$ iff $f_0 f_1 = f_1 f_0$.

(c) Both maps decrease i_x, i.e., $i_y < i_x$, iff $r_0 < r_1 < i_x$ iff $f_0 f_1 < f_1 f_0$ iff $f_0 f_1^2 < f_1^2 f_0$.

(d) Both maps increase i_x, i.e., $i_y > i_x$, iff $r_0 > r_1 > i_x$ iff $f_1 f_0 < f_0 f_1$ iff $f_1^2 f_0 < f_0 f_1^2$.

(e) If f_0 and f_1 both have negative slopes, and if neither fixes i_x, then $|r_0 - r_1| < |r_1 - i_x|$.

(f) Suppose f_0 and f_1 both have negative slopes. If both maps decrease i_x, then if $x \in [r_0, i_x)$, every orbit of x remains inside $(-\infty, i_x)$. If both maps increase i_x, then if $x \in (i_x, r_0]$, every orbit of x remains inside (i_x, ∞).

Proof:

(a) By definition, i_x is the unique value of x such that $f_0 x = f_1 x$.

(b) The first equivalence is immediate from the definition of r_i. For the second equivalence, notice $f_0 f_1 x = a + bc + bdx$ and $f_1 f_0 x = c + ad + bdx$. Then

$$f_0 f_1 = f_1 f_0 \iff a + bc = c + ad \iff a/(1 - b) = c/(1 - d), \tag{36}$$

i.e., $f_0 f_1 = f_1 f_0$ iff $r_0 = r_1$, and this happens iff they both equal i_x since r_0 and r_1 both lie on the line $y = x$.

(c) Both r_0 and r_1 are less than i_x in this case, because the maps contract towards their fixed points, so if $f_i x < x$ then $x > r_i$. We have $i_x = (c - a)/(b - d)$ and $i_y = (bc - ad)/(b - d)$. Remembering that our assumptions imply $b > d$ no matter the sign of each, and observing that neither b nor d equals 1, we have $i_x > i_y$ iff

$$\begin{aligned} \frac{c - a}{b - d} > \frac{bc - ad}{b - d} &\iff c - a > bc - ad \iff c(1 - b) > a(1 - d) \\ &\iff \frac{c}{1 - d} > \frac{a}{1 - b} \iff r_1 > r_0. \end{aligned} \tag{37}$$

Since $r_1 > r_0 \iff c + ad > a + bc \iff f_1 f_0 > f_0 f_1$, this also completes the second equivalence. For the third, note $f_0 f_1^2 x = a + b(c + cd + d^2 x)$ and $f_1^2 f_0 x = c + cd + d^2(a + bx)$. Then

$$\begin{aligned} f_0 f_1^2 x < f_1^2 f_0 x &\iff a + bc + bcd + bd^2 x < c + cd + ad^2 + bd^2 x \\ \iff d(bc - ad) < (c - a) - c(b - d) &\iff c + d\frac{bc - ad}{b - d} < \frac{c - a}{b - d} \\ \iff c + d i_y < i_x &\iff f_1 i_y < i_x \iff f_1^2 i_x < i_x. \end{aligned} \tag{38}$$

The last inequality holds iff $r_1 < i_x$ (as f_1^2 contracts i_x towards its fixed point r_1), which happens iff $i_x > i_y$ by the first equivalence and by part (a).

(d) This follows by swapping ">" with "<" everywhere in the argument for part (c).

(e) By writing $c = r_1(1 - d)$, one can rearrange the formula $i_y = c + di_x$ to obtain $d = (i_y - r_1)/(i_x - r_1)$. Since $|d| \leq 1$, this implies $|i_y - r_1| \leq |i_x - r_1|$. We will finish the proof by showing that when $b < 0$, then in fact $i_x > i_y$ iff $i_y < r_0$ and $i_x < i_y$ iff $i_y > r_0$. That is, depending on whether both maps decrease or increase i_x, we have either $i_y < r_0 < r_1 < i_x$ or $i_y > r_0 > r_1 > i_x$. Then

$$|i_y - r_1| = |i_y - r_0| + |r_0 - r_1| > |r_0 - r_1|, \tag{39}$$

and one gets $|r_0 - r_1| < |i_y - r_1| \leq |i_x - r_1|$.

So, $i_y < r_0$ if and only if

$$\begin{aligned} &\frac{bc - ad}{b - d} < \frac{a}{1 - b} \iff (bc - ad)(1 - b) < a(b - d) \iff bc - b^2c + abd < ab \\ &\iff bc(1 - b) < ab(1 - d) \iff c(1 - b) > a(1 - d) \iff \frac{c}{1 - d} > \frac{a}{1 - b}, \end{aligned} \tag{40}$$

i.e., if and only if $r_1 > r_0$, which is equivalent to $i_x > i_y$. (The change from $<$ to $>$ in the second line is because $b < 0$.) It is clear that one can switch "<" and ">" everywhere in this argument to obtain that $i_y > r_0$ iff $i_x < i_y$, and the proof is complete.

(f) For the first claim, by assumption $i_y < i_x$ and so $r_0 < r_1 < i_x$. Since $f_0 i_x = f_1 i_x$, we have $f_0 f_1 i_x = f_0 f_0 i_x$, which is less than i_x. This implies that r_{01}, the fixed point of $f_0 f_1$, is also less than i_x: if $f_0 f_1$ decreases the value of a point, then that point must be above r_{01}. As $f_0 f_1$ contracts to r_{01}, we have that $f_0 f_1 x < i_x$ whenever $x < i_x$. In other words, if $\rho(w) < i_x$, then $\rho(w^\frown 10) < i_x$ too. The analogous statement holds for $f_1 f_0$, i.e., $\rho(w) < i_x$ implies $\rho(w^\frown 01) < i_x$. Finally, since $|r_0 - r_1| < |r_1 - i_x|$ by part (e), f_1 always sends points in $[r_0, i_x)$ to points below i_x (and of course the same statement is clearly true for f_0). This is clear if $x \in [r_1, i_x)$. If $x \in [r_0, r_1)$, then $|f_1 x - r_1| < |x - r_1| < |r_0 - r_1| < |i_x - r_1|$, so $f_1 x$ is closer to r_1 than i_x is, and must be less than i_x. Overall, then, we have that once an orbit enters $[r_0, i_x)$, it stays below i_x. The second claim can be proven by switching "<" and ">" everywhere in the above argument.

□

An important feature of the analysis in Lemma 4.5 is that under the stated assumptions on f_0 and f_1, they can lie in one of only three possible configurations: intersecting at their shared fixed point, intersecting above both of their fixed points (in which case $r_0 < r_1$), and intersecting below both of their fixed points (in which case $r_1 < r_0$). Since we are assuming the maps do not commute, the first configuration is automatically ruled out, so we have that either both the maps increase i_x or both decrease it—and Lemma 4.5 furnishes each of these cases with very specific information about the behavior of f_0 and f_1. Along with Fact 4.4(3), then, this dichotomy goes a long way towards minimizing the number of possible subcases that must be considered in the proof.

Now begins the main body of the proof of Theorem 4.2. It is split into four cases: the maps do not intersect, they intersect and have positive slope, they intersect and have negative slope, and they intersect with one having positive and the other having negative slope. The last three cases are each split into two further subcases, based on whether the maps both increase or both decrease i_x.

Case 1: f_0 and f_1 do not intersect in $(0, 1)$. Suppose without loss of generality that $f_1x > f_0x$ for all $x \in (0, 1)$, so that $a \leq c$. If both maps have nonnegative slope, it follows that $\rho(1^n)$ is maximal for all n. If both have negative slope, then appending 0 to a maximal probability always leads to a minimal probability, and appending 1 to a minimal probability always leads to a maximal probability. Therefore $(01)^n$ and $1(01)^n$ are maximal for all n, as $f_0x_0 \leq f_1x_0$, with strict inequality if $x_0 \in (0, 1)$. If f_1 has positive and f_0 has negative slope, 1^n is maximal for all n; note the ranges of f_0 and f_1 cannot overlap here, except possibly at one point if $i_x = 0$. And if f_1 has negative and f_0 positive slope, then 0^n1 is maximal for all n: since $f_0 < f_1$ (except possibly at one point if $i_x = 1$), every maximal-probability string must end with a 1, so that its probability is the image under f_1 of a minimal probability. A minimum can only be reached by a string of all 0s.

Strings witnessed in this case: 1^n, $(01)^n$, $1(01)^n$, 0^n1.

Assumption 4.6. From now until the end of the proof of Theorem 4.2, we always make the following assumptions:

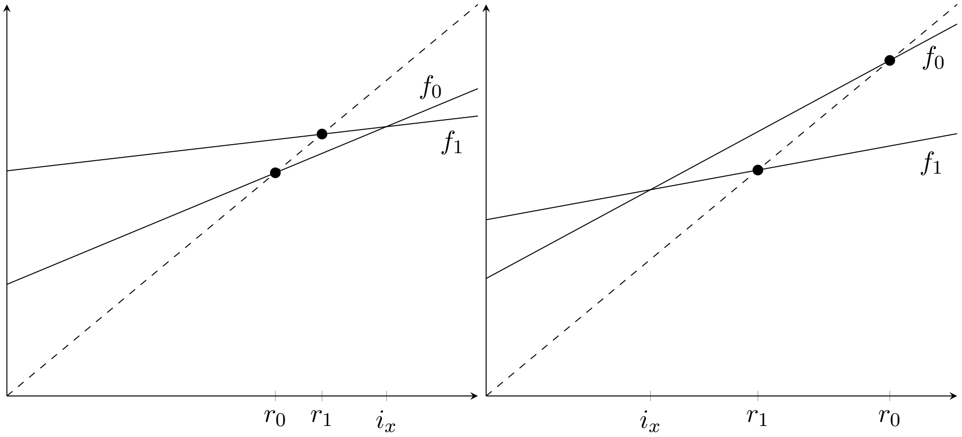
- $a < c$.
- f_0 and f_1 intersect at the unique point $(i_x, i_y) \in (0, 1)^2$.
- f_0 and f_1 do not commute.

As noted in Lemma 4.5, the first two items taken together imply that $b > d$ and that r_0, r_1, and i_x are all distinct. Observe that if $a = c$ then $i_x = 0$, which falls under Case 1 above, and so taking $a < c$ is no loss of generality here since we are working only up to permuting f_0 and f_1.

Case 2: both f_0 and f_1 have nonnegative slopes. Specifically, assume $a, b, c \geq 0$ and $d > 0$ (the case $b = d = 0$ is trivial), as well as Assumption 4.6. There are then two possible subcases of this case, which are illustrated in Figure 3:

(a) Both f_0 and f_1 decrease i_x, or in other words (i_x, i_y) lies below the line $y = x$. Then we must have $r_0 < r_1 < i_x$, and $0^{n_0}1^m$ is witnessed for all m, where $n_0 \geq 0$ is least such that $f_0^{n_0}x_0 < i_x$ (taking f_0^0 to be the identity map). This is because $f_0x > f_1x$ for $x > i_x$, but iterating it will eventually cause the value to drop below i_x, and from that point on, $f_1 > f_0$. We also witness 0^ℓ for $\ell \leq n_0$. If $x_0 < i_x$ then we have $f_1 > f_0$ from the start.

(b) Both f_0 and f_1 increase i_x, i.e., (i_x, i_y) lies above the line $y = x$. Then $i_x < r_1 < r_0$, and $1^{n_0}0^m$ is witnessed for all m, where $n_0 \geq 0$ is least such that $f_1^{n_0}x_0 > i_x$. The reasoning is exactly analogous to that in (a).



(a) Both maps decrease i_x, i.e., (i_x, i_y) lies below the line $y = x$

(b) Both maps increase i_x, i.e., (i_x, i_y) lies above the line $y = x$

Figure 3: Subcases for f_0, f_1 with positive slope (Case 2)

Strings witnessed in the above case: 1^ℓ, $0^n 1^m$, $1^n 0^m$.

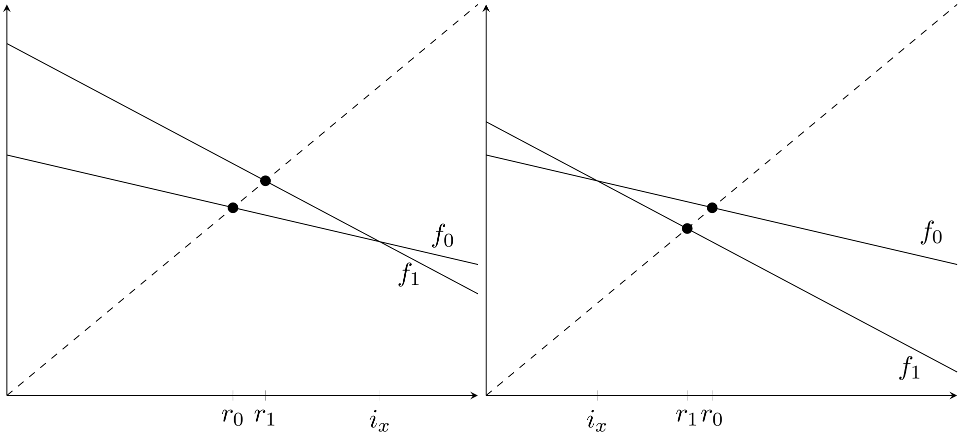
Case 3: both f_0 and f_1 have negative slopes. As before we also work under Assumption 4.6. The special case $b = 0$ and $d < 0$ is discussed under Case 4 below, so assume b and d are both strictly negative here. Recall that having negative slopes means each f_i "flips x over" r_i: if $x < r_i$, then $f_i x > r_i$, and vice versa. If we start an orbit with a maximal probability of length 1, then the orbit can only lead to maximal probabilities for odd-length strings (and this is the only way to witness odd-length strings). This is accomplished by extending a string on even lengths in order to achieve a *minimal* probability. For the same reason, if we start an orbit with a minimal probability, then only even-length strings may have maximal probabilities in that orbit. The two essentially different cases for the configuration of f_0 and f_1 are shown in Figure 4.

(a) Both f_0 and f_1 decrease i_x. Then $r_0 < r_1 < i_x$. Recall that by Lemma 4.5(f), once an orbit enters $[r_0, i_x)$, it stays below i_x forever.

Suppose $x_0 > i_x$ and we start an orbit with the maximal-probability string 0. Then $\rho(0) < i_x$, and $\rho(0^2)$ is minimal. If $\rho(0^2) > i_x$, then $\rho(0^3)$ is maximal, since $f_0 > f_1$ above i_x. Every $(2\ell + 1)$-maximal probability is an image of a 2ℓ-minimal probability, so as long as $\rho(0^{2\ell}) > i_x$, we have that $\rho(0^{2\ell+1})$ is maximal and $\rho(0^{2\ell+2})$ is minimal. Let n_0 be least such that $\rho(0^{2n_0}) < i_x$. Once this happens, since $\rho(0^{2n_0})$ is minimal and $f_0 < f_1$ below i_x, $\rho(0^{2n_0}1)$ is maximal.

$\rho(0^{2n_0})$ is between r_0 and i_x, so we know that its future orbit will always stay below i_x by Lemma 4.5(f). This means that a maximum is always reached by appending 1 to a minimum, and a minimum is always reached by appending 0 to a maximum. Hence, among odd-length strings with length greater than $2n_0 + 1$, we witness $0^{2n_0}1(01)^m$ for all $m \geq 0$.

Next, say $x_0 > i_x$ and we start our orbit with the minimal-probability string 1. Then $\rho(11)$ is maximal. If $\rho(11) > i_x$, then $\rho(1^3)$ is minimal and $\rho(1^4)$ is maximal. So we initially witness $1^{2\ell}$ for



(a) Both maps decrease i_x, i.e., (i_x, i_y) lies below the line $y = x$

(b) Both maps increase i_x, i.e., (i_x, i_y) lies above the line $y = x$

Figure 4: Subcases for f_0, f_1 with negative slope (Case 3)

$\ell \leq n_0$, where n_0 is least such that $\rho(1^{2n_0}) < i_x$. Among longer strings, we then witness $1^{2n_0}(01)^m$ for all m. To see this, one argues in a similar way as when $x_0 < i_x$, since $f_1 f_0 x < i_x$ when $x \in [r_0, i_x)$. The only difference is that once $\rho(1^{2n_0}) < i_x$, the $(2n_0 + 1)$-minimal probability is attained by $\rho(1^{2n_0}0)$, as $f_0 x < f_1 x$ for $x < i_x$. Then appending a 1 gives the $(2n_0 + 2)$-maximum $\rho(1^{2n_0}01)$, and we continue appending 01 to keep the min-max pattern going and get $(2n_0 + 2k)$-maximal probabilities for all k. This concludes the subcase $x_0 > i_x$.

Finally, suppose $x_0 < i_x$. This is analogous to the case $x_0 > i_x$, but with even-odd parity swapped everywhere. In fact, we only need consider the case $x_0 < r_0$, because when $x_0 \in [r_0, i_x)$, we know that all orbits stay below i_x, so for such x_0 we witness $(10)^m$ and $0(10)^m$ for all $m \geq 0$.

