
Article:
Fishman, Lior, Merrill, Keith and Simmons, David orcid.org/0000-0002-9136-6635 (2018) Uniformly de Bruijn sequences and symbolic Diophantine approximation on fractals. ANNALS OF COMBINATORICS. pp. 271-293. ISSN: 0218-0006

https://doi.org/10.1007/s00026-018-0384-2

Ann. Comb.
https://doi.org/10.1007/s00026-018-0384-2

Annals of Combinatorics

Uniformly de Bruijn Sequences and Symbolic Diophantine Approximation on Fractals

Lior Fishman[1], Keith Merrill[2], and David Simmons[3]

[1]Department of Mathematics, University of North Texas, 1155 Union Circle #311430, Denton, TX 76203-5017, USA
lior.fishman@unt.edu

[2]Department of Mathematics, Brandeis University, 415 South Street, Waltham, MA 02454-9110, USA
merrill2@brandeis.edu

[3]Department of Mathematics, University of York, Heslington, York YO10 5DD, UK
David.Simmons@york.ac.uk

Received June 25, 2016

Mathematics Subject Classification: 11J04, 11J83, 05C45

Abstract. Intrinsic Diophantine approximation on fractals, such as the Cantor ternary set, was undoubtedly motivated by questions asked by K. Mahler (1984). One of the main goals of this paper is to develop and utilize the theory of infinite de Bruijn sequences in order to answer closely related questions. In particular, we prove that the set of infinite de Bruijn sequences in $k \geq 2$ letters, thought of as a set of real numbers via a decimal expansion, has positive Hausdorff dimension. For a given k, these sequences bear a strong connection to Diophantine approximation on certain fractals. In particular, the optimality of an intrinsic Dirichlet function on these fractals with respect to the height function defined by symbolic representations of rationals follows from these results.

Keywords: de Bruijn sequences, Diophantine approximation, iterated function systems, Eulerian paths, badly approximable points, height functions, Hausdorff dimension

1. Introduction

In this paper, we give a novel application of combinatorics to the field of Diophantine approximation. Since we do not assume that the reader is familiar with this field, let us first recall some important concepts and ideas. We refer the reader to Section 5 where we rigorously define and discuss these notions.

Classically, the field of Diophantine approximation sought to quantify how well real numbers can be approximated by rationals, weighing the distance to the rational point against some function of its denominator. The inaugural result in the field is Dirichlet's theorem, Theorem 5.2, which states that every irrational real number has

Published online: 27 April 2018

infinitely many rational points p/q that lie within distance $1/q^2$ of it. This result raises the question of whether that function, $1/q^2$, can be improved. That it cannot be, in a sense made precise in Section 5, is due to a result of Liouville, who showed that quadratic irrational numbers, like $\sqrt{2}$, admit no better rate of approximation. In modern terminology, we call such points *badly approximable*.

A more complete description of the set of badly approximable numbers, in this and related contexts, was the subject of much activity in the early-to-mid twentieth century. Via a characterization of badly approximable numbers in terms of continued fraction expansions one can show that the set of badly approximable numbers is uncountable, but it is also relatively easy to show that this set is a Lebesgue null set [5, Theorem 1.9 and Corollary 1.6], so we must turn to other notions of "size". One such notion, particularly well-suited to disntinguishing between sets of measure zero, is that of *Hausdorff dimension*. Jarník showed that despite being a Lebesgue null set, the set of badly approximable real numbers has full Hausdorff dimension, so it is still "large" in some sense.

As discussed further in Section 5, the core questions of Diophantine approximation can be formulated in many diverse contexts, essentially whenever we have a complete metric space X, a countable dense subset $\mathcal{Q}$, and some notion of "height" defined on $\mathcal{Q}$ (this would be the size of the denominator in the classical case above). Over the last decade, a plethora of results regarding Diophantine approximation on fractals have emerged [3, 4, 7, 9–11, 13, 14, 18]. Many of these results were motivated by the following question(s) posed by Mahler in 1984 [16, §2]: "How close can irrational elements of Cantor's set be approximated by rational numbers

(1) in Cantor's set, and
(2) by rational numbers not in Cantor's set?"

In this paper, we will restrict our attention to Mahler's first question; see Section 6 for details. We remark that while in [11], the first- and third-named authors were able to exhibit an optimal Dirichlet function (see Definition 5.3) corresponding to Mahler's second question, it seems that finding an analogous answer to his first question is significantly harder, see, e.g., [4,6,11] for detailed discussions and conjectures regarding this question.

In [11], a new height function was defined on the rational points of the Cantor set (see Section 6), and a Dirichlet-type theorem was proven [11, Corollary 2.2 and its proof]. The purpose of this paper is to demonstrate the optimality of that Dirichlet theorem, and give an estimate on the Hausdorff dimension of the set of "badly approximable" points. This set, as noted in [11], admits a precise combinatorial description, although at the time we had been unable to exhibit any members belonging to it. In the present paper, we focus on a combinatorially defined subset of the set of badly approximable points, the set of *uniformly de Bruijn sequences*. The existence of uniformly de Bruijn sequences demonstrates the optimality of the Dirichlet function (Theorem 6.3), and by estimating from below the Hausdorff dimension of the set of uniformly de Bruijn sequences (Theorem 2.1), we are able to get a positive lower bound for the Hausdorff dimension of the set of badly approximable points (Corollary 6.4), a first step towards a Jarník-type result. See Section 6 for a more nuanced discussion of these points.

2. Finite and Infinite de Bruijn Sequences

Let A be a finite alphabet of cardinality $k \geq 2$. We recall that a (non-cyclic) *de Bruijn sequence of order n* in A is a sequence ω of length $k^n + n - 1$ in the alphabet A that has the property that every sequence of length n in A appears as a consecutive substring of ω exactly once. For example, in the alphabet $\{0, 1\}$, the sequence 00110 is a de Bruijn sequence of order 2 while in the alphabet $\{0, 1, 2\}$, the sequence 000100201101202102211121222 00 is a de Bruijn sequence of order 3. We say that an infinite sequence $\omega \in A^{\mathbb{N}}$ is *infinitely de Bruijn* if the set

$$B_\omega \stackrel{\text{def}}{=} \{n \in \mathbb{N}\colon \text{the initial segment of } \omega \text{ of length } k^n + n - 1 \text{ is a de Bruijn sequence of order } n\} \tag{2.1}$$

is infinite. We say that ω is *totally de Bruijn* if $B_\omega = \mathbb{N}$, and *uniformly de Bruijn* if B_ω has bounded gap sizes. The construction of infinitely de Bruijn sequences goes back to Becher and Heiber [1],* who showed that when $k \geq 3$, totally de Bruijn sequences could be constructed recursively by extending each de Bruijn sequence of order n to a de Bruijn sequence of order $(n+1)$. We shall discuss their method in more detail below. When $k = 2$, it is known that no totally de Bruijn sequence exists, but Becher and Heiber do construct a uniformly de Bruijn sequence such that $B_\omega = 2\mathbb{N}$.

In order to state our main theorem for this section, let us briefly recall the definition and basic properties of the Hausdorff dimension of a fractal† $F \subseteq \mathbb{R}^d$, see, e.g., [8, Chapters 2-3]. Let d denote the standard metric on $\mathbb{R}^d$, and let $\operatorname{diam}(U)$ denote the diameter of a set $U \subseteq \mathbb{R}^d$. Fix $\delta > 0$ and let $F \subseteq \mathbb{R}^d$. We say that a countable collection $\{U_j\colon j \in \mathbb{N}\}$ of subsets of $\mathbb{R}^d$ is a *δ-cover* of F if $F \subseteq \bigcup_{j=1}^{\infty} U_j$ and $\operatorname{diam}(U_j) \leq \delta$ for every j. For each $s \geq 0$, let

$$\mathcal{H}^s_\delta(F) \stackrel{\text{def}}{=} \inf\left\{ \sum_{j=1}^{\infty} \operatorname{diam}(U_j)^s\colon \{U_j\colon j \in \mathbb{N}\} \text{ is a } \delta\text{-cover of } F \right\}.$$

Then the *s-dimensional Hausdorff measure* of F is the number

$$\mathcal{H}^s(F) \stackrel{\text{def}}{=} \lim_{\delta \to 0} \mathcal{H}^s_\delta(F),$$

and the *Hausdorff dimension* of F is the number

$$\dim_H(F) \stackrel{\text{def}}{=} \inf\{s \geq 0\colon \mathcal{H}^s(F) = 0\} = \sup\{s \geq 0\colon \mathcal{H}^s(F) = \infty\}.$$

It is well known that for every $F \subseteq \mathbb{R}^d$ we have $0 \leq \dim_H(F) \leq d$, and that if $\dim_H(F) > 0$, then F is uncountable, but not vice versa.‡

* Note that in [1], the phrase "infinite de Bruijn sequence" has a different meaning; we do not use that meaning in this paper because it makes an ad hoc distinction between the $k = 2$ case and the $k \geq 3$ case.

† The word "fractal" normally has a connotative but not a denotative meaning in mathematics; a set is called a fractal if it is "sufficiently complicated at fine scales". The Cantor ternary set, i.e., the set of all numbers in $[0, 1]$ that can be written in base 3 with only the digits 0 and 2, is a canonical example of a fractal; further examples are given in Subsection 5.2.

‡ The set of Liouville numbers on the real line is a standard example of a comeager (and thus uncountable) set of Hausdorff dimension 0.

