An Embellished Account of Agafonov's Proof of Agafonov's Theorem

Thomas Seiller, Jakob Grue Simonsen

An Embellished Account of Agafonov's Proof of Agafonov's Theorem

Thomas Seiller and Jakob G. Simonsen

Abstract

We give an account of Agafonov's original proof of his eponymous theorem. The original proof was only reported in Russian [11, 1] in a journal not widely available, and the work most commonly cited in western literature is instead the english translation [2] of a summary version containing no proofs [3]. The account contains some embellishments to Agafonov's original arguments, made in the interest of clarity, and provides some historical context to Agafonov's work.

We give an account of Agafonov's original proof of his eponymous theorem. The original proof was only reported in Russian [11, 1] in a journal not widely available, and the work most commonly cited in western literature is instead the english translation [2] of a summary version containing no proofs [3].

The account contains some embellishments to Agafonov's original arguments, made in the interest of clarity:

1. The original proof relies on results of Postnikova [14]. We detail Postnikova's contribution and provide some historic context to her result.

2. The original proof contained a mixture of arguments expressed both via running text and explicit lemmas and theorems. While we have retained the general flow of argumentation from the original, we have used explicit lemmas and propositions for a number of observations occurring in the running text.

3. We have made several arguments explicit and provided detailed arguments in places where Agafonov relied on immediate understanding from his specialist audience, but where we believe that non-expert readers with modern sensibilities might prefer more elaborate explanations. The most pertinent examples are:

   (a) We explicitly prove why it suffices to prove that a connected finite automaton picks out $b \in \{0,1\}^n$ for $n = 1$ with limiting frequency p from any p-distributed sequence (Lemma 13).

   (b) We have appealed directly to probabilistic reasoning (using Chebyshev's Inequality) in the proof that, *par abus de langage*, the probability of deviation from probability p among the symbols selected by a finite automaton from sets of substrings picked from a p-distributed sequence tends to zero with increasing length of the strings (Lemma 16). In [1], this was essentially proved by a reference to the Strong Law of Large Numbers and a statement that the proof was similar to Lemma 3 of Loveland's paper [10].

Acknowledgements. The authors warmly thank Łukasz Czajka and Anastasia Volkova for their help in translating the russian documents.

1 Preliminaries

If $\alpha = a_1 a_1 \cdots$ is a right-infinite sequence over an alphabet $\mathcal{A}$ and N is a positive integer, we denote by $\alpha|_{\leq N}$ the finite string $a_1 a_2 \cdots a_N$.

We denote by $\mathcal{A}^*$ the set of (finite) words over $\mathcal{A}$ and by $\mathcal{A}^+$ the set of finite *non-empty* words over $\mathcal{A}$.

Definition 1. A finite probability map (over an alphabet $\mathcal{A}$) is a map $p : \mathcal{A}^+ \longrightarrow [0,1]$ such that, for all positive integers n, $\sum_{a_1 \cdots a_n \in \mathcal{A}^n} p(a_1 \cdots a_n) = 1$.

A finite probability map p is said to be:

- *Bernoulli* if, for all positive integers n, and all $a_1, \ldots, a_n \in \mathcal{A}$, $p(a_1 \cdots a_n) = \prod_{j=1}^n p(a_j)$.
- *Equidistributed* if, for any string $a_1 \cdots a_n \in \mathcal{A}^n$, $p(a_1 \cdots a_n) = |\mathcal{A}|^{-n}$.

Observe that an equidistributed p is also Bernoulli. For alphabets $|\mathcal{A}| > 1$, any map $g : \mathcal{A} \longrightarrow [0,1]$ with $\sum_{a \in \mathcal{A}} g(a) = 1$ induces a Bernoulli finite probability map p_g by letting $p_g(a_1 \cdots a_n) \triangleq \prod_{j=1}^n g(a_j)$. This map is equidistributed iff $g(a) = |A|^{-1}$ for every $a \in \mathcal{A}$.

The use of the word "Bernoulli" is due to the fact that Bernoulli finite probability maps correspond directly to the measure of cylinders in Bernoulli shifts [20]; in the literature on normal numbers, the word Bernoulli is sometimes used slightly differently, for example Schnorr and Stimm [18] use the term Bernoulli sequences for sequences distributed according to finite probability map that are equidistributed in our terminology.

We are interested in the finite probability maps whose values can be realized as the limiting frequencies of finite words in right-infinite sequences over $\{0,1\}$.

Definition 2. Let $\mathbf{b} = b_1 \cdots b_N$ and $\mathbf{a} = a_1 \cdots a_n$ be finite words over $\mathcal{A}$. We denote by $\#_{\mathbf{a}}(\mathbf{b})$ the number of occurrences of $\mathbf{a}$ in $\mathbf{b}$, that is, the quantity

$$|\{j : b_j b_{j+1} \cdots b_{j+n-1} = a_1 a_2 \cdots a_n\}|$$

Let p be a finite probability map over $\mathcal{A}$, and be α is a right-infinite sequence over $\mathcal{A}$. If the limit

$$\mathrm{freq}_{\mathbf{a}}(\alpha) = \lim_{N \to \infty} \frac{\#_{\mathbf{a}}(\alpha|_{\leq N})}{N}$$

exists and is equal to some real number f, we say that $\mathbf{a}$ occurs in α *with limiting frequency* f. If every $\mathbf{a} \in \mathcal{A}^+$ occurs in α with limiting frequency $p(\mathbf{a})$, we say that α is p-distributed.

Observe that a right-infinite sequence α is normal in the usual sense iff it is p-distributed for (the unique) equidistributed finite probability map p over $\mathcal{A}$. Also observe that it is not all finite probability maps p for which there exists a p-distributed sequence.

An example of a finite probability map that is *not* Bernoulli, but such that there is at least one p-distributed right-infinite sequence, is the map b defined by $b(\alpha) = 1/2$ if α does not contain either of the strings 00 or 11 (note that for each positive integer n, there are exactly two such strings of length n of each length), and $b(\alpha) = 0$ otherwise. Observe that the right-infinite sequence $010101\cdots$ is p-distributed.

In the remaining sections, we will work with the alphabet $\{0,1\}$ unless otherwise specified.

2 Preliminaries and Historical aspects

2.1 Borel

The notion of p-distributed sequences can be traced back to a 1909 paper by Émile Borel [5]. In this work, Borel studies the decimal representation of numbers and introduces the following definitions.

Definition 3 (Borel normality)**.** Consider an integer $b > 1$. Consider a number $0 < a < 1$ and denote by α^b its decimal sequence $a_1^b, \ldots, a_n^b, \cdots \in \{0, 1, \ldots, b-1\}^\omega$ in base b, i.e. $a = \sum_n \frac{a_n^b}{b^n}$. Then x is said to be:

1. *simply normal* w.r.t. the basis b when $\text{freq}_c(\alpha) = \frac{1}{b}$ for all $c \in \{0, 1, \ldots, b-1\}$;
2. *entirely normal* (or just *normal*) w.r.t. the basis b when for all integers n, k the number $b^k x$ is simply normal w.r.t. the basis b^m;
3. *absolutely normal* if it is entirely normal w.r.t. every possible basis b.

Borel already remarks that normality correspond to what we introduced as p-distribution[1]:

> The characterising property of a normal number[2] is the following: considering a sequence of p symbols, denoting by c_n the number of times this sequence is to be found within the n first decimal numbers, we have $\lim_{n\to\infty} \frac{c_n}{n} = \frac{1}{b^p}$.

The main result of Borel on normal numbers is the following theorem.

Theorem 1 (Borel [5])**.** *The probability that a number is absolutely normal is equal to 1, i.e. almost all numbers are absolutely normal.*

As a consequence, the probability that a number is normal, or simply normal, is also equal to 1. In particular, the cardinality of the set of normal numbers is equal to the cardinality of the continuum $2^{\aleph_0}$, and normal numbers are dense in the set of all real numbers.

2.2 von Mises

The notion of p-distributed sequences also appeared in connection with the notion of *kollektiv* introduced by von Mises in order to capture the concept of *random sequence*. The intuition behind von Mises approach it that a random sequence is one that cannot be predicted. I.e. the frequency of each possible outcome is independent from a the choice of a *Spielsystem*, i.e. a way to predict the outcome of successive trials. In other words, a sequence of trials outcomes is *not* random whenever there exists a strategy to select a subsequence of the trials in order to modify the frequency of the outcomes. This is expressed as the second condition in the following definition. As reported by Church [6], a sequence $\alpha = a_1, a_2, \ldots, a_n, \ldots$ in $\{0, 1\}^\omega$ is a *kollektiv* according to von Mises [22, 23] when:

1. $\text{freq}_1(\alpha)$ is defined and equal to p;

[1] The translation is ours, in which we replaced the basis 10 considered by Borel with a parametrised basis b.
[2] I.e. entirely normal w.r.t. the basis b, where $b = 10$ in Borel's original paper.

2. if $\beta = a_{n_1}, a_{n_2}, \ldots$ is any infinite sub-sequence of α formed by deleting some of the terms of the latter sequence according to a rule which makes the deletion or retention of a_n depend only on n and $a_1, a_2, \ldots, a_{n-1}$, then $\text{freq}_1(\beta)$ is defined and equal to p.

However, Church judges this definition to be "too inexact in form to serve satisfactorily as the basis of a mathematical theory" and proposes the following formalisation.

Definition 4 (von Mises kollektiv)**.** Let α be a sequence $a_1, a_2, \ldots, a_n, \ldots$ in $\{0,1\}^\omega$. It is a *kollektiv* (in the sense of von Mises, as formalised by Church) when:

1. $\text{freq}_1(\alpha)$ is defined and equal to p;
2. If φ is any function of positive integers, if[3] $b_1 = 1$, $b_{n+1} = 2b_n + a_n$, $c_n = \varphi(b_n)$, and the integers n such that $c_n = 1$ form in order of magnitude an infinite sequence $n_1, n_2, \ldots$, then the sequence $\beta = a_{n_1}, a_{n_2}, \ldots$ satisfies that $\text{freq}_1(\beta)$ is defined and equal to p.

In this section, several other notions of kollektiv will be discussed and introduced, and we will therefore use the following definitions.

Definition 5 (Strategy)**.** A strategy S is a predicate over the set of finite binary words, i.e. $S \subset \{0,1\}^* = \cup_{i=0}^{\omega} \{0,1\}^i$.

Definition 6 (Selected Subsequence)**.** Given a strategy S and an infinite sequence $\alpha = a_1, a_2, \ldots, a_n, \ldots$ in $\{0,1\}^\omega$, we define the sequence $S(\alpha)$ as follows. Let $i_1, i_2, \ldots, i_k, \ldots$ be the (increasing) sequence of indices j such that $\alpha|_{\leq j-1} \in S$.

$$S(\alpha)_j = a_{i_j}$$

Definition 7 (Kollektiv)**.** A sequence $\alpha = a_1, a_2, \ldots, a_n, \ldots$ in $\{0,1\}^\omega$ is a *kollektiv* w.r.t. a set of strategies $\mathbf{S}$ when:

1. $\text{freq}_1(\alpha)$ is defined and equal to p;
2. for any strategy $S \in \mathbf{S}$, $\text{freq}_1(S(\alpha))$ is defined and equal to p.

2.3 Church

With this definition, the notion of von Mises kollektiv coincides with that of kollektiv w.r.t. the set of all strategies. As discussed by several authors [21, 16, 9, 8], this notion of kollektiv is however inadequate, because it is too restrictive. This is further explained by Church, who explains why no kollektiv can exist if one considers such a strong notion:

> [...] it makes the class of random sequences associated with any probability p other than 0 or 1 an empty class. For the failure of (2) may always be shown by taking $\varphi(x) = a_{\mu(x)}$ where $\mu(x)$ is the least positive integer m such that $2^m > x$: the sequence $a_{n_1}, a_{n_2}, \ldots$ will then consist of those and only those terms of $a_1, a_2, \ldots$ which are 1's[4].

As a consequence, Church introduces a new notion of kollektiv, by factorising in the notion of computability. This choice is furtehr argumented as follows:

[3] Note that the terms b_n are written as follows in binary $b_n = 1a_1a_2\ldots a_{n-1}$.