Now, if $x_0 < r_0$ and we start with the maximal-probability string 1, we at first witness $1^{2\ell+1}$ among odd-length strings, as long as $\rho(1^{2\ell+1}) > i_x$. If n_0 is least such that $\rho(1^{2n_0+1}) < i_x$, then we witness 1^{2n_0+1} and subsequently $1^{2n_0+1}(01)^m$ for all $m \geq 1$. This is because once $\rho(1^{2n_0+1}) < i_x$, then the $(2n_0 + 2)$-minimum is $\rho(1^{2n_0+1}0)$, followed by the $(2n_0 + 3)$-maximum $\rho(1^{2n_0+1}01)$, and continuing to append 01 keeps the min-max pattern going. Starting instead with the minimal-probability 0, we witness $0^{2\ell+2}$ among even-length strings as long as $\rho(0^{2\ell+1}) > i_x$. If n_0 is least such that $\rho(0^{2n_0+1}) < i_x$, then $\rho(0^{2n_0+1})$ is minimal but $\rho(0^{2n_0+2})$ is not maximal. Therefore $\rho(0^{2n_0+1}1)$ must be maximal, and we witness $0^{2n_0+1}1(01)^m$ for all $m \geq 0$. The pattern of appending 01 can be repeated forever to obtain maximal probabilities because $\rho(0^{2n_0+1}1) \in [r_0, i_x)$ and applying $f_1 f_0$ will always stay below i_x, where f_0 is minimal and f_1 is maximal.

Strings witnessed in the above case: 0^{2m}, 1^{2m+1}, $0^{2n}1(01)^m$, $1^{2n}(01)^m$, $0^{2n+1}1(01)^m$.

(b) Both f_0 and f_1 increase i_x. Then $i_x < r_1 < r_0$. By Lemma 4.5(f), once an orbit enters $(i_x, r_0]$, it stays above i_x.

Let $x_0 < i_x$. By starting with the maximal $\rho(1)$, we have that $\rho(1^{2\ell+1})$ is maximal as long as $\rho(1^{2\ell}) < i_x$. (Note that $\rho(1^{2\ell+1})$ is always greater than r_1 and hence also i_x.) If n_0 is least such that $\rho(1^{2n_0}) > i_x$, then $\rho(1^{2n_0})$ is $2n_0$-minimal but $\rho(1^{2n_0+1})$ is not $(2n_0+1)$-maximal. Therefore $\rho(1^{2n_0}0)$ is $(2n_0+1)$-maximal, and since $\rho(1^{2n_0}) \in (i_x, r_0]$, we have that $\rho(1^{2n_0}0(10)^m)$ remains above i_x for all $m \geq 0$ and is therefore maximal. By starting instead with the minimal $\rho(0)$, we have that $\rho(0^{2\ell+2})$ is maximal as long as $\rho(0^{2\ell}) < i_x$. If n_0 is least such that $\rho(0^{2n_0}) > i_x$, then $\rho(0^{2n_0})$ is $2n_0$-maximal but $\rho(0^{2n_0+1})$ is not $(2n_0+1)$-minimal, since $f_0 > f_1$ for $x > i_x$. Therefore $\rho(0^{2n_0}1)$ is minimal, and because $\rho(0^{2n_0}) \in (i_x, r_0]$, its future orbits stay above i_x and we have $\rho(0^{2n_0}(10)^m)$ maximal for all $m \geq 0$.

If $x_0 \in (i_x, r_0]$, then we witness $(10)^m$ and $0(10)^m$ for all $m \geq 0$, as in case (a) when $x_0 \in [r_0, i_x)$. If $x_0 > r_0$ and we start with the maximal $\rho(0)$, then $\rho(0^{2\ell+3})$ is maximal and $\rho(0^{2\ell+2})$ is minimal as long as $\rho(0^{2\ell+1}) < i_x$. If n_0 is least such that $\rho(0^{2n_0+1}) > i_x$, then $\rho(0^{2n_0+1})$ is $(2n_0+1)$-maximal but $\rho(0^{2n_0+2})$ is not $(2n_0+2)$-minimal. Instead, $\rho(0^{2n_0+1}1)$ is $(2n_0+2)$-minimal, and since $\rho(0^{2n_0+1}) \in (i_x, r_0]$, all future orbits stay above i_x, where $f_0 > f_1$. Therefore $\rho(0^{2n_0+1}(10)^m)$ is maximal for all $m \geq 0$. If instead we start with the minimal $\rho(1)$, then $\rho(1^{2\ell+2})$ is maximal as long as $\rho(1^{2\ell+1}) < i_x$. If n_0 is least such that $\rho(1^{2n_0+1}) > i_x$, then $\rho(1^{2n_0+1})$ is $(2n_0+1)$-minimal but $\rho(1^{2n_0+2}) < \rho(1^{2n_0+1}0)$, which is now $(2n_0+2)$-maximal. Since $\rho(1^{2n_0+1}) \in (i_x, r_0]$, all of its future orbits stay above i_x, and thus $\rho(1^{2n_0+1}0(10)^m)$ is maximal for all $m \geq 0$.

Strings witnessed in the above case: 1^{2m+1}, $1^{2n}0(10)^m$, $0^{2n+1}(10)^m$, $1^{2n+1}0(10)^m$.

Case 4: f_0 has positive slope and f_1 has negative slope. The basic possibilities are illustrated in Figure 5. As before, we also make Assumption 4.6.

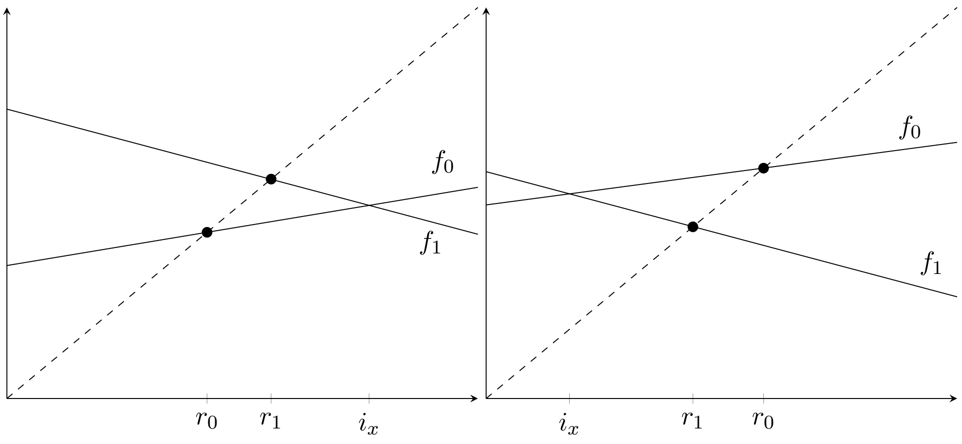
(a) Both f_0 and f_1 decrease i_x. Lemma 4.5 implies that this is equivalent to $f_1 f_0 x > f_0 f_1 x$ for all x, so that appending 01 always gives a higher probability than appending 10 would to the same string. Assume for now that $b > 0$; we will treat the special case $b = 0$ below. The general pattern when $x_0 > i_x$ will follow from the next three claims:

Claim 4.7. There is an n_0 such that $\rho(1^{2n_0+1}) \geq \rho(1^{2n_0-1}0^2)$.

Proof:
The map f_1 contracts to r_1, which is greater than r_0. Therefore $\rho(1^{2n+1}) \geq r_0$ for some n. If that is the case, then either $\rho(1^{2n-1}) \leq r_0$ and so is $\rho(1^{2n-1}0^2)$, or if $\rho(1^{2n-1}) \geq r_0$, then appending 0^2 to 1^{2n-1} decreases the probability towards r_0 while appending 1^2 increases it towards r_1. □

From now on, take n_0 to be the least value as in the previous claim. If $x_0 > i_x$, then 0 is 1-maximal, and it follows that $n_0 \geq 1$. If $\rho(1^{2n_0+1}) = \rho(1^{2n_0-1}0^2)$, then both are minimal, so no further strings will be witnessed as there are no longer unique minima or maxima of any greater length. Hence without loss of generality assume the inequality is strict. The second and third claims will apply to the case $n_0 > 1$; the case $n_0 = 1$ is handled separately afterwards.

Claim 4.8. If $n_0 > 1$, then for all $1 \leq \ell \leq n_0$, we have:



(a) Both maps decrease i_x, i.e., (i_x, i_y) lies below the line $y = x$

(b) Both maps increase i_x, i.e., (i_x, i_y) lies above the line $y = x$

Figure 5: Subcases for f_0 with positive and f_1 with negative slope (Case 4)

(α) $\rho(1^{2\ell})$ is 2ℓ-maximal,

(β) $\rho(1^{2\ell-1}0)$ is 2ℓ-minimal,

(γ) $\rho(1^{2\ell-1}01)$ is $(2\ell+1)$-maximal, and

(δ) if $\ell < n_0$, then $\rho(1^{2\ell+1})$ is $(2\ell+1)$-minimal.

Proof:
By induction on ℓ. For $\ell = 1$, because $\rho(0) > \rho(1)$, the only possible 2-maxima are $\rho(00)$ and $\rho(11)$. If we have $\rho(11) < \rho(00)$, then because $f_1 x < f_1 y$ iff $x > y$ for any x, y, also $f_1^3 x_0 > f_1 f_0^2 x_0$. But from $f_0 f_1 < f_1 f_0$ it follows that $f_1 f_0^2 x_0 > f_0 f_1 f_0 x_0 > f_0^2 f_1 x_0 = \rho(10^2)$, as $f_0 x < f_0 y$ iff $x < y$ for any x, y. (The latter is due to f_0 having positive slope.) Therefore $\rho(1^3) > \rho(10^2)$, or in other words $n_0 = 1$. Since we are assuming $n_0 > 1$, this is a contradiction, hence $\rho(00) < \rho(11)$ and the latter is 2-maximal, which establishes the base case of (α).

Next, because $f_1 f_0 > f_0 f_1$, we have $\rho(01) > \rho(10)$. The latter is less than $\rho(00)$: $\rho(1) < \rho(0)$, so if $\rho(1) < r_0$ then $\rho(10) < r_0 < \rho(00)$. If $\rho(1) \geq r_0$, then appending a 0 moves $\rho(10)$ closer to r_0 than $\rho(00)$ is, i.e., makes it smaller than $\rho(00)$. This implies (β) holds for $\ell = 1$. Then (γ) follows if (α) and (β) hold for any ℓ: the only possible candidates for a $(2\ell+1)$-maximum are $\rho(1^{2\ell}0)$ and $\rho(1^{2\ell-1}01)$, i.e., the image of the 2ℓ-maximum under f_0 and the image of the 2ℓ-minimum under f_1. But $\rho(1^{2\ell}0) < \rho(1^{2\ell-1}01)$ since $f_0 f_1 < f_1 f_0$. For (δ), suppose (α) and (β) are true for ℓ and $\ell < n_0$. Then only $\rho(1^{2\ell-1}0^2)$ or $\rho(1^{2\ell+1})$ could possibly be minima, since they are the images of the 2ℓ-minimum under f_0 and the 2ℓ-maximum under f_1, respectively. And we have $\rho(1^{2\ell+1}) < \rho(1^{2\ell-1}0^2)$ because $\ell < n_0$.

Now suppose all four items hold for some given $\ell < n_0$. Then if (α) and (β) hold for $\ell + 1$, so does (γ) by the above argument, and if $\ell + 1 < n_0$ then additionally (δ) holds for $\ell + 1$. Hence, for the inductive step, it only remains to establish that (α) and (β) hold for $\ell + 1$. For (α), because a $(2\ell + 2)$-maximum is either the image under f_1 of a $(2\ell + 1)$-minimum or the image under f_0 of a $(2\ell + 1)$-maximum, the only possible $(2\ell + 2)$-maxima are $\rho(1^{2\ell+1}1)$ and $\rho(1^{2\ell-1}010)$. But we have $\rho(1^{2\ell-1}010) < \rho(1^{2\ell-1}001)$ because $f_0 f_1 < f_1 f_0$, so $\rho(1^{2\ell-1}010)$ is not maximal. Finally, for (β), $\rho(1^{2\ell+1}0)$ is $(2\ell + 2)$-minimal because the only other possible candidate for a minimum is $\rho(1^{2\ell-1}011)$, and $\rho(1^{2\ell+1}0)$ is less than $\rho(1^{2\ell-1}011)$. The latter follows by Lemma 4.5, as $i_x > i_y$ if and only if $f_0 f_1^2 < f_1^2 f_0$, so that appending 110 always results in a lower probability than appending 011 to the same string. □

Claim 4.9. If $n_0 > 1$, then $\rho(1^{2n_0-1}0^m)$ is $(2n_0 - 1 + m)$-minimal, and hence $\rho(1^{2n_0-1}0^m1)$ is $(2n_0 + m)$-maximal, for all $m \geq 0$.

Proof:
The cases $m = 0$ and $m = 1$ are covered by taking $\ell = n_0 - 1$ and $\ell = n_0$ in the previous claim. If $\rho(1^{2n_0-1}0^m)$ is $(2n_0 - 1 + m)$-minimal, and $\rho(1^{2n_0-1}0^{m-1}1)$ is $(2n_0 - 1 + m)$-maximal, then only $\rho(1^{2n_0-1}0^m0)$ or $\rho(1^{2n_0-1}0^{m-1}11)$ could be $(2n_0 + m)$-minimal. But $f_0 f_1^2 < f_1^2 f_0$ implies $\rho(1^{2n_0-1}0^{m-2}110) < \rho(1^{2n_0-1}0^{m-1}11)$, so the latter is not minimal. Finally, this implies $\rho(1^{2n_0-1}0^{m+1}1)$ is $(2n_0 + m + 1)$-maximal: the only other possibility is $\rho(1^{2n_0-1}0^{m-1}10)$, and this is less than $\rho(1^{2n_0-1}0^{m+1}1)$ because $f_0 f_1 < f_1 f_0$. □

Now suppose $n_0 = 1$. We saw in the proof of Claim 4.8 above that $\rho(11) < \rho(00)$ implies $n_0 = 1$, but a priori both $\rho(11) < \rho(00)$ and $\rho(00) < \rho(11)$ are possible when $n_0 = 1$. Note that $\rho(10)$ is always minimal, however. The possible 3-minima are $\rho(1^3)$ (if $\rho(11)$ is maximal), $\rho(0^21)$ (if $\rho(00)$ is maximal), and $\rho(10^2)$ (in either case). But $\rho(1^3) > \rho(1^20)$ by $n_0 = 1$ and $\rho(0^21) > \rho(10^2)$ because $f_1 f_0^2 > f_0^2 f_1$, as observed in the base case of Claim 4.8. Hence $\rho(10^2)$ is always the 3-minimum when $n_0 = 1$. We now split into two final subcases to finish the argument when $x_0 > i_x$ and $n_0 = 1$. First, assume $\rho(00) < \rho(11)$, so $\rho(11)$ is maximal. For any $m \geq 2$, if $\rho(10^{m-1}1)$ is maximal and $\rho(10^m)$ is minimal, the next maximum is $\rho(10^m1)$ since the other possibility is $\rho(10^{m-1}10)$, which is of the form $f_0 f_1 y$ for some y, and $f_0 f_1 y < f_1 f_0 y$. And in this case the next minimum is $\rho(10^{m+1})$, because the other option is $\rho(10^{m-1}1^2)$, which is of the form $f_1^2 f_0 y$ hence greater than $f_0 f_1^2 y$. So by induction $\rho(10^m1)$ is witnessed for all m if $\rho(11)$ is maximal.

The remaining subcase of $x_0 > i_x$ is $n_0 = 1$ and $\rho(11) < \rho(00)$. In general, it may be that $\rho(0^{\ell-1})$ is maximal for finitely many $\ell \geq 3$, but this cannot be the case for all ℓ (as we assume $b < 1$) because $\rho(0^\ell)$ decreases to r_0 as ℓ increases, while some probabilities of every length will be greater than r_1. Suppose $\rho(0^{\ell-1})$ is maximal and (by induction) $\rho(10^{\ell-2})$ is minimal, for $\ell \geq 3$. Then either $\rho(0^\ell)$ or $\rho(10^{\ell-2}1)$ is ℓ-maximal, and $\rho(10^{\ell-1})$ is always ℓ-minimal (by the same argument as in the last paragraph). If $\rho(10^{\ell-2}1)$ is maximal, then the argument in Claim 4.9 takes over from length $\ell + 1$ onwards. If $\rho(0^\ell)$ is maximal, then $\rho(10^\ell)$ is $(\ell + 1)$-minimal because the other option is $\rho(0^\ell 1) > \rho(0^{\ell-1}10)$. The argument then repeats for $\ell + 1$, and so on, meaning we witness 0^ℓ for finitely many ℓ and then 10^m1 for all large enough m.

This completes the argument when $x_0 > i_x$. If instead $x_0 < i_x$, then something similar happens, but with even-odd parities switched. We state without detailed proofs the three claims (corresponding to those above) that will finish the argument here, as their proofs follow in the same way *mutatis mutandis*. First, there is a least n_0 such that $\rho(1^{2n_0+2}) > \rho(1^{2n_0}0^2)$. The case $n_0 = 0$ is separate and exactly analogous to the case $n_0 = 1$ when $x_0 > i_x$: if $n_0 = 0$ then $\rho(11) > \rho(00)$, so $\rho(00)$ is minimal, $\rho(01)$ is maximal, and in general we have 0^m minimal and $0^{m-1}1$ maximal for all $m \geq 2$. So assume $n_0 > 0$ from now on.

