SKOLEM MEETS BATEMAN-HORN

FLORIAN LUCA, JAMES MAYNARD, ARMAND NOUBISSIE, JOËL OUAKNINE,
AND JAMES WORRELL

Abstract. The Skolem Problem asks to determine whether a given integer linear recurrence sequence has a zero term. This problem arises across a wide range of topics in computer science, including loop termination, formal languages, automata theory, and control theory, amongst many others. Decidability of the Skolem Problem is notoriously open. The state of the art is a decision procedure for recurrences of order at most 4: an advance achieved some 40 years ago, based on Baker's theorem on linear forms in logarithms of algebraic numbers.

A new approach to the Skolem Problem was recently initiated in [LOW21, LOW22] via the notion of a Universal Skolem Set—a set $\mathcal{S}$ of positive integers such that it is decidable whether a given non-degenerate linear recurrence sequence has a zero in $\mathcal{S}$. Clearly, proving decidability of the Skolem Problem is equivalent to showing that $\mathbb{N}$ itself is a Universal Skolem Set. The main contribution of the present paper is to construct a Universal Skolem Set that has lower density at least 0.29. We show moreover that this set has density one subject to the Bateman-Horn conjecture. The latter is a central unifying hypothesis concerning the frequency of prime numbers among the values of systems of polynomials.

1. Introduction

An (integer) linear recurrence sequence (LRS) $\langle u_n \rangle_{n=0}^{\infty}$ is a sequence of integers satisfying a recurrence of the form

$$u_{n+k} = a_0 u_{n+k-1} + \cdots + a_{k-1} u_n \qquad (n \in \mathbb{N}), \tag{1}$$

where the coefficients $a_0, \ldots, a_{k-1}$ are integers. The celebrated theorem of Skolem, Mahler, and Lech [Lec53, Mah35, Sko34] states that the set $\{n \in \mathbb{N} : u_n = 0\}$ of zero terms of an LRS is the union of a finite set and finitely many arithmetic progressions. This result can be refined using the notion of *non-degeneracy* of an LRS. An LRS is non-degenerate if in its minimal recurrence no quotient of distinct characteristic roots is a root of unity. A given LRS can be effectively decomposed as the merge of finitely many non-degenerate sequences, some of which may be identically zero. The core of the Skolem-Mahler-Lech Theorem is the fact that a non-degenerate LRS that is not identically zero has finitely many zero terms. Unfortunately, all

School of Mathematics, University of Witwatersrand, Johannesburg, South Africa
Department of Mathematics, University of Oxford, Oxford, UK
Max Planck Institute for Software Systems, Saarland Informatics Campus, Saarbrücken, Germany
Max Planck Institute for Software Systems, Saarland Informatics Campus, Saarbrücken, Germany
Department of Computer Science, University of Oxford, Oxford, UK
E-mail addresses: Florian.Luca@wits.ac.za, james.alexander.maynard@gmail.com, noubissie@mpi-sws.org, joel@mpi-sws.org, jbw@cs.ox.ac.uk.

known proofs are ineffective—it is not known how to compute the finite set of zeros of a given non-degenerate linear recurrence sequence; equivalently, it is not known how to decide whether an arbitrary given LRS has a zero.

The problem of determining whether a given LRS has a zero is known as the Skolem Problem. This has been memorably characterised by Tao [Tao08] as *"the halting problem for linear automata"*. In fact, the Skolem Problem has been recognised as a fundamental decision problem in a number of different areas of theoretical computer science, including loop termination [OW15], weighted automata and formal power series (see [BR10, Section 6.4], [RS94, Section 4.2], and [SS78, Section III.8]), matrix semigroups [BPS21], stochastic systems and probabilistic programs [AAGT15, BJK20], and control theory [BT00, Section 3]. However, progress towards determining decidability of the problem has been limited. The state of the art is that decidability is known for recurrences of order[1] at most 4 [MST84, Ver85]—an advance made some 40 years ago, based on Baker's celebrated theorem on linear forms in logarithms of algebraic numbers.

Recently [BLN$^+$22] gave a procedure to decide the Skolem Problem for the class of simple LRS (those with simple characteristic roots) of any order, subject to two open conjectures about the exponential function, namely the p-adic Schanuel Conjecture and the exponential local-global principle. The present paper follows a different approach to [BLN$^+$22], via the notion of *Universal Skolem Set.* This is an infinite set $\mathcal{S} \subseteq \mathbb{N}$ for which there is an effective procedure that, given a non-degenerate LRS $\langle u_n \rangle_{n=0}^{\infty}$, outputs the finite set $\{n \in \mathcal{S} : u_n = 0\}$. Evidently, establishing decidability of the Skolem Problem is equivalent to showing that $\mathbb{N}$ is a Universal Skolem Set. Towards this objective, and noting that the class of Universal Skolem Sets is closed under various operations such as finite shifts and finite unions, it is natural to examine diverse means by which to construct Universal Skolem Sets, and particularly such sets of high density.[2] Pursuing this line of research leads in the present paper to new connections between the Skolem Problem and classical questions on the distribution of prime numbers.

The paper [LOW21] introduced the notion of Universal Skolem Set (the terminology is inspired by the notion of Universal Hilbert Set [Bil96]) and exhibited an explicit example of such a set that had density zero. Subsequently [LOW22] produced a set $\mathcal{S}_0 \subseteq \mathbb{N}$ of positive lower density and an effective procedure that, given a non-degenerate *simple* LRS $\langle u_n \rangle_{n=0}^{\infty}$, computes its set of zeros $\{n \in \mathcal{S}_0 : u_n = 0\}$. The present paper further develops ideas introduced in [LOW22] and contains two significant advances. First, we modify the construction of [LOW22] to yield a Universal Skolem Set $\mathcal{S}$ of positive lower density, i.e., such that one can compute the set of zeros $\{n \in \mathcal{S} : u_n = 0\}$ for *any* non-degenerate LRS, not just the simple ones. In fact we give an explicit upper bound for the largest zero in $\mathcal{S}$ of a given LRS. The second main contribution is to show that $\mathcal{S}$ has lower density at least 0.29 unconditionally and density one subject to the Bateman-Horn conjecture in number theory [BH62]. The latter is a central unifying hypothesis concerning the frequency of prime numbers among the values of a system of polynomials. This

[1] The *order* of an LRS is the smallest value of k for which the LRS satisfies a recurrence of the form (1).

[2] The *density* of an infinite set $\mathcal{S}$ of positive integer is the limit (if it exists) of the proportion of elements of $\mathcal{S}$ among all integers from 1 to n as n tends to infinity. The *lower density* of $\mathcal{S}$ is defined analogously, substituting the limit inferior to the limit.

conjecture generalises many classical results and conjectures on the distribution of primes, including Hardy and Littlewood's twin primes conjecture and Schinzel's Hypothesis H [AZFG20, Bai02, Lan96]. We apply the Bateman-Horn conjecture to obtain, given $a_1, a_2, b_1, b_2 \in \mathbb{Z}$, an upper bound on the density of $n \in \mathbb{N}$ such that $a_1 n + b_1$ and $a_2 n + b_2$ are simultaneously prime.

A key ingredient of the present paper are deep results of Schlickewei and Schmidt [SS00] and of Amoroso and Viada [AV09] that yield explicit bounds on the number of solutions of certain polynomial-exponential Diophantine equations. Indeed, it is striking that while there is no known method to elicit the zero set of a given non-degenerate LRS, thanks to the above mentioned results there are fully explicit upper bounds (depending only on the order of the recurrence) on the cardinality of its zero set. Such bounds do not suffice to solve the Skolem Problem, which would require effective bounds on the *magnitude* of the zeros of an LRS. The main idea of our approach is to leverage explicit upper bounds on the number of zeros of polynomial-exponential equations to obtain bounds on the magnitude of the zeros of LRS. Specifically, our Universal Skolem Set $\mathcal{S}$ consists of positive integers n that admit sufficiently many representations of the form $n = Pq + a$, with P, q prime and q, a logarithmic in n. Given an LRS $\langle u_n \rangle_{n=0}^{\infty}$, we associate with the equation $u_n = 0$ a *companion equation* such that each representation of n yields a solution of the companion equation. We then use upper bounds on the number of solutions of the companion equation to derive upper bounds on the magnitude of n. In general, we believe that such a transfer principle is a promising direction to make progress on the Skolem Problem.

