

Approximate NFA Universality and Related Problems Motivated by Information Theory^{*}

Stavros Konstantinidis¹, Mitja Mastnak¹, Nelma Moreira², and Rogério Reis²

¹ Saint Mary's University, Halifax, Nova Scotia, Canada,
s.konstantinidis@smu.ca, mmastnak@cs.smu.ca,

² CMUP & DM, DCC, Faculdade de Ciências da Universidade do Porto,
Rua do Campo Alegre, 4169-007 Porto, Portugal
{nelma.moreira,rogerio.reis}@fc.up.pt

Abstract. In coding and information theory, it is desirable to construct maximal codes that can be either variable length codes or error control codes of fixed length. However deciding code maximality boils down to deciding whether a given NFA is universal, and this is a hard problem (including the case of whether the NFA accepts all words of a fixed length). On the other hand, it is acceptable to know whether a code is ‘approximately’ maximal, which then boils down to whether a given NFA is ‘approximately’ universal. Here we introduce the notion of a $(1 - \varepsilon)$ -universal automaton and present polynomial randomized approximation algorithms to test NFA universality and related hard automata problems, for certain natural probability distributions on the set of words. We also conclude that the randomization aspect is necessary, as approximate universality remains hard for any polynomially computable ε .

1 Introduction

It is well-known that NFA universality is a PSPACE-hard problem and that block NFA universality (whether an NFA of some fixed length words accepts all the words of that length) is a coNP-hard problem. Here we consider polynomial approximation algorithms for these and related NFA problems by considering the concept of an approximate universal NFA, or block NFA, where for instance 95% of all words are accepted by the NFA. In general, for some tolerance $\varepsilon \in (0, 1)$, we assume that we are happy to know that an NFA is at least $(1 - \varepsilon)$ universal. While approximate universality is still hard, it allows us to consider polynomial randomized algorithms that return an incorrect answer with small probability. Inspired from [16, pg 72], we view estimating the universality index of an NFA as the problem of estimating the parameter of some population and then follow the tools of [16] for parameter estimation problems.

Our motivation for defining the concept of approximate universality comes from the problem of generating codes (whether variable length codes, or fixed

^{*} Research supported by NSERC, Canada (Discovery Grants of S.K. and of M.M.) and by CMUP through FCT project UIDB/00144/2020.

length error control codes) that are maximal, where on the one hand the question of deciding maximality is hard, but on the other hand it is acceptable to generate codes that are maximal within a tolerance ε , [6,14]. For infinite languages, we define approximate universality relative to some probability distribution on the set of words. This idea is consistent with our interpretation of languages in the context of coding and information theory where words are in fact abstractions of physical network signals or magnetic polarities, [15,12], and the amount of energy they require should not be exponential.

Our work falls under the general framework of problems about parameter estimation or approximate counting [16,8,2], however, we are not aware of the application of this framework in hard NFA problems, especially in the case where the NFA accepts an infinite language.

Main results and structure of the paper. The next section contains basic notation from formal languages and automata as well concepts of probability distributions on the nonnegative integers, in particular the three distributions: uniform, Lambert and Dirichlet. The Dirichlet distribution is a good substitute for the ‘fictitious’ uniform distribution on the nonnegative integers [9]. **Section 3** discusses what a polynomial randomized approximation (PRAX) algorithm should be for the case of a hard decision problem on NFAs. The necessity for PRAX-like algorithms for NFA universality is demonstrated with (i) the observation that a nonrandomized polynomial approximation (PAX) algorithm might not exist and (ii) the result that $(1 - \varepsilon)$ approximate block NFA universality is hard for every ε that is computable within polynomial time. **Section 4** is about probability distributions on words over some alphabet $A_s = \{0, 1, \dots, s\}$ such that the length sets of these distributions follow the above three distributions on the nonnegatives. **Section 5** considers whether an NFA a is universal relative to a maximum language M (i.e., whether $L(a) = M$), and takes the approach that M is the domain of a probability distribution W on the set of words, in which case the universality index $W(a)$ of a is the probability that a word selected from the distribution W belongs to $L(a)$. Then, a is $p\%$ -universal relative to W if $W(a) \geq p\%$. The section closes with two simple random processes about estimating the universality index of NFAs. **Section 6** gives PRAX algorithms for three hard NFA problems: ADFA.SUBSET_NFA (whether $L(b) \subseteq L(a)$ for given NFA a and acyclic DFA b); UNIV_BNFA (whether $L(a) = A_s^\ell$, for given block NFA a of word length ℓ); and UNIV_MAXLEN_NFA (whether $A_s^{\leq \ell} \subseteq L(a)$, for given NFA a and word length ℓ). **Section 7** defines what a tractable length distribution (on the nonnegatives) is and gives a PRAX algorithm for whether a given NFA is universal relative to any fixed, but arbitrary, tractable word distribution (including the word distributions that are based on the Lambert and Dirichlet length distributions). The section also discusses a possible PAX algorithm for NFA universality relative to a tractable distribution. The last section contains a few concluding results and a short discussion on related hard problems.

2 Basic Notation and Background Information

We use the notation \mathbb{N} for the set of positive integers, \mathbb{N}_0 for the nonnegative integers, and $\mathbb{N}^{>x}$ for the positive integers greater than x , where x is any real number. We assume the reader to be familiar with basics of formal languages and finite automata [18, 11]. Our arbitrary alphabet will be $\mathbf{A}_s = \{0, 1, \dots, s-1\}$ for some positive integer s . Then, we use the following notation

- ε = empty word, $|w|$ = length of word w
- \mathbf{A}_s^ℓ = all words of length ℓ , $\mathbf{A}_s^{\leq \ell}$ = all words of length at most ℓ
- DFA = all DFAs (deterministic finite automata)
- NFA = all NFAs (nondeterministic finite automata)
- ADFA = all acyclic DFAs (accepting finite languages)
- BNFA = all block NFAs, that is, NFAs accepting languages of a fixed word length.
- BNFA[s] = all block NFAs over the alphabet \mathbf{A}_s .
- $|\mathbf{a}|$ = the size of the NFA \mathbf{a} = the number of states plus the number of transitions in \mathbf{a} .
- $L(\mathbf{a})$ = the language accepted by the NFA, or DFA, \mathbf{a} .

Notes: We assume that NFAs have no ε -transitions. It makes no difference in this paper whether a DFA is complete or incomplete.

Next we list some decision problems about automata that are known to be hard, or easily shown to be hard.

- UNIV_NFA = $\{\mathbf{a} \in \text{NFA} : L(\mathbf{a}) = \mathbf{A}_s^*\}$: Deciding whether a given NFA is universal is a PSPACE-complete problem, [11].
- UNIV_BNFA = $\{\mathbf{b} \in \text{BNFA} : L(\mathbf{b}) = \mathbf{A}_s^\ell, \text{ where } \ell \text{ is the word length of } \mathbf{b}\}$: Deciding whether a given block NFA of some word length ℓ accepts all words of length ℓ is a coNP-complete problem, [14].
- UNIV_MAXLEN_NFA = $\{(\mathbf{a}, \ell) : \mathbf{a} \in \text{NFA}, \ell \text{ is unary in } \mathbb{N}, L(\mathbf{A}_s^{\leq \ell}) \subseteq L(\mathbf{a})\}$: Deciding whether $L(\mathbf{A}_s^{\leq \ell}) \subseteq L(\mathbf{a})$, for given $\mathbf{a} \in \text{NFA}$ and **unary** $\ell \in \mathbb{N}_0$, is coNP-complete, [7].
- ADFA_SUBSET_NFA = $\{(\mathbf{a}, \mathbf{b}) : \mathbf{a} \in \text{NFA}, \mathbf{b} \in \text{ADFA}, L(\mathbf{b}) \subseteq L(\mathbf{a})\}$: Deciding whether $L(\mathbf{b}) \subseteq L(\mathbf{a})$, for given $\mathbf{a} \in \text{NFA}$ and $\mathbf{b} \in \text{ADFA}$ is PSPACE-complete—see below Remark 1.
- EMPTY_DFA = $\{(\mathbf{a}_1, \dots, \mathbf{a}_n) : n \in \mathbb{N}, \mathbf{a}_i \in \text{DFA}, \cap_{i=1}^n L(\mathbf{a}_i) = \emptyset\}$: Deciding whether the intersection of given DFAs is empty is PSPACE-complete, [7]. Note that the problem remains hard even if we know that the languages of the given DFAs belong to low levels of the dot-depth or the Straubing-Thérien hierarchies [3].

Remark 1. The problem ADFA_SUBSET_NFA is PSPACE-hard. This follows when we see that UNIV_NFA can be reduced to it using the fact that $L(\mathbf{a}) = \mathbf{A}_s^*$ iff $\mathbf{A}_s^* \subseteq L(\mathbf{a})$. The problem is in PSPACE: as \mathbf{b} is acyclic, one can enumerate all words of $L(\mathbf{b})$, [1], testing whether each one is in $L(\mathbf{a})$; this process works within polynomial space.

Probability distributions. Let X be a countable set. A probability distribution on X is a function $D : X \rightarrow [0, 1]$ such that

$$\sum_{x \in X} D(x) = 1. \quad (1)$$

The domain of D , denoted by $\text{dom } D$, is the subset $\{x \in X \mid D(x) > 0\}$ of X . If $X = \{x_1, \dots, x_\ell\}$, for some $\ell \in \mathbb{N}$, then we write

$$D = (D(x_1), \dots, D(x_\ell)).$$

Following [10], we have the following definition.