The second claim is that when $n_0 > 0$ and $x_0 < i_x$, for all $1 \leq \ell \leq n_0$, $\rho(1^{2\ell-2}01)$ is 2ℓ-maximal; $\rho(1^{2\ell})$ is 2ℓ-minimal; $\rho(1^{2\ell+1})$ is $(2\ell+1)$-maximal; and $\rho(1^{2\ell}0)$ is $(2\ell+1)$-minimal. Both the base case and the inductive step work very similarly as before (here, $\rho(11)$ being minimal relies on $n_0 > 0$). The only possible $(2\ell+2)$-maxima are $\rho(1^{2\ell}01)$ and $\rho(1^{2\ell+1}0)$; the only possible $(2\ell+2)$-minima are $\rho(1^{2\ell+2})$ and $\rho(1^{2\ell}0^2)$; the only possible $(2\ell+3)$-maxima are $\rho(1^{2\ell+3})$ and $\rho(1^{2\ell}010)$; and the only possible $(2\ell+3)$-minima are $\rho(1^{2\ell+2}0)$ and $\rho(1^{2\ell-2}01^2)$. All the alternatives listed can be dispensed with using $f_0f_1 < f_1f_0$, $f_0f_1^2 < f_1^2f_0$, and $\ell < n_0$. (The latter is only needed to show the claim holds for $\ell + 1$ given it holds for ℓ, and so it does hold for $\ell = n_0$ as stated.)

The third and last claim needed is that when $n_0 > 0$ and $x_0 < i_x$, for all $m \geq 1$, we have $\rho(1^{2n_0}0^m)$ minimal and $\rho(1^{2n_0}0^{m-1}1)$ maximal. The case $m = 1$ follows by taking $\ell = n_0$ in the previous claim. The inductive step is again very similar to Claim 4.9 for $x_0 > i_x$, since the other possible minimum is $\rho(1^{2n_0}0^{m-1}1^2)$, which is of the form $f_1^2f_0y$ and hence not minimal since it is greater than $f_0f_1^2y$. The other possible maximum is $\rho(1^{2n_0}0^{m-1}10)$, of the form f_0f_1y, which is less than f_1f_0y and hence not maximal.

To finish the argument for Case 4(a), we dispense with the special case when $b = 0$, i.e., when f_0 is constant. (The subcase where instead f_1 is constant was dealt with in Case 2.) Here $r_0 = a$, and $\rho(w^\frown 01) = f_1a$ for all strings w. If $x_0 > i_x$, then proceeding as above, we see that after applying f_1 some number of times, if n is least such that $\rho(1^{2n+1}) \geq a$, then it is not possible for the probability of any string with length greater than $2n$ to exceed f_1a. This means that maximal probabilities cease to be unique at length $2n+1$, and only finitely many strings can be witnessed. The same holds when $x_0 < i_x$, but now the maxima cease to be unique after $\rho(1^{2n}) \geq a$.

Strings witnessed in the above case: $1^{2n-1}0^m1$, $1^{2n}0^m1$, 0^m, 0^m1.

(b) Both f_0 and f_1 increase i_x. This is equivalent to $f_0f_1x > f_1f_0x$ for all x, so that appending 10 always gives a higher probability than appending 01. It is also equivalent to $f_0f_1^2x > f_1^2f_0x$ for all x (see Lemma 4.5). As before, we put off the special case $b = 0$ for later, and assume for the moment that $b > 0$.

First, say that $x_0 < i_x$. Here we need to split into slightly different subcases than we did in (a). Since $\rho(0) < \rho(1)$, the 2-maximum is always $\rho(10)$ because the only other option is $\rho(01) < \rho(10)$. The possible 2-minima are $\rho(00)$ and $\rho(11)$, and both $\rho(00) < \rho(11)$ and $\rho(11) < \rho(00)$ are possible. Suppose first that $\rho(00) < \rho(11)$. It may be that $\rho(0^\ell)$ is minimal for finitely many ℓ, but eventually this is no longer the case since $\rho(0^\ell)$ increases to r_0 while some other probabilities always stay below r_1. Suppose that for some $\ell \geq 2$, $\rho(0^\ell)$ is minimal and $\rho(10^{\ell-1})$ is maximal. The possible $(\ell+1)$-maxima are $\rho(10^\ell)$ and $\rho(0^\ell1)$, but the latter is of the form f_1f_0y for some y, which is less than f_0f_1y and so not maximal. The possible $(\ell+1)$-minima are $\rho(0^{\ell+1})$ and $\rho(10^{\ell-1}1)$. Either may be the case in general, and if $\rho(0^{\ell+1})$ is minimal then the argument repeats for length $\ell+1$:

now $\rho(10^{\ell})$ is maximal. For large enough ℓ, that is no longer the case, and for such an ℓ we have $\rho(10^{\ell-1})$ maximal and $\rho(10^{\ell-2}1)$ minimal. Once that happens, the $(\ell+1)$-maximum is $\rho(10^{\ell})$ since $\rho(10^{\ell-2}1^2)$ is of the form $f_1^2 f_0 y$ for some y, which is less than $f_0 f_1^2 y$. The $(\ell+1)$-minimum is $\rho(10^{\ell-1}1)$ since the other option is $\rho(10^{\ell-2}10)$, which is of the form $f_0 f_1 y$, and this is greater than $f_1 f_0 y$. It follows by induction that we witness 10^m for all $m \geq 0$ in this case.

For the rest of the argument for $x_0 < i_x$, we assume instead that $\rho(11) < \rho(00)$. The argument follows from a series of three claims, much like in part (a). First, there is a least n_0 such that $\rho(1^{2n_0-1}0^2) > \rho(1^{2n_0+1})$. Then $n_0 \geq 1$. The case $n_0 = 1$ requires special treatment, which we outline before proceeding further. We have $\rho(11)$ minimal and $\rho(10)$ maximal. Since $\rho(1^3) < \rho(10^2)$ when $n_0 = 1$, the 3-maximum is $\rho(10^2)$, and the 3-minimum is $\rho(101)$ because the other option $\rho(1^20)$ is greater than $\rho(01^2)$. Inductively, if for $m \geq 2$ we have $\rho(10^m)$ maximal and $\rho(10^{m-1}1)$ minimal, then the $(m+2)$-maximum is $\rho(10^{m+1})$ since the other option, $\rho(10^{m-1}1^2)$, is of the form $f_1^2 f_0 y$, which is less than $f_0 f_1^2 y$ and so not maximal. And the $(m+2)$-minimum is $\rho(10^m1)$ since the other option is $\rho(10^{m-1}10)$, which is of the form $f_0 f_1 y$, which is greater than $f_1 f_0 y$ and so not minimal. It follows that we witness 10^m for all $m \geq 0$ in this subcase.

Now assume $n_0 > 1$ as well as $\rho(11) < \rho(00)$. The second claim to complete the proof is that for any $1 \leq \ell \leq n_0$, we have that $\rho(1^{2\ell-1}0)$ is 2ℓ-maximal; $\rho(1^{2\ell})$ is 2ℓ-minimal; if $\ell < n_0$, then $\rho(1^{2\ell+1})$ is $(2\ell+1)$-maximal; and $\rho(1^{2\ell-1}01)$ is $(2\ell+1)$-minimal. The induction argument goes exactly as in Claim 4.8 from case (a) where $x_0 > i_x$, except switching the roles of "maximal" and "minimal" everywhere as well as switching the roles of (firstly) $f_0 f_1$ and $f_1 f_0$, and (secondly) $f_0 f_1^2$ and $f_1^2 f_0$. This is because we now have $f_1 f_0 < f_0 f_1$ and $f_1^2 f_0 < f_0 f_1^2$ by Lemma 4.5. The third claim, which completes the picture, is that $\rho(1^{2n_0-1}0^m)$ is maximal and $\rho(1^{2n_0-1}0^{m-1}1)$ is minimal for all $m \geq 0$. The cases $m = 0$ and $m = 1$ follow from taking $\ell = n_0 - 1$ and $\ell = n_0$ in the second claim. For $m = 2$, the $(2n_0+1)$-minimum is $\rho(1^{2n_0-1}01)$ by the second claim again, and the $(2n_0+1)$-maximum is $\rho(1^{2n_0-1}0^2)$ because the other option is $\rho(1^{2n_0+1})$, and this is the lesser value by definition of n_0. The induction can be carried out from here using $f_1^2 f_0 < f_0 f_1^2$ and $f_1 f_0 < f_0 f_1$, finishing the proof for $x_0 < i_x$.

Now suppose $x_0 > i_x$. The proof of this case is split into three claims, as usual. First, there is a least n_0 such that $\rho(1^{2n_0+2}) \leq \rho(1^{2n_0}0^2)$. As before, we first need to consider the case $n_0 = 0$ separately, but fortunately this is equivalent to $\rho(11) < \rho(00)$ so there is no need for a third subcase as with the $x_0 < i_x$ argument. If $n_0 = 0$, then $\rho(00)$ is maximal since $\rho(1) < \rho(0)$, and $\rho(01)$ is minimal by $f_1 f_0 < f_0 f_1$. In general, suppose for any $m \geq 2$ that $\rho(0^m)$ is m-maximal and $\rho(0^{m-1}1)$ is m-minimal. Then $\rho(0^{m+1})$ is $(m+1)$-maximal since the other option is $\rho(0^{m-1}1^2)$, which is of the form $f_1^2 f_0 y$, which is less than $f_0 f_1^2 y$ and so not maximal. And $\rho(0^m1)$ is $(m+1)$-minimal since the other option $\rho(0^{m-1}10)$ is of the form $f_0 f_1 y$, which is greater than $f_1 f_0 y$ and hence not minimal. It follows by induction that 0^m is witnessed in this subcase for all $m \geq 1$.

Assume from now on that instead $n_0 > 0$. The second claim we need to finish the proof is that for $1 \leq \ell \leq n_0$, $\rho(1^{2\ell})$ is 2ℓ-maximal; $\rho(1^{2\ell-2}01)$ is 2ℓ-minimal; $\rho(1^{2\ell}0)$ is $(2\ell+1)$-maximal; and $\rho(1^{2\ell+1})$ is $(2\ell+1)$-minimal. The base case here uses $n_0 > 0$ to show $\rho(11)$ is maximal. The third claim is that $\rho(1^{2n_0}0^m)$ is maximal and $\rho(1^{2n_0}0^{m-1}1)$ is minimal for all $m \geq 2$. Here, for $m = 2$, we have $\rho(1^{2n_0}0^2)$ maximal since by definition of n_0, $\rho(1^{2n_0+2})$ cannot be. And $f_0 f_1 > f_1 f_0$

implies that $\rho(1^{2n_0}01)$ is minimal rather than $\rho(1^{2n_0+1}0)$. The inductive steps of both claims can be shown in a straightforward way using $f_1f_0 < f_0f_1$, $f_1^2f_0 < f_0f_1^2$, and in the first statement of the second claim, $\ell < n_0$. (The latter is used only to show the second claim holds for $\ell + 1$ given it holds for ℓ, so it does hold for $\ell = n_0$ as stated.)

Finally, suppose $b = 0$. As in Case 4(a), only finitely many strings can be witnessed. We have again that $r_0 = a$ and $\rho(w^\frown 01) = f_1a$ for all strings w. If $x_0 < i_x$, and n is large enough that $\rho(1^{2n+1}) \leq a$, then no longer string can have probability greater than f_1a, and this value is never attained uniquely. If $x_0 > i_x$, and n is large enough that $\rho(1^{2n}) \leq a$, the same conclusion holds. Therefore at most finitely many constant strings can be witnessed, and nothing else. This completes the proof of Case 4(b) and of Theorem 4.2.

Strings witnessed in the above case: $1^{2n-1}0^m$, $1^{2n}0^m$.

This proof establishes that for any two-state PFA over a binary alphabet, there is a certain family of strings which must receive maximal probabilities, and goes on to characterize this family. It could be that other strings also receive maximal probabilities, of course; in that case the PFA would not witness the complexity of any string of those lengths. As an aside before continuing to a proof of the other direction of Theorem 4.1, we can undertake an analysis of the above proof to characterize the set of strings with *uniquely* maximal probabilities:

Proposition 4.10. Let M be a two-state PFA reading from the alphabet $\{0, 1\}$, and assume that M witnesses an upper bound for the PFA complexity of infinitely many strings. Then exactly one of the following is true:

(i) M witnesses an upper bound for the PFA complexity of a string of every length. The set of strings witnessed is exactly the set of prefixes of the infinite string $i^n j^{\mathbb{N}}$, where $n \geq 0$ is fixed and $i, j \in \{0, 1\}$.

(ii) M witnesses an upper bound for the PFA complexity of strings of cofinitely many lengths. The set of sufficiently long strings witnessed is exactly $i^n\{j\}^*i^m$ for some fixed $n \geq 0$ and $m \leq 1$.

(iii) M witnesses an upper bound for the PFA complexity of strings of every even length, every odd length, or both. Either

- When infinite, the sets of even- and odd-length strings witnessed are, respectively, the set of odd-length prefixes of $i^{2n_0}j(ij)^{\mathbb{N}}$ and the set of even-length prefixes of $j^{2n_1}(ij)^{\mathbb{N}}$, for n_0, n_1, i, j fixed; or
- Similarly but with the sets of odd-length prefixes of $i^{2n_0+1}(ji)^{\mathbb{N}}$ and even-length prefixes of $j^{2n_1+1}i(ji)^{\mathbb{N}}$; or
- Similarly but with the sets of odd-length prefixes of $i^{\mathbb{N}}$ and even-length prefixes of $j^{\mathbb{N}}$.

Proof:
Suppose M corresponds to the IFS specified by maps $f_0 = a + bx$ and $f_1 = c + dx$ together with the starting value x_0. Assume without loss of generality that $a \leq c$. If $x_0 = i_x$, then $\rho(0) = \rho(1)$,

so no maxima of any length are unique, and so we can take $x_0 \neq i_x$ throughout. First suppose that M falls under Case 1 in the proof of Theorem 4.2, i.e., f_0 and f_1 do not intersect in $(0, 1)$. If f_1 has nonnegative slope, then 1^m is uniquely maximal for all m, which follows immediately from the facts that $f_0 < f_1$ in $(0, 1)$ and that $x \leq y$ iff $f_1 x \leq f_1 y$. Thus M satisfies outcome (i). If f_1 has negative slope and f_0 has positive slope, then the proof of Theorem 4.2 in Case 1 already shows that $\rho(0^m 1)$ is the only possible maximum for each m, and so M satisfies outcome (ii). If both maps have negative slope, then $(01)^m$ and $1(01)^m$ are always maximal as discussed in the proof of Case 1. Hence M turns out to satisfy the first bullet point under outcome (iii); the argument is the same as that used below under the assumption that M is instead in Case 3.

Before examining the other main subcases of the proof of Theorem 4.2, we address what happens when f_0 and f_1 commute. This was not investigated in the proof of Theorem 4.2, since it implies that only constant strings can be uniquely maximal, and so we treat it separately now. Recall from Lemma 4.5 that f_0 and f_1 commuting means that f_0, f_1, and the line $y = x$ intersect at the same point, i.e., at $r_0 = r_1 = i_x$. (In particular, f_0 and f_1 must intersect.) If $a = c$, this forces the lines to coincide, in which case there are no unique maxima at all. Then assume $a < c$. If both maps have nonnegative slope, either $\rho(0^m)$ is uniquely maximal for all m or $\rho(1^m)$ is, depending on whether $x_0 > i_x$ or $x_0 < i_x$, and we are in outcome (i). If both maps have negative slope and $x_0 > i_x$, then $\rho(0^{2m+1})$ and $\rho(1^{2m})$ are maximal for all m, so we are in the third bullet point under outcome (iii). This follows by the same argument used below when M is assumed to be in Case 3 of the proof of Theorem 4.2. If $x_0 < i_x$ then the same is true switching the roles of 0 and 1.

Suppose f_0 and f_1 commute, f_0 has positive slope, and f_1 has negative slope.

- If $x_0 > i_x$, then $\rho(0^{2m+1})$ is uniquely maximal for all m, because the only other options for maxima of odd lengths are not constant. Among even-length strings, either $\rho(0^{2m})$ is maximal for all m or $\rho(1^{2m})$ is. This is because both maps converge to the same point i_x under iteration, and the distance of $\rho(0^{2m})$ to i_x versus $\rho(1^{2m})$ is dictated entirely by the absolute values of the slopes of f_0 and f_1. Hence if (say) f_0 converges more slowly to i_x than f_1 does, then $\rho(0^{2m})$ will always be higher than $\rho(1^{2m})$ for every m. So M must fall under the third bullet in outcome (iii).

- If instead $x_0 < i_x$, then there are no unique maxima of any even length, since the only way to get a value above i_x is with a nonconstant string: any string with an even number of 1s will have probability less than i_x, and $\rho(0^m) < i_x$. Among odd lengths, $\rho(1^{2m+1})$ must be maximal for all m (it is greater than i_x while $\rho(0^{2m+1})$ is not). If $\rho(1^{2m+1})$ fails to be uniquely maximal for any m, then it also fails to be unique for every subsequent m since these values all lie in the same orbit. Hence M again falls under the third bullet in (iii).

Assume for the rest of this proof that f_0 and f_1 do not commute and in particular that Assumption 4.6 holds, so that M falls under one of Case 2, Case 3, and Case 4 in the proof of Theorem 4.2. In Case 2, there is a number n such that either the orbit of $\rho(0^n 1^m)$ consists entirely of maximal probabilities for each $m \geq 0$, or such that the orbit of $\rho(1^n 0^m)$ does. Suppose $\rho(0^n 1^m)$ is maximal for all m and that for some ℓ there are two strings u and w of length ℓ sharing the maximal probability. Then also $\rho(u^\frown 0^{\ell-n} 1^m) = \rho(w^\frown 0^{\ell-n} 1^m)$ is maximal for all m if $\ell < n$, or else $\rho(u^\frown 1^m) = \rho(w^\frown 1^m)$ is maximal for all m if $\ell \geq n$. Either way, once uniqueness of maxima is lost once, it is lost forever, and so outcome (i) holds. The same argument goes through switching 0 and 1 everywhere.

