Kuzmin's Theorem on Algebraic Numbers

Jeff Law

February 19, 2003

Mathematics Department
Princeton University
Princeton, NJ 08544

1 Measure Theory of Continued Fractions

The study of continued fractions is not young, but even after much investigation, there are still only a few interesting approaches to the analysis of the apparatus. We will be investigating these structures through their measure theory. The continued fraction is not an intuitive tool, and the relationships between its properties and its elements must be understood before it can be utilized. Since the properties of real numbers are preserved through addition of an integer, for the sake of simplicity we can, without loss of generality, assume that the numbers in our investigation fall in the unit interval I=(0,1). This means that in the expansion of any real number represented by

$$\alpha = [a_0, a_1, a_2, \dots] = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \dots}} \tag{1.1}$$

where $n \in \mathbb{N}$ we will set the element

$$a_0 = 0 \tag{1.2}$$

and we can simply investigate

$$\alpha - [\alpha] = [a_1, a_2, \dots] = \frac{1}{a_1 + \frac{1}{a_2 + \dots}} \tag{1.3}$$

where [x] is just the greatest integer less than x.

Now consider the continuum on the unit interval. We know that every real number α has a non-trivial unique continued fraction expansion. By non-trivial we mean that

$$\alpha = [a_1, a_2, \dots, a_n] = [a_1, a_2, \dots, a_n - 1, 1] \tag{1.4}$$

only counts as one expansion. Because of this property of uniqueness, by investigating all possible continued fraction expansions both finite, and infinite, we will completely cover the continuum. If we set

$$a_1 = k_1 \tag{1.5}$$

where k_1 can take any natural number value, then we limit the possible real numbers that this expansion can represent to an interval. We will call this an interval of rank 1. For example, if we set $a_1 = k_1$, then the continued fraction expansions of all numbers bound by this are of the form

$$[k_1, r_{n+1}] \tag{1.6}$$

where r_{n+1} is itself a continued fraction which takes all possible real values from 1 to ∞. Then any real number α which is bound by the condition that $a_1 = 1$ is confined to the interval $(\frac{1}{2},1]$. Furthermore, every real number α which is bound by the condition that $a_1 = 1$ exists in the interval $(\frac{1}{2},1]$.

If $a_n = k_n$ and for some subsequence $\{m\}$ of $\{n\}$, where $m < n$, we set $\{a_m = k_m\}$, we will denote the interval by $E_{n,k_n}(\{k_m\})$. If the interval is determined solely by setting $a_n = k_n$ and all other terms a_m are left arbitrary, we will simply denote it by E_{n,k_n}. Then the measure of the set of real numbers in the unit interval for which the element $a_1 = 1$, $E_{1,1}$ is

$$\mu(E_{1,1}) = \frac{1}{2} \tag{1.7}$$

Similarly, if we consider the general case of $a_1 = k_1$, where k_1 is any natural number, then the real numbers represented by these expansions must belong to the interval

$$(\frac{1}{k_1+1}, \frac{1}{k_1}] \tag{1.8}$$

The measure of this set of real numbers, E_{1,k_1} is

$$\mu(E_{1,k_1}) = \frac{1}{k_1(k_1+1)} \tag{1.9}$$

Since the measure of the entire unit interval is $\mu(I) = 1$, and the subsets E are intervals of positive measure, we can treat them as proportions or percentages of the whole unit interval. Therefore we can conclude that the fraction of real numbers in the unit interval for which $a_1 = 1$ is

$$\frac{\mu(E_{1,1})}{\mu(I)} = \frac{1}{2} \tag{1.10}$$

and, more generally, that the fraction of real numbers in the unit interval for which $a_1 = k_1$ is

$$\frac{\mu(E_{1,k_1})}{\mu(I)} = \frac{1}{k_1(k_1+1)} \tag{1.11}$$

Note that, as mentioned earlier, the properties of real numbers are preserved through the addition or subtraction of integers. We can conclude from this that these proportionalities hold not only for the unit interval, but for any interval extending as large as the entire real number line.
Also note that