We also recall that if $b \geq 2$ is an integer, then the *base b expansion* of a number $x \in [0, 1]$ is the series

$$\sum_{i=1}^{\infty} \frac{\omega_i}{b^i},$$

where $\omega_1, \omega_2, \ldots \in \{0, 1, \ldots, b-1\}$ are chosen so that the value of the series is equal to x. This choice is unique unless x is a rational number whose denominator is a power of b, in which case there are exactly two ways in which the infinite word $\omega = \omega_1 \omega_2 \cdots$ can be chosen.

Theorem 2.1. *Fix an integer $b \geq 2$, and let $C(b) = \{0, 1, \ldots, b-1\}$. Fix $A \subseteq C(b)$ such that $k \stackrel{\text{def}}{=} \#(A) \geq 2$. Denote by δ the Hausdorff dimension of the set F consisting of all numbers that can be written in the form $\sum_{i=1}^{\infty} \frac{\omega_i}{b^i}$ with $\omega_i \in A$ for every $i \in \mathbb{N}$, i.e., the set of all numbers in F that have at least one base b expansion composed entirely of digits from A.[§] Then the set S consisting of all elements of F that have at least one base b expansion that is uniformly de Bruijn satisfies*

$$0 < \alpha_k \delta \leq \dim_H(S) \leq \frac{\log(k!)}{k \log(k)} \delta < \delta,$$

where

$$\alpha_k = \begin{cases} 1/49, & k = 2, \\ (8 \cdot (9 \log_4(3) - 1))^{-1}, & k = 3, \\ \frac{\log(k-2)!}{k \log(k)}, & k \geq 4. \end{cases}$$

In particular, S has positive Hausdorff dimension but not full Hausdorff dimension.

Note that for large values of k, Stirling's formula gives $\alpha_k \sim \frac{\log(k!)}{k \log(k)} \sim 1 - \frac{1}{\log(k)}$ (where $x \sim y$ means $(1-x)/(1-y) \to 1$), and in particular $\alpha_k \to 1$ as $k \to \infty$. Thus S gets closer and closer to having full dimension as the number of allowed digits increases.

3. Preliminaries

We begin by recalling some key definitions used in Becher and Heiber's paper, as well as the proof of the well-known BEST[¶] theorem.

Definition 3.1. ([1]) *Given an alphabet A and an integer $n \in \mathbb{N}$, the* de Bruijn graph *of order n on A is the directed graph $G = G_n(A)$ with vertex set $V(G) \stackrel{\text{def}}{=} A^n$ and edge set $E(G) \stackrel{\text{def}}{=} \{(\omega, \tau) \colon \omega_{i+1} = \tau_i \ \forall i \leq n-1\}$. Note that every vertex has in-degree and out-degree both equal to $k \stackrel{\text{def}}{=} \#(A)$, for a total of k^n vertices and k^{n+1} edges.*

[§] It is well known that $\delta = \log(k)/\log(b)$, see Subsection 5.2.

[¶] An acronym after the people who discovered it: de Bruijn, van Aardenne-Ehrenfest, Smith, and Tutte.

If ω is a sequence of length $\ell \geq n$ in A, then the path induced by ω on G *is the path*[‖] $\gamma = \gamma_1 \cdots \gamma_{\ell-n+1}$ *in G defined by the formula*

$$\gamma_i \stackrel{\text{def}}{=} \omega_i \cdots \omega_{i+n-1} \in V(G).$$

Observation 3.2. Let ω be a sequence of length $\ell_\omega = k^m + m - 1$, and let γ be the path induced by ω on $G_n(A)$. Note that the length of γ is $\ell_\gamma = \ell_\omega - n = k^m + m - n - 1$; in particular, $\ell_\gamma = k^{n+1}$ if $m = n+1$, and $\ell_\gamma < k^n$ if $m \leq n$. Moreover,

(I) If $m = n+1$, then ω is de Bruijn if and only if γ is Eulerian.
(II) If $m \leq n$ and ω is de Bruijn, then γ is a simple path.

Remark 3.3. If $m = n$ and ω is de Bruijn, then γ is a simple path that visits each vertex exactly once. However, since γ starts and ends at different vertices, it is not a Hamiltonian cycle, contrary to [1, p. 931, first para.]. In particular, the edge set of γ does not form a regular graph on $V(\Omega)$, as is claimed in [1, Proof of Lemma 3, last para.]. Consequently, the proof given there is technically incorrect; it can be trivially fixed by adding a step where γ is extended to a Hamiltonian cycle; cf. the first two paragraphs of the proof of Corollary 4.3 below. Similar remarks apply to [1, Proof of Lemma 5, last para.].

Now let $X = (V(X), E(X))$ be a directed graph such that for each vertex $x \in V(X)$, the in-degree and out-degree of x are nonzero and equal to each other (though they may depend on x). Fix a vertex $x_0 \in V(X)$, and let $\mathcal{E}$ be the set of Eulerian paths of X that start and end at x_0. Note that, unlike standard convention, we consider two Eulerian paths to be different if they are formally different as sequences of vertices even if they are cyclically equivalent. Let $\mathcal{T}$ be the set of directed spanning trees of X rooted at x_0 with edges pointing towards x_0.

Since both the conclusion of the BEST theorem and its proof will be important for our argument, we recall them now. We once again remind the reader that our statement differs slightly from the usual one because of our convention about counting Eulerian paths: we do *not* consider cyclically equivalent paths to be the same. But the difference is easy to quantify: the number of Eulerian paths in each cyclic equivalence class that start and end at x_0 is equal to the degree of x_0 (we recall that by assumption the in-degree and out-degree are equal). So our count will be off from the conventional one by a factor of $\deg(x_0)$.

Theorem 3.4. (BEST theorem) *We have*

$$\#(\mathcal{E}) = \#(\mathcal{T}) \cdot \deg(x_0) \cdot \prod_{x \in V(X)} [\deg(x) - 1]! \,. \tag{3.1}$$

Proof. Let $T \in \mathcal{T}$ be a directed spanning tree rooted at x_0. For each $x \in V(X)$, let E_x denote the set of edges in X with initial vertex x, and let $T_x = E(T) \cap E_x$, where

[‖] In this paper a "path" in a directed graph is a sequence of vertices such that each pair of consecutive vertices is connected by an edge from the first vertex to the second vertex. The length of a path is the number of such edges, or equivalently, the number of vertices minus one (counting multiplicity in both cases). A path is *simple* if all its vertices are distinct except possible the first and last, and *Eulerian* if it contains each edge exactly once.

$E(T)$ denotes the edge set of T. If $x \neq x_0$, then T_x is a singleton, say, $T_x = \{v_x\}$, while $T_{x_0} = \varnothing$. Now let $\mathrm{Ord}(S)$ denote the set of total orderings of a set S, and note that the cardinality of the set

$$O(T) \stackrel{\text{def}}{=} \prod_{x \in V(X)} \mathrm{Ord}(E_x \setminus T_x)$$

is exactly $\deg(x_0) \cdot \prod_{x \in V(X)} [\deg(x) - 1]!$. Now for each $\mathbf{o} = (o_x)_{x \in V(X)} \in O(T)$ we let $f(T, \mathbf{o})$ be the Eulerian path that starts and ends at x_0 defined recursively as follows: suppose that the points $x_0 = \gamma_0, \gamma_1, \ldots, \gamma_i$ have been defined, and let $x = \gamma_i$. Then the next vertex γ_{i+1} must be chosen so that $\gamma_i\gamma_{i+1} \in E_x$, but $\gamma_i\gamma_{i+1} \neq \gamma_j\gamma_{j+1}$ for all $j < i$. We make this choice so as to minimize $\gamma_i\gamma_{i+1}$ according to the ordering o_x subject to these restrictions. If the edges of $E_x \setminus T_x$ have been exhausted, then if $x \neq x_0$ we choose the vertex v_x, and if $x = x_0$, then we terminate the path. There is some work to do to show that $f(T, \mathbf{o})$ is indeed an Eulerian path, and that every Eulerian path that starts and ends at x_0 can be represented uniquely as $f(T, \mathbf{o})$ for some $T \in \mathcal{T}$ and $\mathbf{o} \in O(T)$, see, e.g., [17, pp. 445–446]. This implies that f is a bijection between $\coprod_{T \in \mathcal{T}} O(T) = \{(T, \mathbf{o})\colon T \in \mathcal{T},\ \mathbf{o} \in O(T)\}$ and $\mathcal{E}$, which completes the proof. ∎

We will also need the following sufficient condition for the right-hand side of (3.1) to be nonzero:

Lemma 3.5. *If X is connected, then there is at least one directed spanning tree rooted at x_0, i.e., $\mathcal{T} \neq \varnothing$.*

Proof. Let T be a maximal directed tree rooted at x_0. By the maximality of T, there is no edge from any vertex not in T to any vertex in T. Since each vertex of X has equal in-degree and out-degree, the number of edges from $V(T)$ to $V(X) \setminus V(T)$ is equal to the number of edges from $V(X) \setminus V(T)$ to $V(T)$, which is equal to zero. Since X is connected, this means that either $V(T) = \varnothing$ or $V(X) \setminus V(T) = \varnothing$. But $x_0 \in V(T)$ by construction, so $V(X) \setminus V(T) = \varnothing$ and thus T is a spanning tree, i.e., $T \in \mathcal{T}$. ∎

4. Proof of Theorem 2.1

4.1. The Upper Bound

We begin by establishing the upper bound of Theorem 2.1. To do this we will use the Hausdorff-Cantelli lemma, a very useful tool for establishing upper bounds on the Hausdorff dimensions of certain sets, see, e.g., [2, Lemma 3.10]. Let $\{U_j\colon j \in \mathbb{N}\}$ be a countable collection of sets in $\mathbb{R}^d$, and let U be the set consisting of those elements of $\mathbb{R}^d$ that belong to infinitely many of the sets U_j ($j \in \mathbb{N}$). In other words,

$$S \stackrel{\text{def}}{=} \limsup_{j \to \infty} U_j = \bigcap_{N=1}^{\infty} \bigcup_{j=N}^{\infty} U_j.$$

Lemma 4.1. (Hausdorff-Cantelli Lemma) *Let $\{U_j\colon j \in \mathbb{N}\} \subseteq \mathbb{R}^d$ be a countable collection of sets, and let $S = \limsup_j U_j$. Fix $s > 0$. If*

$$\sum_{j=1}^{\infty} \mathrm{diam}\,(U_j)^s < \infty, \tag{4.1}$$

then $\mathcal{H}^s(S) = 0$ *and thus* $\dim_H(U) \leq s$.