Definition 1. Let D be a probability distribution on X . For any subset S of X , we define the quantity

$$D(S) = \sum_{x \in S} D(x) \quad (2)$$

and refer to it as *the probability that a randomly selected element from D is in S* . The following notation, borrowed from cryptography, means that x is randomly selected from D :

$$x \xleftarrow{\$} D.$$

Remark 2. Let D be a probability distribution on some countable set X . The next statements follow from (1) and (2).

1. $D(\text{dom } D) = 1$.
2. For any subsets K and L of X , if $K \cap L = \emptyset$ then $D(K \cup L) = D(K) + D(L)$.
3. For any subsets K and L of X , if $K \subseteq L$ then $D(K) \leq D(L)$.

The author of [10] considers three families of probability distributions on \mathbb{N}_0 that are meaningful in information and/or number theory. These distribution families are called uniform, Lambert and Dirichlet, and are defined, respectively, as follows, where $d \in \mathbb{N}_0$, $M \in \mathbb{N}$, $z \in (0, 1)$ and $t \in (1, +\infty)$ are related parameters.

Uniform: $U_M(n) = 1/M$ for $n < M$, and $U_M(n) = 0$ otherwise.

Lambert: $L_{1/z,d}(n) = (1-z)z^{n-d}$ for $n \geq d$, and $L_{1/z,d}(n) = 0$ otherwise.

Dirichlet: $D_{t,d}(n) = (1/\zeta(t))(n+1-d)^{-t}$ for $n \geq d$, where ζ is the Riemann zeta function, and $D_{t,d}(n) = 0$ otherwise.

In fact [10] considers distributions on \mathbb{N} , but here we use \mathbb{N}_0 instead as we intend to apply these distributions to modelling lengths of words, including possibly the empty word ε whose length is 0. We also note that [10] considers $L_{1/z,d}$ and $D_{t,d}$ only for the case where the displacement $d = 1$. We also note that in [9] the same author considers the Dirichlet distribution to be the basis where “many heuristic probability arguments based on the fictitious uniform distribution on the positive integers become rigorous statements.”

Definition 2. We shall call any probability distribution N on \mathbb{N}_0 a *length distribution*. Then, as all values $N(n)$ are numeric, N can be viewed as a random variable, and the *expected value* of a length distribution N is well-defined and denoted by $\mathcal{E}(N)$.

If $t > 2$, the expected value of $D_{t,1}$ is finite and equal to $\zeta(t-1)/\zeta(t)$, [9]. Using standard tools in series manipulation and the fact that $\sum_{i \in \mathbb{N}} iz^i = z/(1-z)^2$, we have the following lemma.

Lemma 1. Let $d \in \mathbb{N}_0, z \in (0, 1)$ and $t \in (2, +\infty)$. We have that

$$\mathcal{E}(L_{1/z,d}) = d + \frac{1}{1/z - 1} \quad \text{and} \quad \mathcal{E}(D_{t,d}) = d + \frac{\zeta(t-1)}{\zeta(t)} - 1.$$

3 Randomized Approximation of $[0,1]$ -value problems

We consider problems for which every instance³ x has a value $v(x) \in [0, 1]$ and we are interested in those instances x for which $v(x) = 1$. Our main set of instances is the set of NFAs (or subsets of that) and the main value function v is the universality index of NFAs, which is defined in Section 5. However for the purposes of this section, our sample set of instances is $\text{BNFA}[2]$ = all block NFAs over the alphabet $\{0, 1\}$, and the $[0, 1]$ -valued function v is such that $v(\mathbf{a}) = |\mathbf{L}(\mathbf{a})|/2^n$, where n is the word length of the block NFA \mathbf{a} . In general, for a fixed but arbitrary $[0, 1]$ -valued function v , we define the language (problem)

$$L_v = \{x : v(x) = 1\}.$$

Deciding whether a given instance x is in L_v might be hard, but we assume that we are happy if we know whether $v(x) \geq 1 - \varepsilon$, for some appropriate *tolerance* $\varepsilon \in (0, 1)$. So we define the following **approximation** language for L_v :

$$L_{v,\varepsilon} = \{x : v(x) \geq 1 - \varepsilon\}.$$

Remark 3. One can verify that $L_v = \bigcap_{\varepsilon \in (0,1)} L_{v,\varepsilon}$; hence L_v can be approximated as close as desired via the languages $L_{v,\varepsilon}$.

Unfortunately deciding $L_{v,\varepsilon}$ can be harder than deciding L_v , as shown in the proof of the next theorem—the proof can be found further below.

Theorem 1. The following problem about block NFAs is *coNP-hard*

$$B_\delta = \{\mathbf{a} \in \text{BNFA}[2] : \frac{|\mathbf{L}(\mathbf{a})|}{2^n} \geq \delta, \text{ where } n = \text{word length of } \mathbf{a}\},$$

for any (fixed) $\delta \in (0, 1)$ that is computable within polynomial time⁴.

³ Following the presentation style of [8, pg 193], we refrain from cluttering the notation with the use of a variable for the set of instances.

⁴ A real $x \in (0, 1)$ is computable if there is an algorithm that takes as input a positive integer n and computes the n -th bit of x . It is computable within polynomial time if the algorithm works in time $O(n^k)$, for some fixed $k \in \mathbb{N}_0$, when the input n is given in unary.

Another idea then is to show that L_v is in the class coRP , that is, there is a polynomial **randomized** algorithm $A(x)$ such that

- if $x \in L_v$ then $A(x) = \text{True}$ (with probability 1), and
- if $x \notin L_v$ then $A(x) = \text{False}$ with probability⁵ at least $3/4$.

However, as L_v can be hard, it is unlikely that it is in the class coRP .

The next idea is to devise an approximating algorithm for L_v via $L_{v,\varepsilon}$. As stated in [8, pg 417], “*The answer to [what constitutes a “good” approximation] seems intimately related to the specific computational task at hand...the importance of certain approximation problems is much more subjective...[which] seems to stand in the way of attempts at providing a comprehensive theory of natural approximation problems.*” It seems that the following approximation method is meaningful. Although our domain of interest involves NFAs, the below definition is given for any set of instances and refers to a fixed but arbitrary $[0,1]$ -valued function v on these instances.

Definition 3. *Let v be $[0,1]$ -valued function. A polynomial approximation (PAX) algorithm for L_v is an algorithm $A(x, \varepsilon)$ such that*

- if $x \in L_v$ then $A(x, \varepsilon) = \text{True}$;
- if $x \notin L_{v,\varepsilon}$ then $A(x, \varepsilon) = \text{False}$;
- $A(x, \varepsilon)$ works within polynomial time w.r.t. $1/\varepsilon$ and the size of x .

Explanation. In the above definition, if $A(x, \varepsilon)$ returns **False** then $x \notin L_v$, that is, $v(x) < 1 - \varepsilon$. If $A(x, \varepsilon)$ returns **True** then $x \in L_{v,\varepsilon}$, that is, $v(x) \geq 1 - \varepsilon$. Thus, whenever the algorithm returns the answer **False**, this answer is correct and exact; when the algorithm returns **True**, the answer is *correct within the tolerance ε* .

It turns out that, in general, there are problems for which no approximation algorithm can do better than the exact algorithms.

Proposition 1. *There is no polynomial approximation algorithm for the problem UNIV_BNFA , unless $P=\text{coNP}$.*

Proof. It is sufficient to consider the subset of the problem for BNFA over the binary alphabet. Given $\mathbf{a} \in \text{BNFA}[2]$, the question of the problem is equivalent to whether $|\mathbf{L}(\mathbf{a})|/2^n = 1$, where n = the word length of \mathbf{a} . If there were a PAX $A(\mathbf{a}, \varepsilon)$ for this problem then we would decide the problem in polynomial time as follows: find out the word length n of the given BNFA \mathbf{a} , compute $\varepsilon = (1 + 2^n)^{-1}$ and run $A(\mathbf{a}, \varepsilon)$ to get the desired answer.

Corollary 1. *There is no polynomial approximation algorithm for the problem UNIV_MAXLEN_NFA , unless $P=\text{coNP}$.*

⁵ Many authors specify this probability to be at least $2/3$, but they state that any value $\geq 1/2$ works [8,2].

Remark 4. Theorem 1 implies that, unless $P=coNP$, block NFA universality over the binary alphabet cannot be approximated by some sequence (B_{δ_n}) , with $\lim \delta_n = 0$ and each δ_n being polynomially computable. Based on this observation and on Proposition 1, we conclude that, *in general, it is necessary to add a randomized aspect to our approximation methods.* We do this immediately below. We also note that there are in fact cases where a PAX algorithm for a hard problem exists—see Section 7.2.

The following definition is inspired from the “approximate” algorithmic solution of [14] for the task of generating an error-detecting code of N codewords, for given N , if possible, or an error-detecting code of less than N codewords which is “close to” maximal.

Definition 4. *Let v be $[0,1]$ -valued function. A polynomial randomized approximation (PRAX) algorithm for L_v is a randomized algorithm $A(x, \varepsilon)$ such that*

- if $x \in L_v$ then $A(x, \varepsilon) = \text{True}$;
- if $x \notin L_{v, \varepsilon}$ then $P[A(x, \varepsilon) = \text{False}] \geq 3/4$;
- $A(x, \varepsilon)$ works within polynomial time w.r.t. $1/\varepsilon$ and the size of x .