$$\sum_{k_1=1}^{\infty} \mu(E_{1,k_1}) = 1 \tag{1.12}$$

Now that we have a good understanding of what continued fractions and the measures of the intervals that are governed by the element a_1, the next obvious step is to investigate the behavior of the intervals due to restrictions on arbitrary elements a_n. If we consider an arbitrary element a_n and set $a_n = k_n$ where k_n is any natural number, then we have to consider many intervals, determined by the values of $a_1, \ldots, a_{n-1}$. Even in the basic case of $a_2 = k_2$, we have the cases of

$$[a_1, a_2, a_3, \ldots] = [1, k_2, r_3], [2, k_2, r_3], [3, k_2, r_3], \ldots, [n, k_2, r_3] \tag{1.13}$$

each with measure

$$\frac{\mu(E_{2,k_2}(k_1))}{\mu(I)} = \frac{1}{(k_1k_2 + 1)(k_1(k_2 + 1) + 1)} \tag{1.14}$$

where $a_1 = k_1$ and $a_2 = k_2$ These intervals of real numbers which are restricted by $a_n = k_n$ are dispersed throughout the unit interval, each with positive measure. We will call these intervals of rank n.

However, how does one determine the measure of the entire set, the union of these disjoint intervals, of real numbers in the unit interval which are bound by the condition that $a_n = k_n$? We must first determine the endpoints of the interval, E_{n,k_n}. We begin by defining the n^{th} element convergent of our real number

$$\alpha = [a_1, a_2, \ldots, a_n, r_{n+1}] = [k_1, k_2, \ldots, k_n, r_{n+1}] \tag{1.15}$$

by the rational number

$$\frac{p_n}{q_n} = [k_1, k_2, \ldots, k_n] \tag{1.16}$$

where

$$p_{n+1} = a_{n+1}p_n + p_{n-1} \tag{1.17}$$

$$q_{n+1} = a_{n+1}q_n + q_{n-1} \tag{1.18}$$

Then writing our real number α in this form we have

$$\alpha = \frac{a_{n+1}p_n + p_{n-1}}{a_{n+1}q_n + q_{n-1}} = \frac{r_{n+1}p_n + p_{n-1}}{r_{n+1}q_n + q_{n-1}} \tag{1.19}$$

where, as above, r_{n+1} ranges from 1 to ∞. Since we are holding all other terms constant, and r_{n+1} is greater than 1, α as a function of r_{n+1} (1.18), is monotonic. Therefore the endpoints of the interval can be determined by considering the two extreme values, the max and min, of r_{n+1}. More clearly, we can find the points $\alpha(r_{n+1})$ when $r_{n+1} = 1$ and when $r_{n+1} \to \infty$ thus determining the span of the interval of possible values of α when $a_n = k_n$. When $r_{n+1} = 1$,

$$\frac{r_{n+1}p_n + p_{n-1}}{r_{n+1}q_n + q_{n-1}} = \frac{p_n + p_{n-1}}{q_n + q_{n-1}} \tag{1.20}$$

When $r_{n+1} \to \infty$

$$\frac{r_{n+1}p_n + p_{n-1}}{r_{n+1}q_n + q_{n-1}} \to \frac{p_n}{q_n} \tag{1.21}$$

Therefore expressions (1.19) and (1.20) are the endpoints of the interval E_{n,k_n} and the measure is

$$\mu(E_{n,k_n}) = \left| \frac{p_n + p_{n-1}}{q_n + q_{n-1}} - \frac{p_n}{q_n} \right| = \frac{1}{q_n(q_n + q_{n-1})} \tag{1.22}$$

Before we move on to the next section, two points must be brought to attention. First note that positive measures have the property such that if a set A is the countable union of disjoint open sets A_i, then the measure of A is the sum of the measures of A_i. More clearly,

$$\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i) \tag{1.23}$$

Also note that the sum of measures of E_{n,k_n} over k_n ranging from $1 \to \infty$ yields a measure of the unit interval. Each n^{th} element ranges from $1 \to \infty$ filling up the entire interval restricted by $a_{n-1} = k_{n-1}$ until every interval in the unit interval has been filled, hence measure 1.