Suppose instead M is in Case 3 of the proof of Theorem 4.2. All maxima of odd lengths and minima of even lengths are contained in the same orbit, and all maxima of even lengths and minima of odd lengths are also contained in the same orbit. (This is true in every subcase of Case 3.) In either of these orbits, if uniqueness of either maxima or minima is lost at any time, then uniqueness of all subsequent maxima and minima is also lost. The proof of Theorem 4.2 in Case 3 then implies that one of the first two bullet points in (iii) holds if there are infinitely many unique maxima. The precise descriptions of the sets of strings witnessed can be gleaned from the four possible subcases of Case 3, namely 3(a) and 3(b) with each being split into further subcases based on whether $x_0 > i_x$ or $x_0 < i_x$.

Finally, let M instead be in Case 4, with f_0 having positive and f_1 negative slope. Through an inspection of that part of the proof, it can be seen that in every one of its many subcases, for all sufficiently long strings, there is some fixed n such that either $\rho(1^n0^m)$ is maximal for all m, or $\rho(1^n0^m)$ is minimal and $\rho(1^n0^m1)$ is maximal for all m, or one of these holds swapping 0 with 1. Either way, there is an orbit which eventually consists entirely of maxima, or which eventually consists entirely of minima whose images under f_1 are maximal. We must be careful to include the word "eventually" here because of the corner cases where finitely many strings of the form 0^ℓ are maximal or minimal before the main pattern settles in. This can happen in Case 4(a) when $x_0 > i_x$, $n_0 = 1$, and $\rho(11) < \rho(00)$, and in Case 4(b) when $x_0 < i_x$ and $\rho(00) < \rho(11)$. In the latter situation, the presence of these irregular minima does not change the fact that $\rho(10^m)$ is maximal for all $m \geq 0$ (the constant prefix must have length 1 here). For the sake of convenience, we will argue below as if the "0^ℓ" subcase does not happen, with the understanding that even if it does, the argument still works for all long enough strings. We will show that M must satisfy outcome (ii), and the latter allows for finitely many exceptions to the overall pattern.

Suppose we have an orbit eventually consisting entirely of maxima. If there is a second $(n+\ell)$-maximal string w for some $\ell > 0$, then $\rho(w^\frown 0^m) = \rho(1^n0^{\ell+m})$ is also maximal for all $m \geq 0$, and so only finitely many strings have uniquely maximal probabilities, a contradiction.

Now assume we instead have an orbit eventually composed of minima x with f_1x maximal. (Note this implies we must specifically be in Case 4(a) of the proof of Theorem 4.2.) This subcase splits into further subcases:

- There is a second $(n+\ell)$-minimal string w for some $\ell > 0$. Then $\rho(w^\frown 0^m) = \rho(1^n0^{\ell+m})$ is also minimal for all $m \geq 0$, meaning that $(n+\ell+m+1)$-maxima are not unique. Thus only finitely many strings have uniquely maximal probabilities, which rules out this possibility.

- All minima in this orbit remain unique but there are nonunique maxima of infinitely many lengths. Let w be a string not of the form 1^n0^m1 such that $\rho(w)$ is $(n+\ell+1)$-maximal for some $\ell > 1$. Since we ignore the possibility that $w = 0^{n+\ell+1}$, as mentioned above, we have $\rho(w) = \rho(1^n0^\ell 1)$. (Otherwise the argument resumes starting at length m for some m large enough so that 0^m is not maximal.)
  - If $w = y^\frown 1$ for some y, then $\rho(y)$ is an $(n+\ell)$-minimum distinct from $\rho(1^n0^\ell)$, and that means $\rho(y^\frown 0^m)$ is also minimal for all m. Thus there are only finitely many unique maxima, so this subcase is also ruled out.
  - If $w = y^\frown 0$, then $\rho(y)$ is also $(n+\ell)$-maximal while y is not of the form 1^n0^m1: it cannot be,

because if it were, then $\rho(w)$ would not be maximal since $\rho(w) = \rho(1^n 0^m 10) < \rho(1^n 0^{m+1} 1)$ by the fact that $f_0 f_1 < f_1 f_0$. The latter holds by Lemma 4.5 as we are in Case 4(a).

If y ends with a 1, then $y \restriction (n + \ell - 1)$ has minimal probability but is not of the form $1^n 0^m$, so maxima cease to be unique after length $n+\ell$, a contradiction. If y ends with a 0, then $y \restriction (n+\ell-1)$ has maximal probability while not being of the form $1^n 0^m 1$, for the same reason as before, and the whole argument repeats with $y \restriction (n + \ell - 1)$ in place of w. It follows by induction that one of two things happens: either there is some ℓ' with $1 < \ell' < n + \ell + 1$ after which length minima (hence maxima) are no longer unique; or, if during the induction we never reach a string y ending with a 1, then all maxima of lengths between n and $n + \ell + 1$ are nonunique. Since we assumed there are infinitely many nonunique maxima, there are infinitely many strings w for which this argument can be carried out, and so in fact there have to be *cofinitely* many nonunique maxima.

In each of these subcases, outcome (ii) must be the case for M if there are infinitely many unique maxima, and we are done.

□

4.3. Proof of Theorem 4.3

We show that for every string w listed in (22), there is an IFS (f_0, f_1, x_0) which falls into the subcase of the proof of Theorem 4.2 which would lead to w being witnessed. This results in a case breakdown into the following seven subfamilies of strings, listed here with the subcases of Theorem 4.2 which they employ. (See also Proposition 4.10 above for a summary of the possible outcomes, although it does not distinguish between even- and odd-length prefixes. Whether the prefixes will have even or odd length depends in part on where x_0 lies relative to i_x.)

- $0^n 1^m$ for a given n and all m – Case 2(a) with $x_0 > i_x$, Proposition 4.11;
- $1^{2n} 0^m 1$ for a given n and all m – Case 4(a) with $x_0 < i_x$, Proposition 4.12;
- $1^{2n-1} 0^m 1$ for a given n and all m – Case 4(a) with $x_0 > i_x$, Proposition 4.13;
- $1^{2n} (01)^m$ for a given n and all m – Case 3(a) with $x_0 > i_x$, Proposition 4.14;
- $1^{2n+1} (01)^m$ for a given n and all m – Case 3(a) with $x_0 < i_x$, Proposition 4.15;
- $1^{2n+1} 0(10)^m$ for a given n and all m – Case 3(b) with $x_0 < i_x$, Proposition 4.16;
- $0^{2n} 1(01)^m$ for a given n and all m – Case 3(a) with $x_0 > i_x$, Proposition 4.17.

The proofs all follow the same basic strategy, which goes roughly as follows. Let $\mathcal{F}_n$ be one of the families of strings listed above, where n corresponds to the fixed length of the constant prefix of each member of $\mathcal{F}_n$. Let $f_0 = a + bx$ and $f_1 = c + dx$ form an IFS with starting value x_0, as in Section 4.2. Given n, derive an inequality equivalent to this IFS satisfying the condition from some subcase of the proof of Theorem 4.2 which results in strings from $\mathcal{F}_n$ being witnessed. This will translate to the requirement that x_0 be chosen inside a certain interval $I_{n,a,b,c,d}$ depending on n and the coefficients of the IFS. For this to be possible, we need $I_{n,a,b,c,d}$ to overlap $[0,1]$, and in particular to overlap either $J=(0,i_x)$ or $J=(i_x,1)$ depending on whether we need $x_0<i_x$ or $x_0>i_x$ in order for strings from $\mathcal{F}_n$ to have maximal probabilities. For any fixed n,a,b,c,d, it will turn out that if $I_{n,a,b,c,d}\cap J\neq\emptyset$ and $\ell\leq n$, then also $I_{\ell,a,b,c,d}\cap J\neq\emptyset$. Hence it suffices to show for infinitely many n that we can find an IFS witnessing the strings $\mathcal{F}_n$. To do this, for each $\mathcal{F}_n$ we will derive an inequality of the form $n<g(a,b,c,d)$, for some function g, which is equivalent to the condition that $I_{n,a,b,c,d}\cap J\neq\emptyset$. Then we let a,b,c,d depend on n and show that $g(a,b,c,d)$ is unbounded as a function of n. More precisely, we let some $e\in\{a,b,c,d\}$ depend on n and let the other three variables depend on e. (Sometimes it is enough to pick suitable constant values for some of the variables.) The dependence of e on n is never explicit and in fact we ignore n for the rest of each proof, showing instead that g tends to ∞ as e tends to either 1 or -1 (depending on $\mathcal{F}_n$). One can then pick a suitable value of e which makes the value of g larger than any given n, so this will complete the proof—as long as we can argue we can always get *unique* maxima, and as long as we make sure that a, b, c, and d can always be chosen so that for any n the IFS remains in the correct subcase from the proof of Theorem 4.2. The latter requirement results in something of a laundry list of conditions on a, b, c, and d that must be met in each proof. All of these conditions must be shown to be consistent with each other as well as with the limit condition on g. It will then follow that the members of $\{a,b,c,d\}\setminus\{e\}$ can be chosen as functions of e which meet the necessary criteria and make g increase without bound. (This is nonconstructive, though in each case it will not be difficult to imagine how one might come up with a suitable set of smooth functions.)

We mentioned in the last paragraph that uniqueness of maximal probabilities is not automatically guaranteed by this approach. However, any IFS which assigns maximal probabilities to strings from $\mathcal{F}_n$ can be perturbed to assign a uniquely maximal probability to any given $w\in\mathcal{F}_n$. To see why, suppose we have $\rho(u)=\rho(w)$ for some $u\neq w$ with $|u|=|w|$. Replacing $\rho(u)$ and $\rho(w)$ with expressions deriving from the maps of the IFS results in an equation of the form $\alpha+\beta x_0=\nu+\eta x_0$, where α, β, ν, and η are polynomials in the coefficients of the IFS. Hence $\alpha-\nu=(\eta-\beta)x_0$. Now, in every one of these proofs, to witness strings from $\mathcal{F}_n$, we are free to choose x_0 to be any value inside some open interval. Hence if $\eta\neq\beta$, we can perturb x_0 as needed as needed to destroy the above equality while maintaining the conditions for $\rho(w)$ to be maximal. Using this new x_0, $\rho(w)$ is now uniquely maximal. Suppose instead that $\eta=\beta$. Expanding these quantities in terms of the coefficients of f_0 and f_1, we will have $\beta=b^kd^\ell$ and $\eta=b^md^n$ for some $k+\ell=m+n=|u|=|w|$. If the values of b and d are perturbed so as to be algebraically independent over $\mathbb{Q}$, the equation $x^ky^\ell-x^my^n=0$ cannot have (b,d) as a solution and so we have $\beta\neq\eta$ in the new IFS. This then boils down to the previous case in which we can change x_0 to achieve unique maximality of $\rho(w)$.

Although all seven subcases follow this outline, the particularities are different enough to warrant separate treatments, albeit with some details omitted. One formula that will be used repeatedly is the following: for any x and n, we have

$$f_0^nx=a\sum_{i=0}^{n-1}b^i+b^nx=a\cdot\frac{1-b^n}{1-b}+b^nx=r_0(1-b^n)+b^nx. \tag{41}$$

An analogous formula holds for f_1^nx.

Proposition 4.11. $A_P(0^n1^m) = 2$ for all $n, m \geq 0$.

Proof:
Let $n \geq 1$ be given (the case $n = 0$ is trivial). The IFS (f_0, f_1, x_0) witnesses the set of strings $\mathcal{F}_n = \{\, 0^n1^m \,:\, m \geq 0 \,\}$ if in Case 2(a) of the proof of Theorem 4.2 with $x_0 > i_x$, and if n is least such that $\rho(0^n) < i_x$, i.e., $f_0^n x_0 < i_x < f_0^{n-1} x_0$. Take $f_0 = bx$ and $f_1 = b/2$ for any $b < 1$ (so $a = d = 0$). Then $f_0^n x_0 = b^n x_0$, and our condition becomes

$$b^n x_0 < i_x < b^{n-1} x_0 \quad \text{or equivalently} \quad x_0 \in J := \left(\frac{i_x}{b^{n-1}}, \frac{i_x}{b^n} \right). \tag{42}$$

Since $b < 1$, we have $i_x/b^n > i_x$ for all $n \geq 1$. In order to be able to choose x_0 to witness 0^n1^m for our given n, we need $i_x/b^{n-1} < 1$, or equivalently

$$\frac{\log i_x}{\log b} + 1 > n. \tag{43}$$

By increasing b arbitrarily close to 1, and setting $c = b/2$ from b, we can make $\log i_x / \log b$ larger than any given n, so that it is possible to choose $x_0 \in (i_x, 1)$ in order for every element of $\mathcal{F}_n$ to receive maximal probability. □

Proposition 4.12. $A_P(1^{2n}0^m1) = 2$ for all $n, m \geq 0$.

Proof:
Let $n \geq 1$ be given (the case $n = 0$ is covered by the previous proposition). The IFS (f_0, f_1, x_0) witnesses the family of strings $\mathcal{F}_n = \{\, 1^{2n}0^m1 \,:\, m \geq 0 \,\}$ if it falls under Case 4(a) of the proof of Theorem 4.2—that is, $b > 0 > d$ and both maps decrease i_x—if $x_0 < i_x$, and if n is least such that $\rho(1^{2n+2}) > \rho(1^{2n}0^2)$, or (in other words) such that

$$f_1^{2n+2} x_0 > f_0^2 f_1^{2n} x_0. \tag{44}$$

Thinking of the left-hand side here as $f_1^2 f_1^{2n} x_0$, as long as $b < |d|$ (or really $b^2 < d^2$), this inequality is equivalent to

$$a + ab + b^2 f_1^{2n} x_0 < c + cd + d^2 f_1^{2n} x_0 \quad \iff \quad F := \frac{f_0^2 0 - f_1^2 0}{d^2 - b^2} < f_1^{2n} x. \tag{45}$$

If n is supposed to be the least number making $F < f_1^{2n} x_0$, then we would like $f_1^{2(n-1)} x_0 < F < f_1^{2n} x_0$. On the one hand,

$$f_1^{2(n-1)} x_0 < F \iff r_1(1 - d^{2(n-1)}) + d^{2(n-1)} x_0 < F \iff x_0 < r_1 - \frac{r_1 - F}{d^{2(n-1)}}, \tag{46}$$

and on the other hand

$$F < f_1^{2n} x_0 \iff x_0 > r_1 - \frac{r_1 - F}{d^{2n}} \tag{47}$$

by a similar calculation. Now, in our situation it will always be the case that $F < r_1 = c/(1-d)$, because

$$\frac{f_0^2 0 - f_1^2 0}{d^2 - b^2} < \frac{c}{1-d} \iff a(1-d+b-bd) < c(1-b^2) \iff \frac{a}{1-b} < \frac{c}{1-d}, \tag{48}$$

i.e., $r_0 < r_1$. As long as we choose a, b, c, d to make $r_0 < r_1$, then, we have $F < r_1$; and as long as we make $b < |d|$, the equivalence of the inequalities in (45) holds. We also need $F > 0$, but this automatically follows from the condition that $f_1^{2(n-1)} x_0 < F$ since the latter LHS is nonnegative for every $n \geq 1$ and x_0.

Putting (46) and (47) together, the IFS witnesses $\mathcal{F}_n$ when we can pick x_0 within the interval

$$J := \left(r_1 - \frac{r_1 - F}{d^{2n}}, r_1 - \frac{r_1 - F}{d^{2(n-1)}} \right). \tag{49}$$

If $r_0 < r_1$, then J is nonempty, since $d^{2n} < d^{2(n-1)}$ and we have shown $r_1 - F > 0$; and both endpoints of J are less than i_x, since they are less than r_1, and $r_1 < i_x$ iff $r_0 < r_1$ by Lemma 4.5. Hence choosing such an x_0 automatically fulfills the requirement that $x_0 < i_x$.

We also need (for a given n) to be able to pick $x_0 > 0$, so at least the right endpoint of J should be positive. For any n,

$$\begin{aligned} r_1 - \frac{r_1 - F}{d^{2(n-1)}} > 0 \iff d^{2(n-1)} > 1 - \frac{F}{r_1} \iff (n-1)\log d^2 > \log\left(1 - \frac{F}{r_1}\right) \\ \iff n < 1 + \frac{\log(1 - F/r_1)}{\log d^2}. \end{aligned} \tag{50}$$

So for arbitrarily large n to be possible, the last RHS must be able to grow arbitrarily large depending on a, b, c, d. To accomplish this we will treat d as a variable, presumed to depend on n, which will decrease to -1 as n increases to infinity. Then we make c a function of d (so that F and r_1 are as well), and require that

$$\lim_{d \to -1^+} \frac{\log(1 - F/r_1)}{\log d^2} = \infty. \tag{51}$$

We need c to be a function of d because c must be greater than $|d|$ for all $d > -1$ if we are to have $c + d > 0$, so c will necessarily approach 1 in the limit. Of course we also need to make sure the logarithm in the numerator is defined for all $d > -1$. If so, then together with the fact that the right endpoint of J is always less than i_x, we will have that for *every* $n \geq 1$ there is a choice of a, b, c, d, x_0 making (f_0, f_1, x_0) witness $1^{2n}0^m1$ for all m.