[4] Indeed, the function defined by Church ensures that $c_n = a_n$.

> the scientist concerned with making predictions or probable predictions of some phenomenon must employ an effectively calculable function : if the law of the phenomenon is not approximable by such a function, prediction is impossible. Thus a Spielsystem should be represented mathematically, not as a function, or even as a definition of a function, but as an effective algorithm for the calculation of the values of a function.

Definition 8 (Church kollektiv)**.** Let α be a sequence $a_1, a_2, \ldots, a_n, \ldots$ in $\{0,1\}^\omega$. It is a *kollektiv* (in the sense of Church) when:

1. $\text{freq}_1(\alpha)$ is defined and equal to p;
2. If φ is any *effectively calculable*[5] function of positive integers, if $b_1 = 1$, $b_{n+1} = 2b_n + a_n$, $c_n = \varphi(b_n)$, and the integers n such that $c_n = 1$ form in order of magnitude an infinite sequence $n_1, n_2, \ldots$, then the sequence $\beta = a_{n_1}, a_{n_2}, \ldots$ satisfies that $\text{freq}_1(\beta)$ is defined and equal to p.

2.4 Admissible sequences

Towards the general purpose of defining mathematically the notion of random sequence, other notions were also considered at the time. For our purpose, the notions of "admissible number" introduced by Copeland [7], also studied by Reichenbach under the name "normal number" [16, 17] will be of interest.

Definition 9 (Copeland-admissible sequence.)**.** Let $\alpha = a_1, a_2, \ldots, a_n, \ldots$ be a sequence in $\{0,1\}^\omega$. For all integers r, n, define the sequence

$$(r/n)\alpha = a_r, a_{r+n}, \ldots, a_{r+kn}, \ldots$$

The sequence α is *admissible* (in the sense of Copeland) if the following are satisfied:

1. For all r, n, $\text{freq}_1((r/n)\alpha)$ is defined and equal to p.
2. $(1/n)\alpha, (2/n)\alpha, \ldots, (n/n)\alpha$ are *independent*[6] numbers.

Note that Copeland remarks that this second item is a consequence of the assumption that the sequence is obtained by independent trials (i.e. "the probability of success is a constant and does not vary from one trial to the next"). Church notes the connection between this notion of normal numbers and that of "completely normal number" by Armand Borel [5]:

> These admissible numbers (to adopt Copeland's term) are closely related to the normal numbers of Borel – indeed an admissible number associated with the probability $\frac{1}{2}$ is the same as a number entièrement normal to the base 2.

[5] Today, one would rather use the terminology "computable".

[6] Independence here is understood in terms of probability theory, as is detailed in Copeland's paper in which he states that two numbers are independent if and only if $p(x \cdot y) = p(x) \cdot p(y)$.

2.5 Postnikov and Pyateskii

Around twenty years after Church's paper, a notion of of *Bernoulli-normal* sequences was introduced by russian mathematicians, Postnikov and Pyateskii [13]. This notion coincides with the notion of p-distributed sequence defined above.

Definition 10 (Bernoulli-normal sequence)**.** Let $\alpha \in \{0,1\}^\omega$ be a sequence $a_1, a_2, \ldots, a_n, \ldots$ and consider for every integer $s > 0$ the s-th *caterpillar* of x:

$$\beta^s = (a_1, \ldots, a_{s-1}), (a_2, \ldots, a_s), \ldots, (a_P, \ldots, a_{P+s-1}), \ldots.$$

The sequence α is *Bernoulli-normal* if for any word $\mathbf{w}$ of length s with j ones, $\mathrm{freq}_{\mathbf{w}}(\beta^s)$ exists and is equal to $p^j(1-p)^{s-j}$.

In subsequent work, Postnikov [12] considers the following alternative definition of admissible sequences. While this differs from Copeland's definition, we provide here a proof that the two notions coincide.

Definition 11 (Postnikov admissible sequence)**.** Let $\alpha = a_1, a_2, \ldots, a_n, \ldots$ be a sequence in $\{0,1\}^\omega$. This sequence is called *admissible* (in the sense of Postnikov) if for any word $\mathbf{w}$ of length m with $r_1, r_2, \ldots, r_k$ the positions of its 1s, the sequence $\beta[\mathbf{w}] = b_0, b_1, \ldots, b_n, \ldots$, defined by

$$b_n = a_{nm+r_1}, a_{nm+r_2}, \ldots, a_{nm+r_k},$$

satisfies that $\mathrm{freq}_{1^k}(\beta[\mathbf{w}])$ exists and is equal to p^k.

Lemma 2. *A sequence $\alpha \in \{0,1\}^\omega$ is admissible in the sense of Copeland if and only if it is admissible in the sense of Postnikov.*

Proof. Let α be a Postnikov-admissible sequence. Let us define the word $\mathbf{u_i^n} = 00\ldots010\ldots0$, of length n with a single 1 at position $1 \leqslant i \leqslant n$. Then $\mathrm{freq}_1(\beta)$ exists and is equal to p. Since $\beta[\mathbf{u_i^n}] = (i/n)\alpha$, this proves α satisfies the first item in Copeland's definition. The second item, namely the independance of the sequence $(1/n)\alpha, (2/n)\alpha, \ldots, (n/n)\alpha$, is obtained by considering words $\mathbf{u_{i,j}^n}$, of length n with 1s exactly at the positions i and j. Indeed, we have $\mathrm{freq}_{11}(\beta[\mathbf{u_{i,j}^n}]) = p^2 = \mathrm{freq}_1(\beta[\mathbf{u_i^n}])\mathrm{freq}_1(\beta[\mathbf{u_j^n}])$, which coincide with Copeland's formalisation of independence.

Conversely, let α be a sequence, $\mathbf{w}$ a word of length m and $r_1, r_2, \ldots, r_k$ the positions of the 1s in $\mathbf{w}$. If α is Copeland-admissible, $\mathrm{freq}_{1^k}(\beta[\mathbf{w}]) = \prod_{i=1}^k \mathrm{freq}_1(\beta[\mathbf{u_{r_i}^n}])$ by the requirement of independence, and therefore

$$\mathrm{freq}_{1^k}(\beta[\mathbf{w}]) = \prod_{i=1}^{k} \mathrm{freq}_1((r_i/n)\alpha) = p^k$$

using that $\beta[\mathbf{u_{r_i}^n}] = (r_i/n)\alpha$ and the first property of Copeland-admissible sequences, namely that $\mathrm{freq}_1((i/n)\alpha) = p$ for all $1 \leqslant i \leqslant n$. □

Postnikov's then shows how the two notions, i.e. Bernoulli-normal and admissibility, coincide. However, the proof of Postnikov's theorem is – to the authors' knowledge – not available in english. As this result is related to the proof of Agafonov's theorem, we expect to include a translation in a later version of this document.

Theorem 3 (Postnikov [12])**.** *A sequence $\alpha \in \{0,1\}^\omega$ is Bernoulli-normal if and only if it is admissible.*

2.6 Postnikova

A few years later, a short and beautiful paper by Postnikova characterises Bernoulli-normal sequences as the sequences for which the distribution of 1s is preserved by selecting strategies depending only on a finite number of preceding bits. In fact, Postnikova's result is the first to introduce finiteness and widely opens the way to Agafonov's theorem. It is stated as a new, restricted, notion of kollektiv.

Definition 12 (Postnikova-kollektiv)**.** Let $\alpha = a_1, a_2, \ldots, a_n, \cdots \in \{0,1\}^\omega$ be a sequence. The sequence α will be called a *kollektiv* (in the sense of Postnikova) if:

1. $\text{freq}_1(\alpha)$ exists and is equal to p;
2. for all word $\mathbf{w}$ of length s, $\mathbf{w}$ occurs in α an infinite number of times, and if a subsequence β is made up consisting of the values immediately following the appearance of $\mathbf{w}$ then $\text{freq}_1(\beta)$ exists and is equal to p.

Note that using our own definition of kollektiv w.r.t. sets of strategies, a Postnikova-kollektiv is a kollektiv w.r.t. the set of strategies defined by a single finite word used as postfix, i.e. strategies $S_{\mathbf{w}}$ defined as

$$\{\mathbf{v} \in \{0,1\}^* \mid \exists \mathbf{u}, \mathbf{v} = \mathbf{u} \cdot \mathbf{w}\}.$$

Theorem 4 (Postnikova)**.** *A sequence $\alpha \in \{0,1\}^\omega$ is Bernoulli-normal if and only if it is a Postnikova-kollektiv.*

The proof of this theorem can be found in the english translation [15] of Postnikova's paper [14]. Note the error in translation in the definition of Postnikov-admissible sequences: the translator mentions "the relative frequency of appearances of ones in the sequence (2)", while it should read *the relative frequency of appearances of the word* 1^k *in the sequence (2).* The confusion comes from the original russian formulation (which can be traced back to Postnikov's work [12]) which is already ambiguous.

2.7 Agafonov

Agafonov's contribution was to relate this to the notion of automata. The main theorem of his original russian paper [11] is stated as follows.

Theorem 5 (Agafonov [11])**.** *A sequence α is normal if and only if it is a kollektiv w.r.t. the set of strategies computable by finite automata, i.e. it satisfies:*

1. $\text{freq}_1(\alpha)$ *exists and is equal to p;*
2. *for all automata M, the subsequence β consisting of the values immediately following the words accepted by M is such that* $\text{freq}_1(\beta)$ *exists and is equal to p.*

In fact, the proof of the implication from right to left in Agafonov's theorem is a consequence of Postnikova's theorem. Agafonov only refers to her work for this part of the proof. Indeed, if a sequence is a kollektiv in the sense of this theorem, it is also a Postnikova-kollektiv. Agafonov's contribution is therefore the proof of the converse implication, namely: if a sequence α is normal, it is a kollektiv w.r.t. the set of strategies computable by finite automata.

However, the notion of normality used by Agafonov is not the notion of p-distributed sequence (Definition 2), or equivalently of Bernoulli-normal sequence (Definition 10). Agafonov uses instead a notion of *normality by blocks.*

Definition 13 (Agafonov normal)**.** Let $\alpha \in \{0,1\}^\omega$ be a sequence $a_1, a_2, \ldots, a_n, \ldots$ and consider for every integer $s > 0$ the s-th *block sequence* of x:

$$\beta^s = (a_1, \ldots, a_{s-1}), (a_s, \ldots, a_{2s-1}), \ldots, (a_{ks}, \ldots, a_{(k+1)s-1}), \ldots.$$

The sequence α is *Agafonov-normal* if for any word $\mathbf{w}$ of length s with j ones, $\mathrm{freq}_{\mathbf{w}}(\beta^s)$ exists and is equal to $p^j(1-p)^{s-j}$.

This definition can be shown to be equivalent to Postnikov-admissibility which, combined with Postnikov's theorem (Theorem 3), proves the notion coincides with the usual notion of normality.

Lemma 6. *A sequence α is Agafonov-normal if and only if it is Postnikov-admissible.*

Proof. In fact, the proof of this appears in the proof of Postnikov theorem, as Agafonov-normality is used as an intermediate notion. The proof of the right-to-left implication is taken from Postnikov's proof [12]. The key observation is that the quantity $\mathrm{freq}_{1^k}(\beta[\mathbf{w}])$ that appears in Postnikov-admissibility corresponds to the frequency of appearance of words in Δ in the sequence of blocks defined from α, where Δ is the set of words $\mathbf{u}$ of length k that have 1s at these positions in which $\mathbf{w}$ has 1s (but which may differ from $\mathbf{w}$ on other bits).

We first show that a Postnikov-admissible sequence α is Agafonov-normal. Let Σ be the set of all length k word with fixed α bits equal to 1 and β bits equal to 0, $\alpha + \beta \leq k$. Write $T_l(\Sigma)$ the number of occurrences of Σ in the sequence of blocks. Then by induction on β, using the definition of admissibility, we obtain:

$$\lim_{l\to\infty} \frac{T_l(\Sigma)}{l} = p^\alpha q^\beta. \tag{1}$$

This gives the result by fixing Σ as a singleton, i.e. $\alpha + \beta = k$.