2 Kuzmin's Theorem

Kuzmin's Theorem addresses the problem of finding the measure of the union of intervals that are bounded by the condition that $a_n = k_n$. As one can see, the construction of measure theory developed in the previous section quickly becomes cumbersome as n grows large. Kuzmin's theorem succeeds in bounding the measure of the union of all intervals such that $a_n = k_n$. His result shows that

$$\left| \mu(E_{n,k_n}) - \frac{\ln(1 + \frac{1}{k_n(k_n+2)})}{\ln 2} \right| < \frac{A}{k_n(k_n+1)} e^{-\lambda\sqrt{n-1}} \tag{2.1}$$

where

$$\frac{A}{k_n(k_n+1)}e^{-\lambda\sqrt{n-1}} \tag{2.2}$$

is obviously the approximation term to the measure of the set E_{n,k_n}.
A and λ are constants which have been proven to exist, but of value still to be determined.
It is clear that as $n \to \infty$, the approximation term

$$\frac{A}{k_n(k_n+1)}e^{-\lambda\sqrt{n-1}} \to 0 \tag{2.3}$$

and

$$\mu(E_{n,k_n}) \to \frac{\ln(1+\frac{1}{k_n(k_n+2)})}{\ln 2} = \log_2\left(1+\frac{1}{k_n(k_n+2)}\right) \tag{2.4}$$

Therefore as n grows sufficiently large, the approximation term decays at a very quick rate and the measure of E_{n,k_n} converges to a term independent of n and solely dependent on the value of k_n

This is an extremely surprising result as it shows that the probability of finding a value k_n in the a_n^{th} place when considering the continuum in the unit interval, does not depend on n at all as n grows large. In our construction of the measure theory of continued fractions, the rank, n, of the interval was essential in determining the measure of it. This result shows that the rank quickly becomes less significant as n grows large and the probability that $a_n = k$ for any k ranging from $1 \to \infty$ depends only on the value of k when investigating the n^{th} convergent of any real number for sufficiently large n, or in other words, sufficient accuracy. From now on, when the value of k does not relate to any particular a_n, we will denote it k_i.

3 Kuzmin's Theorem on Algebraic Numbers

It is clear from the construction of Kuzmin's result (2.1) from the measure theory of continued fractions that it is a necessary condition that the set of numbers being investigated have positive measure. In our development of the measure theory, we confined ourselves to the continuum of the unit interval which has measure 1. We then restricted ourselves to sub-intervals by imposing conditions on the various elements a_n. Since we considered every possible continued fraction expansion, and every real number has a unique continued fraction expansion other than the trivial case of

$$\alpha = [a_1, a_2, \ldots, a_n] = [a_1, a_2, \ldots, a_n - 1, 1] \tag{3.1}$$

we completely covered all of the real numbers in each interval, hence each interval had positive measure. This allowed us to develop a system of determining the fraction of real numbers with element $a_n = k_n$ in the unit interval by considering the fraction of measures

$$\frac{\mu(E_{n,k_n})}{\mu(I)} \tag{3.2}$$

which depends on the fact that the measure of the set of numbers that were considered, I, is finite and positive.

Kuzmin's result holds for all sets of positive measure, but the construction does not provide much insight for the consideration of sets of numbers of measure 0. This is evident since if a set of real numbers has 0 measure, call it A, then each subset of A also has measure 0 and finding proportions of measures would have nothing interesting to offer.

For example, consider the set of rational numbers $Q_I \subset I$. $\mu(Q_I) = 0$. Now consider the elements $q_{m,k_m} \in Q_I$ that have continued fraction expansion with $a_m = k_m$ where m is less than the order, n, of the continued fraction expansion, which is finite since q is rational. The set of $\{q_{m,k_m}\}$ forms a subset of Q_I and thus, is a positive fraction of the total number of elements in Q_I.
However,

$$\mu\{q_{m,k_m}\} = 0 \tag{3.3}$$

and

$$\mu(Q_I) = 0 \tag{3.4}$$

If we depend on measure theory to determine the proportion of $\{q_{m,k_m}\}$ in the whole set of Q_I, then we find

$$\frac{\mu\{q_{m,k_m}\}}{\mu(Q_I)} = \frac{0}{0} \tag{3.5}$$

which tells us nothing.