It turns out to be convenient to consider a collection $\{U_j : j \in \mathbb{N}\}$ that naturally splits up into subcollections, say, $\{U_j : j \in \mathbb{N}\} = \bigcup_m \mathcal{C}_m$ for some sequence of collections $(\mathcal{C}_m)_{m=1}^{\infty}$. In this case, the summability condition (4.1) is equivalent to the condition

$$\sum_{m=1}^{\infty} \operatorname{cost}^s(\mathcal{C}_m) < \infty,$$

where

$$\operatorname{cost}^s(\mathcal{C}_m) \stackrel{\text{def}}{=} \sum_{U \in \mathcal{C}_m} \operatorname{diam}(U)^s$$

is the *s-dimensional cost* of $\mathcal{C}_m$. Note that $\operatorname{cost}^s(\mathcal{C}_m)$ should be distinguished from the expression $\left(\operatorname{cost}^1(\mathcal{C}_m)\right)^s$, which denotes instead the 1-dimensional cost of $\mathcal{C}_m$ raised to the power of s. The set S can be written in terms of the collections $(\mathcal{C}_m)_{m=1}^{\infty}$ as follows:

$$S = \limsup_{m \to \infty} \bigcup_{U \in \mathcal{C}_m} U = \bigcap_{N=1}^{\infty} \bigcup_{m=N}^{\infty} \bigcup_{U \in \mathcal{C}_m} U.$$

In what follows we will abuse terminology somewhat by calling $\operatorname{cost}^s(\mathcal{C}_m)$ the "cost" of the set $S_m \stackrel{\text{def}}{=} \bigcup_{U \in \mathcal{C}_m} U$, although strictly speaking, it depends not only on S_m but also on how it is decomposed.

Proof of upper bound. For each m, let S_m be the set consisting of all elements of F corresponding to base b expansions whose initial segments of length $k^m + m - 1$ are de Bruijn sequences of order m in A. Then the lim sup of the sequence $(S_m)_{m=1}^{\infty}$ consists of those elements of F with infinitely de Bruijn base b expansions. In particular, the set S consisting of those elements of F with uniformly de Bruijn base b expansions satisfies

$$S \subseteq \limsup_{m \to \infty} S_m = \bigcap_{N=1}^{\infty} \bigcup_{m=N}^{\infty} S_m.$$

By the Hausdorff-Cantelli lemma, if we can find an s such that

$$\sum_{m=1}^{\infty} \operatorname{cost}^s(S_m) < \infty, \tag{4.2}$$

then we can conclude that $\dim_H(S) \leq s$. We will show that (4.2) holds for all $s > \delta \frac{\log(k!)}{k \log(k)}$.

For each m, we view S_m as the union of the collection

$$\mathcal{C}_m \stackrel{\text{def}}{=} \left\{ S_m^{\omega} : \omega \text{ is a de Bruijn sequence of order } m \text{ in the alphabet } A \right\},$$

where for each ω, S_m^{ω} is the set of points $x \in F$ corresponding to base b expansions whose initial segments of length $k^m + m - 1$ are equal to ω. Let G be the de Bruijn graph of order $(m-1)$ on A (see Definition 3.1), so that $\#(V(G)) = k^{m-1}$. By Observation 3.2(I), the collection $\mathcal{C}_m$ is in bijection with the set of Eulerian paths on G. Fix a vertex $x_0 \in V(G)$. We can estimate the number of Eulerian paths starting and

ending at x_0 via the BEST theorem. Specifically, we have $\prod_{x\in V(G)}(\deg(x)-1)! = (k-1)!^{\#(V(G))}$, since every vertex $x \in V(G)$ has degree equal to k. The number of spanning trees rooted at x_0 is at most $k^{\#(V(G))-1}$, since an edge must be chosen emanating from each vertex $x \neq x_0$, and each vertex has out-degree k. And for the same reason, $\deg(x_0) = k$. Therefore, the number of Eulerian paths starting and ending at x_0 is at most

$$\begin{aligned}\#(\mathcal{T}) \cdot \deg(x_0) \cdot \prod_{x\in V(G)} [\deg(x)-1]! &\leq k^{\#(V(G))-1} \cdot k \cdot (k-1)!^{\#(V(G))} \\ &= k!^{\#(V(G))} \\ &= k!^{k^{m-1}}.\end{aligned}$$

Since there are $\#(V(G)) = k^{m-1}$ possible choices for x_0, the number of de Bruijn sequences of order m in A is at most $k^{m-1} \cdot k!^{k^{m-1}}$.* Now, if ω is a de Bruijn sequence of order m in A, then the length of ω is $k^m + m - 1$, and thus the diameter of S_m^ω is at most b^{-k^m-m+1}. So the s-dimensional cost of S_m according to the decomposition $\mathcal{C}_m$ is at most

$$k^{m-1} \cdot k!^{k^{m-1}} \cdot \left(b^{-k^m-m+1}\right)^s.$$

Now fix $\varepsilon > 0$ and set

$$s \stackrel{\text{def}}{=} \frac{1}{k}\log_b(k!) + \varepsilon. \tag{4.3}$$

Then

$$\sum_{m=1}^{\infty} \text{cost}^s(S_m) \leq \sum_{m=1}^{\infty} k^{m-1}(k!)^{k^{m-1}} \left(b^{-k^m-m+1}\right)^s.$$

By the ratio test, this series converges as long as $\lim_{m\to\infty} |a_{m+1}/a_m| < 1$, where a_m denotes the mth term. A straightforward computation yields:

$$|a_{m+1}/a_m| = k \cdot b^{-\varepsilon\left(k^{m+1}-k^m\right)} \cdot b^{-s},$$

which tends to 0 as $m \to \infty$.

Thus by Lemma 4.1, we have

$$\dim_H(S) \leq \frac{1}{k}\log_b(k!) = \frac{\log(k!)}{k\log(b)} = \frac{\log(k!)}{k\log(k)}\,\delta,$$

since $\delta = \log(k)/\log(b)$ (see Subsection 5.2).

Since for all $k \geq 2$ we have $k! < k^k$ and thus $\frac{\log(k!)}{k\log(k)} < 1$, we deduce that the Hausdorff dimension of S is strictly less than δ. ∎

* In fact, the exact count for such sequences is known, but we prefer this estimate because it is simpler and yields the same upper bound on the Hausdorff dimension.

4.2. The Lower Bound

The proof of the lower bound is significantly more involved, and will require a few preliminary results. We begin with the following proposition:

Proposition 4.2. *Let X be a k-regular connected directed graph, fix $x_0 \in V(X)$, and let $\mathcal{E}$ be the set of Eulerian paths of X that start and end at x_0. Then there exists $\mathcal{E}' \subseteq \mathcal{E}$ such that*:

(i) $\#(\mathcal{E}') = k \cdot (k-1)!^{\#(V(X))}$;
(ii) *If δ is a path of length ℓ_δ starting at x_0, then the number of paths in $\mathcal{E}'$ that extend δ is at most $k \cdot (k-1)!^{\#(V(X))-\ell_\delta/k}$.*