Conversely, consider given an Agafonov-normal sequence. By definition, we know that the frequency of a word w (with j bits equal to 1) is equal to $p^j q^{k-j}$. We want to sum this frequency over all words that have 1s at the same positions as w but in which some 0s may have become 1s. I.e. we have all combinations of putting 1s in $k-j$ boxes. So the sum can be written as:

$$\mathrm{freq}_{1^k}(\beta[\mathbf{w}]) = p^j \left(\sum_{k-j} \binom{i}{k-j} p^i q^{k-j-i} \right) = p^j (p+q)^{k-j} = p^j. \qquad \square$$

3 Agafonov's original proof: a direct translation

We fix once and for all the alphabet $\Sigma = \{0,1\}$.

Definition 14. Let $\mathbf{a} = a_1 a_2 \ldots a_n$ be a word over Σ. We define:

$$\mu_p(\mathbf{a}) = p^{\#_1(a)}(1-p)^{n-\#_1(a)}$$

Definition 15. For $M \subseteq \{0,1\}^N$, de define $\mu_p(M) = \sum_{\mathbf{w}\in M} \mu_p(\mathbf{w})$.

Definition 16. Let $\alpha = a_1, a_2, \ldots, a_n, \ldots$ be a sequence in $\{0,1\}^\omega$. For all natural number n we define the n-block decomposition of α as the sequence $(\alpha_{(n,r)})_{r\geqslant 1}$ defined by

$$\alpha_{(n,r)} = a_{n(r-1)+1} a_{n(r-1)+2} \ldots x_{nr}$$

Definition 17. Let α be a sequence in $\{0,1\}^\omega$, $\mathbf{w}$ a finite word of length n, and k an integer. We define $\mathrm{freq}_{\mathbf{w}}(\alpha;k) = \frac{1}{k}\mathrm{Card}\{\alpha_{(n,r)} = \mathbf{w} \mid r \leqslant k\}$.

Notice that a sequence α is Agafonov-normal (Definition 13) if and only if for all finite word $\mathbf{w}$ of length n with j bits equal to 1, $\lim_{k\to\infty} \mathrm{freq}_{\mathbf{w}}(\alpha;k)$ exists and is equal to $p^j q^{n-j}$.

Definition 18. Let A be a strongly connected automata with set of states Q. For all $q \in Q$, we write A_q the automata A in which the state q is chosen as initial.

Definition 19. Let A be a strongly connected automata with set of states Q, and $q \in Q$. Let $\mathbf{w} = w_1 w_2 \dots w_n$ be a finite word. We write $A_q[\mathbf{w}]$ the word *picked out* by the automata A_q, i.e. the word $w_{i_1} w_{i_2} \dots w_{i_k}$ where $i_1 < i_2 < \dots < i_k$ is the increasing sequence of indices $1 \leqslant j \leqslant n$ such that $\mathbf{w}|_{\leq j-1}$ is accepted by A_q.

Definition 20. Let A be a strongly connected automata with set of states Q. For all $p \in [0,1]$, $b \in [0,1]$, $n \in \mathbf{N}$ and $\epsilon > 0$, we define the sets:

$$D_n^p(b,\epsilon) = \{\mathbf{w} \in \{0,1\}^n \mid \forall q \in Q, \mathrm{len}(A_q[\mathbf{w}]) > bn, |\frac{\#_1(A_q[\mathbf{x}])}{\mathrm{len}(A_q[\mathbf{x}])} - p| < \epsilon\}$$

Claim 7. *For all $\epsilon > 0$ and all $p \in [0,1]$, $\lim_{n\to\infty} \mu_p(D_n^p(b,\epsilon)) = 1$.*

Proof. This claim is a consequence of Lemma 9 and Lemma 10 below, noting that $D_n^p(b,\epsilon) = \Sigma^n \setminus (E_n(b) \cup G_n(b,\epsilon))$. □

Theorem 8. *Let α be a normal sequence with ratio $p \in [0,1]$, A a strongly connected automata. Then the sequence $\beta = A[\alpha]$ is normal with ratio p.*

Proof. We will show that $\forall \epsilon, \exists L, \forall l \geqslant L, |\frac{1}{l}\sum_{i=1}^{l} \mathbf{y}_i - p| < \epsilon$.

Pick $\delta > 0$ small enough ($\delta < \frac{b\epsilon}{8}$). By Claim 7, we pick $n \in \mathbf{N}$ such that $\mu_p(D_n^p(b,\epsilon)) > 1-\delta$. Now, we consider $\eta < \frac{b\epsilon}{8}$ (i.e. sufficiently small); since α is normal, there exists $S \in \mathbf{N}$ such that $\forall s \geqslant S$, $\forall \mathbf{a} \in \{0,1\}^n$, $|\mathrm{freq}_a(\alpha;s) - \mu_p(\mathbf{a})| < \frac{\eta}{2^n}$, i.e. $\forall M \subseteq \{0,1\}^n$, $|\mathrm{freq}_M(\alpha;s) - \mu_p(M)| < \eta$.

We now consider the sequence $\beta_{[n,r]}$ as the sequence of blocks of $A[\alpha]$ (of changing length between 0 and n) corresponding to the sequence of blocks $\alpha_{(n,r)}$, and write θ the frequency of 1s in the blocks picked out from the blocks in $D_n^p(b,\frac{\epsilon}{2})$. Then $|\theta - p| < \frac{\epsilon}{2}$.

Now let $L = \sum_{i=1}^{s} \mathrm{len}(\beta_{[n,i]})$ and $\ell = \sum_{i\in I} \mathrm{len}(\beta_{[n,i]})$ with $I = \{i \leqslant s \mid \alpha_{(n,i)} \notin D_n^p(b,\frac{\epsilon}{2})\}$. We write $\theta = \frac{\sum_{i\in I} \#_1(\beta_{[n,i]})}{\sum_{i\in I} \mathrm{len}(\beta_{[n,i]})}$ and $\rho = \frac{\sum_{i=1}^{s} \#_1(\beta_{[n,i]})}{L}$. Then $|\rho - \theta| < \frac{\ell}{L}$.

We then show $\frac{\ell}{L} < \frac{\epsilon}{2}$ and deduce that $|\rho - p| < \epsilon$ as follows. We consider a small enough $\delta > 0$ and find S big enough to have

$$\frac{\mathrm{Card}\{i \leqslant S \mid \alpha_{[n,i]} \in D_n^p(b,\frac{\epsilon}{2}))\}}{S} > 1 - \delta - \eta$$

On one hand, for all $\mathbf{w} \in D_n^p(b,\frac{\epsilon}{2}))$ more than bn characters are picked out, therefore we have $L > (1-\delta-\eta)Sbn$. On the other hand, for all $\mathbf{w} \in \{0,1\}^n$ less than n characters are picked out and $\frac{\mathrm{Card}\{i \leqslant S \mid \alpha_{[n,i]} \notin D_n^p(b,\frac{\epsilon}{2}))\}}{S} < \delta + \eta$, thus $\ell < (\delta+\eta)Sn$. Hence $\frac{\ell}{L} < \frac{(\delta+\eta)}{(1-\eta-\delta)b} < \frac{\epsilon}{2}$.

Finally, $|\rho - p| \leqslant |\rho - \theta| + |\theta - p| < \epsilon$. □

Lemma 9. *Define* $E_n(b,q) = \{\mathbf{w} \in \{0,1\}^n \mid A_q[\mathbf{w}] \leqslant bn\}$, *and* $E_n(b) = \cup_{q\in Q} E_n(b,q)$. *Then for all* $p \in [0,1]$ *and for all automaton* A, *there exists* $c, d > 0$ *such that for all* $\epsilon > 0$, *the following holds.*

$$\lim_{n\to\infty} \mu_p(E_n(\frac{c-\epsilon}{d})) = 0$$

Proof. Let us consider $(X, \mathcal{B}, \mu_p)$ the measure space with $X = \{0,1\}^\omega$, $\mathcal{B}$ induced by cylinders, and $\mu_p(\{\alpha \mid \forall j \in \{1,2,\dots,n\}, \alpha_{i_j} = b_j\}) = \mu_p(b_1 b_2 \dots b_n)$.

For a word $\mathbf{v}$, define $C(\mathbf{v}) = \{\alpha \in \{0,1\}^\omega \mid \exists \beta \in \{0,1\}^\omega, \alpha = \mathbf{v}.\beta\}$. If R is a finite (prefix-free[7]) set of words, then

$$\mu_p(\cup_{\mathbf{v}\in R} C(\mathbf{v})) = \mu_p(R). \tag{2}$$

Now, take A a finite automaton $(\{0,1\}, Q, Q^*, \phi)$. This defines a Markov chain of set of states Q:

$$p_{i,j} = \begin{cases} 1 & \text{if } \phi(i,1) = \phi(i,0) = j \\ p & \text{if } \phi(i,1) = j, \phi(i,0) \neq j \\ 1-p & \text{if } \phi(i,1) \neq j, \phi(i,0) = j \\ 0 & \text{otherwise} \end{cases}$$

If A is strongly connected, there exists a smallest $n_{i,j}$ such that $p_{i,j}^{(n_{i,j})} > 0$. Define the period D as the least common multiple of the family $(n_{i,j})_{i,j\in Q^2}$.

Let $Q_0, Q_1, \dots, Q_{D-1}$ be the classes of "periodical states". Given Q_r, we have a Markov chain with probabilities $p_{i,j}^{(D)}$ for $i, j \in Q_r$. For all Q_r, there exists a family $(c_i)_{i\in Q_r}$ such that $\sum_{i\in Q_r} c_i = 1$ and $\lim_{n\to\infty} p_{i,j}^{(Dn)} = c_j$.

Consider $q_{A_j}(\alpha) = q_1 q_2 \dots$ the realisation of the Markov process with α as input and $j \in Q$. We have

$$\mu_{p,A_j}(\{q_{A_j}(\alpha) \mid \alpha \in M\}) = \mu_p(M). \tag{3}$$

Let $\nu_i^{(n)}(\vec{q}) = \text{Card}\{q_j = i \mid j \leqslant n\}$. For all $\epsilon > 0$ and all i, j,

$$\lim_{n\to\infty} \mu_{p,A_j}\{\vec{q} \text{ s.t. } |\frac{d}{n}\nu_i^{(n)}(\vec{q}) - c_i| \geqslant \epsilon\} = 0 \tag{4}$$

by the *law of large numbers for finite regular ergodic Markov chains.*

From Equation (3) and Equation (4), we have

$$\lim_{n\to\infty} \mu_p\{\alpha \text{ s.t. } |\frac{D}{n}\nu_i^{(n)}(q_{A_j}(\vec{x})) - c_i| \geqslant \epsilon\} = 0$$

For a finite word $\mathbf{a}$, write $q_{A_j}(\mathbf{a}) = q_1 q_2 \dots q_n$ ($n = \text{len}(\mathbf{a})$). Using Equation (2),

$$\lim_{n\to\infty} \mu_p\{a_1 a_2 \dots a_n, |\frac{D}{n}\nu_i^{(n)}(q_{A_j}(a_1 a_2 \dots a_n)) - c_i| \geqslant \epsilon\} = 0 \tag{5}$$

If in $q_{A_j}(\mathbf{a})$, there exists $q_i \in Q^*$, then A_j picks out a_j from $\mathbf{a}$. Let $c = \min_{i\in Q^*} c_i$. From Equation (5), for all $j \in Q$, $\lim_{n\to\infty} \mu_p E_n(\frac{c-\epsilon}{D}, j) = 0$.

The lemma then follows from $\mu_p E_n(\frac{c-\epsilon}{D}) \leqslant \sum_{j\in Q} \mu_p E_n(\frac{c-\epsilon}{D}, j)$. □

[7] This precision is added by the authors.

Lemma 10. *Define* $G_n(b,\epsilon,q) = \{\mathbf{w} \in \{0,1\}^n \mid \mathrm{len}(A_q[\mathbf{w}]) > bn, |\frac{\#_1(A_q[\mathbf{w}])}{\mathrm{len}(A_q[\mathbf{w}])} - p| > \epsilon\}$, *and* $G_n(b,\epsilon) = \cap_{q\in Q} G_n(b,\epsilon,q)$. *Then for all* p, b, ϵ *and all automaton* A,

$$\lim_{n\to\infty} \mu_p(G_n(b,\epsilon)) = 0.$$

Proof. (Similar to Lemma 3 from D.W. Loveland, *The Kleene hierarchy classification of recursively random sequences* [10].)