We do not always have to depend on the measure theory to compare cardinalities of sets, but it still remains that Kuzmin's Theorem only offers information on some set A of finite positive measure, up to some subset of A of 0 measure. Kuzmin's theorem does not necessarily exclude sets of measure 0. Just because it does not affirm anything about sets of measure 0 does not mean that the result still might not hold for them. We know that Kuzmin's Theorem holds for sets up to measure 0, but the actual structure of this set is unknown. This paper investigates at what point sets of measure

0 are general enough, unrestricted enough, that they can be treated like the continuum with respect to Kuzmin's predictions. In particular, we will be looking at algebraic numbers of different forms.

Since the set of algebraic numbers is countable, it must have measure 0, and then its complement, the set of transcendentals, must have measure $\mu(X)$. In the case that we have been considering, where our space $X = I$, the measure of the set of transcendentals in I must be 1. Let us consider algebraic numbers, x, of the form

$$x^n - P = 0 \tag{3.6}$$

where n is a natural number and P is prime. Algebraic numbers of this form are convenient to study because of the following property.

Lemma 3.1. *If P is a prime greater than 1, then any n^{th} root of P, $\sqrt[n]{P}$, for $n > 1$, is irrational.*

Proof:
Assume that $(P)^{\frac{1}{n}}$ is rational, hence

$$\sqrt[n]{P} = (P)^{\frac{1}{n}} = \frac{x}{y} \tag{3.7}$$

where $x, y \in \mathbb{Z}$ and $(x, y) = 1$
Then

$$P = \frac{x^n}{y^n} \tag{3.8}$$

Since P is an integer,

$$P = \frac{x^n}{y^n} \in \mathbb{Z} \tag{3.9}$$

Then since $(x, y) = 1$, $(x^n, y^n) = 1$, we can, without loss of generality, assume that $y^n = \pm 1$, and furthermore $y = \pm 1$
Then

$$P = \pm x^n \tag{3.10}$$

and $x|P$.
Since P is prime, $x = 1$ or $x = P$.
If $x = 1$, then

$$P = \frac{x^n}{y^n} = \pm 1 \tag{3.11}$$

which contradicts our initial assumption that $P > 1$
If $x = P$, then

$$P = \pm x^n = \pm P^n \tag{3.12}$$

which also results in $P = \pm 1$.
Therefore, $(P)^{\frac{1}{n}}$ is not rational, hence irrational. This concludes our proof.

Because of this property of irrationality, investigation of algebraic numbers of this form ensures that we do not encounter any rational numbers, since rational numbers have finite continued fraction expansions that will restrict the requirement that n grows sufficiently large in Kuzmin's Theorem.

Another restriction on the form of algebraic numbers that we are investigating is that $n \geq 3$. The trivial case of $n = 1$ simply yields that $x = P$ and this is quite uninteresting. We will show that the case of $n = 2$, or the case of quadratic irrationals, produces only periodic continued fractions, and hence, a set of 0 measure. Consider this lemma first.

Lemma 3.2. [1]

$$\frac{1}{q_n q_{n+2}} < \left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n q_{n+1}}, \tag{3.13}$$

or

$$\frac{1}{q_{n+2}} < \left| p_n - q_n x \right| < \frac{1}{q_{n+1}}. \tag{3.14}$$

Lemma 3.3. [1] *A number x has a peroidic continued fraction if and only if it satisfies a quadratic equation; ie, $\exists A, B, C \in \mathbb{Z}$ such that $Ax^2 + Bx + C = 0$.*

First Direction: If x has a periodic continued fraction, then x satisfies a quadratic equation with integer coefficients.