Explanation. In the above definition, if $A(x, \varepsilon)$ returns **False** then $x \notin L_v$. If $A(x, \varepsilon)$ returns **True** then probably $x \in L_{v, \varepsilon}$, in the sense that $x \notin L_{v, \varepsilon}$ would imply $P[A(x, \varepsilon) = \text{False}] \geq 3/4$. Thus, whenever the algorithm returns the answer **False**, this answer is correct ($x \notin L_v$); when the algorithm returns **True**, the answer is *correct within the tolerance ε* ($x \in L_{v, \varepsilon}$) with probability $\geq 3/4$. The algorithm returns the wrong answer exactly when it returns **True** and $x \notin L_{v, \varepsilon}$, but this happens with probability $< 1/4$.

Use of a PRAX algorithm. The algorithm can be used as follows to determine the approximate membership of a given x in L_v with a probability that can be as high as desired: Run $A(x, \varepsilon)$ k times, for some desired k , or until the output is **False**. If the output is **True** for all k times then $P[A(x, \varepsilon) = \text{True for } k \text{ times} \mid x \notin L_{v, \varepsilon}] < 1/4^k$, that is, the probability of incorrect answer is $< 1/4^k$.

Proof. (Of Theorem 1.) We reduce to B_δ the following known $coNP$ -hard problem $UNIV_BNFA[2] = \{\mathbf{b} \in BNFA[2] : |\mathbf{L}(\mathbf{b})| = 2^\ell, \text{ where } \ell = \text{word length of } \mathbf{b}\}$. We need a reduction that takes any instance \mathbf{b} in $BNFA[2]$, of some word length $\ell \in \mathbb{N}$, and constructs (in polynomial time) an instance \mathbf{a} in $BNFA[2]$, of some word length $n \in \mathbb{N}$, such that

$$|\mathbf{L}(\mathbf{b})| = 2^\ell \quad \text{iff} \quad |\mathbf{L}(\mathbf{a})| \geq 2^n \delta. \quad (3)$$

The main idea is to make a block NFA \mathbf{a} that accepts a language $F \cdot \mathbf{L}(\mathbf{b})$ of $|F| \cdot |\mathbf{L}(\mathbf{b})|$ words of length $k + \ell$, where k and $|F|$ depend on δ and ℓ . If δ is of the form $m/2^k$ for some $m, k \in \mathbb{N}$, then F is any language of m words of length k , and (3) holds. The reduction for the general case of $\delta \neq m/2^k$ is described next, where we use the notation (i) $b_p \triangleq$ the bit at position p in the binary representation of δ , for $p \in \mathbb{N}$; (ii) $m_p \triangleq b_1 2^{p-1} + b_2 2^{p-2} + \dots + b_p$; that is, m_p is the numerator of the fraction $m_p/2^p \in (0, 1)$ that results when we cut from δ all bits after position p .

1. Let $k = p_1 + \ell$, where $p_1 = \min\{p \in \mathbb{N} : b_p = 1\}$.
2. Let $m_k = b_1 2^{k-1} + b_2 2^{k-2} + \dots + b_k$. Note that $2^\ell \leq m_k < 2^k$.
3. Let $\mathbf{b}[m_k]$ be any block NFA accepting a language F of exactly $1 + m_k$ words of length k , such that $\mathbf{b}[m_k]$ has exactly one final state f .
4. Let \mathbf{a} be the block NFA that results by ‘concatenating’ $\mathbf{b}[m_k]$ and \mathbf{b} : change all transitions of $\mathbf{b}[m_k]$ that go to f to go to the start state of \mathbf{b} . Note that \mathbf{a} accepts the language $F \cdot \mathbf{L}(\mathbf{b})$ consisting of $(1 + m_k)|\mathbf{L}(\mathbf{b})|$ words of length $n = k + \ell$.

We need to show that (3) holds and that the above reduction (steps 1–4) can be done within polynomial time with respect to $|\mathbf{b}|$. That (3) holds follows from the below observations.

- For any bit position p of δ , we have $m_p/2^p < \delta < (1 + m_p)/2^p$.
- The above implies that, for any p , there is $x_p \in (0, 1)$ such that $\delta = (1 + m_p - x_p)/2^p$.
- If $|\mathbf{L}(\mathbf{b})| = 2^\ell$ then $|\mathbf{L}(\mathbf{a})| = (1 + m_k)|\mathbf{L}(\mathbf{b})| > 2^k \delta 2^\ell = 2^n \delta$.
- If $|\mathbf{L}(\mathbf{a})| \geq 2^n \delta$ then $(1 + m_k)|\mathbf{L}(\mathbf{b})| \geq 2^k 2^\ell (1 + m_k - x_k)/2^k$ and then

$$|\mathbf{L}(\mathbf{b})| \geq \left(1 - \frac{x_k}{1 + m_k}\right) 2^\ell \quad \Rightarrow \quad |\mathbf{L}(\mathbf{b})| > 2^\ell - 1,$$

where the above follows when we recall that $2^\ell \leq m_k$.

That the above reduction (steps 1–4) is polynomial w.r.t. $|\mathbf{b}|$ follows when we note that (i) p_1 is a constant and $k = O(\ell)$. (ii) ℓ is essentially presented in unary as the length of any accepting path of \mathbf{b} , so $\ell < |\mathbf{b}|$; then ℓ is stored in binary in a variable that can be used to perform arithmetic operations within polynomial time in steps 1–2. (iv) Step 4 can be done in time $O(|\mathbf{b}[m_k]| + |\mathbf{b}|)$. (v) Step 3 can be done in time $O(k^2)$ resulting in $\mathbf{b}[m_k]$ of size $O(k^2)$ as follows:

- Let $m_k = 2^{c_1} + \dots + 2^{c_t}$, where the c_i ’s are the nonzero bit positions in the binary representation of m_k , and such that $t \leq k$ and $c_1 < \dots < c_t < k$.
- For each i , make a ‘straight line’ block NFA \mathbf{b}_i of $k + 1$ states accepting all binary strings $\mathbf{A}_2^{c_i} \mathbf{1}^{k-c_i}$.
- Make the required block NFA $\mathbf{b}[m_k]$ to be the ‘union’ of all \mathbf{b}_i ’s using a single start state s , a single final state f , and connecting s to the second states of the \mathbf{b}_i ’s and connecting the second-last states of the \mathbf{b}_i ’s to f .

4 Word Distributions

A word distribution W is a probability distribution on \mathbf{A}_s^* , that is, $W : \mathbf{A}_s^* \rightarrow [0, 1]$ such that $\sum_{w \in \mathbf{A}_s^*} W(w) = 1$. If \mathbf{a} is an NFA then we use the convention that

$$W(\mathbf{a}) \text{ means } W(\mathbf{L}(\mathbf{a})).$$

The domain and length of W are defined, respectively, as follows:

$$\text{dom } W = \{w \in \mathbf{A}_s^* \mid W(w) > 0\}, \quad \text{len } W = \{|w| \mid w \in \text{dom } W\}.$$

We view $\text{len } W$ as a random variable such that $P[\text{len } W = n] = P[w \in \mathbf{A}_s^n] = W(\mathbf{A}_s^n)$. The expected length of W is the quantity

$$\mathcal{E}(\text{len } W) = \sum_{w \in \mathbf{A}_s^*} W(w)|w|,$$

which could be finite or $+\infty$.

Example 1. For a finite language F , we write U_F to denote the uniform word distribution on F , that is, $U_F(w) = 1/|F|$ for $w \in F$, and $U_F(w) = 0$ for $w \notin F$. Some important examples of uniform word distributions are:

- $U_{\mathbf{A}_s^\ell}$, where ℓ is any word length. Then, $U_{\mathbf{A}_s^\ell}(w) = 1/s^\ell$.
- $U_{\mathbf{A}_s^{\leq \ell}}$, where ℓ is any word length. Then, $U_{\mathbf{A}_s^{\leq \ell}}(w) = 1/t$, where $t = 1 + s + \dots + s^\ell$.
- $U_{L(\mathbf{a})}$, where \mathbf{a} is an acyclic NFA. We also simply write $U_{\mathbf{a}}$ for $U_{L(\mathbf{a})}$.

Definition 5. Let N be a length distribution. Then $\langle N \rangle$ is the word distribution such that

$$\langle N \rangle(w) = N(|w|)s^{-|w|}.$$

Any such word distribution is called a *length-based distribution*.

Remark 5. One can verify that, for any length distribution N , the following statements hold true, where $n \in \mathbb{N}_0$.

1. $\langle N \rangle(\mathbf{A}_s^n) = N(n)$.
2. $\langle N \rangle(\mathbf{A}_s^{>n}) = N(\mathbb{N}^{>n})$.
3. $\mathcal{E}(\text{len } \langle N \rangle) = \mathcal{E}(N)$.