By the "strong law of large numbers", for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mu_p(\cup_{\ell\geqslant n}\{\mathbf{y} \mid |\frac{1}{\ell}\sum_{i=1}^{\ell} y_i - p| \geqslant \epsilon\}) = 0$$

Define $F_n(b,\epsilon) = \cup_{bn<\ell\leqslant n}\{\mathbf{y} \in \{0,1\}^\ell \mid |\frac{1}{\ell}\sum_{i=1}^{\ell} y_i - p| \geqslant \epsilon\}$. And define $R_n(b,\epsilon)$ as the set obtained from $F_n(b,\epsilon)$ by removing the words $\mathbf{w}$ such that there exists a word $\mathbf{u}$ in $F_n(b,\epsilon)$ with $\mathbf{u} \prec \mathbf{w}$.

From the fact that

$$\cup_{\mathbf{w}\in R_n(b,\epsilon)} \mathcal{C}(\mathbf{w}) \subset \cup_{\ell\geqslant bn}\{\mathbf{y} \mid |\frac{1}{\ell}\sum_{i=1}^{\ell} y_i - p| \geqslant \epsilon\}$$

and

$$\mu_p(R_n(b,\epsilon)) = \mu_p(\cup_{\mathbf{w}\in R_n(b,\epsilon)} \mathcal{C}(\mathbf{w}))$$

and the equation above, we deduce that

$$\lim_{n\to\infty} \mu_p R_n(b,\epsilon) = 0.$$

By Lemma 11 and the equality

$$G_n(b,\epsilon,q) = \{\mathbf{w} \in \{0,1\}^n \mid A_q[\mathbf{w}] \in S_n(b,\epsilon)\}$$

we get that $\mu_p(G_n(b,\epsilon,q)) \leqslant \mu_p(R_n(b,\epsilon))$. Consequently, $\lim_{n\to\infty} \mu_p(G_n(b,\epsilon,q)) = 0$ for all $q \in Q$, hence $\lim_{n\to\infty} \mu_p(G_n(b,\epsilon)) = 0$. □

Lemma 11. *Let* S *be a strategy, and* F *a finite subset of* $\{0,1\}^*$. *Let* R *be the set obtained from* F *by removing those words* $\mathbf{w}$ *such that there exists a word* $\mathbf{u} \in F$ *with* $\mathbf{u} \prec \mathbf{w}$ *(i.e.* $\mathbf{u}$ *is a prefix of* $\mathbf{w}$*). Let* M *be the set* $\{\mathbf{w} \in \{0,1\}^n \mid S(\mathbf{w}) \in F\}$. *Then*

$$\mu_p(M) \leqslant \mu_p(R).$$

Proof. It is sufficient to prove that for a given word $\mathbf{w} = a_1 a_2 \dots a_k$ the set $M = \{\mathbf{u} \in \{0,1\}^n \mid \mathbf{w} \preceq S(\mathbf{u})\}$ satisfies $\mu_p(M) \leqslant \mu_p(\mathbf{w})$. This is show by induction on the length k of the word $\mathbf{w}$.

The base case is $\mathbf{w} = a_1 = 1$ (by symmetry – 1 becomes 0, p becomes $1-p$ –, this is sufficient). Let $\alpha = x_1 x_2 \dots x_n$ be a word in M, and write x_f the first symbol picked out by S; in particular $x_f = 1$. Now, one can define $\bar{\alpha} = x_1 x_2 \dots x_{f-1} \bar{x}_f x_{f+1} \dots x_n$, i.e. the word obtained from α by simply flipping the f-th bit. Then $\bar{\alpha} \notin M$. One can then define the set $\bar{M} = \{\bar{\alpha} \mid \alpha \in M\}$. As $\mu_p(\bar{\alpha}) = \frac{1-p}{p}\mu_p(\alpha)$ and $\bar{\cdot}$ defines a one-to-one correspondence between

M and $\bar{M}$, we have $\mu_p(\bar{M}) = \frac{1-p}{p}\mu_p(M)$. Moreover, M and $\bar{M}$ are disjoint subsets of $\{0,1\}^n$, hence $\mu_p(M) \leqslant 1 - \mu_p(\bar{M})$. We can then conclude from these two equations that $\mu_p(M) \leqslant p$.

Now, consider the word $\mathbf{w} = a_1 a_2 \dots a_k a_{k+1}$ with $a_{k+1} = 1$. We have $M = \{\alpha \in \{0,1\}^n \mid a_1 \dots a_k 1 \preceq S(\alpha)\}$. Given $\alpha \in M$, define $\bar{\alpha}$ as the word obtained from α by flipping its $k+1$-th picked out bit, i.e. $\bar{\alpha}$ is the unique word obtained from $\bar{x}$ by flipping a single bit and such that $a_1 \dots a_k 0 \preceq S(\bar{\alpha})$. Define $\bar{M}$ as the set $\{\bar{\alpha} \mid \alpha \in M\}$. Let N be the set $\{\alpha \in \{0,1\}^n \mid a_1 \dots a_k \preceq S(\alpha)\}$. Then N contains both M and $\bar{M}$, and the latter two sets are disjoint. Moreover the induction hypothesis implies that $\mu_p(N) \leqslant \mu_p(a_1 a_2 \dots a_k)$. Hence $\mu_p(M) + \mu_p(\bar{M}) \leqslant \mu_p(a_1 a_2 \dots a_k)$. Since $\mu_{\bar{M}} = \frac{1-p}{p}\mu_p(M)$, we deduce that $\mu_p(M) \leqslant p\mu_p(a_1 a_2 \dots a_k) = \mu_p(a_1 a_2 \dots a_k 1)$. □

4 Adaptation of Agafonov's proof

We now give an embellished, modern account of Agafonov's proof; we have endeavoured to use pedagogical explanations and have extended the treatment to make the text more readily readable the the modern reader.

Definition 21. A *finite-state selector* over $\{0,1\}$ is a deterministic finite automaton $S = (Q, \delta, q_s, Q_F)$ over $\{0,1\}$. A finite-state selector is strongly connected if its underlying directed graph (states are nodes, transitions are edges) is strongly connected. Denote by $L(S)$ the language accepted by the automaton.

If $\alpha = a_1 a_2 \cdots$ is a finite or right-infinite sequence over $\{0,1\}$, the subsequence *selected by* A is the (possibly empty) sequence of letters a_n such that the prefix $a_1 \cdots a_{n-1} \in L(S)$, that is, the automaton when started on the finite word $a_1 \cdots a_{n-1}$ in state q_s ends in an accepting state after having read the entire word.

For two words $\mathbf{u}, \mathbf{v}$, we write $\mathbf{u} \preceq \mathbf{v}$ if $\mathbf{u}$ is a prefix of $\prec v$, and $\mathbf{u} \prec \mathbf{v}$ if $\mathbf{u}$ is a proper prefix of $\mathbf{v}$.

Definition 22. Let $\mathbf{a} = a_1 \cdots a_n$ and $\mathbf{b} = b_1 \cdots b_N$ be finite words over $\{0,1\}$. We denote by $\#_1(\mathbf{a})\mathbf{b}$ the number of occurrences of $\mathbf{a}$ in $\mathbf{b}$, that is, the quantity

$$|\{j : b_j b_{j+1} \cdots b_{j+n-1} = a_1 a_2 \cdots a_n\}|$$

Definition 23. Let $\mathbf{a} = a_1 a_2 \dots a_n$ be a word over $\{0,1\}$, and p a probability distribution on $\{0,1\}$. We define:

$$\mu_p(\mathbf{a}) = \prod_{i=1}^{n} p(a_i)$$

If $M \subseteq \{0,1\}^*$ is finite, we define $\mu_p(M) = \sum_{\mathbf{w} \in M} \mu_p(\mathbf{w})$ (and set $\mu_p(\emptyset) = 0$).

Definition 24. Let $\alpha = x_1 x_2 \dots x_n \cdots$ be a sequence over $\{0,1\}$. We say that α is *p-block-distributed* if, for each $n \geqslant 1$ and every $\mathbf{w} \in \{0,1\}^n$, the n-block decomposition $(\alpha_{(n,r)})_{r \geqslant 1}$ of α satisfies:

$$\lim_{k \to \infty} \frac{|i \leq k : \alpha_{(n,k)} = \mathbf{w}|}{k} = \mu_p(\mathbf{w})$$

As already remarked above, this notion coincides with Agafonov-normality (Definition 13).

Remark 1. Like in Agafonov's original paper, for a finite-state selector A, we *do not* require that all cycles in the underlying directed graph of A contain at least one accepting state. This assumption is occasionally made in modern papers on Agafonov's Theorem to ensure that if $\mathbf{w} \in \{0,1\}^\omega$ is a normal sequence, then $A[\mathbf{w}]$ is infinite as well. But just as in Agafonov's paper, the requirement turns out to be unnecessary (see Lemma 14).

However, in Agafonov's paper, the probability $\mu_p(1)$ of obtaining a 1 was assumed to satisfy $0 < \mu_p(1) < 1$ (i.e., both 0 and 1 occur with positive probability). Without this assumption, there are connected automata that fail to pick out infinite sequences from p-distributed ones. For example, define $A = (\{q_0, q_1\}, \{0,1\}, \delta, q_0, \{q_0\})$ where

$$\begin{array}{ll} \delta(q_0, 0) = q_0 & \delta(q_0, 1) = q_1 \\ \delta(q_1, 1) = q_0 & \delta(q_0, 1) = q_0 \end{array}$$

Define $\mu_p(0) = 1$ and $\mu_p(1) = 0$. Then, $\mathbf{w} = 10^\omega$ is p-distributed, but $A[\mathbf{w}] = 0$, hence is finite.

Motivated by Remark 1, we have the following definition:

Definition 25. A Bernoulli distribution $p : \{0,1\} \longrightarrow [0,1]$ is said to be *positive* if, for all $a \in \{0,1\}$, $p(a) > 0$. The probability map $\mu_p : \{0,1\}^* :\longrightarrow [0,1]$ is positive if p is positive.

Proposition 12 (Finite-State selectors are compositional)**.** *Let A and B be DFAs over the same alphabet. Then there is a DFA C such that, for each sequence $\mathbf{w}$, $C[\mathbf{w}] = B[\mathbf{A}[\mathbf{w}]]$.*

Proof. Let $A = (Q^A, \{0,1\}, \delta^A, q_0^A, F^A)$ and $B = (Q^B, \{0,1\}, \delta^B, q_0^B, F^B)$. Define $Q^C = Q^A \times Q^B$, and set $q_0^C = (q_0^A, q_0^B)$ and $F^C = F^A \times F^B$. For each $q^B \in Q^B$, define the set $D_{q^B} = \{(q, q^B) : q \in Q^A\} \subseteq Q^C$. Observe that $Q^C = \bigcup_{q^B \in Q^B} D_{q^B}$ and that for $q^B, r^B \in Q^B$ with $q^B \neq r^B$, we have $D_{q^B} \cap D_{r^B} = \emptyset$, and thus $\{D_{q^B} : q^B \in Q^B\}$ is a partitioning of Q^C. Hence, the transition relation, δ^C, of C may be defined by defining it separately on each subset D_{q^B}:

$$\delta^C((q, q^B), a) = \begin{cases} (r, q^B) & \text{if } q \notin F^A \text{ and } \delta^A(q, a) = r \\ (r, r^B) & \text{if } q \in F^A \text{ and } \delta^A(q, a) = r \text{ and } \delta^B(q^B, a) = r^B \end{cases}$$

Thus, when C processes its input, it freezes the current state q^B of B (the freezing is represented by staying within D_{q^B}) and simulates A until an accepting state of A is reached (i.e. just before A would select the next symbol); on the next transition, C unfreezes the current state of B and moves to the next state r^B of B and then freezes it and continues with a simulation of A.

Observe that a symbol is picked out by C iff the state is an element of $F^C = F^A \times F^B$ iff the symbol is the next symbol read after simulation of A reaches an accepting state of A when the current frozen state of B is an accepting state of B. □

The following shows that to prove that p-distributedness is preserved under finite-state selection, it suffices to prove that the limiting frequency of each $a \in \{0,1\}$ exists and is equal to $p(a)$.