To sum up thus far: we want to choose numbers a, b and a continuous function $c(d)$ to satisfy the requirements that $|d| < c(d) < 1$, $a < 1 - b$, $b < |d|$, $r_0 < r_1$, $i_x < 1$, $i_y > 0$, and the limit condition (51) holds. This limit condition will imply that x_0 can be chosen inside the interval J as needed. Because we are taking the limit as $d \to -1^+$, we may as well only bother asking for the other requirements to hold in the limit, too. This simplifies things considerably: since $c \to 1$ as $d \to -1$, we have $r_1 \to 1/2$. Then for $r_0 < r_1$ to hold in the limit, it is enough to make $r_0 = a/(1-b) < 1/2$, or in other words to pick positive constants a and b with $2a < 1 - b$. This condition also guarantees

$a < 1 - b$ and hence $a + b \in [0, 1]$, as well as that $b < |d|$. Furthermore, since $c(d)$ will eventually be greater than any fixed $a < 1$, $a < c$ is satisfied in the limit. That $c + d \in [0, 1]$ is automatically implied by the requirement that $|d| < c(d) < 1$.

Only two conditions remain to be checked. Firstly, (51) holds if $1 - F/r_1$ stays strictly between 0 and 1 as $d \to -1^+$: on the one hand, $0 < 1 - F/r_1$ iff $F < r_1$, which as we saw is equivalent to $r_0 < r_1$. On the other hand, $1 - F/r_1 < 1$ iff both F and r_1 are positive, and both of those happen in the limit as noted above. Finally, we need to check that the lines intersect in $[0, 1]^2$. But since $f_1(x) \to 1 - x$ as $d \to -1$, if we make sure to take $a, b > 0$, then f_1 will eventually intersect any line that stays inside $[0, 1]^2$. Hence $i_y > 0$ and $i_x < 1$ hold in the limit as $d \to -1^+$, and we are done. □

The proofs of all but one of the remaining cases are very similar to the above, and we will give a somewhat more streamlined presentation from here on out. The most complicated case we save for last (Proposition 4.17).

Proposition 4.13. $A_P(1^{2n-1}0^m1) = 2$ for all $n \geq 1, m \geq 0$.

Proof:
If $n \geq 1$ is given, then (f_0, f_1, x_0) witnesses $1^{2n-1}0^m1$ for all $m \geq 0$ if in Case 4(a) of the proof of Theorem 4.2 (mixed slopes) with $x_0 > i_x$ and n least such that $\rho(0^{2n-1}0^2) < \rho(1^{2n+1})$, or in other words such that

$$f_0^2 f_1^{2n-1} x_0 < f_1^{2n+1} x_0. \tag{52}$$

We do also need to avoid the corner case in which finitely many strings 0^ℓ can receive a higher probability than those of the form $1^{2n-1}0^m1$. As established in the proof of Case 4(a), this is only possible when $n = 1$, and more specifically when $\rho(11) < \rho(00)$. As before, we will pick $a, b > 0$ constants and c a continuous function of d so that for any given n, there is a d making it possible to choose x_0 so that (52) holds and so that $\rho(11) > \rho(00)$. Taking $n \to \infty$ will correspond to taking $d \to -1^+$. Now, *if* we have that $b < |d|$, then (52) is equivalent to

$$f_1^{2n-3} x_0 < E < f_1^{2n-1} x_0 \quad \text{where} \quad E = \frac{a(1 + b) - c(1 + d)}{d^2 - b^2}, \tag{53}$$

and the inequality in (53) is equivalent to

$$x_0 \in \left(r_1 + \frac{r_1 - E}{|d|^{2n-3}}, r_1 + \frac{r_1 - E}{|d|^{2n-1}} \right). \tag{54}$$

For arbitrarily large n to be possible, we want to pick a, b, c, d so that this interval intersects $(i_x, 1)$, so a suitable x_0 can be chosen. We will see below that this can also be done so that $b < |d|$ and the above equivalences are legitimate. The left endpoint in (54) can be made less than 1 for arbitrarily large n if, in particular,

$$\lim_{d \to -1^+} \frac{\log \dfrac{r_1 - E}{1 - r_1}}{\log d^2} + \frac{3}{2} = \infty. \tag{55}$$

And the right endpoint in (54) is greater than i_x, for a given n, iff

$$\frac{\log \dfrac{r_1 - E}{i_x - r_1}}{\log d^2} + \frac{1}{2} > n. \tag{56}$$

Note that in the limit, E approaches r_0 (as long as $b < 1$). Hence as long as $r_0 < r_1$ in the limit, then eventually $r_1 > E$. If we arrange things so i_x stays below 1, then,

$$\frac{r_1 - E}{i_x - r_1} > \frac{r_1 - E}{1 - r_1}. \tag{57}$$

Also notice that we can take $\frac{r_1 - E}{i_x - r_1} < 1$ in the limit since this is equivalent to $2r_1 < i_x + E$, which in the limit is guaranteed if $2a + b < 1$, as may be checked with a little algebra. Assume that a, b are positive constants with $2a + b < 1$ and $b < |d|$. Since log is increasing, if (55) holds, the LHS of (56) will also approach ∞. This implies that whenever n is such that the left endpoint of (54) is less than 1, for all $n' \leq n$ it is possible to choose $x_0 \in (i_x, 1)$ in order to witness $1^{2n'-1}0^m1$. And the fact that $E \to r_0 < i_x$ means that in the limit, any choice of $x_0 \in (i_x, 1)$ guarantees that $\rho(11) > \rho(00)$, and we thus avoid the issue of finitely many strings of the form 0^ℓ being witnessed instead of the desired ones. This is because under the assumption that $b < |d|$, we have that $\rho(11) < \rho(00)$ is equivalent to $x_0 < E$, so this case is automatically ruled out if $x_0 > i_x$.

So, let $c(d)$ be a continuous function with $|d| < c(d) < 1$ for all $d > -1$, and let a and b be positive constants such that $2a + b < 1$. This immediately implies $a + b, c + d \in [0, 1]$ for all d and that $b < |d|$ in the limit. Since $r_1 \to 1/2$ as $d \to -1$, we have $r_0 < r_1$ in the limit since $r_0 = a/(1 - b) < 1/2$. We also need $E > 0$, which is guaranteed as $d \to -1^+$ since E approaches $r_0 > 0$. Since $c \to 1$ and $d \to -1$, eventually $c > a$ as required. Because $f_1(x) \to 1 - x$ as $d \to -1$, $c + dx$ will eventually intersect $a + bx$ in $[0, 1]^2$, so that $0 < i_y < i_x < 1$. It only remains to check (55). But we already observed that

$$\frac{r_1 - E}{1 - r_1} < \frac{r_1 - E}{i_x - r_1} < 1 \tag{58}$$

as $d \to -1$, and $\frac{r_1 - E}{1 - r_1} > 0$ iff $r_1 > E$, which also holds in the limit. Therefore the logarithm in the numerator of (55) approaches a finite negative number, while $\log d^2$ approaches 0 from below. □

Proposition 4.14. $A_P(1^{2n}(01)^m) = 2$ for all $n, m \geq 0$.

Proof:
For a given n, we witness $1^{2n}(01)^m$ if in Case 3(a) of the proof of Theorem 4.2 (both slopes negative), with $x_0 > i_x$ and n least such that

$$f_1^{2n} x_0 < i_x. \tag{59}$$

We will pick numbers $a > 0$, $b < 0$, and a continuous function $c(d)$ so that as $d \to -1^+$, we have $a + b, c + d \in [0, 1]$, $a < c$, $b > d$, $r_0 < r_1$, and the lines $a + bx$ and $c + dx$ intersecting in $[0, 1]^2$. If $a, b \notin \{0, \pm 1\}$, then the last condition is automatically met as $d \to -1$ since $f_1 \to 1 - x$ and this intersects any line in $[0,1]^2$. The conditions $a < c$ and $b > d$ are also automatically met as $d \to -1$. At the same time, we must (given n) be able to pick

$$x_0 \in \left(r_1 + \frac{i_x - r_1}{d^{2(n-1)}}, r_1 + \frac{i_x - r_1}{d^{2n}} \right) \tag{60}$$

so that $f_1^{2n} x_0 < i_x < f_1^{2(n-1)} x_0$. We need this interval to intersect $(i_x, 1)$ for arbitrarily large n, for suitable choices of a, b, c, d. That the right endpoint is always greater than i_x, for any n, follows from $d^{2n} < 1$, since then $\frac{i_x - r_1}{d^{2n}} > i_x - r_1$. For the left endpoint to be less than 1 for arbitrarily large n we need

$$\lim_{d \to -1^+} \frac{\log \dfrac{i_x - r_1}{1 - r_1}}{\log d^2} = \infty. \tag{61}$$

Pick $a > 0 > b$ with

$$-b < a < \frac{1-b}{2}. \tag{62}$$

Also let $c(d)$ be a continuous function with $|d| < c(d) < 1$ for all $d > -1$. This immediately gives $c + d \in [0,1]$, and (62) implies $a + b \in [0,1]$ too. Next, since $r_1 \to 1/2$ as $d \to -1$ and (62) makes $r_0 = a/(1-b) < 1/2$, we have $r_0 < r_1$ in the limit. Finally, to satisfy (61), we want $\frac{i_x - r_1}{1 - r_1}$ to be strictly between 0 and 1 in the limit. This quantity is automatically positive since $i_x > r_1$ and $1 > r_1$ (both in the limit, again). And because (62) implies $i_x \to \frac{1-a}{b+1} < 1$ as $d \to -1$, the fraction is also less than 1 in the limit. This completes the proof. □

Proposition 4.15. $A_P(1^{2n+1}(01)^m) = 2$ for all $n, m \geq 0$.

Proof:
Given n, take the IFS to be in Case 3(a) of the proof of Theorem 4.2 (both slopes negative) with $x_0 < r_0$ and n such that

$$f_1^{2n+1} x_0 < i_x < f_1^{2n-1} x_0. \tag{63}$$

This is equivalent to

$$x_0 \in \left(r_1 - \frac{i_x - r_1}{|d|^{2n+1}}, r_1 - \frac{i_x - r_1}{|d|^{2n-1}} \right). \tag{64}$$

Since

$$r_1 - \frac{i_x - r_1}{|d|^{2n+1}} < r_0 \iff |d|^{2n+1} < \frac{i_x - r_1}{r_1 - r_0} \tag{65}$$

and the last fraction is greater than 1 by Lemma 4.5(e) while the LHS is less than 1, we have that the left endpoint of (64) is always less than r_0 for all $n \geq 0$. In order to make the right endpoint of (64) greater than 0 for arbitrarily large n (for suitable choice of a, b, c, d), so that an $x_0 \in (0, r_0)$ may be chosen to make the IFS witness exactly $1^{2n+1}(01)^m$, we can arrange for

$$\lim_{d \to -1^+} \frac{\log(i_x/r_1 - 1)}{\log d^2} = \infty. \tag{66}$$

As usual, pick constants $a, b \notin \{0, \pm 1\}$, $a > 0 > b$, and a continuous function $c(d)$ such that $|d| < c(d) < 1$ for all $d > -1$ (so $c + d \in [0, 1]$). To satisfy (66), we want $0 < i_x/r_1 - 1 < 1$ in the limit, or equivalently $r_1 < i_x < 2r_1$. Since i_x converges to $(1-a)/(b+1)$ and $r_1 \to 1/2$, this can achieved (along with $a + b \in [0, 1]$) by making $-b < a < \frac{1-b}{2}$. This implies that $a < c$ and $b > d$ are met in the limit, and again since $f_1 \to 1 - x$ we will eventually have $(i_x, i_y) \in [0, 1]^2$. And $r_0 < r_1$ follows from $r_1 < i_x$. □

Proposition 4.16. $A_P(1^{2n+1}0(10)^m) = 2$ for all $n, m \geq 0$.

Proof:
For this, given n, we take the IFS to be in Case 3(b) of the proof of Theorem 4.2, so that both maps have negative slope and *increase* i_x. We want $x_0 > r_0$ and n to be such that

$$f_1^{2n-1} x_0 < i_x < f_1^{2n+1} x_0. \tag{67}$$

This is equivalent to

$$x_0 \in \left(r_1 + \frac{r_1 - i_x}{|d|^{2n-1}}, r_1 + \frac{r_1 - i_x}{|d|^{2n+1}} \right). \tag{68}$$

Remember that in the present case we have $i_x < r_1 < r_0$. We will pick $a > 0 > b$ with

$$\frac{1-b}{2} < a < 1, \tag{69}$$

and pick $c(d)$ a continuous function with $|d| < c(d) < 1$ for all $d > -1$. Then if we take $d \to -1$, we have $r_1 \to 1/2$ and $r_0 > 1/2$ by choice of a and b, so that $r_1 < r_0$ in the limit. Also $a+b, c+d \in [0, 1]$, $a < c$, and $b > d$ hold in the limit; and as before, $a + bx$ eventually intersects $c + dx$ in $[0, 1]^2$ since $c + dx \to 1 - x$. Now we just need to make sure we can always pick an $x_0 \in (r_0, 1)$ for arbitrarily large n as $d \to -1$. We have

$$r_1 + \frac{r_1 - i_x}{|d|^{2n+1}} > r_0 \iff \frac{r_1 - i_x}{r_0 - r_1} > |d|^{2n+1}. \tag{70}$$

Since the RHS here is less than 1 and the LHS is greater than 1 (by Lemma 4.5(e) again), this always happens for any n. To make the left endpoint of (68) less than 1 for any given n, so that suitable a, b, c, d, x_0 may be chosen to witness the desired string, it suffices to ensure that

$$\lim_{d \to -1^+} \frac{\log \dfrac{r_1 - i_x}{1 - r_1}}{\log d^2} = \infty. \tag{71}$$

Thus we want $0 < \frac{r_1 - i_x}{1 - r_1} < 1$ in the limit, or equivalently $2r_1 - 1 < i_x < r_1$. Since $r_1 \to 1/2$, in the limit the latter inequality becomes

$$0 < \frac{1-a}{b+1} < \frac{1}{2}, \tag{72}$$

which is equivalent to (69). □

Now we arrive at the final and most complex subcase of Theorem 4.3 to prove. The extra difficulty arises because, basically, we will need to take both b and d to -1 while both a and c go to 1. This makes it harder to make certain properties hold "in the limit" as in the previous subcases, and also results in a limit condition in which the limit converges to $\log \frac{0}{0}$. Slightly more delicate handling is needed to get around these problems.

Proposition 4.17. $A_P(0^{2n}1(01)^m) = 2$ for all $n, m \geq 0$.

Proof:
Let n be given. The IFS (f_0, f_1, x_0) witnesses $0^{2n}1(01)^m$ for all m if in Case 3(a) of the proof of Theorem 4.2 (where both maps have negative slope and both decrease i_x), when $x_0 > i_x$ and when n is least such that $f_0^{2n}x_0 < i_x$, i.e.,

$$f_0^{2n}x_0 < i_x < f_0^{2(n-1)}x_0, \tag{73}$$

or equivalently (after rearranging)

$$x_0 \in \left(r_0 + \frac{i_x - r_0}{b^{2n-2}}, r_0 + \frac{i_x - r_0}{b^{2n}}\right). \tag{74}$$

If we can pick a, b, c, d to make $r_0 < i_x$, then this interval is nonempty with positive endpoints. For this n and a, b, c, d, it is possible to choose x_0 to witness the desired family of strings iff (74) intersects with $(i_x, 1)$, that is, iff the left endpoint is less than 1 and the right endpoint is greater than i_x. We now investigate when each of these conditions occurs. First,

$$r_0 + \frac{i_x - r_0}{b^{2n}} > i_x \iff 1 > b^{2n}, \tag{75}$$

which is true for all $n \geq 1$, so if an x_0 can be chosen above i_x for a given n then a suitable x_0 can also be chosen for any $n' \leq n$. And we can choose $x_0 < 1$ iff

$$r_0 + \frac{i_x - r_0}{b^{2n-2}} < 1 \iff \frac{i_x - r_0}{1 - r_0} < b^{2n-2} \iff \frac{\log \dfrac{i_x - r_0}{1 - r_0}}{\log b^2} + 1 > n. \tag{76}$$

This is possible to achieve for any given n if we can make

$$\lim_{b \to -1^+} \frac{\log \dfrac{i_x - r_0}{1 - r_0}}{\log b^2} = \infty. \tag{77}$$

Altogether this means that if $r_0 + (i_x - r_0)/b^{2n} < 1$ for some n and a fixed choice of a, b, c, d, then it is possible for every $1 \leq n' \leq n$ to pick a suitable value of $x_0 > i_x$ making (f_0, f_1, x_0) witness the strings $0^{2n'}1(01)^m$ for every $m \geq 0$. Hence the proof will be complete if we can choose a, c, and d as functions of b such that such that (77) holds and such that the IFS remains in Case 3(a) of the proof of Theorem 4.2 for all $b > -1$. Actually, for technical reasons it will be simpler for now to choose

r_0 as a function of b and then let $a(b) = (1-b)r_0(b)$. This is not a problem because b is never 1, so $r_0(b) = a(b)/(1-b)$ is always well-defined. We will ultimately see that the requirements we impose on $r_0(b)$ do not contradict the behavior of $a(b)$.