In terms of proof techniques, a major difference between the present paper and [LOW22] is that the latter used an existing upper bound of [SS00] on the number of solutions of a certain class of exponential Diophantine equations. To handle the case of non-simple LRS it appears that one cannot use existing results "off the shelf" and must instead adapt the techniques of [AV09, ESS02, SS00] to our setting. This is the subject of Section 3, while Section 4 analyses the density of the set $\mathcal{S}$.

2. Background

In this section we briefly summarise some notions and results in number theory that we will need.

2.1. **Number fields.** Let $\mathbb{K}$ be a finite Galois extension of $\mathbb{Q}$. The ring of algebraic integers in $\mathbb{K}$ is denoted $\mathcal{O}_{\mathbb{K}}$. We denote by $\mathrm{Gal}(\mathbb{K}/\mathbb{Q})$ the group of automorphisms of $\mathbb{K}$. The *norm* of $\alpha \in \mathbb{K}$ is defined by

$$N_{\mathbb{K}/\mathbb{Q}}(\alpha) = \prod_{\sigma \in \mathrm{Gal}(\mathbb{K}/\mathbb{Q})} \sigma(\alpha)\,.$$

The *norm* $N_{\mathbb{K}/\mathbb{Q}}(\alpha)$ is rational for all $\alpha \in \mathbb{K}$ and $N_{\mathbb{K}/\mathbb{Q}}(\alpha)$ is an integer if $\alpha \in \mathcal{O}_{\mathbb{K}}$. Clearly we have $|N(\alpha)| < H^{d_{\mathbb{K}}}$, where $d_{\mathbb{K}}$ is the degree of $\mathbb{K}$ and

$$H := \max_{\sigma \in \mathrm{Gal}(\mathbb{K}/\mathbb{Q})} |\sigma(\alpha)|$$

is the *house* of α. Furthermore, given a rational prime $p \in \mathbb{Z}$ and a prime ideal factor $\mathfrak{p}$ of p in $\mathcal{O}_{\mathbb{K}}$, we have $p \mid N_{\mathbb{K}/\mathbb{Q}}(\alpha)$ for all $\alpha \in \mathfrak{p}$.

We say that $\alpha, \beta \in \mathbb{K}$ are *multiplicatively dependent* if there exist integers k, ℓ, not both zero, such that $\alpha^k = \beta^\ell$. Observe that if $\alpha \in \mathbb{K}$ is not a root of unity then given $\sigma \in \mathrm{Gal}(\mathbb{K}/\mathbb{Q})$, every multiplicative relation $\alpha^k = \sigma(\alpha)^\ell$ is such that $k = \pm\ell$. Indeed, repeatedly applying σ to this relation we deduce that $\alpha^{k^d} = (\sigma^d(\alpha))^{\ell^d}$ for all $d \geq 1$. In particular, choosing d to be the order of σ we get that $\alpha^{k^d} = \alpha^{\ell^d}$ and hence $k = \pm\ell$.

2.2. **Heights.** For a positive integer n, we denote by $h : \mathbb{A}^n(\overline{\mathbb{Q}}) \to [0, \infty)$ the *absolute logarithmic Weil height* on n-dimensional affine space. We refer to [BG06, Chapter 1] for the formal definition. Here we will need only the following properties (see [BG06, Chapter 1] and [Vou96, Corollary 2]).

Theorem 1. *Let $\alpha, \beta, \alpha_1, \ldots, \alpha_s$ be non-zero algebraic numbers for some $s \geq 2$. Then*

(1) $h(\alpha) = \log H$ *if α is an algebraic integer with house H;*
(2) $h(\alpha) = 0$ *if α is a root of unity and if α has degree $d \geq 2$ and is not a root of unity then $h(\alpha) > \frac{2}{d(\log(3d))^2}$;*
(3) $h(\alpha^m) = |m|h(\alpha)$ *for all $n \in \mathbb{Z}$.*
(4) $h(\alpha\beta) \leq h(\alpha) + h(\beta)$*;*
(5) $h(\alpha_1 + \cdots + \alpha_s) \leq h(\alpha_1) + \cdots + h(\alpha_s) + \log s$*;*
(6) $h(\alpha_1, \ldots, \alpha_s) \geq h(\alpha_i/\alpha_j)$ *for $1 \leq i < j \leq s$.*

The following elementary inequality will be useful in relation to calculations involving heights.

Proposition 2. *For all $c, X > 1$, if $\sqrt{\log X} < c \log\log X$ then $X < \exp((4c\log(2c))^2)$.*

2.3. **Linear equations in elements of multiplicative groups.** We now present two results concerning solutions of linear equations with variables in multiplicative groups. These results will play a key role in our construction of a Universal Skolem Set. Throughout this section $\mathbb{K}$ denotes a number field.

The first result that we recall is due to Amoroso and Viada [AV09, Theorem 6.1]. Here, given $n \geq 0$ and elements $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ of $\mathbb{K}^n$, we write $\boldsymbol{x} * \boldsymbol{y} := (x_1y_1, \ldots, x_ny_n)$.

Theorem 3. *Let Γ be a finitely generated subgroup of $\mathbb{K}^n$ of rank r and let $\varepsilon := (8n)^{-6n^3}$. Then the set*

$$\left\{\boldsymbol{x} * \boldsymbol{y} : \boldsymbol{x} \in \Gamma,\, \boldsymbol{y} \in \mathbb{K}^n,\, \boldsymbol{x}^\top \boldsymbol{y} = 1, \text{ and } h(\boldsymbol{y}) \leq \varepsilon(1 + h(\boldsymbol{x}))\right\}$$

is contained in a union of at most $(8n)^{6n^2(n+r)}$ proper linear subspaces of $\mathbb{K}^n$.

The second result, due to Schlickewei and Schmidt [SS00], concerns equations of the form

$$\sum_{i=1}^{s} P_i(\boldsymbol{x})\boldsymbol{\alpha}_i^{\boldsymbol{x}} = 0 \tag{2}$$

in variables $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{Z}^n$, where $P_1, \ldots, P_s \in \mathbb{K}[\boldsymbol{x}]$, and $\boldsymbol{\alpha}_i^{\boldsymbol{x}} = \alpha_{i1}^{x_1} \cdots \alpha_{in}^{x_n}$ with $\alpha_{ij} \in \mathbb{K}^*$ for all i, j. We say that a given solution to Equation (2) is *non-degenerate* if no proper sub-sum vanishes. The following upper bound on the number of non-degenerate solutions appears as [SS00, Theorem 1]:

Theorem 4. *Let δ_i be the total degree of polynomial P_i for $i \in \{1, \ldots, s\}$. Put $A = \sum_{i=1}^{s} \binom{n+\delta_i}{n}$ and $B = \max\{n, A\}$. Suppose that there is no non-zero $\boldsymbol{x} \in \mathbb{Z}^n$ such that $\boldsymbol{\alpha}_i^{\boldsymbol{x}} = \boldsymbol{\alpha}_j^{\boldsymbol{x}}$ for all $i, j \in \{1, \ldots, s\}$. Then Equation (2) has at most $2^{35B^3} d^{6B^2}$ non-degenerate solutions, where d is the degree of the number field $\mathbb{K}$.*

2.4. **Mean Value of Multiplicative Functions.** Let f be a real-valued function with domain the set of positive integers. We say that f is *multiplicative* if $f(1) = 1$ and $f(mn) = f(m)f(n)$ for all coprime m and n. The following result concerning the mean value of a multiplicative function appears as [Ten95, Section I.3,Theorem 11].