Lemma 13. *Let α be a p-distributed sequence. The following are equivalent:*

- *For all connected DFAs A, $A[\alpha]$ is p-distributed.*

- *For all connected DFAs A and all $a \in \{0,1\}$, the limiting frequency of a in $A[\alpha]$ exists and is equal to $p(a)$.*

Proof. If, for all A, $A[\alpha]$ is p-distributed, then in particular the limiting frequency of a in $A[\alpha]$ exists and is equal to $p(a)$ for all A.

Conversely, suppose that, for all DFAs A and all $a \in \{0,1\}$, the limiting frequency of a in $A[\alpha]$ exists and is equal to $p(a)$. We will prove by induction on $k \geq 0$ that the limiting frequency of every $v_1 \cdots v_k v_{k+1} \in \{0,1\}^{k+1}$ exists and equals $p(v_1 \cdots v_k v_{k+1})$.

- $k = 0$: This is the supposition.

- $k \geq 1$. Suppose that the result has been proved for $k-1$. Let $v_1 \cdots v_k \in \{0,1\}^k$; by the induction hypothesis, the limiting frequency of $v_1 \cdots v_k$ in $A[\mathbf{w}]$ is $p(v_1 \cdots v_k)$. We claim that there is a strongly connected DFA B that, from any sequence, selects the symbol after each occurrence of $v_1 \cdots v_k$. To see that such a DFA exists, let there be a state for each element of $\{0,1\}^k$ and assume that the state is the current length-k string in a "sliding window" that moves over $\mathbf{w}$ one symbol at the time; when the window is moved one step, the DFA transits to the state representing the new length-k string in the window, i.e. from the state representing the word $w_1 \cdots w_k$, there are transitions to $w_2 \cdots w_k 0$ and $w_2 \cdots w_k 1$; it is easy to see that each state is reachable from every other state in at most k transitions. The unique final state of B is the state representing $v_1 \vdots v_k$; the start state of B can be chosen to be any state representing a string $w_1 \cdots w_k$ such that there is exactly k transitions to the final state.

  By Proposition 12, there is a connected DFA C such that $C[\mathbf{w}] = B[A[\mathbf{w}]]$.

  For any $a \in \{0,1\}$ and any sufficiently large positive integer N, we have

$$\frac{\#_1(a)C[\mathbf{w}_{\leq \mathbf{N}}]}{|C[\mathbf{w}_{\leq \mathbf{N}}]|} = \frac{\#_1(a)B[\mathbf{A}[\mathbf{w}_{\leq \mathbf{N}}]]}{|B[\mathbf{A}[\mathbf{w}_{\leq \mathbf{N}}]]|} == \frac{\#_1(v_1 \cdots v_k a)A[\mathbf{w}_{\leq \mathbf{N}}]}{\#_1(v_1 \cdots v_k)A[\mathbf{w}_{\leq \mathbf{N}}]}$$

  As C is connected, there is a real number b with $0 < b \leq 1$ such that C selects at least bN symbols from $\mathbf{w}_{\leq N}$, and by the induction hypothesis, for every $\epsilon > 0$, there is an M such that for all $N > M/b$, $\left|\frac{\#_1(a)C[\mathbf{w}_{\leq \mathbf{N}}]}{|C[\mathbf{w}_{\leq \mathbf{N}}]|} - p(a)\right| < \epsilon$ and hence $\left|\frac{\#_1(v_1 \cdots v_k a)A[\mathbf{w}_{\leq \mathbf{N}}]}{\#_1(v_1 \cdots v_k)A[\mathbf{w}_{\leq \mathbf{N}}]} - p(a)\right| < \epsilon$. But for all sufficiently large N, the induction hypothesis furnishes

$$\left|\frac{\#_1(v_1 \cdots v_k)A[\mathbf{w}_{\leq \mathbf{N}}]}{|A[\mathbf{w}_{\leq \mathbf{N}}]|} - p(v_1 \cdots v_k)\right| < \epsilon$$

  But as

$$\frac{\#_1(v_1 \cdots v_k a)A[\mathbf{w}_{\leq \mathbf{N}}]}{|A[\mathbf{w}_{\leq \mathbf{N}}]|} = \frac{\#_1(v_1 \cdots v_k a)A[\mathbf{w}_{\leq \mathbf{N}}]}{\#_1(v_1 \cdots v_k)A[\mathbf{w}_{\leq \mathbf{N}}]} \cdot \frac{\#_1(v_1 \cdots v_k)A[\mathbf{w}_{\leq \mathbf{N}}]}{|A[\mathbf{w}_{\leq \mathbf{N}}]|}$$

  we hence have (as $p(v_1 \cdots v_k)p(a) = p(v_1 \cdots v_k a)$ because p is Bernoulli):

$$\begin{aligned}
&\left|\frac{\#_1(v_1 \cdots v_k a)A[\mathbf{w}_{\leq \mathbf{N}}]}{|A[\mathbf{w}_{\leq \mathbf{N}}]|} - p(v_1 \cdots v_k a)\right| \\
&< \epsilon^2 + \epsilon\left(\frac{\#_1(v_1 \cdots v_k a)A[\mathbf{w}_{\leq \mathbf{N}}]}{\#_1(v_1 \cdots v_k)A[\mathbf{w}_{\leq \mathbf{N}}]} + \frac{\#_1(v_1 \cdots v_k)A[\mathbf{w}_{\leq \mathbf{N}}]}{|A[\mathbf{w}_{\leq \mathbf{N}}]|}\right) \\
&\leq \epsilon^2 + 2\epsilon
\end{aligned}$$

Hence, for all $a \in \{0,1\}$, the limiting frequency of $v_1 \cdots v_k a$ in $A[\mathbf{w}_{\leq \mathbf{N}}]$ exists and equals $pv_1 \cdots v_k a$, as desired.

□

Definition 26. A strategy S is a predicate over the set of finite words, i.e. $S \subseteq \{0,1\}^*$.

Given a strategy S and a right-infinite sequence $\mathbf{x}$ in $\{0,1\}^\omega$, we define the sequence $S(\mathbf{x})$ as follows. Let $i_1, i_2, \ldots, i_k, \ldots$ be the (increasing) sequence of indices j such that $\mathbf{x}_{<j} \in S$ and $S(\mathbf{x})_j = \mathbf{x}_{i_j}$.

Thus, $S(\mathbf{w})$ is simply the subsequence of $\mathbf{w}$ that are "picked out" by applying S to prefixes of $\mathbf{w}$. Note also that if $\mathbf{w} \in S$, then in any word on the form $\mathbf{w} \cdot b \cdot \mathbf{v}$, then S must pick b. Thus, S cannot be made to, for instance, only pick out 0 or 1–it picks out "the next symbol" after any $\mathbf{w} \in S$.

Definition 27. Let $A = (Q, \{0,1\}, \delta, q_0, F)$ be a connected DFA. For all $q \in Q$, we denote by A_q the automaton $(Q, \{0,1\}, \delta, q, F)$, i.e. where the state q is chosen as the initial state.

Definition 28. Let $A = (Q, \{0,1\}, \delta, q_0, F)$ be a connected DFA, and let $q \in Q$. Let α be a right-infinite sequence over $\{0,1\}$. We denote by $A_q[\mathbf{x}]$ the subsequence $\bar{\alpha}$ of α *picked out* by A_q, that is, $w_i \in \bar{\mathbf{w}}$ if and only if $A_q(\mathbf{w}_{<i})$ reaches an accepting state.

For every fixed positive integer n, it is clear that $(\{0,1\}^n, \Pr)$ is a finite probability space where $\Pr(M) = \mu_p(M)$ for every $M \subseteq \{0,1\}^n$.

Definition 29. Let $A = (Q, \{0,1\}, \delta, q_0, F)$ be a strongly connected DFA. For all $p \in [0,1]$, $b \in [0,1]$, $n \in \mathbf{N}$ and $\epsilon > 0$, we define sets $D_n^p(b,\epsilon)$, $E_n(b,q)$ and $G_n(b,\epsilon,q)$ as follows:

$$D_n^p(b,\epsilon,q) = \left\{ \mathbf{w} \in \{0,1\}^n \ : \ |A_q[\mathbf{w}]| > bn \text{ and } \max_{a \in \{0,1\}} \left| \frac{\#_1(a) A_q[\mathbf{w}]}{|A_q[\mathbf{w}]|} - p(a) \right| < \epsilon \right\} \tag{6}$$

$$D_n^p(b,\epsilon) = \bigcap_{q \in Q} D_n^p(b,\epsilon,q) \tag{7}$$

$$E_n(b,q) = \{\mathbf{w} \in \{0,1\}^n : |A_q[\mathbf{w}]| \leq bn\} \tag{8}$$

$$E_n(b) = \bigcup_{q \in Q} E_n(b,q) \tag{9}$$

$$G_n(b,\epsilon,q) = \left\{ \mathbf{w} \in \{0,1\}^n : |A_q[\mathbf{w}]| > bn \text{ and } \max_{a \in \{0,1\}} \left| \frac{\#_1(a) A_q[\mathbf{w}]}{|A_q[\mathbf{w}]|} - p(a) \right| \geq \epsilon \right\} \tag{10}$$

$$G_n(b,\epsilon) = \bigcup_{q \in Q} G_n(b,\epsilon,q) \tag{11}$$

Observe that, for all b, n, ϵ,

$$\{0,1\}^n = E_n(b) \cup D_n^p(b,\epsilon) \cup G_n(b,\epsilon)$$

(but $E_n(b)$ and $G_n(b,\epsilon)$ are not necessarily disjoint).

Lemma 14. *Let $A = (Q, \{0,1\}, \delta, q_0, F)$ be strongly connected, n a positive integer, and b be a real number with $b > 0$. Then there exist real numbers $c, d > 0$ such that for all real numbers $\epsilon > 0$:*

$$\lim_{n \to \infty} \mu_p \left(E_n \left(\frac{c - \epsilon}{d} \right) \right) = 0$$

Proof. Now, the DFA A induces a stochastic $|Q| \times |Q|$ matrix $\mathbf{P}$ by setting

$$\mathbf{P}_{ij} = \sum_{a \in \{0,1\}} \mu_p(a) \cdot [\delta(i,a) = j].$$

Note in particular that $\mathbf{P}_{ij} = 0$ iff there are no transitions from i to j in Q on a symbol $a \in \{0,1\}$ with $\mu_p(a) > 0$. As A is strongly connected, there exists a path from state i to state j for each $i, j \in Q$, and as p is a positive Bernoulli distribution, we have $\mu_p(a) = p(a) > 0, i, j$, whence for each i, j there is an integer n_{ij} such that $\mathbf{P}_{ij}^{n^{ij}} > 0$, that is, $\mathbf{P}$ (and its associated Markov chains) is irreducible. As all states of a finite Markov chain with irreducible transition matrix are positive recurrent, standard results (see, e.g., [19, Thm. 54]) yield that there is a unique positive stationary distribution $\pi : Q \longrightarrow [0,1]$ (s.t., for all $i \in Q$, $\pi(i) > 0$ and $\lambda(i) = \sum_{j \in Q} \lambda(j)\mathbf{P}_{ij}$). Furthermore, the expected return time M_i to state i satisfies $M_i = 1/\pi(i)$ [19, Thm. 54].

Let D be the least common multiple of the set $\{n_{ij} : (i,q) \in Q^2\}$, and let $(X_n)_{n \geq 1} = (X_1, X_2, \ldots)$ be a Markov chain with transition matrix $\mathbf{P}$ and some initial distribution λ on the states.