We proceed by deriving necessary conditions on r_0, c, d to satisfy each requirement, and showing along the way that each new condition is compatible with all the preceding ones. This will imply that functions r_0, c, d satisfying all of them do indeed exist. Our first requirements, which we will take as "atomic" in that they will not reduce to other requirements, are that

$$(1-b)r_0(b) < 1 \quad \text{and} \quad |b| < |d(b)| < c(b) < 1 \tag{78}$$

for all $b > -1$ (with b, d negative). The second of these immediately implies $c + d \in [0,1]$. To guarantee $a + b \in [0,1]$, first note that $a + b = r_0(1-b) + b < 1$ iff $r_0 < 1$, and this follows from the first atomic requirement. Then $a + b > 0$ iff

$$r_0 > -b/(1-b), \tag{79}$$

a new requirement. Actually, (79) will turn out to be a consequence of $i_x, i_y \in [0,1]$, or in other words of f_0 and f_1 intersecting in $[0,1]^2$. We need the latter to happen anyway, so let us now find a sufficient condition for it. Rewriting i_x and i_y in terms of r_0 produces

$$i_x = \frac{c - r_0(1-b)}{b-d} \quad \text{and} \quad i_y = \frac{bc - r_0(1-b)d}{b-d}. \tag{80}$$

If $i_y < i_x$, or equivalently $r_0 < r_1$, then it suffices to make $i_y > 0$ and $i_x < 1$. We will see how to ensure $r_0 < r_1$ in a moment. One can check that

$$i_y > 0 \iff r_0 > \frac{bc}{d(1-b)} \quad \text{and} \quad i_x < 1 \iff r_0 > \frac{c+d-b}{1-b}. \tag{81}$$

Since $c + d > 0$, we have $\frac{c+d-b}{1-b} > \frac{-b}{1-b}$, so that satisfying (81) would automatically result in (79) being satisfied too. Thus (79) is redundant. Next, some more algebra shows that

$$\frac{bc}{d(1-b)} < \frac{c+d-b}{1-b} \iff b > d, \tag{82}$$

an atomic requirement. Hence the first condition in (81) is implied by the second as long as (78) holds, so is also redundant. Then we will have $a + b > 0$, $i_y > 0$, and $i_x < 1$ if we can choose r_0 so that

$$\frac{c+d-b}{1-b} < r_0 < \frac{c}{1-d} = r_1. \tag{83}$$

The latter guarantees that $i_y < i_x$ so that we stay in Case 3(a) of the proof of Theorem 4.2, and also subsumes the second condition in (81), so if (83) holds then (81) is fully redundant. Now, the interval in (83) is nonempty because

$$\frac{c+d-b}{1-b} < \frac{c}{1-d} \iff (c+d-b)(1-d) < c(1-b) \iff (c-1+d)(b-d) < 0, \tag{84}$$

which follows from the second requirement in (78): $b - d > 0$ since $b > d$, and $c - 1 + d < 0$ since $|d| < c < 1$. So (78) makes it possible to choose r_0 to satisfy (83), and together (78) and (83) are enough to ensure we stay in Case 3(a).

It remains to show that the limit requirement (77) is consistent with (78) and (83). We will take r_0, c, and d to be continuously differentiable functions of b. $\log b^2$ approaches 0 from below as $b \to -1^+$, so in order for the limit to reach $+\infty$, one needs the logarithm in the numerator of (77) to stay negative. For this, one must maintain

$$0 < \frac{i_x - r_0}{1 - r_0} < 1 \tag{85}$$

in the limit as $b \to -1^+$, and for this quantity to stay strictly below 1 at $b = -1$. Now, $d(b) \to -1^+$ as $b \to -1^+$ since d is always less than b, and $c(b) \to 1$. Then after some more algebra, we have that

$$\frac{i_x - r_0}{1 - r_0} = \frac{c - r_0(1 - d)}{(b - d)(1 - r_0)} \to \frac{0}{0} \quad \text{as } b \to -1. \tag{86}$$

An application of L'Hôpital's Rule shows that the limit is equal to

$$\lim_{b \to -1^+} \frac{c' - r_0'(1 - d) + r_0 d'}{(1 - r_0)(1 - d') - r_0'(b - d)} = \frac{2c'(-1) - 4r_0'(-1) + d'(-1)}{1 - d'(-1)}. \tag{87}$$

(The calculation follows since r_0', c', and d' are bounded everywhere by assumption, and $r_0 \to 1/2$.) Since d decreases to -1 as b decreases to -1, $d'(-1) \geq 0$, and we will need $d'(-1) \neq 1$ for (87) to be well-defined. If we take $0 < d'(-1) < 1$, then the denominator of the limit in (87) is positive. Hence the limit in (77) will tend to $\frac{-\infty}{0^-} = +\infty$, as needed, if

$$0 < \frac{2c'(-1) - 4r_0'(-1) + d'(-1)}{1 - d'(-1)} < 1. \tag{88}$$

If $L(b) = \frac{c+d-b}{1-b}$ is the lower bound in (83), then one can calculate

$$L'(-1) = \frac{2c'(-1) + 2d'(-1) - 1}{4}, \quad r_0'(-1) = \frac{2a'(-1) + 1}{4}, \quad \text{and} \quad r_1'(-1) = \frac{2c'(-1) + d'(-1)}{4}. \tag{89}$$

Using these expressions we see that (88) is equivalent to

$$L'(-1) < r_0'(-1) < r_1'(-1). \tag{90}$$

Our final objective is to show (90) is consistent with the other requirements (78) and (83), which will complete the proof since that means (77), (78), and (83) can all be satisfied simultaneously. Actually, under the above assumption that $0 < d'(-1) < 1$, and up to possibly perturbing r_0, c, and d, (90) is equivalent to (83) holding in the limit. This follows because for any continuously differentiable functions $f(x), g(x)$ having the same limit as $x \to C^+$, where C is some constant, then for any $\varepsilon > 0$, $f(x) > g(x)$ on $(C, C + \varepsilon)$ iff $f'(x) > g'(x)$ on $(C, C + \varepsilon)$. Then since L, r_0, and r_1 all tend to $1/2$ as $b \to -1$, we have that (83) holding in a right neighborhood of $b = -1$ is equivalent to $L' < r_0' < r_1'$ holding in the same neighborhood. By smoothly perturbing r_0, c, and d if necessary, as long as $0 < d'(-1) < 1$ is maintained, we can ensure strict inequality between the derivatives holds at $b = -1$, i.e., that (90) holds. (A bit more formally, one could say that these strict inequalities are all open conditions in the C^1 topology.) Thus (90) implies (83) holds near $b = -1$, and conversely, (83) implies that r_0, c, and d may be taken to satisfy (90) and hence (77). In particular, (90) and (78) are also consistent with each other.

So to sum up, there are continuously differentiable functions $r_0(b)$, $c(b)$, and $d(b)$ (and consequently $a(b) = (1-b)r_0(b)$) satisfying (78), (83), and $0 < d'(-1) < 1$. We have established that all of this suffices to be able to choose, given any n, values of x_0 and b which result in the IFS (f_0, f_1, x_0) witnessing the strings $0^{2n}1(01)^m$ for all $m \geq 0$. This finishes the proof of the final subcase of Theorem 4.3, and at last the proof of Theorem 4.1 is complete. □

4.4. Further remarks

The proof of Theorem 4.3 appears to explicitly rely on the use of IFSs derived from PFAs reading from a two-letter alphabet, and a priori does not extend to show that, e.g., $A_P(0^n1^n) = 2$ may be witnessed by an IFS over $\{0, 1, 2\}$, for which another map f_2 must be specified. However, if one defines $f_0x = a + bx$ and $f_1x = c + dx$ as in any of the proofs in the last section, and lets $f_jx = \frac{a+c}{2} + \frac{b+d}{2}x$ for all other $j \in \Sigma$, then f_jx is strictly between f_0x and f_1x except at $x = i_x$, and so a string containing a j can have neither minimal nor maximal probability. Hence we have

Corollary 4.18. If $|\Sigma| = 2$, then $A_P(w, \Sigma) = 2$ implies $A_P(w, \Sigma') = 2$ for every $\Sigma' \supset \Sigma$ and $w \in \Sigma^*$.

Theorem 4.1 immediately implies that the set of binary strings with $A_P = 2$ is a regular language. More particularly, Proposition 4.10 has the following consequence, which is somewhat intriguing given that stochastic languages—which are defined by fixed probability thresholds (the cut-point)—are not generally regular, or even recursively enumerable, although Rabin did show that a stochastic language defined by an isolated cut-point is regular [12].

Corollary 4.19. For every two-state PFA M over a binary alphabet, the language of strings whose complexity is witnessed by M is regular.

Proof:
The set of such strings is either finite, or is fully characterized as one of the cases in Proposition 4.10. Each of these cases can be described by a regular expression. □

Another consequence of the classification is that we can save an arbitrarily high number of states by switching from NFAs to PFAs to describe a given (binary) string:

Corollary 4.20. The quantity $A_N(w) - A_P(w)$ may be arbitrarily large among binary w.

Proof:
The statement follows if we can show $A_N(0^n1^n)$ is unbounded in n,[3] since $A_P(0^n1^n) = 2$ for all n by Theorem 4.1. Suppose $A_N(0^n1^n) \leq K$ for all n and some constant K. For any w, $A_N(w)$ can be witnessed by an NFA whose unique accepting path of length $|w|$ uses every edge. Hence by the pigeonhole principle, there is some NFA M with at most K states such that for infinitely many n, there is a unique path of length $2n$ which accepts 0^n1^n and uses every edge of M. We show this is impossible. First, if the digraph of M has fewer than two distinct cycles, then at most one string of the form 0^n1^n is accepted. Then we can assume there are distinct cycles of lengths a and b, respectively. For any string w accepted by M, the portion of w which was read while traversing these cycles has length $\ell = ax + by$ for some $x, y \in \mathbb{N}$. If such an ℓ is greater than $2ab - a - b$, then there are at least two different pairs of natural numbers (x, y) and (x', y') with $ax + by = ax' + by' = \ell$ (see, e.g., [4, Lemma 11]). In terms of M, this means for all large enough m such that M accepts a word of length m with a path that uses both cycles, there are at least two distinct accepting paths of length m—corresponding to traversing the cycles x and y times on the one hand, and x' and y' times on the other. In particular, the accepting path for 0^n1^n uses both cycles for infinitely many n such that $A_N(0^n1^n)$ is witnessed by M, and so for all but finitely many of these n there are two different accepting paths of length $2n$, a contradiction. □

Of course, the 2-state PFA describing 0^n1^n may have to be somewhat complicated, a problem we briefly return to in Section 6 below.

As remarked earlier, no evidence has yet appeared to suggest that A_P is unbounded, or even that any string has complexity greater than 3. All binary strings of length 10 or less have complexity at most 3, and witnesses with three states have been found for a number of longer strings as well. Therefore, we may pose the following questions, the first being restated from the introduction:

Question 1.4. Is A_P unbounded? If not, what is its maximum value? Similarly when restricted to a given finite alphabet, and similarly for $A_{P,\gamma}$.

Question 4.21. What is a tight upper bound for $A_P(w)$ as a function of $|w|$?

Lastly, one may call a string *random* for a measure of complexity if its complexity is the maximum possible for its length. For example, a string is random for Kolmogorov complexity if its complexity is equal to its length, up to an additive constant not depending on the string. For A_N, the string w is random if $A_N(w) = \lfloor |w| /2 \rfloor + 1$, and this is known to be tight (except over a binary alphabet; see [10, Theorem 9] and [9]). But without a general asymptotic upper bound, it is unclear what strings could be considered random for A_P, and so we ask:

Question 4.22. Is there a suitable notion of a string being random with respect to A_P? If so, then asymptotically, how many strings are random in this sense?

[3] [4, Theorem 12] establishes that $A_D(0^n1^n) \geq \sqrt{n} - 1$ for all n, but the proof does not quite go through for NFAs. Probably a similar explicit lower bound on A_N can be found.

5. Computability of probabilistic automatic complexity

A primary motivation for introducing the DFA and NFA complexities was that they are computable, unlike the Kolmogorov complexity, as we mentioned earlier. Hence it is natural to ask whether the PFA complexity A_P is also computable, along with its parametrized variant $A_{P,\gamma}$. In this section, we give a strong positive answer to this question by establishing the following:

Theorem 5.1. For every finite alphabet Σ and every $\gamma \in [0, 1)$, the function $w \mapsto A_{P,\gamma}(w, \Sigma)$ is γ-computable.

That is, there is an oracle Turing machine which can compute the value of $A_{P,\gamma}(w, \Sigma)$ when given input w and an oracle encoding the number γ. This theorem is really two theorems in one, and to prove them we will need to study the computability of $A_{P,\gamma}$ from two different points of view. The difference centers on how the number γ is represented, and what is more specifically meant by "γ-computable" depends on that representation. If γ is algebraic and one has a finitary description of it as the root of some polynomial with integer coefficients, then Tarski's classical theorem on the decidability of real closed fields can be applied to show that there is an algorithm computing $A_{P,\gamma}$ as a function from Σ^* to $\mathbb{N}$ (Theorem 5.2). If γ is arbitrary, one can instead represent γ as a rapidly converging sequence of rational numbers and view $A_{P,\gamma}(w)$ as a function both of w and of such a sequence, i.e., as a function from $[0, 1) \times \Sigma^*$ to $\mathbb{N}$. This function is not everywhere continuous, and so cannot be computable—but we show in Theorem 5.3 that it *is* computable where continuous, except possibly at $\gamma = 0$, by a uniform algorithm which works for any point in its domain of continuity (aside from those with $\gamma = 0$). To achieve this, we topologize the space of k-state PFAs for each k and argue about the computability of certain real-valued quantities defined using that space (in particular (97)). Then the proof of Theorem 5.1 is completed by showing that all γ at which this two-variable function is discontinuous are definable in the language of real closed fields, so that Theorem 5.2 applies to them.

We will assume Σ to be fixed in advance for the rest of this section, and moreover that $|\Sigma| > 1$ (since the statement of Theorem 5.1 becomes trivial otherwise: the complexity of every string would be 1).

5.1. Computability for definable γ

The proof of the following theorem was suggested to the author by Bjørn Kjos-Hanssen.

Theorem 5.2. The function $w \mapsto A_{P,\gamma}(w)$ is computable whenever γ is first-order definable in the language $\mathcal{L}$ of real closed fields.

This $\mathcal{L}$ consists of constant symbols 0 and 1 together with a binary relation symbol $<$ and binary function symbols $+$, $-$, and $\cdot$. The precise definition of a real closed field is not relevant for our purposes and may be found, e.g., in [25, §3.3]. The important point for us is that $(\mathbb{R}; 0, 1, +, -, \cdot, <)$ is a real closed field with these symbols being given their usual meaning. Tarski proved that the first-order theory of real closed fields is decidable, i.e., there is an algorithm which decides from a given $\mathcal{L}$-sentence whether the theory of real closed fields proves that sentence [25, Corollary 3.3.16]. Tarski's theorem implies that to prove Theorem 5.2, it suffices to show that for a given k, w, and definable number γ, the relation $A_{P,\gamma}(w) \leq k$ is equivalent to an $\mathcal{L}$-sentence, and that (a Gödel number for) such a sentence can be uniformly computed from k and w if one is given a formula defining γ. The proof is uniform in (a Gödel number for) the formula defining γ as well. (Note that by quantifier elimination, the $\mathcal{L}$-definable numbers are exactly the algebraic numbers.)

Proof:
In what follows, tuples $\bar{a}$ and $\bar{b}$ will always be elements of $\mathbb{R}^{k(bk+2)}$ where $|\Sigma| = b$ is fixed. For each k, define

$$\mathrm{ispfa}_k(\bar{a}) \equiv \bigwedge_{i=1}^{k(bk+2)} (0 \leq a_i \leq 1) \wedge \bigwedge_{n=1}^{bk+1} \left(\sum_{j=nk+1}^{(n+1)k} a_j = 1 \right) \wedge \bigwedge_{i=k(bk+1)+1}^{k(bk+2)} (a_i = 0 \vee a_i = 1). \quad (91)$$

In words, this means that $\bar{a}$ can be split into $bk + 1$ stochastic vectors followed by a 0-1 vector, all of length k. If M is the PFA defined by the tuple $\bar{x}$, write $p_w(\bar{x})$ for $\rho_M(w)$; this is a polynomial in the entries of $\bar{x}$ and hence an $\mathcal{L}$-term which can be uniformly computed from w. Next, given w, write $\Sigma^{|w|} \setminus \{w\}$ as $\{w_1, \ldots, w_n\}$ (with n depending on w and each word appearing only once). Let

$$\mathrm{isgap}_{k,w}(g, \bar{a}) \equiv \bigvee_{i=1}^{n} \left((p_w(\bar{a}) - p_{w_i}(\bar{a}) = g) \wedge \bigwedge_{j \neq i} \big(p_w(\bar{a}) - p_{w_j}(\bar{a}) \geq g\big) \right). \quad (92)$$

Thus $\mathrm{isgap}_{k,w}(g, \bar{a})$ holds if and only if $\mathrm{gap}_M(w) = g$ where M is the k-state PFA defined by $\bar{a}$. It is clear that one can uniformly compute the $\mathcal{L}$-formulas $\mathrm{ispfa}_k(\bar{a})$ and $\mathrm{isgap}_{k,w}(g, \bar{a})$ from k and w. If γ is defined by the formula $\varphi(x)$ (which may have extra parameters), then $A_{P,\gamma}(w) \leq k$ if and only if

$$\exists \bar{a} \exists g_1 \exists g_2 \ \big(\mathrm{ispfa}_k(\bar{a}) \wedge \mathrm{isgap}_{k,w}(g_1, \bar{a}) \wedge \varphi(g_2) \wedge (g_1 > g_2)\big). \quad (93)$$

This completes the proof of Theorem 5.2. □

5.2. Computability for arbitrary γ

We now switch our approach, as mentioned earlier, in order to speak meaningfully of the computability of $A_{P,\gamma}$ for arbitrary γ. Recall that we now view $A_{P,\gamma}(w)$ as a two-variable function of both w and γ, where γ is represented by any sequence of rational numbers rapidly converging to it (a Cauchy name, formally defined below). This subsection is devoted to proving the following:

Theorem 5.3. For any finite Σ, the function from $[0, 1) \times \Sigma^*$ to $\mathbb{N}$ given by $(\gamma, w) \mapsto A_{P,\gamma}(w, \Sigma)$ is

- Continuous everywhere on $[0, 1) \times \Sigma^*$ except on a countably infinite set enumerable by a single algorithm;

- Computable on $(0, 1) \times \Sigma^*$ where it is continuous.