Example 2. Using the Lambert length distribution $L_{s,d}(n) = (1 - 1/s)(1/s)^{n-d}$, we define the Lambert, or **geometric**, word distribution $\langle L_{s,d} \rangle$ on \mathbf{A}_s^* such that $\langle L_{s,d} \rangle(w) = 0$ if $|w| < d$ and, for $|w| \geq d$,

$$\langle L_{s,d} \rangle(w) = (1 - 1/s)(1/s)^{2|w|-d}.$$

Then, for all $n, d \in \mathbb{N}_0$ with $n \geq d$, we have

$$\langle L_{s,d} \rangle(\mathbf{A}_s^n) = (1 - 1/s)(1/s)^{n-d}, \langle L_{s,d} \rangle(\mathbf{A}_s^{>n}) = (1/s)^{n+1-d}, \mathcal{E}(\text{len } \langle L_{s,d} \rangle) = d + 1/(s-1).$$

In particular, for the alphabet $\mathbf{A}_2 = \{0, 1\}$, we have that $\langle L_{2,1} \rangle(\mathbf{A}_2) = 1/2$, $\langle L_{2,1} \rangle(\mathbf{A}_2^2) = 1/2^2$, etc.

Example 3. Let $t \in (2, +\infty)$. Using the Dirichlet length distribution $D_{t,d}(n) = (1/\zeta(t))(n+1-d)^{-t}$, we define the Dirichlet word distribution $\langle D_{t,d} \rangle$ on \mathbf{A}_s^* such that $\langle D_{t,d} \rangle(w) = 0$ if $|w| < d$ and, for $|w| \geq d$,

$$\langle D_{t,d} \rangle(w) = (1/\zeta(t))(|w| + 1 - d)^{-t} s^{-|w|}.$$

Then, for all $n, d \in \mathbb{N}_0$ with $n \geq d$, we have $\langle D_{t,d} \rangle(\mathbf{A}_s^n) = (1/\zeta(t))(n+1-d)^{-t}$,

$$\langle D_{t,d} \rangle(\mathbf{A}_s^{>n}) = 1 - (1/\zeta(t)) \sum_{i=1}^{n+1-d} i^{-t}, \quad \mathcal{E}(\text{len}\langle D_{t,d} \rangle) = d + \zeta(t-1)/\zeta(t) - 1.$$

In particular, for $t = 3, d = 1$ and alphabet $\mathbf{A}_2 = \{0, 1\}$, we have that $\langle D_{2,1} \rangle(\mathbf{A}_2^n) = (1/\zeta(3))n^{-3}$.

Selecting a word from a distribution. We are interested in word distributions W for which there is an efficient (randomized) algorithm that returns a randomly selected element from W . We shall assume available (randomized) algorithms as follows.

- **tossCoin**(p): returns 0 or 1, with probability p or $1-p$, respectively, where $p \in [0, 1]$, and the algorithm works in constant time for most practical purposes—this is a reasonable assumption according to [2, pg 134].
- **selectUnif**(s, ℓ): returns a uniformly selected word from \mathbf{A}_s^ℓ , and the algorithm works in time $O(\ell)$.

Remark 6. As in [2, pg 126], we assume that basic arithmetic operations are performed in constant time. Even if we relax this assumption and we account for a parameter q for arithmetic precision, the arithmetic operations would require a polynomial factor in q .

The next lemma seems to be folklore, but we include it here for the sake of clarity and self-containment.

Lemma 2. *There is a polynomial randomized algorithm **selectFin**(D), where D is a finite probability distribution $(D(x_1), \dots, D(x_n))$ on some set $\{x_1, \dots, x_n\}$, that returns a randomly selected x_i with probability $D(x_i)$. In fact the algorithm works in time $O(n)$ using the assumption of constant cost of **tossCoin** and of arithmetic operations.*

Proof. The algorithm works as follows: perform up to $n-1$ coin tosses such that

- in the i -th coin toss, the outcome 0 means to return the element x_i and terminate, and the outcome 1 means to continue to the next coin toss (or return x_n if $i = n-1$);
- each coin toss i uses the algorithm **tossCoin**(p_i), where $p_1 = D(x_1)$ and $p_{i+1} = D(x_{i+1})/((1-p_1) \cdots (1-p_i))$.

We have that p_i is the probability that coin toss i is 0 given that all previous tosses (when $i > 1$) are all 1. The outcome O of the algorithm is such that $P[O = x_1] = p_1 = D(x_1)$ and $P[O = x_{i+1}] = p_{i+1} \cdot (1-p_1) \cdots (1-p_i)$.

Augmented word distributions. Selecting a word from a distribution W with infinite domain $\text{dom } W$ could return a very long word, which can be intractable. For this reason we would like to define distributions on $\mathbf{A}_s^* \cup \{\perp\}$, where ‘ \perp ’ is a symbol outside of \mathbf{A}_s , which could select the outcome ‘ \perp ’ (no word). These could be versions of word distributions in which there is a bound on the length of words they can select.

Definition 6. An *augmented word distribution* is a probability distribution on $\mathbf{A}_s^* \cup \{\perp\}$. Let W be a word distribution and let $M \in \mathbb{N}_0$. We define the augmented distribution W^M such that

$$W^M(w) = W(w), \text{ if } |w| \leq M; \quad W^M(\perp) = W(\mathbf{A}_s^{>M}).$$

Remark 7. The probability that W^M selects a word longer than M is zero. We have that $\text{dom } W^M = (\text{dom } W \cap \mathbf{A}_s^{\leq M}) \cup \{\perp\}$. Moreover, the following facts about W^M and any language L are immediate

$$W(L \cap \mathbf{A}_s^{\leq M}) = W^M(L \cap \mathbf{A}_s^{\leq M}), \quad W(\mathbf{A}_s^{>M}) = W^M(\perp). \quad (4)$$

Remark 8. The proof of Lemma 2 uses a general formula for computing the quantities p_i . However, these quantities can be computed in a much simpler way for specific distributions. For the augmented Lambert distribution $\mathbf{L}_{s,d}^M$, for instance, we have that each $p_i = 1 - 1/s$.

5 Universality Index of NFAs

Here we intend to define mathematically the informal concept of an “approximately universal NFA” with respect to a certain fixed language \mathbf{M} . Our motivation comes from coding theory where the codes of interest are subsets of \mathbf{M} , and it is desirable that a code is a maximal subset of \mathbf{M} . Two typical cases are (i) $\mathbf{M} = \mathbf{A}_s^*$, when variable-length codes are considered, such as prefix or suffix codes; and (ii) $\mathbf{M} = \mathbf{A}_s^n$ for some $n \in \mathbb{N}$, when error control codes are considered. Testing whether a regular code C is a maximal subset of \mathbf{M} is a hard problem and, in fact, this problem normally reduces to whether a certain NFA that depends on C accepts \mathbf{M} —see e.g., [6, 14]. In practice, however, it could be acceptable that a code is “close” to being maximal, or an NFA is “close” to being universal.

Our approach here assumes that the maximum language \mathbf{M} is equal to $\text{dom } W$, where W is the word distribution of interest.

Definition 7. Let W be a word distribution, let \mathbf{a} be an NFA, and let $p \in [0, 1]$.

- We say that \mathbf{a} is *universal relative to W* , if $\mathbf{L}(\mathbf{a}) = \text{dom } W$.
- We say that \mathbf{a} is *p -universal relative to W* , if $W(\mathbf{a}) \geq p$. We call the quantity $W(\mathbf{a})$ the *universality index of \mathbf{a} (relative to W)*.

Example 4. Let \mathbf{b} be a block NFA. If $|\mathbf{L}(\mathbf{b})|/s^\ell \geq p$, where ℓ is the word length of \mathbf{b} , then \mathbf{b} is p -universal relative to the uniform distribution on \mathbf{A}_s^ℓ and the quantity $|\mathbf{L}(\mathbf{b})|/s^\ell$ is the universality index of \mathbf{b} .

Remark 9. The universality index $W(\mathbf{a})$ represents the probability that a randomly selected word from W is accepted by \mathbf{a} —see Definition 1. When $W(\mathbf{a})$ is close to 1 then \mathbf{a} is close to being universal, that is, $\mathbf{L}(\mathbf{a})$ is close to $\text{dom } W$. The concept of a p -universal NFA formalizes the loose concept of an approximately universal NFA—see also the next lemma. Thus, for example, we can talk about a 98%-universal block NFA with respect to the uniform distribution on \mathbf{A}_s^ℓ , where ℓ is the word length of the NFA.

Remark 10. The method of [13] embeds a given \mathbf{t} -code⁶ K into a maximal one by successive applications of a language operator $\mu_{\mathbf{t}}$ on K which yields supersets K_i of K until these converge to a maximal \mathbf{t} -code. The operation $\mu_{\mathbf{t}}$ on each K_i (represented as an NFA) can be expensive to compute and one can simply stop at a step where the current superset K_i is close to maximal, or according to the concepts of this paper, when the NFA for $(\mathbf{t}(K_i) \cup \mathbf{t}^{-1}(K_i) \cup K_i)$ is close to universal.

Lemma 3. *If L is universal relative to W , that is $L = \text{dom } W$, then $W(L) = 1$, for any word distribution W . Conversely, if there is a word distribution W such that $W(L) = 1$, then L is universal relative to $\text{dom } W$.*

Proof. Immediate.