Theorem 5. *Let f be a multiplicative function with values in $[0, 1]$. Write*

$$M_f := \prod_{p \text{ prime}} (1 - p^{-1}) \sum_{\nu=0}^{\infty} f(p^\nu) p^{-\nu} ,$$

where the infinite product is considered to be zero when it diverges. Then, for Y tending to infinity, we have

$$\sum_{n \leq Y} f(n) = Y(M_f + o(1)) .$$

2.5. **Distribution of primes.** Consider linear forms $f_1(t) := a_1 t + b_1$ and $f_2(t) = a_2 t + b_2$ for integers a_1, a_2, b_1, b_2, with $a_1, a_2 > 0$. We say that $f := f_1 f_2 \in \mathbb{Z}[x]$ is *admissible* if it does not vanish identically modulo any prime. Note that if f is not admissible then f_1 and f_2 are simultaneously prime only at most twice. For a prime p, let $\omega_f(p)$ denote the number of $x \in \mathbb{F}_p$ such that $f(x) = 0$. We have the following special case of the Bateman-Horn conjecture:

Conjecture 6 (Bateman-Horn). *Let f_1, f_2 be a pair of linear forms such that $f = f_1 f_2$ is admissible. Then*

(3)

$$\#\{x \leq X : f_1(x), f_2(x) \text{ prime}\} \sim \frac{C_f X}{(\log X)^2}, \quad \text{where } C_f := \prod_{p \text{ prime}} \frac{p(p - \omega_f(p))}{(p-1)^2} .$$

In particular, the infinite product that defines C_f converges.

The general formulation of the Bateman-Horn conjecture concerns the set of positive integers over which a family $f_1, \ldots, f_k$ of polynomials is simultaneously prime. The version above is the special case that $k = 2$ and the f_i have degree one.

The only case of the Bateman-Horn conjecture that has been proven is the prime number theorem for arithmetic progressions (the case $k = 1$ and f_1 has degree one). Notably the Hardy-Littlewood twin prime conjecture (the case $k = 2$, $f_1(x) = x$, and $f_2(x) = x + 2$) remains open. Bateman and Horn [BH62] show an upper bound that resembles (3). Specifically, they apply the Brun sieve to establish the following bound for the value $\kappa := 8$:

$$\#\{x \leq X : f_1(x), f_2(x) \text{ prime}\} \leq \frac{\kappa C_f X}{(\log X)^2} . \tag{4}$$

Later, Wu used the large sieve to show that Inequality (4) holds for $\kappa := 3.418$ (see [Ten95, Section I.4.6]).

If we assume that $\Delta := |a_1 a_2(a_1 b_2 - a_2 b_1)|$ is non-zero then (4) implies that

$$\#\{x \le X : f_1(x), f_2(x) \text{ prime}\} \ll \frac{\Delta}{\varphi(\Delta)} \frac{X}{(\log X)^2}\,, \tag{5}$$

where $\ll$ is the Vinogradov symbol ($f(x) \ll g(x)$ iff there is a constant M such that, for all x sufficiently large, $|f(x)| \le M g(x)$), φ is the Euler totient function, and the implied constant is absolute (see, e.g., [HR74, Chapter 2.6, Theorem 2.3]).

3. A Universal Skolem Set

For $x > 1$ and a positive integer $k \ge 1$, we inductively define the iterated logarithm function $\log_k x$ as follows: $\log_1 x := \log x$, and for $k \ge 2$ we set $\log_k x := \max\{1, \log_{k-1}(\log x)\}$. Thus, for x sufficiently large, $\log_k x$ is the k-fold iterate of log applied to x. We omit the subscript when $k = 1$.

Fix a positive integer parameter X. We define disjoint intervals

$$A(X) := \left[\log_2 X, \sqrt{\log X}\right] \quad \text{and} \quad B(X) := \left[\frac{\log X}{\sqrt{\log_3 X}}, \frac{2\log X}{\sqrt{\log_3 X}}\right].$$

We further define a *representation* of an integer $n \in [X, 2X]$ to be a triple (q, P, a) such that $q \in A(X)$, $a \in B(X)$, P and q are prime, and $n = Pq + a$. We say that two representations $n = Pq + a$ and $n = P'q' + a'$ are *correlated* if

$$q \ne q',\ a \ne a' \quad \text{and} \quad |(a + \eta q) - (a' + \eta q')| < \sqrt{\log X}$$

for some $\eta \in \{\pm 1\}$.

We denote by $r(n)$ the number of representations of n. Finally we put

$$\mathcal{S}(X) := \{n \in [X, 2X] \ :\ r(n) > \log_4 X \text{ and no two representations of } n \text{ are correlated}\}$$

and we define

$$\mathcal{S} := \bigcup_{k \ge 10} \mathcal{S}(2^k)\,.$$

The following result shows that $\mathcal{S}$ is a Universal Skolem Set and furthermore gives an explicit upper bound on the largest element of $\mathcal{S}$ that is a zero of a given non-degenerate LRS.

Theorem 7. *Let $\boldsymbol{u} = \langle u_n \rangle_{n=0}^{\infty}$ be a non-degenerate LRS of order $k \ge 2$ given by*

$$u_{n+k} = a_0 u_{n+k-1} + \cdots + a_{k-1} u_n$$

for $n \ge 0$, with given initial terms $u_0, \ldots, u_{k-1}$ not all zero. If $u_n = 0$ and $n \in \mathcal{S}$, then

$$n < \max\{\exp_3(A^2), \exp_5(10^{10} k^6)\}, \quad \textit{where } A := \max\{10, |u_i|, |a_i| : 0 \le i \le k-1\}.$$

The rest of this section is devoted to the proof of Theorem 7, which is divided into 5 steps. Our goal is to show that for all X, if $u_n = 0$ for some $n \in \mathcal{S}(X)$ then

$$2X \le \max\{\exp_3(A^2), \exp_5(10^{10} k^6)\}\,. \tag{6}$$

The theorem follows since $n \le 2X$.

Step 1: Rescaling. We rescale $\boldsymbol{u}$ so that all the coefficients of the polynomials in its closed form representation are algebraic integers. To this end, let

$$\Psi(x) := x^k - a_0 x^{k-1} - \cdots - a_{k-1} = \prod_{i=1}^{s} (x - \alpha_i)^{\nu_i}$$

be the characteristic polynomial of $\boldsymbol{u}$ and let $\mathbb{K} := \mathbb{Q}(\alpha_1, \dots, \alpha_s)$ be the splitting field of Ψ, which has degree at most $k!$ over $\mathbb{Q}$. If $|\alpha_i| > 1$, then

$$|\alpha_i| = \left| a_0 + \frac{a_1}{\alpha_i} + \cdots + \frac{a_{k-1}}{\alpha_i^{k-1}} \right| < kA\,,$$

for A as in the statement of Theorem 7. Writing $\rho := \max |\alpha_i|_{i=1}^s$, we have $\rho < kA$.

The sequence $\boldsymbol{u}$ admits a closed-form solution $u_n = \sum_{i=1}^s Q_i(n)\alpha_i^n$, where the coefficients of the polynomials $Q_i(x)$ are computed from the initial values $u_0, \dots, u_{k-1}$ by solving a system of linear equations. By Cramer's rule, each of the coefficients of $Q_i(x)$ is the quotient of an algebraic integer by the determinant

$$\Delta := \begin{vmatrix} 1 & \dots & 0 & 1 & \dots & 0 & 1 & \cdots \\ \alpha_1 & \dots & \alpha_1 & \alpha_2 & \dots & \alpha_{s-1} & \alpha_s & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_1^{k-1} & \dots & (k-1)^{\nu_1 - 1}\alpha_1^{k-1} & \alpha_2^{k-1} & \dots & (k-1)^{\nu_s - 1}\alpha_{s-1}^{k-1} & \alpha_s^{k-1} & \dots \end{vmatrix}.$$

The length of each column vector above is at most

$$\sqrt{k(k-1)^{2(k-1)}\rho^{2k}} < k^k(kA)^k = k^{2k}A^k.$$

Thus, by the Hadamard inequality, $\Delta^2 < (k^{2k^2}A^{k^2})^2 = (k^2A)^{2k^2}$.