Proof. Since X is connected, by Lemma 3.5 there exists a directed spanning tree T rooted at x_0. Let $\mathcal{E}'$ be the set of Eulerian paths δ that start and end at x_0 such that for all $xy \in E(X)$ and $xz \in E(T)$ with $y \neq z$, the edge xy appears in δ before xz does. Equivalently, $\mathcal{E}' = \{f(T, \mathbf{o}) \colon \mathbf{o} \in O(T)\}$ where the notation is as in the proof of the BEST theorem. Then the proof of the BEST theorem implies that $\#(\mathcal{E}') = \#(O(T)) = k \cdot (k-1)!^{\#(V(X))}$. Now let δ be a path starting at x_0 that has at least one extension in $\mathcal{E}'$. For each $\mathbf{o} \in O(T)$, the path $f(T, \mathbf{o})$ is an extension of δ if and only if the algorithm described in the proof of the BEST theorem produces δ on input $\mathbf{o}$. Equivalently, $f(T, \mathbf{o})$ is an extension of δ if for each edge xy of δ, the rank of xy according to o_x is the same as its rank according to its location in δ. The number of elements $\mathbf{o} \in O(T)$ satisfying this condition is given by the formula

$$\begin{aligned} N_\delta &= \prod_{x \in V(X)} [\#(E_x \setminus (E(\delta) \cup E(T))]! \\ &= \#(E_{x_0} \setminus E(\delta)) \cdot \prod_{x \in V(X)} [\#(E_x \setminus E(\delta)) - 1]! \\ &\leq k \cdot \prod_{x \in V(X)} [\#(E_x \setminus E(\delta)) - 1]!, \end{aligned}$$

where E_x denotes the set of edges with initial vertex x, and $E(\delta)$ denotes the edge set of δ. Here we use the convention $(-1)! = 1$, since if $E_x \setminus E(\delta) = \varnothing$, then there is exactly one ordering o_x satisfying the appropriate condition, namely, the ordering determined by δ, and by hypothesis the element v_x comes last in this ordering. Now since

$$(i-1)! \leq (k-1)!^{i/k}, \qquad \forall\, i = 0, \ldots, k,$$

we have

$$N_\delta \leq k \cdot (k-1)!^{M/k},$$

where

$$M \stackrel{\text{def}}{=} \sum_{x \in V(X)} \#(E_x \setminus E(\delta)) = \#(E(X) \setminus E(\delta)) = k\#(V(X)) - \ell_\delta. \qquad \blacksquare$$

The next result will furnish the lower bound for $k \geq 4$. Although it is valid for $k = 3$, it provides no useful information in this case since 0 is always a (trivial) lower bound on the dimension.

Corollary 4.3. *Let the notation be as in Theorem 2.1, and let S be the set of numbers in F with totally de Bruijn base b expansions. Assume that $k \geq 4$. Then the Hausdorff dimension of S is bounded below by $\alpha_k \delta > 0$, where δ is the Hausdorff dimension of F $\big($and equals $\log(k)/\log(b)\big)$, and*

$$\alpha_k = \frac{\log(k-2)!}{k\log(k)}. \tag{4.4}$$

Before we turn to the proof, we recall the so-called *Mass Distribution Principle*, an extremely useful tool for bounding the Hausdorff dimension from below.

Lemma 4.4. ([8, Principle 4.2]) *Let F be a metric space, and let μ be a measure on F such that $0 < \mu(F) < \infty$. Fix $s, \varepsilon > 0$, and suppose that there exists $C > 0$ such that $\mu(U) \leq C \cdot \operatorname{diam}(U)^s$ for every set $U \subseteq F$ such that $\operatorname{diam}(U) \leq \varepsilon$. Then*

$$\dim_H(F) \geq s.$$

Proof of Corollary 4.3. Fix $n \in \mathbb{N}$, and let $\omega = \omega_1 \cdots \omega_{k^n+n-1}$ be a de Bruijn sequence of order n in A. Since the path induced by ω on $G_{n-1}(A)$ is an Eulerian path in a directed graph in which each vertex has equal in-degree and out-degree, it must start and end at the same vertex, which means that the first $(n-1)$ letters of ω are the same as the last $(n-1)$ letters, i.e., $\omega_{k^n+i} = \omega_i$ for all $i = 1, \ldots, n-1$.* Now let $\omega_{k^n+n} = \omega_n$ and $\omega' = \omega_1 \cdots \omega_{k^n+n}$. Then the first n letters of ω' are the same as the last n letters, but no other block of n letters is repeated in ω'.

Let $G = G_n(A)$ be the de Bruijn graph of order n on A, and let $\gamma = \gamma_1 \cdots \gamma_{k^n+1}$ be the path induced by ω' on G. Then γ is a Hamiltonian cycle (i.e., a simple path traversing each vertex once). The collection of de Bruijn sequences of order $(n+1)$ that extend ω' is isomorphic to the collection of Eulerian paths on G that extend γ.

Let $x_0 \stackrel{\text{def}}{=} \gamma_1 = \gamma_{k^n+1}$ be the common initial and terminal vertex of γ. Then the collection of Eulerian paths of G that extend γ is isomorphic to the set of Eulerian paths of $X_\omega \stackrel{\text{def}}{=} G \setminus E(\gamma)$ that start and end at x_0, which we denote by $\mathcal{E}(\omega)$. Since X_ω is a $(k-1)$-regular connected directed graph whose vertex set has size k^n (see the proof of [1, Lemma 3] for connectedness), we may use Proposition 4.2 to extract a subset $\mathcal{E}'(\omega) \subseteq \mathcal{E}(\omega)$. Pulling this subset back via the appropriate correspondences gives us a set $S'(\omega)$, contained in the set of all de Bruijn sequences of order $(n+1)$ extending ω' (and thus also extending ω), with the following properties:

(i) $\#(S'(\omega)) = (k-1) \cdot (k-2)!^{k^n}$.

(ii) If τ is a sequence of length ℓ_τ extending ω, then the number of sequences in $S'(\omega)$ that extend τ is at most $(k-1) \cdot (k-2)!^{k^n - (\ell_\tau - \ell_\omega - 1)/k}$, where $\ell_\omega = k^n + n - 1$ is the length of ω.

* This phenomenon is related to the fact that we consider non-cyclic de Bruijn sequences instead of cyclic ones: each cyclic de Bruijn sequence $\omega = \omega_1 \cdots \omega_{k^n}$ corresponds to a non-cyclic de Bruijn sequence $\omega_1 \cdots \omega_{k^n} \omega_1 \cdots \omega_{n-1}$ that is longer but has the same number of consecutive substrings. This correspondence makes it obvious that the first $(n-1)$ letters of a non-cyclic de Bruijn word are expected to be the same as the last $(n-1)$ letters. However, by itself this is not a proof, because our definition of non-cyclic de Bruijn sequences did not assume that they were constructed from cyclic ones.

Now we proceed to define a probability measure μ on $F \equiv E^{\mathbb{N}}$ via a random algorithm: start with a fixed de Bruijn sequence $\omega^{(1)}$ of order 1, and if $\omega^{(n)}$ is a de Bruijn sequence of order n, then let $\omega^{(n+1)} \in S'(\omega^{(n)})$ be chosen randomly with respect to the uniform measure on $S'(\omega^{(n)})$, independent of all previous selections. Let ω be the unique infinite sequence that extends all of the finite sequences $\omega^{(n)}$ ($n \in \mathbb{N}$). Then ω is a base b expansion of a unique point $\pi(\omega) \in F$. (The point $\pi(\omega)$ may have a base b expansion other than ω, but there is no other point with base b expansion ω.) We let μ be the probability measure describing the distribution of the random variable $\pi(\omega)$. (The existence of such a μ can be guaranteed, e.g., by the Kolmogorov extension theorem.)

To demonstrate that μ satisfies the hypotheses of the mass distribution principle, we first estimate the measure of cylinder sets of a certain length, then arbitrary cylinder sets, then balls. Here a *cylinder set* is a set of the form $[\tau] = \{\pi(\omega)\colon\ \omega_i = \tau_i\ \ \forall\, i = 1, \ldots, \ell_\tau\}$, where $\tau = \tau_1 \cdots \tau_{\ell_\tau}$ is a finite sequence in the alphabet A. Our first estimate is easy: if $\ell_\tau = k^{n+1} + n$ for some n, then $[\tau]$ is precisely the set of $\pi(\omega)$ in the above construction such that $\omega^{(n+1)} = \tau$, so $\mu([\tau])$ is just the probability that $\omega^{(n+1)} = \tau$, i.e.,

$$\mu([\tau]) = \prod_{i=1}^{n} \frac{1}{\#(S'(\tau^{(i)}))} = \prod_{i=1}^{n} \frac{1}{(k-1)\cdot(k-2)!^{k^i}} \le (k-2)!^{-(k^n + k^{n-1} + n/k)} \tag{4.5}$$

if it is possible that $\omega^{(n+1)} = \tau$, and $\mu([\tau]) = 0$ otherwise. Now consider the more general case where the length of τ satisfies $k^n + n - 1 < \ell_\tau \le k^{n+1} + n$ for some n. Then by (ii) above, $[\tau]$ contains at most $(k-1)\cdot(k-2)!^{k^n - (\ell_\tau - (k^n+n))/k}$ cylinders of length $k^{n+1} + n$. Combining with (4.5) shows that

$$\begin{aligned}\mu([\tau]) &\le (k-1)\cdot \exp_{(k-2)!}\left(k^n - (\ell_\tau - (k^n+n))/k) - \left(k^n + k^{n-1} + n/k\right)\right) \\ &= (k-1)\cdot(k-2)!^{-\ell_\tau/k}.\end{aligned}$$

Here and hereafter we use the notation $\exp_x(y) \stackrel{\text{def}}{=} x^y$.