Consider the case where \mathbf{b} is a block NFA of length ℓ and W is the uniform word distribution on \mathbf{A}_s^ℓ . In this work, we view estimating the universality index of \mathbf{b} as a *parameter estimation problem* for finite populations [16, pg 72]: let p be an unknown population parameter (ratio of elements having some attribute over the cardinality of the population). Select n elements from the population (here, n words from \mathbf{A}_s^ℓ) and compute c , the number of these elements having the attribute of interest (here, words that are in $L(\mathbf{b})$). Then, c/n is an estimate for the population parameter p (here, the estimate is for $W(\mathbf{b})$) in the sense that the expected value of the random variable c/n is equal to p and

$$\mathbb{P}[|c/n - p| > \varepsilon] < e^{-n\varepsilon^2/2} + e^{-n\varepsilon^2/3}, \quad (5)$$

where $\varepsilon > 0$ is the acceptable estimation error. The above inequality is given in [?] and follows from Chernoff bounds. Here we extend the idea of parameter estimation to various distributions on languages. Moreover, we use the simpler Chebysev inequality for bounding the error probability, as it gives in practice a smaller bound than the one in the above inequality. Let X be a random variable and let $a > 0$. The Chebyshev inequality is as follows

$$\mathbb{P}[|X - \mathcal{E}(X)| \geq a] \leq \sigma^2/a^2,$$

where σ^2 is the variance of X . When X is the binomial random variable with parameters $n = \text{'number of trials'}$ and $p = \text{'probability of success in one trial'}$, then $\mathcal{E}(X) = np$ and $\sigma^2 = np(1-p)$. For $p \in [0, 1]$, the maximum value of $p(1-p)$ is $1/4$; therefore, the above inequality becomes as follows:

$$\mathbb{P}[|X - \mathcal{E}(X)| \geq a] \leq n/(4a^2). \quad (6)$$

Lemma 4. *Let \mathbf{a} be an NFA, let W be a word distribution, and let $p, g \in [0, 1]$ with $p > g$. Consider the random process $\text{UnivIndex}_W(\mathbf{a}, n)$ in Fig. 1, and let Cnt be the random variable for the value of cnt when the algorithm returns. If $W(\mathbf{a}) < g$ then $\mathbb{P}[\text{Cnt}/n \geq p] \leq \frac{1}{4n(p-g)^2}$.*

⁶ Depending on \mathbf{t} , which is a transducer, one can have prefix codes, suffix codes, infix codes, error control codes.

```

UnivIndexW(a, n)
  cnt := 0;
  i := 0;
  while (i < n):
     $w \xleftarrow{\$} W$ ;
    i := i+1;
    if ( $w \in L(\mathbf{a})$ ) cnt := cnt+1;
  return cnt / n;

```

Fig. 1. This random process refers to a particular word distribution W . It returns an estimate of the universality index $W(\mathbf{a})$ of the given NFA \mathbf{a} .

Proof. First note that Cnt is binomial: the number of successes (words in $L(\mathbf{a})$) in n trials. Thus, $\mathcal{E}(\text{Cnt}) = nW(\mathbf{a})$. Now assume that $W(\mathbf{a}) < g$. We have:

$$\begin{aligned}
\mathbb{P}[\text{Cnt}/n \geq p] &= \mathbb{P}[\text{Cnt} - nW(\mathbf{a}) \geq np - nW(\mathbf{a})] \\
&\leq \mathbb{P}[|\text{Cnt} - nW(\mathbf{a})| \geq np - nW(\mathbf{a})] \\
&\leq \mathbb{P}[|\text{Cnt} - nW(\mathbf{a})| \geq np - ng] \leq \frac{1}{4n(p-g)^2},
\end{aligned}$$

where we have used inequality (6).

In Section 6 we give a polynomial randomized approximation algorithm (PRAX) for testing universality of block NFAs, which is based on the random process in Fig. 1. That process, however, cannot lead to a PRAX for the universality of NFAs accepting infinite languages, as the selection $w \xleftarrow{\$} W$ could produce a word of exponential length. In Fig. 2 we modify that process so that a selected word cannot be longer than a desired $M \in \mathbb{N}_0$ —in Section 7 we investigate how this can lead to a PRAX for the universality of any NFA relative to tractable word distributions.

```

UnivIndexMaxLenW(a, n, M)
  cnt := 0;
  i := 0;
  while (i < n):
     $w \xleftarrow{\$} W^M$ ;
    i := i+1;
    if ( $w = \perp$  or  $w \in \mathbf{A}_s^{\leq M} \cap L(\mathbf{a})$ ) cnt := cnt+1;
  return cnt / n;

```

Fig. 2. This random process refers to a particular word distribution W . It returns an estimate of $W^M(\perp) + W^M(\mathbf{A}_s^{\leq M} \cap L(\mathbf{a}))$, which is equal to $W(\mathbf{A}_s^{>M}) + W(\mathbf{A}_s^{\leq M} \cap L(\mathbf{a}))$ —see (4). When M is chosen such that $W^M(\perp)$ is small enough, then the returned quantity cnt/n can be an acceptable estimate of $W(\mathbf{a})$.

Lemma 5. *Let \mathbf{a} be an NFA, let W be a word distribution, let $M \in \mathbb{N}_0$, and let $p, g \in [0, 1]$ such that $p > g + W(\mathbf{A}_s^{>M})$. Consider the random process $\text{UnivIndexMaxLen}_W(\mathbf{a}, n, M)$ in Fig. 2, and let Cnt be the random variable whose value is equal to the value of cnt when the algorithm returns. If $W(\mathbf{a}) < g$ then*

$$\mathbb{P}[\text{Cnt}/n \geq p] \leq \frac{1}{x} + \frac{1}{4n(p - g - xW(\mathbf{A}_s^{>M}))^2}, \quad \text{for all } x \in (1, \frac{p - g}{W(\mathbf{A}_s^{>M})}).$$

Proof. Referring to the n selections $w \stackrel{\$}{\leftarrow} W^M$ in $\text{UnivIndexMaxLen}_W(\mathbf{a}, n, M)$, let \mathbf{I} be the random variable for the number of selections that are in $\mathbf{L}(\mathbf{a}) \cap \mathbf{A}_s^{\leq M}$, and let \mathbf{B} be the random variable for the number of selections equal to \perp . Then, $\text{Cnt} = \mathbf{I} + \mathbf{B}$. Now note that (i) \mathbf{I} is binomial: the number of successes (words in $\mathbf{L}(\mathbf{a}) \cap \mathbf{A}_s^{\leq M}$) in n trials, and (ii) \mathbf{B} is binomial: the number of successes (selections \perp) in n trials. Thus, using (4), we have

$$\mathcal{E}(\mathbf{I}) = nW(\mathbf{L}(\mathbf{a}) \cap \mathbf{A}_s^{\leq M}) \leq nW(\mathbf{a}), \quad \mathcal{E}(\mathbf{B}) = nW(\mathbf{A}_s^{>M}).$$

Now assume that $W(\mathbf{a}) < g$, and let x be a number with $1 < x < (p - g)/W(\mathbf{A}_s^{>M})$. We have:

$$\begin{aligned} \mathbb{P}[\text{Cnt}/n \geq p] &= \mathbb{P}[\mathbf{I} + \mathbf{B} \geq np] = \mathbb{P}[\mathbf{I} \geq np - \mathbf{B}] \\ &= \mathbb{P}[\mathbf{I} \geq np - \mathbf{B} \text{ and } \mathbf{B} > xnW(\mathbf{A}_s^{>M})] + \mathbb{P}[\mathbf{I} \geq np - \mathbf{B} \text{ and } \mathbf{B} \leq xnW(\mathbf{A}_s^{>M})] \\ &\leq \mathbb{P}[\mathbf{B} > xnW(\mathbf{A}_s^{>M})] + \mathbb{P}[\mathbf{I} \geq np - \mathbf{B} \text{ and } np - \mathbf{B} \geq np - xnW(\mathbf{A}_s^{>M})] \\ &\leq \mathbb{P}[\mathbf{B} > xnW(\mathbf{A}_s^{>M})] + \mathbb{P}[\mathbf{I} \geq np - xnW(\mathbf{A}_s^{>M})] \\ &\leq \mathcal{E}(\mathbf{B})/(xnW(\mathbf{A}_s^{>M})) + \mathbb{P}[\mathbf{I} - \mathcal{E}(\mathbf{I}) \geq np - \mathcal{E}(\mathbf{I}) - xnW(\mathbf{A}_s^{>M})] \\ &\leq 1/x + \mathbb{P}[|\mathbf{I} - \mathcal{E}(\mathbf{I})| \geq np - ng - xnW(\mathbf{A}_s^{>M})] \\ &\leq \frac{1}{x} + \frac{1}{4n(p - g - xW(\mathbf{A}_s^{>M}))^2}, \end{aligned}$$

where we have used Markov's inequality " $\mathbb{P}[\mathbf{B} > a] < \mathcal{E}(\mathbf{B})/a$, for all $a > 0$ ", as well as inequality (6).

6 Randomized Approximation of NFA problems relative to Uniform Distributions

In this section we consider polynomial randomized approximation algorithms for the problems ADFA_SUBSET_NFA , UNIV_BNFA , UNIV_MAXLEN_NFA . As discussed below, the latter two problems are essentially special cases of the problem ADFA_SUBSET_NFA , but they can also be answered using a couple of more standard tools leading to more efficient algorithms.