In particular, for every w and Σ, $A_{P,\gamma}(w, \Sigma)$ is computable for all but at most $A_D(w) - 2$ many values of γ, and is continuous at $\gamma = 0$.

The reader should note that the theorem makes no claim one way or the other as to the computability of this function at $\gamma = 0$. While $A_{P,0} = A_P$ is computable by Theorem 5.2, the number 0 in the present context is specified by a Cauchy name rather than being defined by a finitary formula, and the proof of Theorem 5.3 does not extend to this case. Then we may ask

Question 5.4. Is the set $\{0\} \times \Sigma^*$ contained within the domain of computability of the function $(\gamma, w) \mapsto A_{P,\gamma}(w)$?

To prove the theorem we need some machinery from computable analysis, and we introduce the needed background in the next subsection before proceeding to the proof. Afterwards in Section 5.3, we show that Theorem 5.2 can be used to compute $A_{P,\gamma}$ at every one of its discontinuities, which completes the proof of Theorem 5.1.

5.2.1. Background in computable analysis

Our approach is standard and can be found in, e.g., [26]. For a separable metric space (X, d), suppose we are given an enumeration $\alpha \colon \mathbb{N} \to X$ of a dense subset of X. Fix some enumeration $(q_i)_{i\in\mathbb{N}}$ of $\mathbb{Q}$. Then we say X is a *computable metric space* if $d \colon X \times X \to \mathbb{R}$ is computable when restricted to the range of α, in the sense that the set

$$\{ (i, j, n, m) \in \mathbb{N}^4 : q_i < d(\alpha(n), \alpha(m)) < q_j \} \tag{94}$$

is computably enumerable. The function α gives rise to a canonical computable enumeration of a basis for the topology on X, namely

$$\langle i, j \rangle \mapsto B_{q_j}(\alpha(i)), \tag{95}$$

where $B_q(x)$ is the open ball of radius q centered at $x \in X$. We will from now on refer to the sets in this canonical enumeration as *basic open balls*. We may refer to a procedure as "outputting an open ball" or "listing open balls" when we really mean that it produces an index $\langle i, j \rangle$ for a basic open ball, or a list of such indices.

A *name* for a point $x \in X$ is a list N_x^X (in any order) of all basic open balls in X containing x. If (X, d_X) and (Y, d_Y) are two computable metric spaces, a function $f \colon X \to Y$ is *computable* if there is a Turing functional which sends N_x^X to $N_{f(x)}^Y$ for all $x \in X$. A *Cauchy name* for a point x is a sequence $(x_n) \subset D$ converging to x such that for all n, $d(x_n, x_{n+1}) < 2^{-n}$. One can compute a Cauchy name for x from N_x^X by first finding a subsequence of basic open balls listed in N_x^X with exponentially decreasing radii, then taking their centers. Conversely, one can compute a name N_x^X from a Cauchy name: if (x_n) is a Cauchy name for x and $B_q(y)$ is any basic open ball, then $d(x, y) < q$ iff $d(x_n, y) < q - 2^{-n}$ for some n, and the latter will be witnessed in finite time since by assumption $d(x_n, y)$ is computable in the sense given above. Neither algorithm depends on x, and so if f is computable in the above sense, then there is also a uniform computable procedure mapping a Cauchy name for x to a Cauchy name for $f(x)$ for all x. Every computable function is continuous.

The real line $\mathbb{R}$ is a computable metric space with the usual Euclidean metric, taking $D = \mathbb{Q}$. A computable real number is a number having a computable Cauchy name, viewed as an element of Baire space. If $f, g\colon X \to \mathbb{R}$ are computable functions, then so are $f+g$, $f-g$, fg, $\max\{f,g\}$, and $\min\{f,g\}$. In particular, by taking both f and g to be the identity map on $\mathbb{R}$, we get that the function $(x,y) \mapsto \max\{x,y\}$ is computable. If given $x \neq y$, one can also decide in finite time from their Cauchy names which is larger.

A computable metric space X is *computably compact* if there is a computable function which outputs a finite open cover of X by basic open balls of radius at most 2^{-n}, given input n. If $f\colon X \to \mathbb{R}$ is computable and X is computably compact, then $\sup_{x\in X} f(x)$ and $\inf_{x\in X} f(x)$ are computable numbers, and this is uniform in f (identifying f with an index for an oracle Turing machine mapping $x \mapsto f(x)$).

5.2.2. Proof of Theorem 5.3

For any $k \geq 2$, let $\mathscr{A}_k$ denote the space of k-state PFAs over a fixed finite alphabet Σ, identified with $\{0,\dots,b-1\}$. To be precise, define

$$\begin{aligned}\mathscr{A}_k = \big\{(\vec{\pi}, P_0, P_1, \dots, P_{b-1}, \vec{\eta}) : \vec{\pi} \in [0,1]^k \text{ is a probability vector,}\\ \text{each } P_a \text{ is a } k\times k \text{ stochastic matrix, and } \vec{\eta} \in \{0,1\}^k\big\} \subset [0,1]^{2k+bk^2}.\end{aligned} \tag{96}$$

If $A \in \mathscr{A}_k$, write the components of A as $\vec{\pi}^A, P_0^A, \dots, P_{b-1}^A$, and $\vec{\eta}^A$. Also write M^A for the vector $(\vec{\pi}^A, P_0^A, \dots, P_{b-1}^A)$. We give $\mathscr{A}_k$ the uniform (maximum) distance $d(\cdot,\cdot)$, i.e., that induced from the product topology on $[0,1]^{2k+bk^2}$. (The euclidean distance would work just as well.) Then $\mathscr{A}_k$ is a computably compact metric space. There are several easy ways to see this, but we give a direct proof for convenience. Let Q_k be the set of rational k-state PFAs, that is, the set of $A \in \mathscr{A}_k$ such that all entries of M^A are rational, given as quotients of natural numbers. Clearly Q_k has a computable enumeration and is dense in $\mathscr{A}_k$, and $d(A,B)$ is computable for any $A, B \in Q_k$, hence $\mathscr{A}_k$ is a computable metric space. Then for any fixed n, one can enumerate all $A \in Q_k$ such that every entry of M^A is equal to $j2^{-n-1}$ for some $j \in \{0,\dots,2^{n+1}\}$. The set of $B_{2^{-n}}(A)$ for all such A is a finite open cover of $\mathscr{A}_k$ by basic open balls of radius at most 2^{-n}, so $\mathscr{A}_k$ is computably compact by definition.

The function $(A,w) \mapsto \rho_A(w)$ is computable, because it is a polynomial in the entries of A resulting from multiplication of $\vec{\pi}^A$, $\vec{\eta}^A$, and the matrices P_a^A in an order determined by w. Therefore $(A,w) \mapsto \operatorname{gap}_A(w)$ is the minimum of finitely many computable functions and hence itself computable, as is $A \mapsto \operatorname{gap}_A(w)$ for any fixed w. Now let

$$\Gamma_k(w) = \max_{A\in\mathscr{A}_k} \operatorname{gap}_A(w). \tag{97}$$

For each k and w, $\Gamma_k(w)$ is a computable real number, because it is equal to the supremum of the computable function $A \mapsto \operatorname{gap}_A(w)$ over the computably compact space $\mathscr{A}_k$. And since the procedure to compute $\operatorname{gap}_A(w)$ is uniform in w, the function $(k,w) \mapsto \Gamma_k(w)$ is computable. Finally, let

$$E = \{\, (\Gamma_k(w), w) \colon 2 \leq k \leq A_D(w) - 1,\ w \in \Sigma^*, 0 < \Gamma_k(w) < 1 \,\} \subset (0,1) \times \Sigma^*. \tag{98}$$

This will turn out to be exactly the set of discontinuities of $A_{P,\gamma}(w)$, and it can clearly be enumerated by a single algorithm by definition. Proposition 3.1(ii) implies that $A_{P,\gamma}(w)$ is continuous at $(0, w)$ for all w. Continuity on the remainder of the complement of E will follow from the computability argument below.

That E is countably infinite is a consequence of the following fact of potential independent interest, whose proof establishes that in some sense, a 2-state PFA giving a gap of 1 to even a single word (with more than three letters) behaves much like a DFA as far as A_P is concerned.

Lemma 5.5. For any w with $|w| \geq 4$, $\Gamma_2(w) = 1$ iff w is constant.

Proof:
The right-to-left implication is immediate, since then $A_D(w) = 1$ (or technically 2 if considering DFAs over Σ with $|\Sigma| \geq 2$, but this does not change the statement). For the other direction, assume for sake of contradiction that w is nonconstant, and that $\Gamma_2(w) = 1$ is witnessed by the IFS with starting value x_0 and maps $f_j x = a_j + b_j x$ for each letter $j \in \Sigma$. Then $\rho(w) = 1$ and $\rho(y) = 0$ for every other y of length $|w|$, and if $w = z^\frown i$ where $i \in \Sigma$, then in particular $f_i \rho(z) = 1$ and $f_j \rho(z) = 0$ for all $j \neq i$. Now, if the range of f_i omits the value 0, then $\rho(i^n) > 0$ for all n, regardless of the value of x_0. Then either w is constant or $\text{gap}(w) < 1$, a contradiction, and we may thus assume the range of f_i includes both 0 and 1. By drawing a picture, one sees that only $f_i x = x$ and $f_i x = 1 - x$ are possible. If $f_i x$ is the identity then only constant strings may be witnessed, so we can assume that $f_i x = 1 - x$.

If $f_i x = 1 - x$, then $\rho(z) = f_i^{-1}(1) = 0$, so $f_j 0 = 0$ and thus $f_j x = b_j x$ for all $j \neq i$. We can take $b_j < 1$, as otherwise f_j is the identity map and only constant strings can be witnessed. If $b_j = 0$ for some j, so that $f_j \equiv 0$, then every string ending in ji has probability 1, thus maximal probabilities are nonunique starting at length 3, a contradiction. Then $0 < b_j < 1$ for all $j \neq i$, and this means once an orbit leaves $\{0, 1\}$ it can never return to either value. In particular, $x_0 \in \{0, 1\}$. If $x_0 = 0$, then for all $n \geq 1$ we have $\rho(ji^{2n-1}) = \rho((ji)^{2n}) = 1$ among even-length strings and $\rho(i^{2n+1}) = \rho(j^2 i^{2n-1}) = 1$ among odd-length strings. If $x_0 = 1$, then for all $n \geq 1$ we have $\rho(i^{2n}) = \rho(ij^{2n-2}i) = 1$ among even-length strings and $\rho(ij^{2n-1}i) = \rho(iji^{2n-1}) = 1$ among odd-length strings. Either way, uniqueness of maxima is lost starting at length at most 4, so $\text{gap}(w) < 1$ and by contradiction the proof is complete. □

There are infinitely many nonconstant w with $|w| \geq 4$ and $A_P(w) = 2$, of course, by Theorem 4.1. For such w, the lemma implies that $0 < \Gamma_2(w) < 1$, so that $(\Gamma_2(w), w) \in E$ and in particular E is infinite.

We now show that $A_{P,\gamma}(w)$ is discontinuous on E and computable on the complement of E, minus the points with $\gamma = 0$. Endow Σ^* with the discrete topology in its standard metrization, i.e., $d(x, y) = 1$ iff $x \neq y$. Then we give $[0, 1) \times \Sigma^*$ the product metric, that is, $d((\alpha, x), (\beta, y)) = \max\{|\alpha - \beta|, d_{\Sigma^*}(x, y)\}$. The codomain $\mathbb{N}$ of $A_{P,\gamma}(w)$ also has the discrete topology as a subset of $\mathbb{R}$. Now, $A_{P,\gamma}(w)$ is continuous at (γ, w) iff for all $\varepsilon > 0$ there is an $\eta > 0$ such that $d((\gamma, w) - (\gamma', w')) < \eta$ implies $|A_{P,\gamma'}(w) - A_{P,\gamma}(w)| < \varepsilon$—so that actually $|\gamma - \gamma'| < \eta$ implies $A_{P,\gamma'}(w) = A_{P,\gamma}(w)$ (since Σ^* and $\mathbb{N}$ both have the discrete topology). If $\gamma = \Gamma_k(w)$ for some k and w, then by definition of Γ_k there is no $\gamma' < \gamma$ such that $A_{P,\gamma'}(w) = A_{P,\gamma}(w)$, because there is a k-state PFA having a gap greater than γ' for w but not one having a gap greater than γ. Hence $A_{P,\gamma}(w)$ is discontinuous at every point of E.

Finally, let $(\gamma, w) \notin E$ be given with $\gamma \neq 0$. Under these hypotheses, for any $k \geq 2$, we have $\gamma > \Gamma_k(w)$ if and only if $A_{P,\gamma}(w) > k$, because in this case there is no $A \in \mathscr{A}_k$ exhibiting the required gap. Conversely, $\gamma < \Gamma_k(w)$ if and only if $A_{P,\gamma}(w) \leq k$. To compute $A_{P,\gamma}(w)$, then, decide for each $k = 2, 3, \ldots, A_D(w)$ whether γ or $\Gamma_k(w)$ is greater. The least k such that $\gamma < \Gamma_k(w)$ is exactly equal to $A_{P,\gamma}(w)$. It is clear that this procedure does not depend on γ or w, and the proof of Theorem 5.3 is complete. □

5.3. Proof of Theorem 5.1

We are now ready to finish the proof of Theorem 5.1 by showing that γ is definable in the language $\mathcal{L}$ whenever $(\gamma, w) \in E$, where $\mathcal{L}$ was defined after the statement of Theorem 5.2 and E is given by (98), so that Theorem 5.2 applies to every such γ. It will then follow that for every γ, either $A_{P,\gamma}$ is computable outright, or it is oracle computable from a Cauchy name for γ. Let

$$\begin{aligned} \operatorname{ismaxgap}_{k,w}(g) \equiv {} & \big(\exists \bar{a}\, \big[\operatorname{ispfa}_k(\bar{a}) \wedge \operatorname{isgap}_{k,w}(g, \bar{a})\big]\big) \\ & \wedge \left(\forall \bar{b} \left[\operatorname{ispfa}_k(\bar{b}) \to \bigvee_{i=1}^{n} \big(p_w(\bar{b}) - p_{w_i}(\bar{b}) \leq g\big)\right]\right), \end{aligned} \tag{99}$$

where all notation is as in the proof of Theorem 5.2. In words, this says there is a k-state PFA giving w gap g and that all k-state PFAs give w gap at most g. Hence $\operatorname{ismaxgap}_{k,w}(g)$ holds iff $g = \Gamma_k(w)$. A Gödel number for $\operatorname{ismaxgap}_{k,w}$ can evidently be uniformly computed from k and w, and so we are done. □

Remark 5.6. The proof of Theorem 5.2 can be adapted to show that many variants on A_P and $A_{P,\gamma}$ are computable. For example, one might change $A_{P,\gamma}$ to $A_{P,\geq\gamma}$ by requiring a witness M to satisfy $\operatorname{gap}_M(w) \geq \gamma$ rather than $> \gamma$, and then replace $g_1 > g_2$ in (93) by $g_1 \geq g_2$ to show $A_{P,\geq\gamma}$ is computable for definable γ. The proof of Theorem 5.3 goes through verbatim for $A_{P,\geq\gamma}$, and one thus recovers Theorem 5.1. Something similar can also be done for the variants discussed in Section 6.1 below.

6. Other approaches to probabilistic complexity

6.1. Relaxing the definition of a PFA

We saw earlier that A_P shares the property of A_D that the complexity of a string is not necessarily equal to that of its reversal. In addition, there are strings whose PFA complexity is known to be witnessed by a PFA with dead states. One might try to solve these problems by relaxing the definition of a PFA to directly generalize an NFA (rather than a DFA). NFAs are allowed to have rows of all zeros in their transition matrices, and also have the property that different out-transitions from the same state and for the same letter are not weighted differently—they are simply all possible. The same applies to the initial state distribution.

To directly translate these properties to a generalization of a PFA, one would need to require that all nonzero entries of $\vec{\pi}$ are equal, as well as the same for $\vec{\eta}$, and that all nonzero entries of all the matrices P_a are equal to the same number (which may result in the row sums being different). One can see that the proof of the first part of Proposition 3.4 can be recovered for the class $\mathscr{N}$ of such automata. Since $\mathscr{N}$ is closed under switching $\vec{\pi}$ and $\vec{\eta}$ and taking the transpose of every transition matrix, if $\tilde{A}_P$ is the corresponding complexity notion where $\tilde{A}_P(x)$ must be witnessed by a member of $\mathscr{N}$, then one has $\tilde{A}_P(\overleftarrow{x}) = \tilde{A}_P(x)$ for all x. (Moreover $\tilde{A}_P(x) \leq A_N(x)$.) However, this is not a very natural class of automata to consider and it is certainly not a generalization of a PFA.

Instead of trying to design a specific class of automata in an attempt to recover properties of A_N, it might make more sense to define a unified complexity notion which takes as parameter a family of automata and study its properties in general. In [27], Turakainen introduced *generalized (probabilistic) finite automata* (GPFAs), which are finite-state automata whose operation is described as follows:

- The initial state of the machine is an arbitrary real row vector.
- Transitions between states are described by multiplication of arbitrary real square matrices.
- The final state of the machine is again an arbitrary real column vector.