To apply the mass distribution principle (Lemma 4.4), we now need to relate this measure to the diameter of the cylinder $[\tau]$. Since elements of $[\tau]$ have the first ℓ_τ digits of their base b expansions fixed, the diameter of $[\tau]$ is approximately $b^{-\ell_\tau}$ (to be precise, it is $c \cdot b^{-\ell_\tau}$ for some constant $0 < c \le 1$). Thus

$$\operatorname{diam}([\tau])^{\alpha_k \delta} = c^{\alpha_k \delta} \exp_b\left(-\ell_\tau \frac{\log(k-2)!}{k\log(k)} \frac{\log(k)}{\log(b)}\right) = c^{\alpha_k\delta} \cdot (k-2)!^{-\ell_\tau/k},$$

so

$$\mu([\tau]) \le C \cdot \operatorname{diam}([\tau])^s,$$

where $C = (k-1)\cdot c^{-\alpha_k\delta}$ and $s = \alpha_k \delta$. But any subset of F can be covered by at most two cylinder sets with comparable diameter, so a similar formula holds for arbitrary sets. Thus by Lemma 4.4, we have $\dim_H(S) \ge s = \alpha_k \delta$. ∎

As is evident from Corollary 4.3, we now have to deal with the cases $k = 2$ and $k = 3$ separately, since in those cases the formula (4.4) gives $\alpha_2 = \alpha_3 = 0$, which is not a useful bound. Note that the Cantor ternary set falls into the case $k = 2$, since its set of admissible numerators is $A = \{0, 2\}$.

Proposition 4.5. *If $k = 2$ and ω is a de Bruijn sequence of order $(n-2)$ in A, then the number of de Bruijn sequences of order $(n+1)$ that extend ω is at least $2^{2^{n-2}}$.*

In the case where $k = 3$ and ω is a de Bruijn sequence of order $(n-1)$ in A, then the number of de Bruijn sequences of order $(n+1)$ that extend ω is at least $4^{3^{n-1}}$.

Proof. For convenience, we let $\Delta = 2$ if $k = 3$, and $\Delta = 3$ if $k = 2$; then ω is a de Bruijn sequence of order $(n-\Delta+1)$. The first paragraph of Corollary 4.3 shows that the first $(n-\Delta)$ letters of ω are the same as the last $(n-\Delta)$ letters. So if we extend ω to a word ω' of length $k^{n-\Delta+1}+n$ by letting $\omega_{k^{n-\Delta+1}+i} = \omega_i$ for $i = n-\Delta+1, \dots, n$, then the first n letters of ω' are the same as the last n letters, but no other block of n letters is repeated.

Let G be the de Bruijn graph of order n on A, and let γ be the path induced by ω' on G. The length of γ is $\ell_\gamma = k^{n-\Delta+1}$, and γ is a simple path that starts and ends at the same vertex x_0. As in the proof of Corollary 4.3, we let $X = X_\omega = G \setminus E(\gamma)$, where $E(\gamma)$ is the edge set of γ. The collection of de Bruijn sequences of order $(n+1)$ that extend ω is isomorphic to the collection of Eulerian paths on G that extend γ, which in turn is isomorphic to the collection of Eulerian paths on X_ω that start and end at x_0. By the BEST theorem, the cardinality of this collection is

$$\mathcal{N} \stackrel{\text{def}}{=} \#(\mathcal{T}) \cdot \deg(x_0; X_\omega) \cdot \prod_{x \in V(G)} [\deg(x; X_\omega) - 1]!.$$

If $k = 3$, we complete the proof with the following calculation:

$$\begin{aligned}
\mathcal{N} &\geq \prod_{x \in V(G)} [\deg(x; X_\omega) - 1]! \\
&= \exp_2(\#\{x \in V(G) \colon \deg(x; X_\omega) = 3\}) \\
&= \exp_2\left(\#(V(G)) - \ell_\gamma\right) \\
&= \exp_2\left(3^n - 3^{n-1}\right) \\
&= 4^{3^{n-1}}.
\end{aligned}$$

In the first inequality, we have used Lemma 3.5 and the proof of [1, Lemma 3] to deduce that $\#(\mathcal{T}) \geq 1$.

For the remainder of the proof, we assume that $k = 2$. In this case, the strategy of the above calculation cannot work, since we have $[\deg(x; X_\omega) - 1]! = 1$ for all $x \in V(G)$ and thus $N \leq 2\#(\mathcal{T})$. Instead we must estimate the number of spanning trees in X_ω.

Let S be the set of sequences of length $(n-1)$ that do not occur in ω, and note that $\#(S) = 2^{n-1} - 2^{n-2} = 2^{n-2}$. For each $\tau \in S$, let $E_\tau = \{a\tau b \colon a, b \in A\} \subseteq E(X_\omega)$, where $a\tau b$ is shorthand for $(a\tau)(\tau b)$, the edge from the vertex $a\tau$ to the vertex τb. Note that the sets E_τ ($\tau \in S$) are disjoint.

Lemma 4.6. *If T is a directed spanning tree and $\tau \in S$, then there exists a directed spanning tree $T' \neq T$ such that $T' \setminus E_\tau = T \setminus E_\tau$.*

Proof. By contradiction, suppose that the conclusion of the lemma is false, i.e., that there exists no such spanning tree T'.

Denote the partial order on $V(G)$ induced by the tree T by $<$, i.e., write $x < y$ if there is a path in T from x to y, and write $x \le y$ if either $x < y$ or $x = y$. We write $x <^* y$ if x is a direct descendant of y, i.e., if $xy \in E(T)$. For each $a \in A$, let $f(a) \in A$ be chosen to satisfy $a\tau f(a) \in E(T)$, and let $g(a) = \sigma(f(a))$, where $\sigma\colon A \to A$ is the permutation that swaps the two elements of A. Consider the graph $T' = T \cup \{a\tau g(a)\} \setminus \{a\tau f(a)\}$. Then $T' \neq T$ and $T' \setminus E_\tau = T \setminus E_\tau$, so by the contradiction hypothesis, T' is not a directed spanning tree, which implies that $\tau g(a) \le a\tau$. On the other hand, we have $a\tau <^* \tau f(a)$ since $a\tau f(a) \in T$. Now write $A = \{a, b\}$, $c = f(a)$, and $d = \sigma(c) = g(a)$. Then either $f(b) = c$ or $f(b) = d$, and thus we have one of the following two diagrams:

$$\tau d \le a\tau <^* \tau c >^* b\tau \ge \tau d \quad \text{or} \quad \tau d \le a\tau < \tau c \le b\tau < \tau d.$$

Both diagrams are impossible for directed trees: the left-hand diagram is impossible because if $a\tau$ and $b\tau$ are siblings, then they have no common descendants, while the right-hand diagram is disjoint because it is a nontrivial directed loop. This is the desired contradiction. ∎

It follows from Lemma 4.6 that there exists a function $\phi\colon \mathcal{T} \times S \to \mathcal{T}$ such that for all $T \in \mathcal{T}$ and $\tau \in S$, we have $\phi(T, \tau) \neq T$ and $\phi(T, \tau) \setminus E_\tau = T \setminus E_\tau$.

Now by Lemma 3.5 and the proof of [1, Lemma 5], X has a directed spanning tree T_0 rooted at x_0. Let $(\tau_i)_{i=1}^N$ be an indexing of S, where $N = 2^{n-2}$. Given $\omega \in \{0,1\}^N$, we define recursively

$$T_{\omega,0} = T_0, \qquad T_{\omega,i} = \begin{cases} T_{\omega,i-1}, & \omega_i = 0, \\ \phi(T_{\omega,i-1}, \tau_i), & \omega_i = 1. \end{cases}$$

Then the map $\{0,1\}^N \ni \omega \to T_{\omega,N} \in \mathcal{T}$ is injective. Thus $\mathcal{N} \ge \#(\mathcal{T}) \ge \#\big(\{0,1\}^N\big) = 2^{2^{n-2}}$, which completes the proof. ∎

Corollary 4.7. *Let the notation be as in Theorem 2.1. Suppose that $k \le 3$, and let*

$$\alpha_k = \begin{cases} 1/49, & \text{if } k = 2, \\ (8 \cdot (9\log_4(3) - 1))^{-1}, & \text{if } k = 3. \end{cases} \qquad \Delta = \begin{cases} 3, & \text{if } k = 2, \\ 2, & \text{if } k = 3. \end{cases}$$

Then the Hausdorff dimension of the set

$$\{\pi(\omega) \in F\colon B_\omega \text{ contains an arithmetic progression with gap size } \Delta\}$$

is at least $\alpha_k \delta$.

Proof. Let $B = 2$ if $k = 2$ and $B = 4$ if $k = 3$. Then

$$\alpha_k = \frac{1}{(k^\Delta - 1) \cdot (k^\Delta \log_B(k) - 1)},$$

and Proposition 4.5 can be expressed uniformly as follows: if ω is a de Bruijn sequence of order n in A, then the number of de Bruijn sequences of order $n+\Delta$ that extend ω is at least $\exp_B(k^n)$. We denote the set of all such extensions by $S'(\omega)$.

As in the proof of Corollary 4.3, we define a probability measure μ by a random algorithm: let $\omega^{(1)}$ be a fixed de Bruijn sequence of order Δ, and if $\omega^{(n)}$ is a de Bruijn sequence of order $n\Delta$, then let $\omega^{(n+1)}$ be chosen randomly with respect to the uniform measure on $S'\left(\omega^{(n)}\right)$, independent of all previous selections. As before we let $\omega \in A^{\mathbb{N}}$ be the unique common extension, we let $\pi(\omega) \in F$ be the unique number for which ω is a base b expansion, and we let μ be the probability measure describing the distribution of $\pi(\omega)$.