Lemma 6. *Selecting uniformly at random an accepting word of a given ADFA \mathbf{a} can be done in polynomial time.*

<pre> ADFASubsetNFA(a, b, ε) $n := \lceil 1/\varepsilon^2 \rceil$; $i := 0$; while ($i < n$): $w := \text{selectUnif}(\mathbf{b})$; $i := i + 1$; if ($w \notin L(\mathbf{a})$) return False; return True; </pre>	<pre> ADFASubsetNFA(a, b, ε) $n := \lceil 1/\varepsilon^2 \rceil$; $i := 0$; $\text{cnt} := 0$; while ($i < n$): $w := \text{selectUnif}(\mathbf{b})$; $i := i + 1$; if ($w \in L(\mathbf{a})$) $\text{cnt} := \text{cnt} + 1$; if ($\text{cnt} < n$) return False else return True; </pre>
---	--

Fig. 3. On the left is the PRAX algorithm for the problem ADFA_SUBSET_NFA: whether the language of the given acyclic DFA \mathbf{b} is a subset of the language of the given NFA \mathbf{a} . This is equivalent to whether $L(\mathbf{b}) \subseteq L(\mathbf{a}) \cap L(\mathbf{b})$. The function $\text{selectUnif}(\mathbf{b})$ returns a uniformly selected word from $L(\mathbf{b})$. The version on the right is logically equivalent; it mimics the process in Fig. 1 and is intended to give a more clear explanation of correctness.

Proof. The statement can be shown using results from [4]. However, we give here a simple self-contained presentation. **First**, let $N(q)$ be the number of words accepted by \mathbf{a} from the start state s to state q . We have that $N(s) = 1$ and then, for each state q in breadth-first order, $N(q)$ is the sum of $N(p)$ for all transitions (p, σ, q) leading into q , where each computed value $N(q)$ is recorded so that it can be reused. Let F be the set of final states of \mathbf{a} and let $N_F = \sum_{f \in F} N(f)$, which is equal to $|L(\mathbf{a})|$. **Then**, selecting a word $w \in L(\mathbf{a})$ can be done in two steps. The first step is to use selectFin to select a final state f from the distribution $(N(f)/N_F)_{f \in F}$. The second step is to select a word w accepted by \mathbf{a} at the final state f . Each symbol σ of w is selected starting from the last one, as follows. Let T_f be the set of transitions leading to state f . Use again selectFin to select one transition $(p, \sigma, f) \in T_f$ from the distribution that consists of the values $N(p)/N(f)$ for all $(p, \sigma, f) \in T_f$. Then, the last symbol of w is σ . Repeat the same process, for $f \leftarrow p$, selecting the next symbol of w , until the start state s is encountered.

Theorem 2. *Algorithm ADFASubsetNFA($\mathbf{a}, \mathbf{b}, \varepsilon$) is a polynomial randomized approximation algorithm for ADFA_SUBSET_NFA.*

Proof. First we note that ADFA_SUBSET_NFA can be expressed as follows as a $[0,1]$ -value problem

$$\text{ADFA_SUBSET_NFA} = \left\{ (\mathbf{a}, \mathbf{b}) : \mathbf{a} \in \text{NFA}, \mathbf{b} \in \text{ADFA}, v(\mathbf{a}, \mathbf{b}) = 1 \right\},$$

where $v(\mathbf{a}, \mathbf{b}) = \frac{|L(\mathbf{a}) \cap L(\mathbf{b})|}{|L(\mathbf{b})|}$; therefore the problem $\text{ADFA_SUBSET_NFA}_\varepsilon$ is well-defined. For brevity we write $A(\mathbf{a}, \mathbf{b}, \varepsilon)$ to refer to $\text{ADFASubsetNFA}(\mathbf{a}, \mathbf{b}, \varepsilon)$. We consider the three conditions of Definition 4. The **third** condition about the time complexity follows when we note that (i) testing whether a word w is accepted by

an NFA \mathbf{a} can be done in time $O(|w||\mathbf{a}|)$; and (ii) selecting uniformly at random a word from an acyclic DFA \mathbf{b} can be done in polynomial time (see Lemma 6). For the **first** condition of Definition 4, if $L(\mathbf{b}) \subseteq L(\mathbf{a})$ then every selected word w is in $L(\mathbf{a})$, so the algorithm will return **True**. For the **second** condition, assume that $|L(\mathbf{a}) \cap L(\mathbf{b})|/|L(\mathbf{b})| < 1 - \varepsilon$. Consider the version of the algorithm on the right and the random process in Lemma 4 and assume that it selects exactly the same words w as $A(\mathbf{a}, \mathbf{b}, \varepsilon)$ does. Then, algorithm $A(\mathbf{a}, \mathbf{b}, \varepsilon)$ returns **True** if and only if the random variable **Cnt** in Lemma 4 takes the value n . Moreover, using $p = 1$ and $g = 1 - \varepsilon$ in Lemma 4, we have

$$\begin{aligned} P[A(\mathbf{a}, \mathbf{b}, \varepsilon) = \text{True}] &= P[\text{Cnt} = n] = P[\text{Cnt}/n \geq 1] \\ &\leq \frac{1}{4n(1 - (1 - \varepsilon))^2} = \frac{1}{4n\varepsilon^2} \leq 1/4. \end{aligned}$$

The next corollaries follow from the above theorem; however, using a more self-contained choice of tools we get more efficient algorithms with estimates of their time complexity.

$\text{UnivBlockNFA}(\mathbf{a}, \varepsilon)$	$\text{UnivMaxLenNFA}(\mathbf{a}, \ell, \varepsilon)$
$\ell := \text{the word length of } L(\mathbf{a});$	$t := 1 + s + \dots + s^\ell;$
$n := \lceil 1/\varepsilon^2 \rceil;$	$N := (1/t, s/t, \dots, s^\ell/t);$
$i := 0;$	$n := \lceil 1/\varepsilon^2 \rceil;$
while ($i < n$):	$i := 0;$
$w := \text{selectUnif}(s, \ell);$	while ($i < n$):
$i := i+1;$	$k := \text{selectFin}(N);$
if ($w \notin L(\mathbf{a})$) return False ;	$w := \text{selectUnif}(s, k);$
return True ;	$i := i+1;$
	if ($w \notin L(\mathbf{a})$) return False ;
	return True ;

Fig. 4. UnivBlockNFA decides approximate block NFA universality (see Corollary 2) and UnivMaxLenNFA decides approximate up to a maximum length NFA universality (see Corollary 3).

Corollary 2. *Algorithm $\text{UnivBlockNFA}(\mathbf{a}, \varepsilon)$ in Fig. 4 is a polynomial randomized approximation algorithm for block NFA universality and works in time $O(\ell |\mathbf{a}| (1/\varepsilon)^2)$, where ℓ is the word length of \mathbf{a} .*

Proof. The existence of a polynomial randomized approximation algorithm for block NFA universality follows from the algorithm $\text{ADFASubsetNFA}(\mathbf{a}, \mathbf{b}, \varepsilon)$ of Theorem 2 when we note that given block NFA \mathbf{a} of some word length ℓ , one can construct in time $O(\ell)$ a block (hence, acyclic) DFA \mathbf{b} accepting the language A_s^ℓ . Here however, step $\text{selectUnif}(\mathbf{b})$ of ADFASubsetNFA can be replaced by the simpler process of selecting uniformly a word of length ℓ .

Use of the algorithm $\text{UnivBlockNFA}(\mathbf{a}, \varepsilon)$. Suppose that we want to test whether a block NFA \mathbf{a} of some word length ℓ is universal relative to the uniform distribution on \mathbf{A}_s^ℓ , and that we allow a 2% approximation tolerance, that is, we consider it acceptable to say that \mathbf{a} is universal when it is in fact 98%-universal. Then we run the algorithm using $\varepsilon = 0.02$. If \mathbf{a} is universal, then the algorithm correctly returns **True**. If \mathbf{a} is not 98%-universal, then the probability that the algorithm returns **True** is at most $1/4$. Note that for this choice of arguments, the loop would iterate at most 2500 times.

Corollary 3. *Algorithm $\text{UnivMaxLenNFA}(\mathbf{a}, \ell, \varepsilon)$ in Fig. 4 is a polynomial randomized approximation algorithm for UNIV_MAXLEN_NFA . In fact the algorithm works in time $O(\ell |\mathbf{a}| (1/\varepsilon)^2)$ under the assumption of constant cost of `tossCoin` and of arithmetic operations.*

Proof. The existence of a polynomial randomized approximation algorithm for UNIV_MAXLEN_NFA follows from the algorithm $\text{ADFASubsetNFA}(\mathbf{a}, \mathbf{b}, \varepsilon)$ of Theorem 2 when we note that given ℓ in unary, one can construct in time $O(\ell)$ an acyclic DFA \mathbf{b} accepting the language $\mathbf{A}_s^{\leq \ell}$. Here however, step `selectUnif(b)` of ADFASubsetNFA can be replaced by the process of selecting uniformly a word length $k \in \{0, 1, \dots, \ell\}$ according to the distribution

$$(|\mathbf{A}_s^0|/t, |\mathbf{A}_s^1|/t, \dots, |\mathbf{A}_s^\ell|/t)$$

and then selecting uniformly a word of length k .