Solving with Cramer's rule for the coefficients of $Q_i(x)$ gives, via the Hadamard inequality again, that they are bounded by $kA|\Delta|$. Thus, replacing $\boldsymbol{u}$ by $\Delta\boldsymbol{u}$, we have that

$$(7) \qquad Q_i(x) := \sum_{j=0}^{\nu_i - 1} c_{i,j}x^j, \quad \text{where } |c_{i,j}| \le (k^2A)^{2k^2+1} \quad \text{and} \quad c_{i,j} \in \mathcal{O}_{\mathbb{K}}\,.$$

From Inequality (7) and Theorem 1(1) we furthermore have

$$(8) \quad h(c_{i,j}) \le (2k^2+1)\log(k^2A) \quad \text{for all } i \in \{1, \dots, s\} \text{ and } j \in \{0, \dots, \nu_i - 1\}\,.$$

Step 2: Reduction modulo P**.** Fix $n \in \mathcal{S}(X)$ such that $u_n = 0$ and consider a representation $n = qP + a$. Let $\mathfrak{p}$ be a prime ideal factor of P in $\mathcal{O}_{\mathbb{K}}$ and let $\sigma \in \mathrm{Gal}(\mathbb{K}/\mathbb{Q})$ be the Frobenius automorphism corresponding to $\mathfrak{p}$, such that $\sigma(\alpha) \equiv \alpha^P \bmod \mathfrak{p}$ for all $\alpha \in \mathcal{O}_{\mathbb{K}}$. From $u_n = 0$ and $n = qP + a$ we have

$$(9) \qquad \sum_{i=1}^{s} Q_i(a)\alpha_i^a\sigma(\alpha_i)^q \equiv 0 \pmod{\mathfrak{p}}\,.$$

Since our goal is to establish (6) we may freely assume that $X > \exp(10k^2\log k)$, which gives $a \ge \log X/\sqrt{\log_3 X} > 4k^2 + 3$. It follows that $a^k < k^a$. Noting also that $q \le a$, the absolute value of the left-hand side of (9) is at most

$$\begin{aligned} (k^2A)^{2k^2+1}ka^k\rho^{2a} &\le (k^2A)^{2k^2+1}ka^k(kA)^{2a} \\ &< (k^2A)^{2k^2+1}k(k^a)(kA)^{2a} \\ &\le (kA)^{4k^2+3}(kA)^{4a} \\ &= (kA)^{4k^2+3+4a} < (kA)^{5a}\,. \end{aligned}$$

Suppose that the left-hand side of (9) is non-zero. Then it is a non-zero algebraic integer of degree at most $k!$, all of whose conjugates have absolute value at most $(kA)^{5a}$, and which is divisible by $\mathfrak{p}$. This implies that P divides an integer of size at most $(kA)^{5ak!}$. Since $P \geq \frac{X-a}{q} \geq \frac{X-\log X}{\sqrt{\log X}} > \sqrt{X}$ for $X > X_0 := 100$, and $a \leq 2\log X/\sqrt{\log_3 X}$, taking logs we have

$$5k!\log(kA)\frac{2\log X}{\sqrt{\log_3 X}} > \frac{\log X}{2}.$$

It follows that $\sqrt{\log_3 X} < 20k!\log(kA)$ and so $X < \exp_3((20k!\log(kA))^2)$. But this last inequality implies Inequality (6), by the following case analysis:

(1) if $\dfrac{A}{\log A} > 40k!$ then $X < \exp_3(A^2)$,

(2) if $\dfrac{A}{\log A} < 40k!$ then $A < 80k!\log(40k!)$ and so $X < \max\{\exp_4(14), \exp_4(10k\log k)\}$.

Step 3: Companion equation. In Step 2 we have proved Inequality (6) under the assumption that the left-hand side of (9) is non-zero for some representation of n. Now suppose, on the contrary, that the left-hand side of (9) is zero for each of the $r(n) > \log_4 X$ representations of n. Of these representations, at least $(\log_4 X)/k!$ have the same Frobenius automorphism σ. For this choice of σ we have that the *companion equation* (the equational analog of the congruence (9))

$$\sum_{i=1}^{s} Q_i(a)\alpha_i^a \sigma(\alpha_i)^q = 0 \tag{10}$$

has least $(\log_4 X)/k!$ solutions in integer variables q, a. The remainder of the proof is dedicated to deriving an upper bound on the number of solutions of (10) that arise from representations of n. From this we obtain (6), as desired.

Every solution of (10) has a non-degenerate vanishing sub-sum and we focus on bounding the number of such sub-sums. The following claim, proven in Section A.1, considers sub-sums that involve only terms from a single summand $Q_i(a)\alpha_i^a\sigma(\alpha_i)^q$ of (10).

Claim 8. *Suppose that there exists $i \in \{1,\ldots,s\}$ such that $R_i(a)\alpha_i^a\sigma(\alpha_i)^q = 0$, where R_i a sub-polynomial of Q_i (meaning that every monomial in R_i appears in Q_i). Then $X < \max\{\exp_4(14), \exp_3(A), \exp_3(4k\log k)\}$.*

Since the upper bound on X in Claim 8 entails Inequality (6), it remains to bound the total number of non-degenerate solutions of each of the at most 2^k sub-equations of the form

$$\sum_{i\in I} R_i(a)\sigma(\alpha_i)^q\alpha_i^a = 0\,, \tag{11}$$

of (10), where $I \subseteq \{1,\ldots,s\}$ contains at least two elements, and where $R_i(x)$ is a sub-polynomial of $Q_i(x)$ for all $i \in I$. For this task, a key structure is the group $\mathcal{P}$ of $\boldsymbol{z} = (z_1, z_2) \in \mathbb{Z}^2$ such that

$$\sigma(\alpha_i)^{z_1}\alpha_i^{z_2} = \sigma(\alpha_j)^{z_1}\alpha_j^{z_2} \quad \text{for all} \quad i,j \in I\,.$$

For $\boldsymbol{z} = (z_1, z_2) \in \mathcal{P}$ we have $\sigma(\alpha_i/\alpha_j)^{z_1} = (\alpha_j/\alpha_i)^{z_2}$ for all $i, j \in I$. As shown in Section 2, since α_i/α_j is not a root of unity, this entails that $z_1 = z_2$ or $z_1 = -z_2$. There are thus three possibilities for $\mathcal{P}$: either $\mathcal{P} = \{\mathbf{0}\}$, $\mathcal{P}$ is parallel to $(1,1)$, or $\mathcal{P}$ is parallel to $(1,-1)$.

Step 4: The easy case. The easy case is that $\mathcal{P} = \{\mathbf{0}\}$. Recall that in Equation (11) polynomial R_i has degree at most ν_i. Now Theorem 4 shows that if we put

$$A := \sum_{i \in I} \binom{2+\nu_i-1}{2} \quad \text{and} \quad B = \max\{2, A\}$$

then the number of solutions (a,q) of (11) is at most $2^{35B^3}(k!)^{6B^2}$. But

$$A \leq \sum_{i \in I} \binom{\nu_i+1}{2} = \sum_{i \in I} \frac{\nu_i(\nu_i+1)}{2} \leq \sum_{i \in I} \nu_i^2 \leq k^2,$$

so the number of solutions of (11) is at most $2^{35k^6}(k!)^{6k^4}$. After multiplying the above bound by 2^k to account for the number of non-degenerate sub-sums, the resulting quantity is greater than the number $(\log_4 X)/k!$ of solutions of the companion equation. In other words,

$$\log_4 X < 2^k k!\, 2^{35k^6}(k!)^{6k^4},$$

from which we obtain

$$X < \max\{\exp_5(10^{10}), \exp_5(25k^6)\}\,, \tag{12}$$

which yields Inequality (6).

Step 5: The hard case. We are left with the case $\mathcal{P} \neq \{\mathbf{0}\}$, where we cannot apply Theorem 4. In this case $\mathcal{P}$ is either a subgroup of $\{(z,z) : z \in \mathbb{Z}\}$ or a subgroup of $\{(z,-z) : z \in \mathbb{Z}\}$. It follows that either $\sigma(\alpha_i)\alpha_i$ takes the same value for all $i \in I$ or $\sigma(\alpha_i)/\alpha_i$ takes the same value for all $i \in I$. Cancelling the common value of $(\sigma(\alpha_i)\alpha_i)^q$ or $(\sigma(\alpha_i)/\alpha_i)^q$ in (11) we have

$$\sum_{i \in I} R_i(a)\alpha_i^{a+\eta q} = 0 \quad \text{for some} \quad \eta \in \{\pm 1\}\,. \tag{13}$$

Similar to Step 4, we will establish the bound (6) by giving an upper bound on the number of solutions of (13) and hence on the number of representations of n. In lieu of Theorem 4, we use a bespoke argument that uses ideas of [AV09, ESS02, SS00], but which is greatly simplified by exploiting the assumption that no two representations of n are correlated.