Let x have a periodic continued fraction:

$$\begin{aligned} x &= [a_0, a_1, \ldots, a_{L-1}, a_L, \ldots, a_{L+k-1}, a_L, \ldots, a_{L+k-1}, a_L, \ldots] \\ &= [a_0, a_1, \ldots, a_{L-1}, a'_L], \end{aligned} \tag{3.15}$$

where

$$\begin{aligned} a'_L &= [a_L, a_{L+1}, a_{L+2}, \ldots] \\ &= [a_L, a_{L+1}, \ldots, a_{L+k-1}, a'_L] \end{aligned} \tag{3.16}$$

As

$$a'_L = \frac{p' a'_L + p''}{q' a'_L + q''}, \tag{3.17}$$

a'_L solves the quadratic equation

$$q'(a'_L)^2 + (q'' - q')a'_L - p'' = 0. \tag{3.18}$$

As

$$x = [a_0, a_1, \dots, a_{L-1}, a'_L] \tag{3.19}$$

upon solving we obtain

$$x = \frac{p_{L-1}a'_L + p_{L-2}}{q_{L-1}a'_L + q_{L-2}}, \tag{3.20}$$

which gives

$$a'_L = \frac{p_{L-2} - xq_{L-2}}{q_{L-1}x - p_{L-1}}. \tag{3.21}$$

Substituting the above for a'_L in Equation 3.18 yields

$$q'\left(\frac{p_{L-2} - xq_{L-2}}{q_{L-1}x - p_{L-1}}\right)^2 + (q'' - q')\left(\frac{p_{L-2} - xq_{L-2}}{q_{L-1}x - p_{L-1}}\right) - p'' \;=\; 0. \tag{3.22}$$

Multiplying through by $(q_{L-1}x - p_{L-1})^2$, we find that x solves a quadratic equation with non-zero quadratic coefficient.

Second Direction: If x satisfies a quadratic equation with integer coefficients, then x has a periodic continued fraction. Assume x solves

$$ax^2 + bx + c = 0. \tag{3.23}$$

Further, we may assume the quadratic equation is irreducible over $\mathbb{Z}$ or $\mathbb{Q}$ (if not, then x would satisfy a linear equation). Thus, Equation 3.23 has no rational roots.
We must show x has a periodic continued fraction expansion.
We write $x = [a_0, a_1, \dots]$. We may write $x = [a_0, \dots, a_{n-1}, a'_n]$ and we get

$$x = \frac{p_{n-1}a'_n + p_{n-2}}{q_{n-1}a'_n + q_{n-2}}. \tag{3.24}$$

Substitute this value of x for x in Equation 3.23. Clear denominators by multiplying through by $(q_{n-1}a'_n + q_{n-2})^2$. We find a'_n satisfied the following quadratic equation

$$A_n(a'_n)^2 + B_n a'_n + C_n = 0, \tag{3.25}$$

where a messy (but straightforward) calculation gives

$$\begin{aligned} A_n &= ap_{n-1}^2 + bp_{n-1}q_{n-1} + cq_{n-1}^2 \\ B_n &= 2ap_{n-1}q_{n-2} + b(p_{n-1}q_{n-2} + p_{n-2}q_{n-1}) + 2cq_{n-1}q_{n-2} \\ C_n &= ap_{n-2}^2 + bp_{n-2}q_{n-2} + cq_{n-2}^2. \end{aligned} \tag{3.26}$$

Remark 3.4. $A_n \neq 0$.

Proof:
If $A_n = 0$, then dividing the expression for A_n by q_{n-1}^2 gives $\frac{p_{n-1}}{q_{n-1}}$ satisfies Equation 3.23; however, Equation 3.23 has no rational solutions.
Thus, we have

$$A_n y^2 + B_n y + C_n = 0, \quad y = a'_n \text{ is a solution, } A_n \neq 0. \tag{3.27}$$

The discriminant of the above quadratic is (another messy but straightforward calculation)

$$\Delta = b_n^2 - 4A_nC_n = b^2 - 4ac. \tag{3.28}$$

By Lemma 3.2,

$$xq_{n-1} - p_n = \frac{\delta_{n-1}}{q_{n-1}}, \ |\delta_{n-1}| < 1. \tag{3.29}$$