As before, we first estimate the measure of special cylinders, then arbitrary cylinders, then balls. For ease of notation we fix $k=3$ in this proof; for the case $k=2$ one can apply the substitutions $9 \mapsto 8$, $8 \mapsto 7$, $4 \mapsto 2$, $3 \mapsto 2$, and $2 \mapsto 3$. Fix $n \in \mathbb{N}$ and let τ be a sequence of length 9^n+2n-1 in A. Then

$$\mu([\tau]) \leq \prod_{i=1}^{n-1} \frac{1}{\#(S'(\tau^{(i)}))} \leq \prod_{i=1}^{n-1} \frac{1}{\exp_4(3^{2i})} = \exp_4\left(-\frac{9^n-9}{9-1}\right).$$

Now let τ be an arbitrary sequence of length $9^n+2n-1 < \ell_\tau \leq 9^{n+1}+2(n+1)-1$. There are two ways that we could bound $\mu([\tau])$:

1. Since $[\tau] \subseteq \left[\tau^{(n)}\right]$, we have

$$\mu([\tau]) \leq \mu\left(\left[\tau^{(n)}\right]\right) \leq \exp_4\left(-\frac{9^n-9}{8}\right).$$

2. Since $[\tau]$ can be written as the union of at most $\exp_3\left(9^{n+1}+2(n+1)-1-\ell_\tau\right)$ cylinder sets corresponding to de Bruijn sequences of order $2(n+1)$, we have

$$\mu([\tau]) \leq \exp_3\left(9^{n+1}+2(n+1)-1-\ell_\tau\right) \cdot \exp_4\left(-\frac{9^{n+1}-9}{8}\right).$$

Which of these bounds is better depends on the value of ℓ_τ. Now, as in the proof of Corollary 4.3, we have $\operatorname{diam}([\tau]) = c \cdot b^{-\ell_\tau}$ for some constant c. Fix $0 < s < \alpha_3\delta$. To apply the mass distribution principle, we need to show that

$$\mu([\tau]) \leq C \cdot \operatorname{diam}([\tau])^s,$$

for some constant C. It is enough to show that

$$\mathbf{min}\left(4^{-9^n/8}, \exp_3\left(9^{n+1}+2n-\ell_\tau\right) \cdot 4^{-9^{-n+1}/8}\right) \leq C \cdot b^{-s\ell_\tau} = C \cdot 3^{-t\ell_\tau},$$

possibly with a different value of C, where $t = s/\delta < \alpha_3 < 1$. Equivalently, we need to show that

$$\mathbf{min}\left(4^{-9^n/8} \cdot 3^{t\ell_\tau}, 3^{9^{n+1}} \cdot 9^n \cdot 4^{-9^{n+1}/8} \cdot 3^{(t-1)\ell_\tau}\right) \leq C.$$

Now the first input to the binary operator **min** is an increasing function of ℓ_τ, while the second input is a decreasing function of ℓ_τ. It follows that the largest value the left-hand side can attain is the value attained when the two inputs to **min** are equal, i.e., when

$$4^{-9^n/8} = 3^{9^{n+1}} \cdot 9^n \cdot 4^{-9^{n+1}/8} \cdot 3^{-\ell_\tau},$$

at which point the left-hand side is

$$4^{-9^n/8} \cdot \left(3^{9^{n+1}} \cdot 9^n \cdot 4^{9^n/8 - 9^{n+1}/8}\right)^t.$$

We need this expression to be bounded as $n \to \infty$. Applying the change of variables $x = 9^n$, we need to show that

$$\limsup_{x\to\infty} 4^{-x/8} \cdot \left(3^{9x} \cdot x \cdot 4^{x/8 - 9x/8}\right)^t < \infty.$$

This is true if and only if

$$4^{-1/8} \cdot (3^9 \cdot 4^{-1})^t < 1,$$

which in turn is true if and only if $t < \alpha_3$. This proves that the hypothesis of the mass distribution principle holds for cylinder sets. As in the proof of Corollary 4.3, any subset of F can be covered by at most two cylinder sets with comparable diameter, so the hypothesis of the mass distribution principle holds for arbitrary sets as well. ∎

Combining Corollaries 4.3 and 4.7 yields Theorem 2.1.

Remark 4.8. Either of the strategies used in this proof, the (simpler) strategy for the $k = 3$ case or the (more complicated) strategy for the $k = 2$ case, could have been used (after minor modification) in the case $k \geq 4$ as well, but the resulting bound would have been significantly worse, measured by the fact that the analogues of α_k would not have tended to 1. Similarly, the strategy for the $k = 2$ case could have been used for the $k = 3$ case, again resulting in a worse bound. In general, the principle is that whatever techniques work for one value of k will also work for higher values of k, but may not give very good estimates for higher values of k.

5. Intrinsic Diophantine Approximation

5.1. Diophantine Approximation — a Brief Survey

We first recall some definitions and state some well-known classical theorems:

Definition 5.1. *Let $H\colon \mathbb{Q} \to \mathbb{R}_{>0}$ be a function. We think of H as a "height function", and for all $p \in \mathbb{Z}$ and $q \in \mathbb{N}$, we define the* height *of p/q to be the number $H(p/q)$. We say that a function $\psi\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is a* Dirichlet function (*with respect to the height function H*) *if for every $x \in \mathbb{R} \setminus \mathbb{Q}$ there exist infinitely many rationals p/q such that*

$$\left|x - p/q\right| < \psi(H(p/q)).$$

Historically speaking, the only height function considered on the unit interval $[0, 1]$ was the function $H_{\text{std}}(p/q) = q$, where p and q are chosen in reduced form, i.e., $\gcd(p, q) = 1$. We will refer to this as the *standard* height. It is readily verified that, for example, $\psi_0(q) = 1$ and $\psi_1(q) = 1/q$ are Dirichlet functions with respect to the standard height function and using the terminology of Definition 5.1, Dirichlet's approximation theorem may be stated as follows:

Theorem 5.2. (Dirichlet) $\psi_2(q) = \frac{1}{q^2}$ *is a Dirichlet function with respect to the standard height function.*[*]

For our purposes, although of interest in its own right, an improvement of a Dirichlet function by a multiplicative constant is not significant. More precisely:

Definition 5.3. *We say that a Dirichlet function ψ is* optimal *if there does not exist a Dirichlet function ϕ for which* $\lim_{q\to\infty} \frac{\phi(q)}{\psi(q)} \to 0$.

It is clear that Dirichlet's theorem implies that the Dirichlet functions ψ_0 and ψ_1 defined above are not optimal. The optimality of the function $\psi_2(q) = 1/q^2$ was demonstrated by Liouville, who proved that quadratic irrationals are badly approximable. A real number x is called *badly approximable* if there exists $c(x) > 0$ such that

$$\left|x - p/q\right| > \frac{c(x)}{q^2}, \quad \text{for all } p/q \in \mathbb{Q}.$$

Liouville's result was later significantly improved by Jarník, who proved that the Hausdorff dimension of the set of badly approximable numbers is 1.

5.2. Iterated Function Systems, Limit Sets, and Hausdorff Dimension

Let $k \geq 2$ be an integer. In what follows, we shall consider a finite famiily $(S_i)_{i=1}^k$ of contracting similarities on the unit interval $I = [0, 1]$. This means that for every $1 \leq i \leq k$, the map $S_i \colon I \to I$ satisfies

$$|S_i(x) - S_i(y)| = c_i|x - y|, \qquad \forall x, y \in I,$$

for some $0 < c_i < 1$. We shall call such a family of similarities an *Iterated Function System* or IFS. A nonempty compact set $F \subseteq I$ is said to be the *attractor* or the *limit set* of the IFS if

$$F = \bigcup_{i=1}^{k} S_i(F).$$

It is well known (see, e.g., [8, Chapter 9]) that the attractor F exists and is unique. Furthermore, if there exists a bounded nonempty open set U such that

$$\bigcup_{i=1}^{k} S_i(U) \subseteq U,$$

[*] In fact, Dirichlet's theorem furnishes a similar result for all dimensions d. It was recently pointed out to us by Y. Bugeaud that the one-dimensional version of this result is actually much older, coming directly from the theory of continued fractions (see, e.g., [15, displayed equation on p. 28]). Nevertheless, we call the theorem "Dirichlet's theorem" so as to conform to usual practice.

with the union disjoint, then the IFS is said to satisfy the *open set condition*. In this case, the Hausdorff dimension of the attractor is equal to the unique solution $s > 0$ of the equation

$$\sum_{i=1}^{k} c_i^s = 1. \tag{5.1}$$

We say that the IFS $(S_i)_{i=1}^k$ satisfies the *strong separation condition* if

$$S_i(F) \cap S_j(F) = \varnothing,$$

for all $i \neq j$, where F is the attractor.[†]

A particularly important example of an iterated function system is the system

$$S_i(x) = \frac{i+x}{b}, \quad i \in C(b) \stackrel{\text{def}}{=} \{0, \ldots, b-1\}, \tag{5.2}$$

where $b \geq 2$ is fixed. This system satisfies the open set condition $\big($with $U = (0,1)\big)$ but not the strong separation condition, and its attractor is the entire interval I. In some sense this IFS encodes the base b expansion(s) of any number in the interval $[0,1]$, since the number

$$x = \pi(\omega) = 0.\omega_1\omega_2\cdots \text{ (base } b) \ = \sum_{i=1}^{\infty} \frac{\omega_i}{b^i}$$

can be written as

$$x = \lim_{n\to\infty} S_{\omega_1} \circ \cdots \circ S_{\omega_n}(0).$$

By looking at subsystems of the system (5.2), we can find IFSes whose limit sets can be described in terms of base b expansions. Fix $A \subseteq C(b)$, and consider the subsystem of (5.2) consisting of the similarities $(S_i)_{i\in A}$. We call such a subsystem a *base b IFS*. Its limit set is precisely the set of all numbers in $[0,1]$ that have at least one base b expansion whose digits all lie in A, i.e.,

$$F = \left\{ x \in [0,1] \colon \exists \omega \in A^{\mathbb{N}} \text{ with } x = \sum_{i=1}^{\infty} \frac{\omega_i}{b^i} \right\}. \tag{5.3}$$

For example, if $b = 3$ and $A = \{0,2\}$, then F is the standard Cantor ternary set, i.e., the set of all numbers in $[0,1]$ that have at least one base 3 expansion containing only the digits 0 and 2.