Consider, for each $i \in Q$, the variable $V(i)$ where V_i, where

$$V_i(n) = \sum_{k=0}^{n-1} 1_{X_k = i}$$

As $\mathbf{P}$ is irreducible, the Ergodic Theorem for Markov chains (see, e.g., [19, Thm. 75]) yields that

$$\lim_{n \to \infty} \Pr\left(\left| \frac{V_i(n)}{n} - \pi(i) \right| \geq \epsilon \right) = \lim_{n \to \infty} \Pr\left(\left| \frac{V_i(n)}{n} - \frac{1}{M_i} \right| \geq \epsilon \right) = 0 \tag{12}$$

Let $\alpha \in \{0,1\}^n$ and let $q_{A_j}(\alpha) = q_1 \cdots q_{n-1}$ be the sequence of states visited when A is given α as input starting from state j (i.e., $q_1 = j$). Observe that the probability of observing the state sequence $q_1 \cdots q_{n-1}$ in a Markov chain with transition matrix $\mathbf{P}$ is $\Pr(q_1 \cdots q_{n-1}) = \mu_p(\{\alpha : q_{A_j}(\alpha) = q_1 \cdots q_{n-1}).$ and thus:

$$\Pr\left(q_1 \cdots q_{n-1} : \left| \frac{\sum_{k=0}^{n-1} [q_k = i]}{n} - \pi(i) \right| \geq \epsilon \right) = \tag{13}$$

$$\mu_p\left(\alpha : q_{A_j}(\alpha) = q_1 \cdots q_n \wedge \left| \frac{\sum_{k=0}^{n-1} [q_k = i]}{n} - \pi(i) \right| \geq \epsilon \right) = \tag{14}$$

$$\mu_p\left(\alpha : q_{A_j}(\alpha) = q_1 \cdots q_n \wedge \left| \frac{V_i(n)}{n} - \pi(i) \right| \geq \epsilon \right) \tag{15}$$

Hence, by (Equation (12)) and the above, we have

$$\lim_{n \to \infty} \mu_p\left(\alpha : q_{A_j}(\alpha) = q_1 \cdots q_n \wedge \left| \frac{V_i(n)}{n} - \pi(i) \right| \geq \epsilon \right) = 0 \tag{16}$$

If $q_{A_j}(\mathbf{w}) = q_1 \cdots q_{n-1}$ and one of the states $q_i \in \{q_1, \ldots, q_{n-1}\}$ is an element of F, then A_j picks out w_i. Set $c = \min_{q_j \in F} \pi(i)$. Then, for all $j \in Q$, (Equation (16)) yields that $\lim_{n \to \infty} \mu_p(E_n(c - \epsilon)) = 0$. The result now follows from $\mu_p(E_n(c-\epsilon)) \leq \sum_{j \in Q} \mu_p(E_n(c-\epsilon), j)$. □

Remark 2. In Lemma Lemma 14, the assumption that the DFA A is strongly connected can be omitted if we make the assumption that every cycle of A contains an accepting state.

Let k be the maximal number of non-accepting states in any path in A from one accepting state to another that does not contain any other accepting states than the start and end states of the path. As every cycle of A contains an accepting state, k is well-defined. If $\mathbf{w} = w_1 w_2 \cdots \in \{0,1\}^\omega$ and $A[\mathbf{w}]$ is infinite, then, by construction, $|A_q[w_1 \cdots w_n]| \geq d$ where $n = d(k+1) + r$ and $0 \leq r < k+1$. As $d = (n-r)/(k+1) > n/(k+1) - 1 \geq n/(k+2)$ for $n > 2(k+1)$, we have $|A_q[w_1 \cdots w_n]| \geq n/(k+2)$. Hence, for $n > 2(k+1)$, $A_q[w_1 \cdots w_n] > n/(k+2)$, whence $E_n(1/(k+2), q) = \emptyset$ for $n > 2(k+1)$, and thus $\mu_p(E_n(1/(k+2), q)) = 0$; setting $c = 1$ and $d = 1/(k+2)$ then proves the lemma).

The assumption that every cycle of A contains an accepting state is occasionally made in the modern literature on Agafonov's Theorem, e.g. [4]. The reason for not making this assumption is that it is unnecessary for strongly connected automata

Lemma 15. *Let S be a strategy, and let F be finite subset of $\{0,1\}^*$. Let $R = F \setminus \{\mathbf{w} : \exists \mathbf{u} \in F. \mathbf{u} \prec \mathbf{w}\}$ be the set obtained from F by removing words $\mathbf{w}$ that already have a proper prefix in F. Define, for each positive integer n, the set $M_n = \{\mathbf{w} \in \{0,1\}^n : S(\mathbf{w}) \in F\}$. Then, $\mu_p(M_n) \leqslant \mu_p(R)$.*

Proof. Observe that

$$M_n = \bigcup_{\mathbf{u} \in R} \{\mathbf{w} : S(\mathbf{w}) \in F \wedge \mathbf{u} \preceq S(\mathbf{w})\}$$

and thus

$$\mu_p(M_n) = \sum_{\mathbf{u} \in R} \mu_p(\{\mathbf{w} : S(\mathbf{w}) \in F \wedge \mathbf{u} \preceq S(\mathbf{w})\}) \leq \sum_{\mathbf{u} \in R} \mu_p(\{\mathbf{w} : \mathbf{u} \preceq S(\mathbf{w})\})$$

Thus, if, for any word $\mathbf{u} = a_1 a_2 \cdots a_k$, the set $M_{\mathbf{u}} = \{\mathbf{w} \in \{0,1\}^n : \mathbf{u} \preceq S(\mathbf{w})\}$ satisfies $\mu_p(M_{\mathbf{u}}) \leqslant \mu_p(\mathbf{u})$, it follows that

$$\mu_p(M_n) \leq \sum_{\mathbf{u} \in R} \mu_p(\{\mathbf{w} : \mathbf{u} \preceq S(\mathbf{w})\}) \leq \sum_{\mathbf{u} \in R} \mu_p(\mathbf{u}) = \mu_p(R)$$

as desired. We thus proceed to prove $\mu_p(M_{\mathbf{u}}) \leqslant \mu_p(\mathbf{u})$ by induction on $k = |\mathbf{u}|$.

- Base case: $\mathbf{u} = a \in \{0,1\}$, so $\mu_p(\mathbf{u}) = \mu_p(a) = p(a)$. Let $\alpha = x_1 x_2 \ldots x_n$ be a word in $M_{\mathbf{u}}$ and let $x_f \in \{0,1\}$ be the first symbol selected by S when applied to α; as $\alpha \in M_{\mathbf{u}}$, we have $x_f = a$. Now, for each $b \in \{0,1\} \setminus \{x_f\}$, define $\bar{\alpha}_b = x_1 x_2 \cdots x_{f-1} b x_{f+1} \ldots x_n$, that is, $\bar{\alpha}_b$ is the word obtained from α by changing the fth symbol to b. Then, $\bar{\alpha}_b \notin M_{\mathbf{u}}$.

  We define the set $\bar{M}_{\mathbf{u}} = \{\bar{\alpha}_b : \alpha \in M_{\mathbf{u}}, b \in \{0,1\} \setminus \{a\}\}$. Observe that $\mu_p(\bar{\alpha}_b) = \mu_p(\alpha)p(b)/p(a)$, and hence:

$$\begin{aligned}\mu_p(\bar{M}_{\mathbf{u}}) &= \sum_{\alpha \in M_{\mathbf{u}}} \sum_{b \in \{0,1\} \setminus \{a\}} \mu_p(\alpha) \frac{p(b)}{p(a)} = \sum_{\alpha \in M_{\mathbf{u}}} \frac{\mu_p(\alpha)}{p(a)} \left(\sum_{b \in \{0,1\} \setminus \{a\}} p(b) \right) \\ &= \frac{1-p(a)}{p(a)} \sum_{\alpha \in M_{\mathbf{u}}} \mu_p(\alpha) = \frac{1-p(a)}{p(a)} \mu_p(M_{\mathbf{u}})\end{aligned}$$

Furthermore, as $\bar{\alpha}_b \notin M_{\mathbf{u}}$ for any $b \in \{0,1\} \setminus \{a\}$, we have $M_{\mathbf{u}} \cap \bar{M}_{\mathbf{u}} = \emptyset$, whence $\mu_p(M_{\mathbf{u}}) + \mu_p(\bar{M}_{\mathbf{u}}) \leq \mu_p(\{0,1\}^n) = 1$ and therefore $\mu_p(M_{\mathbf{u}}) \leq 1 - \mu_p(\bar{M}_{\mathbf{u}})$. Thus,

$$\mu_p(M_{\mathbf{u}}) \leq 1 - \mu_p(M_{\mathbf{u}})\frac{1-p(a)}{p(a)}$$

that is,

$$\mu_p(M_{\mathbf{u}}) \leq \frac{1}{1+\frac{1-p(a)}{p(a)}} = p(a) = \mu_p(\mathbf{u})$$

as desired.

- Inductive case: $\mathbf{u} = a_1 a_2 \ldots a_k a_{k+1}$ with $a_{k+1} = a$ for some $a \in \{0,1\}$. We have $M_{\mathbf{u}} = \{\alpha \in \{0,1\}^* : a_1 \cdots a_k a \preceq S(\alpha)\}$. Given $\alpha \in M_{\mathbf{u}}$, let for each $b \in \{0,1\} \setminus \{a\}$, $\bar{\alpha}_b$ be the word obtained from α by changing the $k+1$th symbol selected by S to b. Observe that $\neg(\bar{\alpha}_b \preceq \mathbf{u})$. Define $\bar{M}_{\mathbf{u}}$ to be the set $\{\bar{\alpha}_b : \alpha \in M, b \in \{0,1\} \setminus \{a\}\}$, and note that $M_{\mathbf{u}} \cap \bar{M}_{\mathbf{u}} = \emptyset$, and that $\mu_p(\bar{\alpha}_b) = \mu_p(\alpha)p(b)/p(a)$ and thus, as above, $\mu_p(\bar{M}_{\mathbf{u}}) = \mu_p(M_{\mathbf{u}})(1-p(a))/p(a)$.

  Let $N_{\mathbf{u}}$ be the set $\{\alpha \in \{0,1\}^* : a_1 \ldots a_k \preceq S(\alpha)\}$. Then $N_{\mathbf{u}}$ contains as subsets both $M_{\mathbf{u}}$ and $\bar{M}_{\mathbf{u}}$, whence $\mu_p(M_{\mathbf{u}}) + \mu_p(\bar{M}_{\mathbf{u}}) \leq \mu_p(N_{\mathbf{u}})$. The induction hypothesis furnishes that $\mu_p(N_{\mathbf{u}}) \leq \mu_p(a_1 a_2 \ldots a_k)$, and thus $\mu_p(M_{\mathbf{u}}) + \mu_p(\bar{M}_{\mathbf{u}}) \leq \mu_p(a_1 a_2 \ldots a_k)$. As $\mu_p(\bar{M}_{\mathbf{u}}) = \mu_p(M_{\mathbf{u}})(1-p(a))/p(a)$, we deduce that

$$\mu_p(M_{\mathbf{u}}) \leq \mu_p(a_1 a_2 \ldots a_k) - \mu_p(\bar{M}_{\mathbf{u}}) = \mu_p(a_1 a_2 \ldots a_k) - \mu_p(M_{\mathbf{u}})\frac{1-p(a)}{p(a)}$$

  and thus that

$$\mu_p(M_{\mathbf{u}}) \leq \frac{\mu_p(a_1 a_2 \ldots a_k)}{1+\frac{1-p(a)}{p(a)}} = \mu_p(a_1 a_2 \ldots a_k)p(a) = \mu_p(a_1 a_2 \ldots a_k a)$$

  as desired.