Thus,

$$A_n = a\left(xq_{n-1} + \frac{\delta_{n-1}}{q_{n-1}}\right)^2 + bq_{n-1}\left(xq_{n-1} + \frac{\delta_{n-1}}{q_{n-1}}\right) + cq_{n-1}^2. \tag{3.30}$$

Taking absolute values and remembering that $ax^2 + bx + c = 0$ gives

$$\left|(ax^2+bx+c)q_{n-1}^2 + 2ax\delta_{n-1} + a\frac{\delta_{n-1}^2}{q_{n-1}^2} + b\delta_{n-1}\right| \leq 2\cdot|a|\cdot|x| + |b| + |a|. \tag{3.31}$$

As $C_n = A_{n-1}$ we find that

$$\begin{aligned} |C_n| &\leq 2\cdot|a|\cdot|x| + |b| + |a| \\ B_n^2 - 4A_nC_n &= b^2 - 4ac \\ B_n &\leq \sqrt{|4A_nC_n| + |b^2 - 4ac|} < \sqrt{4\Big(2|a|\cdot|x| + |b| + |a|\Big)^2 + |b^2 - 4ac|}. \end{aligned} \tag{3.32}$$

We have shown

Lemma 3.5. [1] *There is an M such that, for all n,*

$$|A_n|, |B_n|, |C_n| < M. \tag{3.33}$$

Thus, by Dirichlet's Box Principle, we can find three triples such that

$$(A_{n_1}, B_{n_1}, C_{n_1}) = (A_{n_2}, B_{n_2}, C_{n_2}) = (A_{n_3}, B_{n_3}, C_{n_3}). \tag{3.34}$$

We get three numbers a'_{n_1}, a'_{n_2} and a'_{n_3} which all solve the same quadratic equation (Equation 3.25), and the polynomial *is not* the zero polynomial as $A_n \neq 0$.

As any non-zero polynomial has at most two distinct roots, two of the three a_{n_i}s are equal. Without loss of generality, assume $a'_{n_1} = a'_{n_2}$. This implies periodicity because

$$\begin{aligned} [a_{n_1}, a_{n_1+1}, \ldots, a_{n_2}, \ldots] &= [a_{n_1}, a_{n_1+1}, \ldots, a'_{n_2}] \\ &= [a_{n_1}, a_{n_1+1}, \ldots, a'_{n_1}]. \end{aligned} \tag{3.35}$$

We know that the measure of a set of periodic continued fractions is 0 and no information can be gained through Kuzmin's Theorem. This is made more apparent by the fact that there are only a finite number, m, of values k_n so that the expansion will look like

$$[a_1, a_2, \ldots, a_m, a_{m+1}, a_{m+2}, \ldots, a_j, \ldots] = [k_1, k_2, \ldots, k_m, k_1, k_2, \ldots, k_m, \ldots] \tag{3.36}$$

Kuzmin's Theorem, however, predicts that each and every set of numbers E_{n,k_i} for all values of k_i in $[1, \infty)$, has positive measure. In other words, any value of k_i has positive probability of occurring. This is impossible for the case of quadratic irrationals, or in terms of our investigation, numbers x which satisfy $x^2 - P = 0$, because they are periodic and periodic continued fractions only have a finite number of values k_i. Since $\{k_i\} \subset [1, \infty)$, this leaves an infinite number of values which k_i never takes, or takes with 0 probability.

Therefore in every interval arbitrarily small, the property of periodicity prohibits that every real number in that interval be represented, and the measure of the set of periodic continued fractions is 0.
We then restrict ourselves to algebraic numbers of the form

$$x^n - P = 0 \tag{3.37}$$

where $n \geq 3$.

The question now is: Even though sets of algebraic numbers of this form have 0 measure, are they "general" enough for Kuzmin's Theorem, which deals with sets of positive measure, to apply?