The argument is by induction on the number of summands $|I| \leq k$. To get started, we set $\ell := |I|$, relabel the roots so that $I = \{1, \ldots, \ell\}$, and restate Equation (13) as follows:

$$\sum_{i=1}^{\ell} R_i(a)\alpha_i^{a+\eta q} = 0\,.$$

We dehomogenise the above equation by dividing through by the first summand, yielding

$$1 = \sum_{i=2}^{\ell} (-R_i(a)/R_1(a))(\alpha_i/\alpha_1)^{a+\eta q}. \tag{14}$$

Our goal is to apply Theorem 3 in order to find a homogeneous linear relation among the summands on the right-hand side of (14). This will yield an equation similar to (13) but with strictly fewer summands. To this end, let Γ be the rank-one multiplicative subgroup of $(\mathbb{K}^*)^{\ell-1}$ generated by $\vec{\gamma} := (\alpha_2/\alpha_1, \ldots, \alpha_\ell/\alpha_1)$. Then Equation (14) can be written $\boldsymbol{x}^\top \boldsymbol{y} = 1$, where $\boldsymbol{y} = -(R_2(a)/R_1(a), \ldots, R_\ell(a)/R_2(a))$

and $\boldsymbol{x} = \vec{\gamma}^{a+\eta q}$. Write $\varepsilon := (8k)^{-6k^3}$ and recall that h denotes absolute logarithmic Weil height. Then we have the following claim, which is proven in Section A.2.

Claim 9. *If* $2X > \max\{\exp_3(A^2), \exp_5(10^{10}k^6)\}$ *then* $h(\boldsymbol{y}) \leq (1 + h(\boldsymbol{x}))\varepsilon$.

Since the inequality $2X \leq \max\{\exp_3(A^2), \exp_5(10^{10}k^6)\}$ implies the desired bound (6), by Claim 9 we may assume without loss of generality that $h(\boldsymbol{y}) \leq (1 + h(\boldsymbol{x}))\varepsilon$. This height inequality allows us to apply Theorem 3 to conclude that there is a collection of at most $(8k)^{6k^3(k+1)}$ vectors $\mathbf{A} = (A_2, \dots, A_\ell) \in \overline{\mathbb{Q}}^{\ell-1}$ such that each solution of (14) satisfies

$$\sum_{i=2}^{\ell} A_i R_i(a) \alpha_i^{a+\eta q} = 0 \tag{15}$$

for one of these vectors $\mathbf{A}$. We will use such linear relations to proceed by induction.

Fix a vector $\mathbf{A} = (A_2, \dots, A_\ell)$ among the $(8k)^{6k^3(k+1)}$ possibilities and consider the corresponding version of Equation (15). Assume that at least three of the A_i's are nonzero and re-index so that the non-zero A_i's have indices $i = 2, 3, \dots, \ell'$, where $\ell' \leq \ell$. We dehomogenise the above equation to get

$$1 = \sum_{i=3}^{\ell'} (-A_i/A_2)(R_i(a)/R_2(a))(\alpha_i/\alpha_2)^{a+\eta q}.$$

We take now $\Gamma \subseteq (\mathbb{K}^*)^{\ell'-2}$ to be the rank-two multiplicative subgroup generated by the vectors $(\alpha_3/\alpha_2, \dots, \alpha_{\ell'}/\alpha_2)$ and $((-A_3/A_2), \dots, (-A_{\ell'}/A_2))$. The above equation is again of the form $\boldsymbol{x}^\top \boldsymbol{y} = 1$, where now

$$\boldsymbol{x} = ((-A_3/A_2)(\alpha_3/\alpha_2)^{a+\eta q}, \dots, (-A_{\ell'}/A_2)(\alpha_{\ell'}/\alpha_2)^{a+\eta q})),$$

and $\boldsymbol{y} = (R_3(a)/R_2(a), \dots, R_{\ell'}(a)/R_2(a))$. To continue the induction we wish to establish the height bound

$$h(\boldsymbol{y}) \leq (1 + h(\boldsymbol{x}))\varepsilon\,, \tag{16}$$

for the same ε as in Claim 9. The challenge is that the components of $\mathbf{A}$ (arising from the application of Theorem 3) are not known. But in the case at hand this is not a problem thanks to the following lemma, which is proved in Section A.3:

Claim 10. *If* $X > \exp_5(10^{10}k^6)$ *then there is at most one value of* $a + \eta q$ *such that* (16) *fails for the corresponding* $\boldsymbol{x}$.

Given that our ultimate goal is to prove Inequality (6) we may freely apply Claim 10 to argue that there is at most solution of (15) for which the corresponding vector $\boldsymbol{x}$ fails to satisfy (16). This allows us to continue the induction for all but one solution of (15).

In summary, at step one, the group Γ had rank 1 and the application of Theorem 3 led to a homogeneous equation in at most $k-1$ unknowns whose coefficients had unknown heights. At most one solution of the equation failed to satisfy the height bound (16) for the induction step, for which we had now a group Γ of rank 2 yielding an equation in at most $k-2$ of the unknowns. At each step, when the rank of Γ was r then the number of equations was at most $2^k \cdot (8k)^{6k^3(k+r)}$ and at each step there was at most one solution violating the height bound (16). So, if we have at least

$$\sum_{j=1}^{k-2} 2^k \prod_{i=1}^{j} (8k)^{6k^3(k+i)} < k2^{k(k-2)}(8k)^{6k^3(k-2)+6k^3(k-1)(k-2)/2} < (8k)^{3k^5}$$

solutions of the original Equation (13), then we arrive at a two-dimensional equation that has at least two solutions. That is, we have

$$A_i R_i(a)\alpha_i^{a+\eta q} + A_j R_j(a)\alpha_j^{a+\eta q} = 0,$$

for some $i \neq j$ and some A_i, A_j nonzero, $\eta \in \{\pm 1\}$, and the same equation with (q, a) replaced by (q', a'). Dividing these two equations yields

$$\frac{R_i(a)}{R_i(a')} \frac{R_j(a')}{R_j(a)} = \left(\frac{\alpha_j}{\alpha_i}\right)^{(a+\eta q)-(a'+\eta q')} . \tag{17}$$

We now prepare to take heights in Equation (17). As shown in the proof of Claim 9, we have $h(R_i(a)/R_i(a')) \leq 4k \log_2 X$ for $i = 1, \dots \ell$. Meanwhile Item 2 of Theorem 1 implies that

$$h(\alpha_i/\alpha_j) > \frac{2}{k^2(\log(6k^2))^3} .$$

Since $|(a + \eta q) - (a' + \eta q')| > \sqrt{\log X}$, we have

$$8k \log_2 X > \sqrt{\log X}\, h(\alpha_i/\alpha_j) > \frac{2\sqrt{\log X}}{k^2(\log(6k^2))^3} .$$

The above inequality can be rewritten $\sqrt{\log X} < c \log_2 X$ where $c := 4(k \log(3k^2))^3$. But then by Proposition 2 we have $X < \exp((4c \log(2c))^2)$, which implies Inequality (6), our ultimate goal.

It remains to consider the case that we have fewer than $(8k)^{3k^5}$ solutions of Equation (13). In this case we have

$$\frac{\log_4 X}{k!} < (8k)^{3k^5} \quad \text{and so} \quad X < \exp_5(13k^5 \log k) ,$$

which again implies Inequality (6). This concludes the proof of Theorem 7.

4. The Density of $\mathcal{S}$

This section is devoted to a proof of the following result.