It follows directly from (5.1) that the Hausdorff dimension of the base b IFS corresponding to an alphabet $A \subseteq C(b)$ is precisely $\log \#(A) / \log(b)$.

We remark that it is easy to check whether a base b IFS satisfies the strong separation condition:

Observation 5.4. The base b IFS defined by the alphabet $A \subseteq C(b)$ satisfies the strong separation condition if and only if at least one of the following is true:

(1) $0 \notin A$.

[†] Note that the strong separation condition implies (but is not implied by) the open set condition.

(2) $b-1 \notin A$.

(3) A does not contain any pair of consecutive integers.

If a base b IFS satisfies the strong separation condition, then every element of its limit set F has exactly one base b expansion whose digits come from A. In this case, there is no ambiguity about talking about "the base b expansion" of a number in F, since we understand that if there is more than one base b expansion, then we are talking about the one whose digits come from A.

5.3. Intrinsic Approximation on Limit Sets

Let $F \subseteq \mathbb{R}$ be a closed set, which we will think of as a fractal. The field of *intrinsic Diophantine approximation* is concerned with finding rational approximations to an irrational number $x \in F$ by rational numbers that lie *on* the fractal F. Thus Mahler's first question is about intrinsic approximation on the Cantor set. More generally, one may ask about intrinsic approximation on the attractor of any similarity IFS. This leads to the following definition:

Definition 5.5. *Let $F \subseteq \mathbb{R}$ be a closed set, and let $H\colon F \cap \mathbb{Q} \to \mathbb{R}_{>0}$ be a height function. We say that a function $\psi\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is an* intrinsic Dirichlet function on F *(with respect to the height function H) if for every $x \in F \setminus \mathbb{Q}$ there exist infinitely many rationals $p/q \in F \cap \mathbb{Q}$ such that*

$$\left|x - p/q\right| < \psi(H(p/q)).$$

Optimality of intrinsic Dirichlet functions can be defined in the same way as in Definition 5.3.

We have the following result:

Proposition 5.6. ([4, Corollary 2.2]) *Let F be the limit set of a base b IFS, and let δ be the Hausdorff dimension of F. Then for all $x \in F$, there exist infinitely many rational numbers $p/q \in F$ ($p \in \mathbb{Z}$, $q \in \mathbb{N}$) such that*

$$\left|x - p/q\right| < \frac{1}{q(\log_b q)^{1/\delta}}.$$

In other words, the function $\psi_*(q) = \left(q \cdot (\log_b q)^{1/\delta}\right)^{-1}$ is an intrinsic Dirichlet function on F for the standard height function.

6. The Symbolic Height Function

Let F be the limit set of a base b IFS satisfying the strong separation condition, and fix a rational number $r \in F \cap \mathbb{Q}$. It is well known that the base b expansion of r is preperiodic, i.e.,

$$r = 0.\omega_1 \cdots \omega_i \overline{\omega_{i+1} \cdots \omega_{i+j}} \quad (\text{base } b), \tag{6.1}$$

for some $i \geq 0$, $j \geq 1$, and $\omega_1, \ldots, \omega_{i+j} \in A$. Here the bar indicates that the string $\omega_{i+1} \cdots \omega_{i+j}$ is infinitely repeated. Rewriting the right-hand side as a sum of fractions yields

$$\begin{aligned} r &= \frac{\omega_1 \cdots \omega_i}{b^i} + \sum_{m=1}^{\infty} \frac{\omega_{i+1} \cdots \omega_{i+j}}{b^{i+mj}} \\ &= \frac{\omega_1 \cdots \omega_i}{b^i} + \frac{\omega_{i+1} \cdots \omega_{i+j}}{b^i} \cdot \frac{1/b^j}{1 - 1/b^j} \\ &= \frac{\omega_1 \cdots \omega_i}{b^i} + \frac{\omega_{i+1} \cdots \omega_{i+j}}{b^i} \cdot \frac{1}{b^j - 1}, \end{aligned}$$

where $\omega_1 \cdots \omega_i$ and $\omega_{i+1} \cdots \omega_{i+j}$ are integers that have been written in base b. Adding the two resulting fractions together, we end up with a (complicated) expression whose denominator is $b^i\left(b^j - 1\right)$. Further cancellations may or may not be possible, but we can *always* write the rational number as a fraction of two integers, the denominator of which is $b^i\left(b^j - 1\right)$.

This fact leads to a natural height function on $F \cap \mathbb{Q}$ related to the base b structure of the fractal F:

$$H_{\text{sym}}(r) \stackrel{\text{def}}{=} b^i \cdot \left(b^j - 1\right), \tag{6.2}$$

where the indices i and j are the smallest integers such that r can be written in the form (6.1). The function H_{sym} is called the *symbolic* height function. It was studied in a more general context in [11]. Notice the symbolic height of a rational number may not be the same as its standard height (i.e., its denominator in reduced form). For example, the rational number $0.\overline{20}_3$ in the Cantor ternary set is equal to $\frac{3}{4}$, so its standard height is 4. Nonetheless, the symbolic height of $0.\overline{20}_3$ is $3^0 \cdot \left(3^2 - 1\right) = 8$. It should be thought of as the denominator resulting from the following calculation:

$$\begin{aligned} 0.\overline{20}_3 &= \frac{20_3}{3^0} \sum_{m=1}^{\infty} \left(\frac{1}{3^2}\right)^m \\ &= \frac{6}{1} \cdot \frac{1/3^2}{1 - 1/3^2} \\ &= \frac{6}{1} \cdot \frac{1/9}{8/9} \\ &= \frac{6}{8}. \end{aligned}$$

Although more cancellation is possible at the end of this calculation, this will not always be the case,[‡] so in a principled way we have stopped reducing the fraction here. The calculation illustrates the fact that the symbolic height of a rational number r can be thought of as a "symbolic denominator", i.e., the denominator of a certain representation of r as the quotient of two integers. The numerator of this representation can be thought of as a "symbolic numerator" (in the above example the symbolic

[‡] For example, the fraction at the end of the calculation $0.2\overline{70}_9 = \frac{2_9}{9} + \frac{70_9}{9} \cdot \frac{1}{9^2-1} = \frac{2 \cdot 80 + 7 \cdot 9}{9 \cdot 80} = \frac{223}{720}$ is already in reduced form.

numerator would be 2), but as usual, for purposes of Diophantine approximation it is simpler to just work with the denominator. Note that the standard height is by definition smaller than the symbolic one, since we have $p_{\text{std}}/q_{\text{std}} = p_{\text{sym}}/q_{\text{sym}}$, but the left-hand side is in reduced form.

We remark that heuristically, if we are given two rational numbers r_1 and r_2, and we are told that r_1 lies in the limit set of a base b IFS, but we are not told anything about r_2, then we should expect the (multiplicative) discrepancy between the standard height and the symbolic height to be smaller for r_1 than for r_2. This is because if we choose the numerator and denominator of a rational randomly, then the numbers i and j satisfying (6.1) may be comparable to the standard height of the rational (meaning that the symbolic height is an exponential function of the standard height), but the number would be exceedingly unlikely to lie in any base b limit set, since its digits would essentially be random. By contrast, if we choose the digits of a rational randomly out of a fixed alphabet A (with a fixed period and preperiod), then the amount of cancellation we expect to see in the symbolic representation of the rational will be much smaller, so the standard height and symbolic height will be relatively close. More heuristics regarding the relation between the symbolic height function and the standard one were discussed in [11].

One reason the symbolic height function is interesting is that it naturally shows up in the proofs of results regarding the standard height function. For example, the proof of Proposition 5.6 can easily be modified to bound $|x - p/q|$ in terms of the symbolic height of p/q rather than the standard height:

Proposition 6.1. ([4, Proof of Corollary 2.2]) *Let F be the limit set of a base b IFS, and let δ be the Hausdorff dimension of F. Then for all $x \in F$, there exist infinitely many rational numbers $r = p_{\text{sym}}/q_{\text{sym}} \in F$ such that*

$$\left|x - p/q\right| < \frac{1}{q_{\text{sym}} \left(\log_b q_{\text{sym}}\right)^{1/\delta}}.$$

In other words, the function $\psi_*(q) = \left(q \cdot (\log_b q)^{1/\delta}\right)^{-1}$ is an intrinsic Dirichlet function on F for the symbolic height function.

In fact, the proof of [4, Corollary 2.2] essentially proceeds by first proving Proposition 6.1 and then using the inequality $H_{\text{std}} \leq H_{\text{sym}}$ to deduce Proposition 5.6. It appears extremely difficult to prove any improvement (either for all points or only for some) of Proposition 5.6 for the standard height without just proving the same bound for the symbolic height. So in some way, the symbolic height is measuring the "strength of our techniques".