□

Lemma 16. *Let S be a strategy, $a \in \{0,1\}$, b, ϵ be real numbers with $0 < b \leq 1$ and $\epsilon > 0$, and define, for all positive integers n:*

$$\begin{aligned} H_n(b,\epsilon) &= \left\{\mathbf{w} \in \{0,1\}^n : |S(\mathbf{w})| > bn \wedge \left|p(a) - \frac{\#_1(a)S(\mathbf{w})}{|S(\mathbf{w})|}\right| \geq \epsilon\right\} \\ &= \bigcup_{bn<\ell\leq n} \left\{\mathbf{w} \in \{0,1\}^n : S(\mathbf{w}) \in \{0,1\}^\ell \wedge \left|p(a) - \frac{\#_1(a)S(\mathbf{w})}{\ell}\right| \geq \epsilon\right\} \end{aligned}$$

Then:

$$\lim_{n\to\infty} \mu_p(H_n(b,\epsilon)) = 0$$

Proof. Define

$$F_n(b,\epsilon) = \bigcup_{bn<\ell\leq n} \left\{ \mathbf{y} \in \{0,1\}^\ell : \left| p(a) - \frac{\#_1(a)\mathbf{y}}{\ell} \right| \geq \epsilon \right\}$$

Observe that $H_n(b,\epsilon) = \{\mathbf{w} \in \{0,1\}^n : S(\mathbf{w}) \in F_n(b,\epsilon)\}$. Let $R_n(b,\epsilon) \subseteq \{0,1\}^{\leq n}$ be the set obtained by removing from $F_n(b,\epsilon)$ all $\mathbf{w}$ such that there is $\mathbf{u} \in F_n(b,\epsilon)$ with $\mathbf{u} \prec \mathbf{w}$ (i.e., remove all words from $F_n(b,\epsilon)$ that already have a prefix in $F_n(b,\epsilon)$). Lemma Lemma 15 yields that $\mu_p(H_n(b,\epsilon)) \leq \mu_p(R_n(b,\epsilon))$, and thus that $\lim_{n\to\infty} \mu_p(R_n(b,\epsilon)) = 0$.

Consider the stochastic variable X_a that is 1 when a is picked from $\{0,1\}$ with probability $p(a)$, and 0 otherwise. Then, the mean of X_a is $p(a)$ and the variance of X_a is $p(a)(1-p(a))$. Now consider performing $\ell \geq 1$ independent Bernoulli trials drawn according to X_a. Define $q(1) = p(a)$, $q(0) = 1-p(a)$, and $q(1c) = p(a)q(c)$ and $q(0c) = (1-p(a))q(c)$ for $c \in \{0,1\}^+$, and consider the probability distribution $\bar{q} : \{0,1\}^\ell \longrightarrow [0;1]$ on $\{0,1\}^\ell$. Now, for any $\mathbf{v} \in \{0,1\}^\ell$, $\bar{q}(\mathbf{v})$ is the probability of obtaining $\mathbf{v}$ by performing ℓ repeated Bernoulli trials as above.

Define the stochastic variable $X_a^\ell = X_a + X_a + \cdots + X_a$ (ℓ times). Then, X^ℓ counts the number of occurrences of a by performing ℓ Bernoulli trials as above. By Chebyshev's inequality, X_a^ℓ satisfies:

$$\Pr\left(\left| p(a) - \frac{X_a^\ell}{\ell} \right| \geq \epsilon \right) \leq \frac{p(a)^2(1-p(a))^2}{\ell\epsilon^2} \tag{17}$$

Define the map $g : \{0,1\} \longrightarrow \{0,1\}$ by $g(a) = 1$ and $g(b) = 0$ for all $b \in \{0,1\} \setminus \{a\}$; g obviously extends homomorphically to a map $\tilde{g} : \{0,1\}^\ell \longrightarrow \{0,1\}^\ell$ by setting $\tilde{g}(c_1c_2\cdots c_\ell) = g(c_1)g(c_2)\cdots g(c_\ell)$.

Observe that, for any $\mathbf{y} \in \{0,1\}^\ell$, we have:

$$|p(a) - \#_1(1)\tilde{\mathbf{g}}(\mathbf{y})/\ell| \geq \epsilon \quad \text{iff} \quad |p(a) - \#_1(a)\mathbf{y}/\ell| \geq \epsilon \tag{18}$$

For any $\mathbf{u} \in \{0,1\}^\ell$,

$$\bar{q}(\mathbf{u}) = p(a)^{\#_1(1)\mathbf{u}}(1-p(a))^{\ell-\#_1(1)\mathbf{u}} = \sum_{\mathbf{y}\in\{0,1\}^\ell : \tilde{g}(\mathbf{y})=\mathbf{u}} \mu_p(\mathbf{y}) = \mu_p(\{\mathbf{y} \in \{0,1\}^\ell : \tilde{g}(\mathbf{y}) = \mathbf{u}\}))$$

Hence, for any event $\mathcal{U} \subseteq \{0,1\}^\ell$, we have:

$$\begin{aligned} \Pr(\mathcal{U}) &= \sum_{\mathbf{u}\in\mathcal{U}} \bar{q}(\mathbf{u}) = \sum_{\mathbf{u}\in\mathcal{U}} \mu_p(\{\mathbf{y} \in \{0,1\}^\ell : \tilde{g}(\mathbf{y}) = \mathbf{u}\})) \\ &= \mu_p\left(\left\{ \mathbf{y} \in \{0,1\}^\ell : \tilde{g}(\mathbf{y}) \in \mathcal{U} \right\} \right) \end{aligned} \tag{19}$$

The event $|p(a) - X_a^\ell/\ell| \geq \epsilon$ is shorthand for the set

$$\left\{ \mathbf{u} \in \{0,1\}^\ell : \left| p(a) - \frac{\sum_{j=1}^\ell u_j}{\ell} \right| \geq \epsilon \right\} = \left\{ \mathbf{u} \in \{0,1\}^\ell : \left| p(a) - \frac{\#_1(1)\mathbf{u}}{\ell} \right| \geq \epsilon \right\}$$

We thus obtain:

$$\begin{aligned}
\Pr\left(\left|p(a)-\frac{X_a^\ell}{\ell}\right|\geq\epsilon\right) &= \Pr\left(\left\{\mathbf{u}\in\{0,1\}^\ell : \left|p(a)-\frac{\#_1(1)\mathbf{u}}{\ell}\right|\geq\epsilon\right\}\right)\\
&= \mu_p\left(\left\{\mathbf{y}\in\{0,1\}^\ell : \left|p(a)-\frac{\#_1(1)\tilde{\mathbf{g}}(\mathbf{y})}{\ell}\right|\geq\epsilon\right\}\right) \quad \text{by (\textit{Equation} (19))}\\
&= \mu_p\left(\left\{\mathbf{y}\in\{0,1\}^\ell : \left|p(a)-\frac{\#_1(a)\mathbf{y}}{\ell}\right|\geq\epsilon\right\}\right) \quad \text{by (\textit{Equation} (18))}\\
&= \sum_{\mathbf{y}\in\{0,1\}^\ell}\mu_p\left(\left\{\mathbf{y} : \left|p(a)-\frac{\#_1(a)\mathbf{y}}{\ell}\right|\geq\epsilon\right\}\right) \qquad (20)
\end{aligned}$$

Observe that:

$$\begin{aligned}
\mu_p\left(R_n(b,\epsilon)\right) &= \mu_p\left(\bigcup_{bn<\ell\leq n}\left\{\mathbf{y}\in\{0,1\}^\ell\cap R_n(b,\epsilon) : \left|p(a)-\frac{\#_1(a)\mathbf{y}}{\ell}\right|\geq\epsilon\right\}\right)\\
&= \sum_{bn<\ell\leq n}\mu_p\left(\left\{\mathbf{y}\in\{0,1\}^\ell\cap R_n(b,\epsilon) : \left|p(a)-\frac{\#_1(a)\mathbf{y}}{\ell}\right|\geq\epsilon\right\}\right) \qquad (21)
\end{aligned}$$

But as $\mu_p(a_1\cdots a_\ell)\geq\mu_p(a_1\cdots a_\ell a_{\ell+1})$ for any $a_1,\ldots,a_\ell,a_{\ell+1}\in\{0,1\}$ and no element of $R_n(b,\epsilon)$ is a prefix of any other element, we have

$$\begin{aligned}
&\sum_{bn<\ell\leq n}\mu_p\left(\left\{\mathbf{y}\in\{0,1\}^\ell\cap R_n(b,\epsilon) : \left|p(a)-\frac{\#_1(a)\mathbf{y}}{\ell}\right|\geq\epsilon\right\}\right)\\
&\leq\ \mu_p\left(\left\{\mathbf{y}\in\{0,1\}^{\lfloor bn\rfloor} : \left|p(a)-\frac{\#_1(a)\mathbf{y}}{\lfloor bn\rfloor}\right|\geq\epsilon\right\}\right) \qquad (22)
\end{aligned}$$

We thus have:

$$\begin{aligned}
\mu_p(R_n(b,\epsilon) &\leq \mu_p\left(\left\{\mathbf{y}\in\{0,1\}^{\lfloor bn\rfloor} : \left|p(a)-\frac{\#_1(a)\mathbf{y}}{\ell}\right|\geq\epsilon\right\}\right) \quad \text{by (\textit{eq.} (21)) and (\textit{section} 4)}\\
&= \Pr\left(\left|p(a)-\frac{X_a^{\lfloor bn\rfloor}}{\lfloor bn\rfloor}\right|\geq\epsilon\right) \quad \text{by (\textit{Equation} (20))}\\
&\leq \frac{p(a)^2(1-p(a))^2}{\lfloor bn\rfloor\epsilon^2} \quad \text{by (\textit{Equation} (17))}
\end{aligned}$$

Thus, $\lim_{n\to\infty}\mu_p(R_n(b,\epsilon))=0$, as desired. □

Corollary 16.1. *Let b,ϵ be real numbers with $0<b\leq 1$ and $\epsilon>0$. Then,*

$$\lim_{n\to\infty}\mu_p(G_n(b,\epsilon))=0$$

Proof. By Lemma Lemma 16 with $S=A_q$, we obtain $\lim_{n\to\infty}\mu_p(G_n(b,\epsilon,q))=0$ and as $G_n(b,\epsilon)=\bigcup_{q\in Q}G_n(b,\epsilon,q)$, we have $\mu_p(G_n(b,\epsilon))\leq\sum_{q\in Q}\mu_p(G_n(b,\epsilon,q))$. As Q is finite, we hence obtain $\lim_{n\to\infty}\mu_p(G_n(b,\epsilon))=0$. □

Lemma 17. *There is a real number b with $0 < b \leq 1$ such that for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mu_p(D_n^p(b,\epsilon)) = 1.$$

Proof. Observe that, for all b with $0 < b \leq 1$:

$$\begin{aligned}\{0,1\}^n \setminus D_n^p(b,\epsilon)) &= \{\mathbf{w} \in \{0,1\}^n : \exists q \in Q.|A_q[\mathbf{w}]| \leq bn\} \\ &\cup \left\{\mathbf{w} \in \{0,1\}^n : \exists q \in Q.|A_q[\mathbf{w}]a| > bn \wedge \max_{a\in\{0,1\}} \left|\frac{\#_1(a)A_q[\mathbf{w}]}{|A_q[\mathbf{w}]|} - p(a)\right| \geq \epsilon\right\} \\ &= \left(\bigcup_{q\in Q} E_n(b,q)\right) \cup \left(\bigcup_{q\in Q} G_n(b,\epsilon,q)\right)\end{aligned}$$

and thus,

$$\begin{aligned}\mu_p(\{0,1\}^n \setminus D_n^p(b,\epsilon)) &\leq \mu_p\left(\bigcup_{q\in Q} E_n(b,q)\right) + \mu_p\left(\bigcup_{q\in Q} G_n(b,\epsilon,q)\right) \\ &= \mu_p(G_n(b,\epsilon)) + \mu_p(E_n(b))\end{aligned}$$

Choose, by Lemma Lemma 14 real numbers c, d such that $\lim_{n\to\infty} \mu_p(E_n(\frac{c-\epsilon}{d})) = 0$, and set $b = (c-\epsilon)/d$. By Corollary Corollary 16.1, we obtain $\lim_{n\to\infty} G_n(b,\epsilon) = 0$, and thus $\lim_{n\to\infty} \mu_p(\{0,1\}^n \setminus D_n^p(b,\epsilon)) = 0$. The result now follows by $\mu_p(D_n^p(b,\epsilon))) = 1 - \mu_p(\{0,1\}^n \setminus D_n^p(b,\epsilon))$. □

Theorem 18. *Let α be a p-block-distributed right-infinite sequence, and A a strongly connected DFA. Then the sequence $\beta = A[\alpha]$ is p-distributed.*

Proof. By Lemma Lemma 13 it suffices to show that, for all $a \in \{0,1\}$, the limiting frequency of a in $A[\alpha]$ exists and is equal to $p(a)$.