Theorem 11. *The set $\mathcal{S}$ has lower density at least 0.29 unconditionally and has density one subject to the Bateman-Horn conjecture.*

Recall from Section 3 that we exclude from $\mathcal{S}$ all $n \in \mathbb{N}$ that have two correlated representations. In Section 4.1 we show that the set of numbers thus excluded has density zero. In Section 4.2 we show, assuming the Bateman-Horn conjecture, that

$$\#\{n \in [X, 2X] : r(n) > \log_4 X\} = (1 + o(1))X . \tag{18}$$

We conclude that $\mathcal{S}$ has density one. We also show unconditionally that

$$\#\{n \in [X, 2X] : r(n) > \log_4 X\} \geq ((1/\kappa) + o(1))X , \tag{19}$$

for κ the absolute constant in Inequality (4), which entails that $\mathcal{S}$ has lower density at least $1/\kappa \geq 0.29$.

In this section the indices p, q, P, P' in summations and products run over positive primes.

4.1. **Counting correlated representations.** We will need the following simple fact:

Proposition 12. $\sum_{q\in A(X)} \frac{1}{q} \sim \log_3 X$

Proof. Since $A(X) = [\log_2 X, \sqrt{\log X}]$, the result follows from Merten's theorem: $\sum_{q\leq n} \frac{1}{q} - \log\log n = O(1)$. □

We now have:

Lemma 13. *The set of $n \in [X, 2X]$ with two correlated representations $n = Pq + a = P'q' + a'$ (namely such that*

$$q \neq q',\ a \neq a' \quad \text{and} \quad |(a+\eta q) - (a' + \eta q')| < \sqrt{\log X}$$

for some $\eta \in \{\pm 1\}$), is of cardinality $O(X/(\log X)^{1/3})$.

Proof. We fix $q \neq q' \in A(X)$ and $a \neq a' \in B(X)$ and count the number of pairs of primes P and P' such that

$$qP + a = q'P' + a' \in [X, 2X]\,. \tag{20}$$

A general solution of Equation (20) in nonnegative integers P and P' can be written in the form $P = P_0 + q't$ and $P' = P_0' + qt$, where t is a nonnegative integer parameter and P_0, P_0' is a particular solution (chosen to minimise P_0 and P_0' simultaneously). The condition that $qP + a \leq 2X$ implies that $P \leq 2X/q$ and hence that $t \leq \frac{2X}{qq'}$. We can apply Inequality (5) with

$$\Delta := |qq'(qP_0 - q'P_0')| = |qq'(a - a')| \neq 0$$

to deduce that the number of $t \leq \frac{2X}{qq'}$ such that $P_0 + q't$ and $P_0' + qt$ are both prime is

$$\ll \frac{X}{qq'(\log X)^2}\,\frac{\Delta}{\varphi(\Delta)} \ll \frac{X}{qq'(\log X)^2}\,\frac{|a-a'|}{\varphi(|a-a'|)} \ll \frac{X \log_3 X}{qq'(\log X)^2}\,,$$

where we have used the inequality $m/\varphi(m) \ll \log\log m$ in the last step.

We next sum up the number of solutions of (20) over the different choices of $q \neq q' \in A(X)$ and $a \neq a' \in B(X)$ such that $|(a + \eta q) - a' + \eta q')| < \sqrt{\log X}$. Note here that since $q, q \leq \sqrt{\log X}$, the condition $|(a + \eta q) - a' + \eta q')| < \sqrt{\log X}$ implies that $|a - a'| < 2\sqrt{\log X}$ and hence a' is determined in at most $2\sqrt{\log X}$ different ways by the choice of a. Moreover, since $a \in B(X)$, there at most $\frac{2\log X}{\sqrt{\log_3 X}}$ choices of a. We thus get a count of

$$\frac{X\log_3 X}{(\log X)^2}\left(\sum_{q\leq\sqrt{\log X}} \frac{1}{q}\right)^2 \frac{\log X\sqrt{\log X}}{\sqrt{\log_3 X}} \ll \frac{X(\log_3 X)^{2.5}}{\sqrt{\log X}}\,,$$

where we use the inequality $\displaystyle\sum_{q\in A(X)} \frac{1}{q} \ll \log_3 X$ from Proposition 12. This is a count on the number of sextuples (q, q', a, a', P, P') satisfying Equation (20), so the number of n's arising as $Pq{+}a$ from such a sextuple is also $O(X/(\log X)^{1/3})$. □ □

4.2. **Counting all representations.** In view of Lemma 13, to show that $\mathcal{S}$ has density one it suffices to establish Equation (18). We will use the moment method. To set this up, for $i \in \{0,1,2\}$ write

$$M_i(X) = \sum_{\substack{n\in[X,2X]\\ r(n)>\log_4 X}} r(n)^i \,.$$

We estimate the first moment $M_1(X)$ as follows. We have

$$\begin{aligned}\sum_{n\in[X,2X]} r(n) &= \sum_{\substack{q\in A(X)\\ a\in B(X)}} \sum_{\frac{X-a}{q}\leq P\leq\frac{2X-a}{q}} 1\\ &= (1+o(1)) \sum_{\substack{q\in A(X)\\ a\in B(X)}} \frac{X}{q\log X} \quad \text{(by the Prime Number Theorem)}\\ &= (1+o(1))X\sqrt{\log_3 X} \quad \text{(by Proposition 12).} \qquad (21)\end{aligned}$$

It follows immediately that $M_1(X) = (1+o(1))X\sqrt{\log_3 X}$.

Turning to the second moment $M_2(X)$, If we were able to show that

$$\sum_{n\in[X,2X]} r(n)^2 = (1+o(1))X\log_3 X \qquad (22)$$

then it would follow that $M_2(X) = (1+o(1))X\log_3 X$ and hence, by the Cauchy-Schwarz inequality $M_0(X)M_2(X) \geq M_1(X)^2$, that $M_0(X) = (1+o(1))X$, which would establish (18).

It thus suffices to establish Equation (22). This is out of the reach of unconditional techniques, but it follows quite quickly from standard conjectures. In particular, we show that the Bateman-Horn conjecture implies that for any given $a \neq a' \in B(X)$ and $q \neq q' \in A(X)$, if $\gcd(a-a', qq') = 1$ and $2|(a-a')$ then the number of pairs of primes P, P' such that

$$qP + a = q'P' + a' \in [X, 2X] \qquad (23)$$

is given by

$$(C+o(1))\frac{X}{qq'(\log X)^2} g(|a-a'|)\,, \qquad (24)$$

where

$$C := 2\prod_{p>2}\frac{p(p-2)}{(p-1)^2} \approx 1.32 \quad \text{and} \quad g(m) := \prod_{\substack{p|m\\ p>2}}\frac{p-1}{p-2}\,.$$

Indeed, as explained in the proof of Lemma 13, there exist two linear forms $f_1(t) := P_0 + q't$ and $f_2(t) := P_0' + qt$ such that the number of solutions of Equation (23) in primes P and P' is equal to the number of values t in the range $\frac{X-o(X)}{qq'} \leq t \leq \frac{2X}{qq'}$ such that $f_1(t)$ and $f_2(t)$ are both prime. The assumptions $\gcd(a-a', qq') = 1$ and $2|(a-a')$ guarantee that $f := f_1 f_2$ is admissible in the sense of Conjecture 6—specifically we have that f does not vanish identically modulo 2, q, or q'. If one of the previous two assumptions fails then there are no solutions of (23) in primes P and P'. We can thus apply Conjecture 6 to obtain the estimate (24) of the number of solutions on (23). Note here that the constant C_f in Conjecture 6 becomes $Cg(|a-a'|)$ in (24).