7 Randomized Approximation of NFA Universality

Here we present an analogue to the uniform distribution algorithms for the case where the NFA accepts an infinite language and universality is with respect to some word distribution $\langle T \rangle$. The approximation algorithm of this section is based on the random process in Fig. 2 and requires that the distribution $\langle T \rangle$ be tractable, which loosely speaking means that words longer than a certain length $M = M(\varepsilon)$ have low probability and can be ignored when one wants to approximate the universality index of the given NFA *within a given tolerance* ε —recall, this approach is consistent with our interpretation of languages in the context of coding and information theory.

Definition 8. *A length distribution T is called **tractable**, if the following conditions hold true.*

1. *For all $\varepsilon \in (0, 1)$, there is $M \in \mathbb{N}_0$ such that $T(\mathbb{N}^{>M}) \leq \varepsilon$, M is of polynomially bounded magnitude w.r.t. $\log(1/\varepsilon)$, that is, $M = O((\log \frac{1}{\varepsilon})^k)$ for some $k \in \mathbb{N}_0$, and there is an algorithm $\text{maxLen}_T(\varepsilon)$ that returns such an M and works within polynomial time w.r.t. $1/\varepsilon$.*
2. *There is an algorithm $\text{prob}_T(m)$, where $m \in \mathbb{N}_0$, that returns the value $T(m)$ and works within polynomial time w.r.t. m .*

Lemma 7. *Let T be a tractable length distribution.*

1. There is $k \in \mathbb{N}_0$ such that

$$T(\mathbb{N}^{>(\log \frac{1}{\varepsilon})^k}) \leq \varepsilon, \quad \text{for all } \varepsilon \in (0, 1). \quad (7)$$

2. The expected length of $\langle T \rangle$ is finite.

Proof. The first statement follows when we note that, as T is tractable, there is $k \in \mathbb{N}_0$ such that, for any $\varepsilon \in (0, 1)$, there is $M \leq (\log(1/\varepsilon))^k$ such that $T(\mathbb{N}^{>M}) \leq \varepsilon$. For the second statement, we have the following about the expected length of $\langle T \rangle$.

$$\mathcal{E}(\text{len}\langle T \rangle) = \mathcal{E}(T) = \sum_{i \in \mathbb{N}_0} iT(i) \leq \sum_{i \in \mathbb{N}_0} iT(\mathbb{N}^{>i-1}). \quad (8)$$

Using $\varepsilon = 1/2^{(i-1)^{1/k}}$ in (7), we have that $T(\mathbb{N}^{>i-1}) \leq 1/2^{(i-1)^{1/k}}$, which implies that the series in (8) is finite.

<p>UnivNFA_T(\mathbf{a}, ε)</p> <pre> $\varepsilon := \min(\varepsilon, 1/6);$ $n := \lceil 5/(\varepsilon - 5\varepsilon^2)^2 \rceil;$ $M := \text{maxLen}_T(\varepsilon^2);$ for each $\ell = 0, \dots, M$ $t_\ell := \text{prob}_T(\ell);$ $D := (t_0, \dots, t_M, 1 - \sum_{\ell=0}^M t_\ell);$ $i := 0;$ while ($i < n$): $\ell := \text{selectFin}(D);$ if ($\ell \neq \perp$) $w := \text{selectUnif}(s, \ell);$ $i := i+1;$ if ($\ell \neq \perp$ and $w \notin L(\mathbf{a})$) return False; return True;</pre>	<p>UnivNFA_T(\mathbf{a}, ε)</p> <pre> $\varepsilon := \min(\varepsilon, 1/6);$ $n := \lceil 5/(\varepsilon - 5\varepsilon^2)^2 \rceil;$ $M := \text{maxLen}_T(\varepsilon^2);$ for each $\ell = 0, \dots, M$ $t_\ell := \text{prob}_T(\ell);$ $D := (t_0, \dots, t_M, 1 - \sum_{\ell=0}^M t_\ell);$ $i := 0;$ cnt := 0; while ($i < n$): $\ell := \text{selectFin}(D);$ if ($\ell \neq \perp$) $w := \text{selectUnif}(s, \ell);$ $i := i+1;$ if ($\ell = \perp$ or $w \in L(\mathbf{a})$) cnt := cnt+1; if (cnt < n) return False else return True;</pre>
---	--

Fig. 5. On the left is the PRAX for NFA universality with respect to a certain tractable word distribution $\langle T \rangle$ —see Theorem 3. The value $1/6$ in $\min(\varepsilon, 1/6)$ can be replaced with any value $< 1/5$. The version of the algorithm on the right is logically equivalent; it mimics the process in Fig. 2 and is intended to give a more clear explanation of correctness.

Theorem 3. Let T be a tractable word distribution. Algorithm UnivNFA_T(\mathbf{a}, ε) in Fig. 5 is a polynomial randomized approximation algorithm for NFA universality relative to $\langle T \rangle$.

Proof. For brevity we use $A(\mathbf{a}, \varepsilon)$ to refer to $\text{UnivNFA}_T(\mathbf{a}, \varepsilon)$. The algorithm needs to be able to select repeatedly either a word w of length $\leq M$ from $\langle T \rangle$ or the outcome ‘ \perp ’. The finite probability distribution D refers to the outcomes $\{0, 1, \dots, M, \perp\}$; that is, a length $\ell \leq M$ or ‘ \perp ’. Statement $w \xleftarrow{\$} W^M$ of the process in Fig. 2 corresponds, for $W = \langle T \rangle$, to the first two statements of the while loop: First, select ℓ to be either a length $\leq M$ or ‘ \perp ’ using $\text{selectFin}(D)$ of Lemma 2. If a length ℓ is selected then use $\text{selectUnif}(s, \ell)$ to get a word from \mathbf{A}_s^ℓ .

Next we need to verify the three conditions about $A(\mathbf{a}, \varepsilon)$ in Definition 4. For the **first** one, suppose that \mathbf{a} is universal with respect to $\text{dom}\langle T \rangle$, that is, $\langle T \rangle(\mathbf{a}) = 1$, equivalently $\mathbf{L}(\mathbf{a}) = \text{dom}\langle T \rangle$. Then, every selection w from $\langle T \rangle^M$ is either \perp or a word in $\mathbf{L}(\mathbf{a})$, so the algorithm will return **True**. For the **second** condition, we assume that $\langle T \rangle(\mathbf{a}) < 1 - \varepsilon$. As T is tractable and $M = \text{maxLen}_T(\varepsilon^2)$, we have that $T(\mathbb{N}^{>M}) \leq \varepsilon^2$. Consider the version of the algorithm $A(\mathbf{a}, \varepsilon)$ on the right and the random process in Fig. 2 and assume that it selects exactly the same words w as $A(\mathbf{a}, \varepsilon)$ does. Then, algorithm $A(\mathbf{a}, \varepsilon)$ returns **True** if and only if the random variable **Cnt** in Lemma 5 takes the value n . Let $x = 5$. Then, using $p = 1$ and $g = 1 - \varepsilon$ in Lemma 5, we have $(p - g)/T(\mathbf{A}_s^{>M}) \geq \varepsilon/\varepsilon^2 = (1/\varepsilon)^{2-1} > x$ and

$$\mathbb{P}[A(\mathbf{a}, \varepsilon) = \text{True}] = \mathbb{P}[\text{Cnt} = n] = \mathbb{P}[\text{Cnt}/n \geq 1] \leq \frac{1}{x} + \frac{1}{4n(\varepsilon - x\varepsilon^2)^2} \leq \frac{1}{4}.$$

For the **third** condition, first note that $n = O(1/\varepsilon^2)$. As T is tractable, the magnitude of M and the running times of $\text{maxLen}_T(\varepsilon^2)$ and $\text{prob}_T(M)$ are polynomially bounded as required. Testing whether w is in $\mathbf{L}(\mathbf{a})$ can be done in time $O(|w||\mathbf{a}|)$, which is also polynomially bounded, as $|w| \leq M$. Thus, $A(\mathbf{a}, \varepsilon)$ runs within polynomial time w.r.t. $|\mathbf{a}|$ and $1/\varepsilon$, as required.

7.1 PRAX for the Lambert and Dirichlet Distributions

We apply next Theorem 3 to the Lambert and Dirichlet Distributions.

Corollary 4. *There is a polynomial randomized approximation algorithm for NFA universality relative to the Lambert distribution. In fact the algorithm works in time $O(|\mathbf{a}|(1/\varepsilon)^2 \log(1/\varepsilon))$ under the assumption of constant cost of `tossCoin` and of arithmetic operations⁷.*

Proof. First we need to show that the Lambert distribution is tractable. We have that $\mathbf{L}_{s,d}(\mathbb{N}^{>M}) \leq \varepsilon$ when

$$M \geq \log_s(1/\varepsilon) + d - 1$$

and the smallest such M is of magnitude $O(\log(1/\varepsilon))$. Computing the M and each value $\text{prob}_{\mathbf{L}_{s,d}}(\ell) = (1 - 1/s)(1/s)^{\ell-d}$, for $\ell \geq d$, can be done within polynomial time. Under the assumption of constant costs, computing M has constant cost and computing each $(1 - 1/s)(1/s)^{\ell-d}$ has cost $O(\ell)$. Hence, the time of the algorithm in Fig. 5 is $O(M^2 + n \times (M + |\mathbf{a}|M))$, where recall $n = O(1/\varepsilon^2)$.

⁷ If the precision q , say, of arithmetic needs to be accounted for then a polynomial in q term would be factored in.