Consider the sequence $(\beta_{(n,r)})$ of blocks of $A[\mathbf{x}]$ corresponding to the sequence of blocks $(\alpha_{(n,r)})$, that is $\beta_{(n,r)}$ is the sequence of symbols picked out from block $\alpha_{(n,r)}$ when A is applied to α; note that each $\beta_{[n,r]}$ has length between 0 and n.

For each positive integer m, define $L_m = \sum_{i=1}^m |\beta_{[n,i]}|$, and for each $a \in \{0,1\}$, write $\rho_a^m = \frac{\sum_{i=1}^m \#_1(a)\beta_{(n,i)}}{L}$. Observe that, to prove the theorem, it suffices to show that, for any real number ϵ with $0 < \epsilon < 1$ and sufficiently large m, that $|\rho_a - p(a)| < \epsilon$.

Furthermore, set $I_m = \{i \leqslant m : \alpha_{(n,i)} \notin D_n^p(b, \frac{\epsilon}{2})\}$, and set $\ell_m = \sum_{i\in I_m} |\beta_{(n,i)}|$.

Now, define θ_a^m by:

$$\theta_a^m = \frac{\sum_{i\in\{1,\ldots,m\}\setminus I_m} \#_1(a)\beta_{(n,i)}}{\sum_{i\in\{1,\ldots,m\}\setminus I_m} |\mathbf{y}_{(n,i)}|} = \frac{\sum_{i\in\{1,\ldots,m\}\setminus I_m} \#_1(a)\beta_{(n,i)}}{L_m - \ell_m}$$

That is, θ_a^m is the frequency of occurrences of as when the blocks $\beta_{(n,i)]}$ picked out from blocks $\alpha_{(n,r)} \in D_n^p(b, \frac{\epsilon}{2})$ are concatenated. Observe that, by definition of D_n^p, we have $|\theta_a^m - p(a)| < \frac{\epsilon}{2}$.

We have:

$$\begin{aligned}
\rho_a^m - \theta_a^m &= \frac{\sum_{i=1}^m \#_1(a)\beta_{(n,i)}}{L_m} - \frac{\sum_{i\in\{1,\ldots,m\}\setminus I} \#_1(a)\beta_{(n,i)}}{L_m - \ell_m} \\
&= \left(\frac{\sum_{i\in I_m} \#_1(a)\beta_{(n,i)}}{L_m} + \frac{\sum_{i\in\{1,\ldots,m\}\setminus I_m} \#_1(a)\beta_{(n,i)}}{L_m}\right) - \frac{\sum_{i\in\{1,\ldots,m\}\setminus I} \#_1(a)\beta_{(n,i)}}{L_m - \ell_m} \\
&\overset{(\dagger)}{=} \frac{\sum_{i\in\{1,\ldots,m\}\setminus I_m} \#_1(a)\beta_{(n,i)}}{L_m} - \frac{\sum_{i\in\{1,\ldots,m\}\setminus I_m} \#_1(a)\beta_{(n,i)}}{L_m - \ell_m} + \frac{\sum_{i\in I_m} \#_1(a)\beta_{(n,i)}}{L_m} \\
&\leq \frac{\sum_{i\in I_m} \#_1(a)\beta_{(n,i)}}{L_m} \leq \frac{\sum_{i\in I_m} |\beta_{(n,i)}|}{L_m} = \frac{\ell_m}{L_m} \qquad (23)
\end{aligned}$$

where the penultimate inequalities in the last line above follows because $L_m \geq L_m - \ell_m$ implies $\frac{\sum_{i\in\{1,\ldots,m\}\setminus I} \#_1(a)\beta_{(n,i)}}{L} - \frac{\sum_{i\in\{1,\ldots,m\}\setminus I} \#_1(a)\beta_{[n,i]}}{L-\ell} \leq 0$, and the final inequality follows because $\sum_{i\in I} \#_1(a)\beta_{[n,i]} \leq \sum_{i\in I} |\beta_{[n,i]}| = \ell_m$.

By basic algebra, we have:

$$\frac{\sum_{i\in\{1,\ldots,m\}\setminus I_m} \#_1(a)\beta_{(n,i)}}{L_m} - \frac{\sum_{i\in\{1,\ldots,m\}\setminus I_m} \#_1(a)\beta_{(n,i)}}{L_m - \ell_m} = \frac{-\ell_m \sum_{i\in\{1,\ldots,m\}\setminus I} \#_1(a)\beta_{(n,i)}}{L_m(L_m - \ell_m)}$$

and as

$$\sum_{i\in\{1,\ldots,m\}\setminus I_m} \#_1(a)\beta_{(n,i)} \leq \sum_{i\in\{1,\ldots,m\}\setminus I_m} \leq \sum_{i\in\{1,\ldots,m\}\setminus I_m} |\beta_{(n,i)}| \leq L_m - \ell_m$$

we conclude that

$$\frac{-\ell \sum_{i\in\{1,\ldots,m\}\setminus I} \#_1(a)\beta_{[n,i]}}{L(L-\ell)} \geq -\frac{\ell_m}{L_m}$$

and thus by (†) that

$$\rho_a^m - \theta_a^m + \frac{\sum_{i\in I_m} \#_1(a)\beta_{(n,i)}}{L_m} \geq -\frac{\ell_m}{L_m}$$

whence $-\ell_m/L_m \leq \rho_a - \theta_a$, which combined with (Equation (23)) yields $|\rho_a - \theta_a| \leq \ell/L$.

By Lemma 17 pick a b such that such that for all $\epsilon > 0$, we have $\lim_{n\to\infty} \mu_p(D_n^p(b,\epsilon)) = 1$. Choose $\delta > 0$ with $\delta < \frac{b\epsilon}{8}$. Pick $n \in \mathbf{N}$ such that $\mu_p(D_n^p(b,\epsilon)) > 1 - \delta$. Now, pick $\alpha < \frac{b\epsilon}{8}$. Because α is p-block-distributed, there exists $M \in \mathbf{N}$ such that for all $k \geq M$ and all $\mathbf{G} \subseteq \{0,1\}^n$, the prefix $\alpha_{\leq kn}$ of α of length kn satisfies:

$$\left|\frac{|\{i \leq k : \alpha_{(n,i)} \in \mathbf{G}\}|}{k} - \mu_p(\mathbf{G})\right| < \alpha$$

In the particular case $\mathbf{G} = D_n^p(b, \epsilon/2)$, we thus have:

$$\left|\frac{|\{i \leq k : \alpha_{(n,i)} \in D_n^p\left(b, \frac{\epsilon}{2}\right)\}|}{k} - \mu_p\left(D_n^p\left(b, \frac{\epsilon}{2}\right)\right)\right| < \alpha$$

and thus

$$\begin{aligned}
1 - \delta - \frac{|\{i \leq k : \alpha_{(n,i)} \in D_n^p(b, \frac{\epsilon}{2})\}|}{k} &\leq \mu_p\left(D_n^p\left(b, \frac{\epsilon}{2}\right)\right) - \frac{|\{i \leq k : \alpha_{(n,i)} \in D_n^p(b, \frac{\epsilon}{2})\}|}{k} \\
&< \alpha
\end{aligned}$$

and thus

$$\left|\left\{i \leq k : \alpha_{(n,i)} \in D_n^p\left(b, \frac{\epsilon}{2}\right)\right\}\right| > k(1-\delta-\alpha) \tag{24}$$

By definition of $D_n^p(b, \frac{\epsilon}{2}))$, every $\alpha_{(n,i)} \in D_n^p(b, \frac{\epsilon}{2}))$ satisfies $|A[\alpha_{(\mathbf{n},\mathbf{i})}]| > bn$, and we thus have, whence

$$L_m = \sum_{i=1}^{m} |\mathbf{y}_{(n,i)}| = \sum_{i=1}^{m} |A[\alpha_{(\mathbf{n},\mathbf{i})}]| \geq \left|\left\{i \leq m : \alpha_{(n,i)} \in D_n^p\left(b, \frac{\epsilon}{2}\right)\right\}\right| bn > m(1-\delta-\alpha)bn$$

Furthermore, by definition of I_m and (Equation (24)),

$$\begin{aligned} |I_m| &= \left|\left\{i \leqslant m : \alpha_{(n,i)} \notin D_n^p\left(b, \frac{\epsilon}{2}\right)\right\}\right| = m - \left|\left\{i \leq m : \alpha_{(n,i)} \in D_n^p\left(b, \frac{\epsilon}{2}\right)\right\}\right| \\ &< m - m(1-\delta-\alpha) = m(\delta+\alpha) \end{aligned}$$

But then,

$$\ell_m = \sum_{i \in I_m} |\mathbf{y}_{(i,n)}| \leq |I_m| n < mn(\delta+\alpha)$$

and thus:

$$\frac{\ell_m}{L_m} < \frac{mn(\delta+\alpha)}{m(1-\delta-\alpha)bn} = \frac{\delta+\alpha}{b(1-\delta-\alpha)} < \frac{\frac{b\epsilon}{8}+\frac{b\epsilon}{8}}{b\left(1-\frac{b\epsilon}{8}-\frac{b\epsilon}{8}\right)} < \frac{\frac{\epsilon}{8}}{1-\frac{1}{4}} < \frac{\epsilon}{2}$$

where we have used that $b\epsilon < 1$ in the penultimate inequality.

We now finally have

$$|\rho_a - p(a)| \leq |\rho_a^m - \theta_a^m| + |\theta_a - p(a)| < \frac{\ell_m}{L_m} + \frac{\epsilon}{2} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

concluding the proof. □

References

[1] V. N. Agafonov. Normal sequences and and finite automata. *Problemy Kibernetiki*, 20:123–129, 1968. In Russian.

[2] V. N. Agafonov. Normal sequences and finite automata. *Sov. Math., Dokl.*, 9:324–325, 1968. Originally published in Russian (vol. 179:2, p. 255-266).

[3] V. N. Agafonov. Normal sequences and finite automata. *Dokl. Akad. Nauk SSSR*, 179(2):255–256, 1968.

[4] V. Becher and . A. Heiber. Normal numbers and finite automata. *Theoretical Computer Science*, 477:109–116, 2013.

[5] E. Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rend. Circ. Matem. Palermo*, 27:247–271, 1909.

[6] A. Church. On the concept of a random sequence. *Bulletin of the American Mathematical Society*, 46(2):130–135, 1940.

[7] A. H. Copeland. Admissible numbers in the theory of probability. *American Journal of Mathematics*, 50(4):535–552, 1928.

[8] A. H. Copeland. Point set theory applied to the random selection of the digits of an admissible number. *American Journal of Mathematics*, 58(1):181–192, 1936.

[9] E. Kamke. Über neuere begründungen der wahrscheinlichkeitsrechnung. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 42:14–27, 1933.

[10] D. W. Loveland. The kleene hierarchy classification of recursively random sequences. *Transactions of the American Mathematical Society*, 125(3):497–510, 1966.

[11] . . . , 179(2):255–256, 1968.

[12] . . . , 57:3–84, 1960.

[13] . and . . . , 21(4):501–514, 1957.

[14] . . . , 6(2):232–234, 1961.

[15] L. Postnikova. On the connection between the concepts of collectives of Mises-Church and normal Bernoulli sequences of symbols. *Theory of Probability & Its Applications*, 6(2):211–213, 1961. translation of [14] by Eizo Nishiura.

[16] H. Reichenbach. Axiomatik der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 34(1):568–619, 1932.

[17] H. Reichenbach. Les fondements logiques du calcul des probabilités. In *Annales de l'institut Henri Poincaré*, volume 7, pages 267–348, 1937.

[18] C. Schnorr and H. Stimm. Endliche Automaten und Zufallsfolgen. *Acta Inf.*, 1:345–359, 1972.

[19] R. Serfozo. *Basics of Applied Stochastic Processes*. Probability and Its Applications. Springer-Verlag, 2009.

[20] P. Shields. *The Theory of Bernoulli Shifts*. Univ. Chicago Press, 1973.

[21] E. Tornier. Wahrscheinlichkeitsrechnung und zahlentheorie. erste mitteilung. *Journal für die reine und angewandte Mathematik*, 1929(160):177–198, 1929.

[22] R. Von Mises. Grundlagen der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5(191):52–99, 1919.

[23] R. von Mises. *Wahrscheinlichkeit Statistik und Wahrheit*. Springer, 1936.