So GPFAs are like PFAs except that the entries of $\vec{\pi}$, $\vec{\eta}$, and each P_a can be any real numbers. Turakainen proved the remarkable fact that GPFAs have in a sense the same descriptive power as PFAs: if one also allows a cut-point in the context of a GPFA to be any real number, then the class of languages accepted by GPFAs is exactly the class of stochastic languages.

This suggests that it is not too unreasonable to throw the gates open and consider a version of A_P that allows any GPFA. Let $\mathscr{G}$ be the set of all GPFAs. For any family $\mathscr{F} \subseteq \mathscr{G}$, let $\mathscr{F}_k$ be the set of members of $\mathscr{F}$ having k states. Then define the *$\mathscr{F}$-automatic complexity* of a word $x \in \Sigma^*$ to be

$$A_{\mathscr{F}}(x) = \min\{\, k : \exists F \in \mathscr{F}_k \text{ such that } \operatorname{gap}_F(x) > 0 \,\}. \tag{100}$$

For example, A_P as defined before coincides with $A_{\mathscr{A}}$ if $\mathscr{A} \subset \mathscr{G}$ is the set of all PFAs. One can also define $A_{\mathscr{F},\gamma}$ for any $\gamma \geq 0$ by analogy with $A_{P,\gamma}$. We have that for all x,

$$A_{\mathscr{E}}(x) \leq A_{\mathscr{F}}(x) \quad \text{whenever} \quad \mathscr{E} \supseteq \mathscr{F}, \tag{101}$$

so that in particular $A_{\mathscr{G}}(x) \leq A_{\mathscr{F}}(x)$ for every $\mathscr{F}$ and x. We have not investigated $A_{\mathscr{F}}$ in general, and it is unclear how coarse of a measurement it might be. As a motivating question, we could ask

Question 6.1. Is $A_{\mathscr{G}}(x) \leq 2$ for all binary strings x?

It is at least true that $A_{\mathscr{G}}(x) = A_{\mathscr{G}}(\overleftarrow{x})$ for all x, by the same observation we made above for $\tilde{A}_P$—and indeed $A_{\mathscr{G},\gamma}(x) = A_{\mathscr{G},\gamma}(\overleftarrow{x})$ for all x and $\gamma \geq 0$, since one can simply switch $\vec{\pi}$ and $\vec{\eta}$ and replace P_a with P_a^T for all a to change a witness for $A_{\mathscr{G},\gamma}(x)$ into one for $A_{\mathscr{G},\gamma}(\overleftarrow{x})$ while preserving the acceptance probability of every word. Together with Proposition 3.3, whose proof goes through verbatim for $A_{\mathscr{G},\gamma}$, this implies that $A_{\mathscr{G},\gamma}(xyz) \geq A_{\mathscr{G},\gamma}(y)$ for all x, y, and z, like A_D and A_N.[4] Theorem 5.2 also holds for $A_{\mathscr{G}}$ simply by making $\mathrm{ispfa}_k(\bar{a})$ true for all $\bar{a}$ and all k.

A potentially helpful observation here is that the ability to have unbounded real entries does not really confer any advantage as far as the complexity of individual strings is concerned. For any GPFA M, if C is the largest absolute value of any entry of $\vec{\pi}^M$, $\vec{\eta}^M$, and the matrices P_a^M, then one could divide all these matrices and vectors by C to obtain a GPFA M' with entries in $[-1,1]$ such that

$$\rho_M(x) < \rho_M(y) \iff \rho_{M'}(x) < \rho_{M'}(y) \tag{102}$$

whenever $|x| = |y|$. Hence if $\mathscr{S}$ is the set of GPFAs whose entries are all in $[-1,1]$, we have $A_{\mathscr{S}}(x) = A_{\mathscr{G}}(x)$ for all x. In addition, the direct analogue of Theorem 5.3 holds for $A_{\mathscr{S},\gamma}$, because $\mathscr{S}_k$ is now a computably compact metric space for each k (and each fixed alphabet), unlike $\mathscr{G}_k$.

One advantage of A_P that appears to be immediately lost in passing to $A_{\mathscr{G}}$ or $A_{\mathscr{S}}$ is the dimension reduction of the IFS approach, and the dynamical analysis made more tractable by it. Since the correspondence between PFAs and IFSs relies explicitly on the transition matrices being stochastic, perhaps one could allow only generalized stochastic transition matrices, with any real entries permitted as long as each row sums to 1. This notion would for example allow us to describe 0100 in two states via

$$P_0 = \begin{pmatrix} -1 & 2 \\ 1/2 & 1/2 \end{pmatrix}, \quad P_1 = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}, \quad \vec{\pi} = (0,1), \quad \vec{\eta} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{103}$$

whereas $A_P(0100) = 3$, so strictly greater compression is achieved. This automaton is equivalent to the IFS with $f_0(x) = \frac{1}{2} - \frac{3}{2}x$, $f_1(x) = 1 - \frac{1}{2}x$, and $x_0 = 0$. (Other strings with $A_P = 3$ which have complexity 2 according to this notion include 01000, 01011, and 01100.) Unfortunately, uniformly rescaling the transition matrices as with $\mathscr{S}$ no longer works here, so the set of allowed transition probabilities is unbounded and we lose uniform computability of the analogue of the two-variable function $(w,\gamma) \mapsto A_{P,\gamma}(w)$, i.e., the proof of Theorem 5.3 cannot be recovered. Of course, the proof of Theorem 5.2 can still be easily modified to fit this case by simply dropping the requirement that probabilities lie between 0 and 1 from the formula ispfa_k.

6.2. Gap structure function

We saw in the proof of Theorem 5.3 that the function $\Gamma_k(w)$ mapping w to the maximum value of $\mathrm{gap}_M(w)$ among all k-state M is computable. It could be interesting to study $w \mapsto \Gamma_k(w)$ as a parametrized complexity measure in itself. We have $0 \leq \Gamma_2(w) \leq \Gamma_3(w) \leq \cdots \leq \Gamma_{A_D(w)}(w) = 1$, with $\Gamma_k(w) = 0$ if and only if $k < A_P(w)$. The number $\Gamma_k(w)$ is never negative since one can always make a PFA accepting every word with the same probability by setting all transition matrices to the identity matrix. Furthermore, Proposition 3.3 implies $\Gamma_k(z) \geq \Gamma_k(wz)$ for all w, z, and k. This comes close to justifying the empirical observation made in Section 3 that gaps tend to decrease for longer words. A result to the effect that $\Gamma_k(w) \geq \Gamma_k(wz)$ would put the observation on fully rigorous ground:

[4] This result can in fact already be recovered if one merely weakens the definition of a PFA to allow $\vec{\eta}$ to be an arbitrary probability vector, like $\vec{\pi}$. Proceed in a similar way as in the proof of Proposition 3.3, except making M' have final vector $\vec{\eta}' = P_M(y)\vec{\eta}$ while keeping its initial vector the same. Then one gets the analogue of $A_{P,\gamma}(xy) \geq A_{P,\gamma}(x)$ for all x, y, γ.

Question 6.2. Is $A_{P,\gamma}(wz) \geq A_{P,\gamma}(w)$ for all w, z, and γ?

Intuitively, $\Gamma_k(w)$ measures how well w is described by the model class of k-state PFAs—the widest margin of probability by which w can be recognized by any such PFA. This relates it at least philosophically to the Kolmogorov structure function, which measures the minimal size of a set of strings containing w which can be described by a Turing machine of size at most k, and hence captures in a sense how well w can be singled out by such machines. Similar functions have also been considered by Kjos-Hanssen [8], who introduced both a structure function and a dual structure function for the NFA complexity. His dual structure function is in part motivated by having a simple domain and complicated range, rather than the other way around as with his regular structure function for A_N. This is even more true for $\Gamma_k(w)$ in contrast with its dual $A_{P,\gamma}(w)$: it maps a string and natural number to a Cauchy name for a real number, rather than mapping a string and some representation of a real to a natural number.

6.3. Least number of bits of a witness

Heuristically it appears that witnesses for the PFA complexity of many strings are relatively complicated; this certainly seems to be the case for most strings with $A_P = 2$, as pointed out below. If one is interested solely in compression, it might make the most sense to measure the complexity of w as the least number of bits required to describe an M having $\mathrm{gap}_M(w) > 0$, or perhaps $\mathrm{gap}_M(w) > \gamma$ for a parameter γ. One potential drawback of this approach is that it is not obvious whether this measure is computable, although this depends on the precise definition used. The least number of bits also depends on the choice of encoding, and so this measure would only be defined up to an additive constant, like the Kolmogorov complexity. Not only that, but it could well be that the simplest PFAs achieving a positive gap are very often DFAs, and in that case one could argue it is hardly a satisfying notion of PFA complexity.

6.4. Measure of the set of witnesses

We conclude by mentioning one more idea for modifying A_P and $A_{P,\gamma}$, with the aim of refining the numerical measurement itself. In the proof of for example Proposition 4.11, we saw that although all strings $0^n 1^m$ have complexity 2, as n increases, x_0 must be chosen in a narrower and narrower range in order for the IFS to witness $0^n 1^m$. The coefficient b must also be made arbitrarily close (but not equal) to 1. Something similar is true of the other subcases of the proof of Theorem 4.3. Thus it is in a sense more complicated to witness the complexity of a string the longer its prefix is. So, we could introduce a real-valued complexity measure that accounts for that difference as follows. Let μ be a Borel probability measure with full support on $\mathscr{A}_k$, the space of k-state PFAs. Let $G^k(x) = \mathrm{gap}_\bullet(x)^{-1}((0,1])$ be the set of k-state witnesses for $A_P(x) \leq k$, and let

$$A_\mu(x) = A_P(x) + 1 - \mu(G^k(x)). \tag{104}$$

We can also let $G^k_\gamma(x) = \mathrm{gap}_\bullet(x)^{-1}((\gamma,1])$ and define

$$A_{\mu,\gamma}(x) = A_{P,\gamma}(x) + 1 - \mu(G^k_\gamma(x)). \tag{105}$$

Since gap is a computable function on a computably compact metric space, $G^k(x)$ and $G^k_\gamma(x)$ are c.e. open, meaning the indices of all basic open balls contained in each of them can be computably enumerated. In particular, all these sets have positive μ-measure if nonempty. Thus $A_\mu(x)$ assigns x a value strictly between $A_P(x)$ and $A_P(x)+1$, and $A_{\mu,\gamma}(x)$ is strictly between $A_{P,\gamma}(x)$ and $A_{P,\gamma}(x)+1$. Moreover, in at least the case of binary strings with $A_P(x) = 2$, if x is a string such as 0^n101 which can only be witnessed by a PFA that also witnesses $0^n1(01)^m$ for all m, then those strings receive exactly the same value of A_μ as x does. This makes sense, because these extensions of x in a sense do not require any further effort to find a witness. The latter observation holds for any measure μ. The goal in defining A_μ (or $A_{\mu,\gamma}$) would then be to find a suitable μ which gives sets like $G^k(x)$ (or $G^k_\gamma(x)$) large measure for strings like $x = (01)^n$ which are easy to witness, while giving smaller measure to $G^k(x)$ (or $G^k_\gamma(x)$) for strings whose witnessing automata require a more precise configuration.

A natural choice would be for μ to be induced by Lebesgue measure on the unit simplex in $\mathbb{R}^{k-1}$, identifying k-state PFAs with $(k-1)$-dimensional affine IFSs as in Section 4.1. If $|\Sigma| = b$, one could take the $(bk+1)$-fold product of the $(k-1)$-dimensional Lebesgue measure with itself, one for each stochastic row vector in an element of $\mathscr{A}_k$, and average it over the $2^k - 2$ connected components of $\mathscr{A}_k$ corresponding to nontrivial choices of $\vec{\eta}$. The result is a fully supported computable probability measure μ on $\mathscr{A}_k$.

Question 6.3. Does this μ lead to satisfactory, and in particular computable, complexity measures A_μ and $A_{\mu,\gamma}$?

Question 6.4. How should the definitions of A_μ and $A_{\mu,\gamma}$ account for the fact that lower-complexity strings are also witnessed by members of $\mathscr{A}_k$? How should one deal with the likely problem of the sets $G^k(x)$ generally having high measure when $A_P(x) < k$, which would make A_μ clustered near $k+1$ among strings having $A_P(x) = k$?

References

[1] Gill K. Two studies in complexity. Ph.D. thesis, Penn State University, 2023.

[2] Diwan AA. A new combinatorial complexity measure for languages. Technical report, Computer Science Group, Tata Institute, 1986.

[3] Bannai H, Hirayama M, Hucke D, Inenaga S, Jez A, Lohrey M, Reh CP. The smallest grammar problem revisited. *IEEE Trans. Inf. Theory*, 2021. **67**:317–328. doi:10.1109/TIT.2020.3038147. arXiv:1908.06428.

[4] Shallit J, Wang Mw. Automatic complexity of strings. *J. Autom. Lang. Comb.*, 2001. **6**(4):537–554. doi:10.25596/JALC-2001-537.

[5] Hyde K. Nondeterministic finite state complexity. Master's thesis, University of Hawai'i, Manoa, 2013.

[6] Hyde K, Kjos-Hanssen B. Nondeterministic automatic complexity of overlap-free and almost square-free words. *Electronic J. Comb.*, 2015. **22**(3):Paper 3.22. doi:10.37236/4851. arXiv:1402.3856 (updated version, 2020).

[7] Kjos-Hanssen B. An incompressibility theorem for automatic complexity. *Forum of Mathematics, Sigma*, 2021. **9**:paper e62. doi:10.1017/fms.2021.58. arXiv:1908.10843.

[8] Kjos-Hanssen B. Kolmogorov structure functions for automatic complexity. *Theoret. Comput. Sci.*, 2015. **607**:435–445. doi:10.1016/j.tcs.2015.05.052. `arXiv:1409.0584`.

[9] Kjos-Hanssen B. Maximal automatic complexity and context-free languages. In: Aspects of Computation and Automata Theory with Applications, pp. 335–352. 2023. doi:10.1142/9789811278631_0013. `arXiv:2206.10130`.

[10] Kjos-Hanssen B. On the complexity of automatic complexity. *Theory Comput. Syst.*, 2017. **61**:1427–1439. doi:10.1007/S00224-017-9795-4. `arXiv:1607.06106`.

[11] Kjos-Hanssen B. Automatic complexity: A computable measure of irregularity. De Gruyter, 2024. doi:10.1515/9783110774870.

[12] Rabin MO. Probabilistic automata. *Inform. and Control*, 1963. **6**:230–245. doi:10.1016/S0019-9958(63)90290-0.

[13] Chadha R, Sistla AP, Viswanathan M. Probabilistic automata with isolated cut-points. In: Chatterjee K, Sgall J (eds.), MFCS 2013 (LNCS, vol. 8087), pp. 254–265. Springer, 2013. doi:10.1007/978-3-642-40313-2_24.

[14] Carlyle JW. Reduced forms for stochastic sequential machines. *J. Math. Anal. Appl.*, 1963. **7**:167–175. doi:10.1016/0022-247X(63)90045-3.

[15] Vidal E, Thollard F, Higuera Cdl, Casacuberta F, Carrasco RC. Probabilistic finite-state machines—Part I & II. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005. **27**(7):1013–1039. doi:10.1109/TPAMI.2005.147.

[16] Calude CS, Salomaa K, Roblot TK. Finite state complexity. *Theoret. Comput. Sci.*, 2011. **412**:5668–5677. doi:10.1016/j.tcs.2011.06.021.

[17] Gill K. WeightedAutomata. doi:10.5281/zenodo.13821868. URL `https://github.com/nowheredense/weightedautomata`.

[18] Paz A. Introduction to probabilistic automata. Academic Press, 1971. doi:10.1016/C2013-0-11297-4.

[19] Barnsley MF, Hurd LP. Fractal image compression. AK Peters, 1993. ISBN 978-1568810003.

[20] Culik II K, Dube S. Rational and affine expressions for image description. *Discrete Appl. Math.*, 1993. **41**:85–120. doi:10.1016/0166-218X(93)90031-I.

[21] Sprott JC. Automatic generation of iterated function systems. *Comput. & Graphics*, 1994. **18**(3):417–425. doi:10.1016/0097-8493(94)90042-6.

[22] Culik II K, Dube S. Affine automata and related techniques for generation of complex images. *Theoret. Comput. Sci.*, 1993. **116**:373–398. doi:10.1016/0304-3975(93)90329-R.

[23] Rystsov IK. Affine automata and classical fractals. *Cybernet. Systems Anal.*, 2018. **54**(1). doi:10.1007/s10559-018-0003-6.

[24] Kocić LM, Simoncelli AC. Cantor dust by AIFS. *Filomat*, 2001. **15**:265–276. URL `http://www.jstor.org/stable/26453430`.

[25] Marker D. Model theory: An introduction, volume 217 of *Graduate Texts in Mathematics*. Springer, 2002.

[26] Downey RG, Melnikov AG. Computably compact metric spaces. *Bull. Symb. Log.*, 2023. **29**(2):170–263. doi:10.1017/bsl.2023.16. URL `https://homepages.ecs.vuw.ac.nz/~melnikal/compcomp(BSL).pdf`.

[27] Turakainen P. Generalized automata and stochastic languages. *Proc. Amer. Math. Soc.*, 1969. **21**:303–309. doi:10.2307/2036989.