Although the symbolic height function is motivated in terms of the standard height function, it can also be analyzed on its own terms. For example, we can ask whether the intrinsic Dirichlet function ψ_* appearing in Proposition 6.1 is optimal for the symbolic height function. This is the same (cf. [12, §2.1]) as asking whether there exist any points in F that are badly symbolically approximable with respect to ψ_*:

Definition 6.2. (Special case of [11, Definition 4.7]) *Let F be a base b limit set, and let δ denote the Hausdorff dimension of F. A number $x \in F$ is called* badly

symbolically approximable (with respect to ψ_*) *if there exists* $\kappa > 0$ *such that for every* $r = p_{\mathrm{sym}}/q_{\mathrm{sym}} \in F \cap \mathbb{Q}$*, we have*

$$|x - r| \geq \frac{\kappa}{q_{\mathrm{sym}}(\log_b q_{\mathrm{sym}})^{1/\delta}}. \tag{6.3}$$

Theorem 6.3. (Corollary of [11, Lemma 4.9]; or see below) *Let F be the limit set of a base b IFS satisfying the strong separation condition. Then any $x \in F$ whose base b expansion is uniformly de Bruijn is badly symbolically approximable.*

Combining with Theorem 2.1 gives:

Corollary 6.4. *With F as above, the set of badly symbolically approximable points has dimension at least $\alpha_k \delta > 0$, where*

$$\alpha_k = \begin{cases} 1/49, & k = 2, \\ (8 \cdot (9\log_4(3) - 1))^{-1}, & k = 3, \\ \frac{\log(k-2)!}{k\log(k)}, & k \geq 4. \end{cases}$$

In particular, the intrinsic Dirichlet function ϕ_* appearing in Proposition 6.1 is optimal.

We remark that the optimality assertion follows directly from combining Theorem 6.3 with [1, Corollary 7]; Theorem 2.1 is not needed.

In contrast to Proposition 6.1, Theorem 6.3 and Corollary 6.4 are weaker than their (unproven) analogues for the standard height function. This is because while Proposition 6.1 is about finding good approximations to points, in Theorem 6.3 and Corollary 6.4 we show that for certain points, good approximations cannot exist. But the inequality $H_{\mathrm{std}} \leq H_{\mathrm{sym}}$ means that the quality of an approximation is better according to the standard height than according to the symbolic height, which yields the appropriate implications.

We remark that Theorem 6.3 is only a one-way implication: there may be (and almost certainly are) badly symbolically approximable numbers whose base b expansions are not uniformly de Bruijn. A combinatorial characterization of the base b expansions of badly symbolically approximable numbers was given in [11, Lemma 4.9]. As a consequence of the one-sidedness of the implication, Theorem 6.3 yields a lower bound on the dimension of the set of badly symbolically approximable points but not an upper bound. In fact, we believe that there is no nontrivial upper bound: we conjecture that the Hausdorff dimension of the set of badly symbolically approximable points of any base b limit set F is equal to the Hausdorff dimension of F. This conjecture is motivated by other situations in Diophantine approximation where the dimension of the set of badly approximable points has always turned out to be full. However, Theorem 2.1 shows that this conjecture cannot be proven using uniformly de Bruijn sequences.

Although Theorem 6.3 is a consequence of the much more general result [11, Lemma 4.9], we prove it here for completeness and ease of exposition.

Proof of Theorem 6.3. Let $x \in F$ be a number whose base b expansion, which we denote by ω, is uniformly de Bruijn. Let ℓ denote the size of the largest gap in the set

B_ω defined by (2.1). Fix $r \in F \cap \mathbb{Q}$, and let the representation $r = 0.\tau_1 \cdots \tau_i \overline{\tau_{i+1} \cdots \tau_j}$ be chosen so as to minimize i and j. Then the symbolic height of r, as defined in (6.2), is $q_{\text{sym}} = b^i\big(b^{j-i} - 1\big) \leq b^j$. Since the IFS defining F is assumed to satisfy the strong separation condition, the distance between x and r is comparable to b^{-m}, where m is the largest index for which $\omega_i = \tau_i$ for all $i \leq m$. In fact, a careful analysis shows that $|x - r| \geq b^{-(m+2)}$, though the precise constant factor is not relevant. We claim that if $j \geq \ell$, then

$$b^{-m} \geq \frac{b^{-\ell}}{b^j j^{1/\delta}}, \tag{6.4}$$

which demonstrates that (6.3) holds with $\kappa = b^{-(\ell+2)}$. We now separate into two cases:

Case 1: $m \leq j + \ell$. In this case, we have

$$b^{-m} \geq b^{-j} b^{-\ell} \geq \frac{b^{-\ell}}{b^j j^{1/\delta}},$$

as required.

Case 2: $m > j + \ell$. In this case, by the mth letter, the sequence τ will have already begun to repeat. The longest repeated string in the sequence $\tau_1 \cdots \tau_m$ is $\tau_{i+1} \cdots \tau_{m-(j-i)}$ $= \tau_{j+1} \cdots \tau_m$. Note that although the two sides of this equation represent distinct instances of the same string as a substring of $\tau_1 \cdots \tau_m$, the two instances may overlap with each other; this happens if and only if $m > 2j - i$. For the purposes of our calculations, it does not matter whether these two instances overlap or not.

By the definition of m, we have $\omega_1 \cdots \omega_m = \tau_1 \cdots \tau_m$, so ω also has a repeated string $\omega_{i+1} \cdots \omega_{m-(j-i)} = \omega_{j+1} \cdots \omega_m$ of length $(m - j)$ occurring in the first m letters. On the other hand, by the definition of ℓ, there exists $m - j - \ell < n \leq m - j$ such that $n \in B_\omega$, which implies that ω has no repeated string of length n occurring in the first $k^n + n - 1$ letters of ω. Since $n \leq m - j$, it follows that $m > k^n + n - 1$, and thus

$$k^n \leq m - n < j + \ell \leq 2j.$$

Since $k \geq 2$ and $n \geq m - j - \ell + 1$, this implies

$$k^{m-j-\ell} \leq j.$$

Raising both sides to the power of $1/\delta$ gives

$$b^{m-j-\ell} \leq j^{1/\delta},$$

and rearranging gives (6.4). ∎

Acknowledgments. The first-named author was supported in part by the Simons Foundation grant #245708. The third-named author was supported in part by the EPSRC Programme Grant EP/J018260/1. The authors would like to thank Joseph Kung for valuable comments on an earlier version of the paper, and Jonah Ostroff for introducing us to the notion of de Bruijn sequences. The authors thank the anonymous referee for valuable comments.

References

1. Becher, V., Heiber, P.: On extending de Bruijn sequences. Inform. Process. Lett. 111(18), 930–932 (2011)
2. Bernik, V., Dodson, M.: Metric Diophantine Approximation on Manifolds. Cambridge Tracts in Math., Vol. 137. Cambridge University Press, Cambridge (1999)
3. Broderick, R., Fishman, L., Kleinbock, D., Reich, A., Weiss, B.: The set of badly approximable vectors is strongly C^1 incompressible. Math. Proc. Cambridge Philos. Soc. 153(2), 319–339 (2012)
4. Broderick, R., Fishman, L., Reich, A.: Intrinsic approximation on Cantor-like sets, a problem of Mahler. Mosc. J. Combin. Number Theory 1(4), 291–300 (2011)
5. Bugeaud, Y.: Approximation by Algebraic Numbers. Cambridge Tracts in Mathematics, Vol. 160. Cambridge University Press, Cambridge (2004)
6. Bugeaud, Y., Durand, A.: Metric Diophantine approximation on the middle-third Cantor set. J. Eur. Math. Soc. (JEMS) 18(6), 1233–1272, (2016)
7. Einsiedler, M., Fishman, L., Shapira, U.: Diophantine approximations on fractals. Geom. Funct. Anal. 21(1), 14–35 (2011)
8. Falconer, K.: Fractal Geometry: Mathematical Foundations and Applications. John Wiley & Sons, Ltd., Chichester (1990)
9. Fishman, L.: Schmidt's game on fractals. Israel J. Math. 171(1), 77–92 (2009)
10. Fishman, L., Simmons, D.: Intrinsic approximation for fractals defined by rational iterated function systems - Mahler's research suggestio. Proc. Lond. Math. Soc. (3) 109(1), 189–212 (2014)
11. Fishman, L., Simmons, D.: Extrinsic Diophantine approximation on manifolds and fractals. J. Math. Pures Appl. (9) 104(1), 83–101 (2015)
12. Fishman, L., Simmons, D., Urbański, M.: Diophantine approximation in Banach spaces. J. Théor. Nombres Bordeaux 26(2), 363–384 (2014)
13. Kleinbock, D., Lindenstrauss, E., Weiss, B.: On fractal measures and Diophantine approximation. Selecta Math. 10(4), 479–523 (2004)
14. Kleinbock, D., Weiss, B.: Badly approximable vectors on fractals. Israel J. Math. 149, 137–170 (2005)
15. Legendre, A.-M.: Essai sur la théorie des nombres (Essay on number theory). Second Edition. Cambridge University Press, Cambridge (2009)
16. Mahler, K.: Some suggestions for further research. Bull. Austral. Math. Soc. 29(1), 101–108 (1984)
17. Martin, A.: A Course in Enumeration. Graduate Texts in Mathematics, Vol. 238. Springer, Berlin (2007)
18. Weiss, B.: Almost no points on a Cantor set are very well approximable. R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. 457(2008), 949–952 (2001)