Taking stock, we see that

$$\begin{aligned}\sum_{n\in[X,2X]} r(n)^2 &= \sum_{\substack{a,a'\in B(X)\\ q,q'\in A(X)}} \sum_{\substack{P,P'\\ qP+a=q'P'+a'\in[X,2X]}} 1\\ &= \sum_{\substack{a\neq a'\in B(X)\\ q\neq q'\in A(X)\\ 2|(a-a')\\ \gcd(a-a',qq')=1}} (C+o(1))\frac{X}{qq'(\log X)^2}g(|a-a'|)+O\left(X\sqrt{\log_3 X}\right).\end{aligned}$$

Here the $O(X\sqrt{\log_3 X})$ term comes from bounding the contribution of those summands for which $q=q'$ or $a=a'$. Applying Proposition 12 twice to evaluate the inner sum over $q\neq q'\in A(X)$, the equation above yields

(25)
$$\sum_{n\in[X,2X]} r(n)^2 = (C+o(1))\frac{X(\log_3 X)^2}{(\log X)^2}\sum_{\substack{a\neq a'\in B(X)\\ 2|(a-a')}} g(|a-a'|)+O\left(X\sqrt{\log_3 X}\right).$$

Next we simply the expression on the right-hand side of Equation (25). Since the function g is multiplicative, by Theorem 5 we have that for Y tending to infinity:

(26)
$$\sum_{\substack{n\leq Y\\ 2|n}} g(n) = \sum_{n\leq \frac{Y}{2}} g(n) = \frac{Y+o(Y)}{2}\prod_{p>2}\Big(1+\frac{g(p)-1}{p}\Big) = \frac{Y+o(Y)}{2}\prod_{p>2}\Big(1+\frac{1}{p(p-2)}\Big).$$

The first equality follows from the observation that $g(2n)=g(n)$ for all n; the second equality is an application of Theorem 5; the final equality follows from the fact that $g(p)=(p-1)/(p-2)$. Clearly the infinite product on the right-hand side converges to a finite value.

Recalling that the interval $B(X)$ has length $\frac{\log X}{\sqrt{\log_3 X}}$, we deduce from Equation (26) that

$$\begin{aligned}\sum_{\substack{a\neq a'\in B(X)\\ 2|(a-a')}} g(|a-a'|) &= \frac{1+o(1)}{2}\Big(\frac{\log X}{\sqrt{\log_3 X}}\Big)^2\prod_{p>2}\Big(1+\frac{1}{p(p-2)}\Big)\\ &= \frac{1+o(1)}{C}\,\frac{(\log X)^2}{\log_3 X}.\end{aligned}$$

Substituting the above equation into Equation (25) we deduce Equation (22). This is what we wanted to prove, and we conclude that $\mathcal{S}$ has density one subject to the Bateman-Horn conjecture.

If, in place of the Bateman-Horn conjecture, one uses the (unconditional) upper bound from Inequality (4), then the derivation of (22) shows *mutatis mutandis* that

$$\sum_{n\in[X,2X]} r(n)^2 \leq \kappa(1+o(1))X\log_3 X\,, \tag{27}$$

where κ is the absolute constant mentioned in (4). The application of the Cauchy-Schwarz inequality now yields $M_0(X) \geq ((1/\kappa) + o(1))X$. This establishes Equation (19), which shows that the density of $\mathcal{S}$ is at least $1/\kappa \geq 0.29$.

Appendix A. Deferred Proofs

A.1. **Proof of Claim 8.**

Claim 8. *Suppose that there exists $i \in \{1, \dots, s\}$ such that $R_i(a)\alpha_i^a \sigma(\alpha_i)^q = 0$, where R_i a sub-polynomial of Q_i (meaning that every monomial in R_i appears in Q_i). Then $X < \max\{\exp_4(14), \exp_3(A), \exp_3(4k \log k)\}$.*

Proof. Suppose that $R_i(a)\alpha_i^a \sigma(\alpha_i)^q = 0$. Write

$$R_i(x) = b_{i_0} x^{i_0} + b_{i_1} x^{i_1} + \cdots + b_{i_t} x^{i_t},$$

where $t \geq 1$, $0 \leq i_0 < i_1 < \cdots < i_t \leq \nu_i - 1$ and $b_{i_0}, \dots, b_{i_t}$ are nonzero algebraic integers. Simplifying across by x^{i_0}, we may assume that $x = a$ is a root of

$$b_{i_0} + b_{i_1} x^{i_1 - i_0} + \cdots + b_{i_t} x^{i_t - i_0}.$$

But then a divides the norm of b_{i_0}, a nonzero integer of size at most $(k^2 A)^{(2k^2+1)k!}$. Since $a \in B(X)$, this implies that

$$\frac{\log X}{\sqrt{\log_3 X}} < (kA)^{4(k+2)!},$$

and so

$$\log X < 2(kA)^{4(k+2)!} \log((kA)^{2(k+2)!}) < (kA)^{8(k+2)!} < \exp(8(k+2)^{k+2} \log(kA)) .$$

This implies that

$$X < \exp_3((k+2)\log(k+2) + \log(8\log(kA)))$$

and this in turn yields the upper bound stated in the claim. Indeed, either

$$A > 10k \log k, \quad \text{and then the above gives } \; X < \max\{\exp_4(14), \exp_3 A\}$$

or

$$A < 10k \log k, \quad \text{in which case } \; X < \max\{\exp_4(14), \exp_3(4k \log k)\}).$$

$\square$ $\square$

A.2. **Proof of Claim 9.**

Claim 9. *If $2X > \max\{\exp_3(A^2), \exp_5(10^{10}k^6)\}$ then $h(\boldsymbol{y}) \leq (1 + h(\boldsymbol{x}))\varepsilon$.*

Proof. We make a case distinction on the value of $h(a)$. First, assume that $h(a) \geq 3k^2 \log(k^2 A)$. Using the height bound on the coefficients of the R_i given in (8) together with general properties of the height function in Theorem 1, we have that for some $i \in \{2, \dots, \ell\}$,

$$\begin{aligned} h(\boldsymbol{y}) &\leq h(R_i(a)) + h(R_1(a)) \leq 2kh(a) + 2k\log 2 + 2k \log((k^2A)^{2k^2+1}) \\ &< 2kh(a) + 2k(2k^2+2)\log(k^2A) < 2kh(a) + 6k^3 \log(k^2 A) < 4kh(a). \end{aligned}$$

We deduce that

$$h(\boldsymbol{y}) < 4kh(a) = 4k \log a < 4k \log_2 X.$$

Next we give a lower bound on $h(\boldsymbol{x})$. Let α be an element of maximum height in $\{\alpha_i/\alpha_j : 1 \le i < j \le s\}$. Then

$$h(\boldsymbol{x}) \ge (a-q)h(\boldsymbol{\alpha}) \ge \left(\frac{\log X}{2\sqrt{\log_3 X}}\right) h(\alpha) \ge \sqrt{\log X}\, h(\alpha) \quad \text{for} \quad X > 55,$$

since $\log X/(2\sqrt{\log_3 X}) > \sqrt{\log X}$ for $X > 55$. Since α is not a root of unity (by the non-degeneracy of $\boldsymbol{u}$) it follows from Item 2 of Theorem 1 that

$$h(\alpha) \ge \frac{2}{k^2(\log(3k^2))^3}\,. \tag{28}$$

Suppose that $h(\boldsymbol{y}) > (1+h(\boldsymbol{x}))\varepsilon$. Then, combining the upper bound on $h(\boldsymbol{y})$ and lower bound on $h(\boldsymbol{x})$, we get that

$$4(k\log(3k^2))^3(8k)^{6k^3}\log_2 X > \sqrt{\log X}\,.$$

Putting $c := 4(k\log(3k^2))^3(8k)^{6k^3}$ and applying Proposition 2 we deduce that $X < \exp((4c\log(2c))^2)$. Thus, the hypothesis $X > \exp_5(10^6k^6)$ of the claim implies that $h(\boldsymbol{y}) \le (1+\varepsilon)h(\boldsymbol{x})$.