Corollary 5. *There is a polynomial randomized approximation algorithm for NFA universality relative to the Dirichlet distribution. In fact the algorithm works in time $O(|\mathbf{a}|(1/\varepsilon)^2)$ under the assumption of constant cost of `tossCoin` and of arithmetic operations.*

Proof. First we need to show that the Dirichlet distribution is tractable. We need to find an appropriate M such that $D_{t,d}(\mathbb{N}^{>M}) \leq \varepsilon$, or equivalently, $1 - 1/\zeta(t) \sum_{n=d}^M (n+1-d)^{-t} \leq \varepsilon$. As each term $(n+1-d)^{-t}$ of the sum is $\geq (M+1-d)^{-t}$, it is sufficient to find an appropriate M such that $1 - 1/\zeta(t)(M+1-d)^{-(t-1)} \leq \varepsilon$. The required M is the largest integer such that

$$M \leq {}^{t-1}\sqrt{\frac{1}{\zeta(t)(1-\varepsilon)}} + d - 1$$

Using the fact that $1/(1-\varepsilon) = (1/\varepsilon)/(1/\varepsilon - 1)$, we get $M \in O\left({}^{t-1}\sqrt{\frac{1/\varepsilon}{1/\varepsilon - 1}}\right)$, which implies that the above M is of magnitude $O(\log(1/\varepsilon))$ and, in fact, M is of magnitude $O(1)$. Computing M and each $\text{prob}_{D_{t,d}}(\ell) = (1/\zeta(t))(\ell+1-d)^{-t}$, for $\ell \geq d$, can be done within polynomial time. Under the assumption of constant costs, computing M has constant cost and computing each $(1/\zeta(t))(\ell+1-d)^{-t}$ also has constant cost. Hence, the time of the algorithm in Fig. 5 is $O(M + n \times (M + |\mathbf{a}|M))$, where recall $n = O(1/\varepsilon^2)$.

7.2 A possible PAX for NFA Universality

Using the concept of a tractable distribution T , which is assumed fixed, we define below a simple PAX for NFA universality.

```

UnivNFA2 $_T$ ( $\mathbf{a}, \varepsilon$ )
   $M := \text{maxLen}_T(\varepsilon)$ ;
  for each  $\ell = 0, \dots, M$ 
    for each  $w \in \mathbf{A}_s^\ell$ 
      if ( $w \notin L(\mathbf{a})$ ) return False;
  return True;

```

Fig. 6. The algorithm tests whether all words of length up to M are accepted by the given NFA \mathbf{a} , that is, whether $\mathbf{A}_s^{\leq M} \subseteq L(\mathbf{a})$. If yes then $\langle T \rangle(\mathbf{A}_s^{>M}) \leq \varepsilon$ implies $\langle T \rangle(L(\mathbf{a})) \geq 1 - \varepsilon$, as required.

There are two problems with the above idea. Processing all possible words of length up to M would be inefficient in practice. So although in theory the above algorithm is a PAX, the PRAX version would be faster in practice. The second problem is that, for certain word distributions like the Lambert and Dirichlet distributions, one could in fact allow the minimum length d be part of the input

to the algorithm, in which case computing all words of length d would be an exponential time task, so the algorithm would not be a PAX.

What about the case where the alphabet is **unary**, namely $A_1 = \{0\}$? In this case, we have that the NFA universality problem is NP-complete, and we note that the inner loop of the above algorithm can be omitted, so the PAX algorithm could be faster than the PRAX one in Theorem 3. Of course the case of unary alphabets normally falls outside the context of coding and information theory so the value of the PAX algorithm is not clear.

8 Concluding Remarks

The concept of approximate maximality of a block code introduced in [14] leads naturally to the concept of approximately universal block NFAs and also of approximately universal NFAs in general relative to a desirable probability distribution on words. These concepts are meaningful in coding theory where the languages of interest are finite or even regular and can be represented by automata, [15,19,5,14].

Algorithm UnivNFA can be used to decide approximate universality (relative to tractable distributions) of any context-free language, or even any polynomially decidable language $L(\mathbf{a})$, where now \mathbf{a} would be a context-free grammar, or a polynomial Turing machine. Of course universality of context-free grammars (or Turing machines) is undecidable! However, extending our approach to grammars, or Turing machines, is outside of our motivation from coding and information theory and we cannot tell whether it could lead to any meaningful results.

Our approach can possibly be used to address other similar hard problems. For example, consider the empty DFA intersection problem EMPTY_DFA. Let $p \in [0, 1]$. We say that a DFA \mathbf{a} is p -empty relative to a word distribution W , if $W(\mathbf{a}) \leq p$. For example, a block DFA \mathbf{b} of word length ℓ is p -empty relative to the uniform distribution on A_s^ℓ , if $|L(\mathbf{b})|/s^\ell \leq p$. Let \mathbf{a}^c denote the complement of the DFA \mathbf{a} relative to W , that is, the DFA accepting $\text{dom } W - L(\mathbf{a})$. In particular, here we assume that $\text{dom } W = A_s^*$. Then, \mathbf{a}^c can be constructed from \mathbf{a} in linear time.

Remark 11. A DFA \mathbf{a} is p -empty relative to W if and only if \mathbf{a}^c is $(1-p)$ -universal relative to W .

As stated already in [17], given DFAs $\mathbf{a}_1, \dots, \mathbf{a}_m$, deciding whether their intersection is empty is equivalent to deciding whether the union of $\mathbf{a}_1^c, \dots, \mathbf{a}_m^c$ accepts A_s^* . Note here that, in linear time, one can compute an NFA \mathbf{a} accepting that union. The question of whether the intersection of $\mathbf{a}_1^c, \dots, \mathbf{a}_m^c$ is p -empty (relative to some W) is equivalent to whether the NFA \mathbf{a} is $(1-p)$ -universal (relative to W). Thus, Corollary 2 or Theorem 3 can be used to give a randomized approximate answer to the p -emptiness problem for DFA intersection.

Another hard problem that can possibly be approximated via a tractable distribution T is whether two languages are approximately equal (or two NFAs

are approximately equivalent). In analogy to the universality index of a language, here one can define the overlap index of two languages to be the probability that a word selected from T is not in the symmetric difference of the two languages.

In closing we note that every coNP language L can be expressed as a $[0, 1]$ -value language L_v and, therefore, it can be approximated by languages $L_{v,\varepsilon}$. However, the study of this generalization is outside the scope of the present paper, so we leave it as a topic for future research.

References

1. Margareta Ackerman and Jeffrey O. Shallit. Efficient enumeration of words in regular languages. *Theor. Comput. Sci.*, 410(37):3461–3470, 2009.
2. Sanjeev Arora and Boaz Barak. *Computational Complexity – a modern approach*. Cambridge University Press, New York, 2009.
3. Emmanuel Arrighi, Henning Fernau, Stefan Hoffmann, Markus Holzer, Ismaël Jecker, Mateus de Oliveira Oliveira, and Petra Wolf. On the complexity of intersection non-emptiness for star-free language classes. *CoRR*, abs/2110.01279, 2021.
4. Olivier Bernardi and Omer Giménez. A linear algorithm for the random sampling from regular languages. *Algorithmica*, 62(1-2):130–145, 2012.
5. Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata*. Cambridge University Press, 2009.
6. Krystian Dudzinski and Stavros Konstantinidis. Formal descriptions of code properties: decidability, complexity, implementation. *International Journal of Foundations of Computer Science*, 23:1:67–85, 2012.
7. Henning Fernau and Andreas Krebs. Problems on finite automata and the exponential time hypothesis. *Algorithms*, 10, 2017.
8. Oded Goldreich. *Computational complexity - a conceptual perspective*. Cambridge University Press, 2008.
9. Solomon W. Golomb. A class of probability distributions on the integers. *Journal of Number Theory*, 2:189–192, 1970.
10. Solomon W. Golomb. Probability, information theory, and prime number theory. *Discrete Mathematics*, 106/107:219–229, 1992.
11. John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to automata theory, languages, and computation, 2nd Edition*. Addison-Wesley-Longman, 2001.
12. Helmut Jürgensen. Complexity, information, energy. *Int. J. Found. Comput. Sci.*, 19(4):781–793, 2008.
13. Stavros Konstantinidis and Mitja Mastnak. Embedding rationally independent languages into maximal ones. *J. Automata, Languages and Combinatorics*, 21(4):311–338, 2016.
14. Stavros Konstantinidis, Nelma Moreira, and Rogério Reis. Randomized generation of error control codes with automata and transducers. *RAIRO - Theoretical Informatics and Applications*, 52:169–184, 2018.
15. Brian H. Marcus, P. Siegel, and R. Roth. Constrained systems and coding for recording channels. In *Handbook of Coding Theory*, pages 1635–1764. Elsevier, 1998. See also 2001 version at <http://www.math.ubc.ca/~marcus/Handbook/>.

16. Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge Univ. Press, 2nd edition, 2017.
17. Narad Rampersad, Jeffrey Shallit, and Zhi Xu. The computational complexity of universality problems for prefixes, suffixes, factors, and subwords of regular languages. *Fundamenta Informaticae*, 116:223–236, 2012.
18. Grzegorz Rozenberg and Arto Salomaa, editors. *Handbook of Formal Languages, Vol. I*. Springer-Verlag, Berlin, 1997.
19. Alexander Vardy. Trellis structure of codes. In *Handbook of Coding Theory*, pages 1989–2117. Elsevier, 1998.