It remains to consider the case that $h(a) < 3k^2\log(k^2A)$. Here, since $a \ge \frac{\log X}{\sqrt{\log_3 X}}$, we have

$$\log X < a\sqrt{\log_3 X} \le \exp(3k^2\log(k^2A))\sqrt{\log_3 X}.$$

This yields

$$\log X < 6k^2\log(k^2A)\exp(3k^2(\log k^2A)) < \exp(6k^2\log(3k^2A)),$$

and so

$$X < \exp_2(6k^2\log(3k^2A)).$$

If $A > k\log k$, then $X < \max\{\exp_2(100), \exp_2(A^4)\}$ and if $A \le k\log k$, then $X < \max\{\exp_2(100), \exp_2(k^4)\}$. But both these upper bounds on X contradict the lower bound on X in the hypothesis of the claim and so the assumption $h(a) < 3k^2\log(k^2A)$ leads to a contradiction, i.e., the second case of the proof is vacuous. □ □

A.3. **Proof of Claim 10.**

Claim 10. *If $X > \exp_5(10^{10}k^6)$ then there is at most one value of $a+\eta q$ such that (16) fails for the corresponding $\boldsymbol{x}$.*

Proof. Suppose that (16) fails for both $(q,a) \ne (q',a')$. For the vectors $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{x},' \boldsymbol{y}'$ respectively corresponding to (q,a) and (q',a') we have $h(\boldsymbol{y}) > (1+\varepsilon)h(\boldsymbol{x})$ and $h(\boldsymbol{y}') > (1+\varepsilon)h(\boldsymbol{x}')$. Adding these two inequalities we get

$$\begin{aligned} 8k\log_2 X &> 4k\log a + 4k\log a' > h(\boldsymbol{y}) + h(\boldsymbol{y}') > (2+h(\boldsymbol{x})+h(\boldsymbol{x}'))\varepsilon \\ &\ge (2+h(\boldsymbol{x}/\boldsymbol{x}'))\varepsilon. \end{aligned} \tag{29}$$

For the right–most inequality above see equation (7.6) in [ESS02]. But in $\boldsymbol{x}/\boldsymbol{x}'$, the unknown vector $\mathbf{A}$ is gone and

$$h(\boldsymbol{x}/\boldsymbol{x}') = h((\alpha_3/\alpha_2)^{(a+\eta q)-(a'+\eta q')}, \ldots, (\alpha_{\ell'}/\alpha_2)^{(a+\eta q)-(a'+\eta q')}).$$

In particular, by (28) and the fact that $|(a+\eta q)-(a'+\eta q')| > \sqrt{\log X}$, we have

$$\begin{aligned} h(\boldsymbol{x}/\boldsymbol{x}') &\geq |(a+\eta q)-(a'+\eta q')| \min\{h(\alpha_i/\alpha_j) : i \neq j \in \{2,3,\ldots,\ell'\}\} \\ &\geq \frac{2\sqrt{\log X}}{k^2(\log(3k^2))^3} . \end{aligned}$$

So the estimate (29) leads to

$$8k\varepsilon^{-1}\log_2 X > \frac{2\sqrt{\log X}}{k^2(\log(3k^2))^3} ,$$

and hence $\sqrt{\log X} < c\log_2(X)$ for $c := 4(k\log(3k^2))^3(8k)^{6k^3}$. Now Proposition 2 yields $X < \exp((4c\log(2c))^2)$, which contradicts the assumption $X \geq \exp_5(10^{10}k^6)$. □

Acknowledgements

Florian Luca is also affiliated with the Centro de Ciencias Matemáticas UNAM, Morelia, Mexico. Joël Ouaknine is also affiliated with Keble College, Oxford as `emmy.network` Fellow; he was supported by DFG grant 389792660 as part of TRR 248 (see `https://perspicuous-computing.science`). James Worrell is supported by UKRI Fellowship EP/X033813/1.

References

[AAGT15] M. Agrawal, S. Akshay, B. Genest, and P. S. Thiagarajan. Approximate verification of the symbolic dynamics of Markov chains. *J. ACM*, 62(1):2:1–2:34, 2015.

[AV09] F. Amoroso and E. Viada. Small points on subvarieties of a torus. *Duke Mathematical Journal*, 150(3), 2009.

[AZFG20] S. Aletheia-Zomlefer, L. Fukshansky, and S. Ramon Garcia. The Bateman–Horn conjecture: Heuristic, history, and applications. *Expositiones Mathematicae*, 38(4):430–479, 2020.

[Bai02] Stephan Baier. On the bateman–horn conjecture. *Journal of Number Theory - J NUMBER THEOR*, 96, 10 2002.

[BG06] E. Bombieri and W. Gubler. *Heights in Diophantine Geometry.* Number 4 in New Mathematical Monographs. Cambridge University Press, Cambridge, 2006.

[BH62] P. T. Bateman and R. A. Horn. A heuristic asymptotic formula concerning the distribution of prime numbers. *Mathematics of Computation*, 16:363–367, 1962.

[Bil96] Y Bilu. A note on universal hilbert sets. *Journal für die reine und angewandte Mathematik*, 479:195–204, 1996.

[BJK20] G. Barthe, C. Jacomme, and S. Kremer. Universal equivalence and majority of probabilistic programs over finite fields. In *LICS'20: 35th Annual ACM/IEEE Symposium on Logic in Computer Science*, pages 155–166. ACM, 2020.

[BLN$^+$22] Y. Bilu, F. Luca, J. Nieuwveld, J. Ouaknine, D. Purser, and J. Worrell. Skolem meets schanuel. In *47th International Symposium on Mathematical Foundations of Computer Science, MFCS*, volume 241 of *LIPIcs*, pages 20:1–20:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[BPS21] P. Bell, I. Potapov, and P. Semukhin. On the mortality problem: From multiplicative matrix equations to linear recurrence sequences and beyond. *Inf. Comput.*, 281:104736, 2021.

[BR10] J. Berstel and C. Reutenauer. *Noncommutative Rational Series with Applications.* Cambridge University Press, 2010.

[BT00] V. Blondel and J. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, 2000.

[ESS02] J.-H. Evertse, H. P. Schlickewei, and W. M. Schmidt. Linear equations in variables which lie in a multiplicative group. *Annals of Mathematics*, 155(3):807–836, 2002.

[HR74] H. Halberstam and H.-E. Richert. *Sieve methods.* LMS Monographs. 1974.

[Lan96] S. Lang. La conjecture de Bateman-Horn. *Gaz. Math*, 67:82–84, 1996.

[Lec53] C. Lech. A note on recurring series. *Ark. Mat.*, 2, 1953.

[LOW21] F. Luca, J. Ouaknine, and J. Worrell. Universal Skolem Sets. In *36th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS*, pages 1–6. IEEE, 2021.

[LOW22] F. Luca, J. Ouaknine, and J. Worrell. A Universal Skolem Set of Positive Lower Density. In *47th International Symposium on Mathematical Foundations of Computer Science, MFCS*, volume 241 of *LIPIcs*, pages 73:1–73:12. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[Mah35] K. Mahler. Eine arithmetische Eigenschaft der Taylor Koeffizienten rationaler Funktionen. *Proc. Akad. Wet. Amsterdam*, 38, 1935.

[MST84] M. Mignotte, T. Shorey, and R. Tijdeman. The distance between terms of an algebraic recurrence sequence. *J. für die reine und angewandte Math.*, 349, 1984.

[OW15] J. Ouaknine and J. Worrell. On linear recurrence sequences and loop termination. *ACM SIGLOG News*, 2(2):4–13, 2015.

[RS94] G. Rozenberg and A. Salomaa. *Cornerstones of Undecidability.* Prentice Hall, 1994.

[Sko34] T. Skolem. Ein Verfahren zur Behandlung gewisser exponentialer Gleichungen. In *Comptes rendus du congrès des mathématiciens scandinaves*, 1934.

[SS78] A. Salomaa and M. Soittola. Automata-theoretic aspects of formal power series. In *Texts and Monographs in Computer Science*, 1978.

[SS00] H.P. Schlickewei and W.P. Schmidt. The number of solutions of polynomial-exponential equations. *Compositio Mathematica*, 120:193–225, 01 2000.

[Tao08] T. Tao. *Structure and randomness: pages from year one of a mathematical blog.* American Mathematical Society, 2008.

[Ten95] G. Tenenbaum. *Introduction to Analytic and Probabilistic Number Theory.* Number 46 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

[Ver85] N. K. Vereshchagin. The problem of appearance of a zero in a linear recurrence sequence (in Russian). *Mat. Zametki*, 38(2), 1985.

[Vou96] P. Voutier. An effective lower bound for the height of algebraic numbers. *Acta Arith.*, 74(1):81–95, 1996.