4 Test of Kuzmin's Theorem

The four values of n tested were $n = 3, n = 4, n = 5$, and $n = 103$. $n = 3$ was tested to find out the lowest $n > 2$ such that Kuzmin's Theorem would hold. $n = 4$ and $n = 5$ were tested to support that $n = 3$ didn't have any exclusive properties, and $n = 103$ was tested to ensure that roots of lower order were not special cases that satisfied Kuzmin's Theorem. A program was written to count the occurances of each value of k_i in the $50,001^{th}$ to $100,000^{th}$ elements a_n, or digits of the continued fraction expansions for each x satisfying $x^n - P = 0$ for the first 1,000 primes. This was done to ensure that similiarities in earlier digits due to numbers being close to each other was avoided. It also focused the investigation on elements a_n for n large, which is criteria for the convergence to the predicted measure in Kuzmin's Theorem.

The program executed this count by method of a characteristic function thus defined:

$$A_{n,k_i}(y) \; = \; \begin{cases} 1 & \text{if } a_n(y) = k \\ 0 \text{if } a_n(y) \neq k_i \end{cases} \tag{4.1}$$

Then

$$\text{num}_{M,N;k_i}(y) = \sum_{n=M}^{N-1} A_{n,k_i}(y) \tag{4.2}$$

is the number of digits n of x which equal k_i with $M \leq n < N$.
In our investigation, $M = 50,001$ and $N = 100,000$.

Kuzmin states that up to a small error, $\Theta(e^{-\lambda})$, and for large n, the probability $\phi(k_i)$ that $a_n = k_i$ is

$$\phi(k_i) = \log_2 \left(1 + \frac{1}{k_i(k_i+1)} \right) + \Theta(e^{-\lambda}) \tag{4.3}$$

Since our characteristic function $A_{n,k_i}(y)$ takes only 0 and 1 as values, the expected value of A_{n,k_i} is $E[A_{n,k_i}] = \phi(k_i)$. Since Kuzmin's Theorem states that for large n, the probability $\phi(k_i)$ converges to a term (4.3) that is independent of n, and the expected value of a sum is the sum of expected values, the expected value is linear. More clearly,

$$\begin{aligned} E\left[\text{num}_{M,N;k}(x)\right] &= E\left[\sum_{n=M}^{N-1} A_{n,k}(x)\right] \\ &= \sum_{n=M}^{N-1} E\left[A_{n,k}(x)\right] \\ &= \sum_{n=M}^{N-1} \left(q_k + \Theta(e^{-\lambda})\right) \\ &\approx (N-M)q_k + \Theta(e^{-\lambda}) \cdot (N-M). \end{aligned} \tag{4.4}$$

So the approximate number of times we expect to find k_i in $N-M$ digits is simply

$$(N-M)\log_2\left(1+\frac{1}{k_i(k_i+1)}\right) \tag{4.5}$$

Or, more concretely, in the case of $k_i = 1$ in digits 50,001 to 100,000, as in our investigation,

$$E\left[\text{num}_{M,N;k}(x)\right] = (N-M)\phi(k_i) = (50,000)\log_2\left(1+\frac{1}{3}\right) \approx 20750 \tag{4.6}$$

For each $P^{\frac{1}{n}}$, for fixed P, and fixed n, a value of k_i was set and the number of times it occurred in those $50,000$ digits was counted first for each integer value of k_i from 1 to 100. Then for values of k_i greater than 100, a range of values instead of specific integers was set, since as Kuzmin predicts, the probability of k_i occurring rapidly drops as k_i grows large. k_i was set to simply be any number in the range 101 to 1,000, and the number of times any of these values occurred in the 50,000 digits was recorded. The same was done for k_i taking any values between 1,001 and 10,000, and likewise, with 10,001 and 100,000.

This process was performed for the first thousand primes P, for fixed n. And then performed again for the first thousand primes P for each of the other values of n. As a sample, the next page includes tables comparing the averages of the data collected with Kuzmin's results for k_i values: 1, 2, 3, 4, and 5.

In the tables below, "average", "stdev", and "Kuzmin" represent the average value of all k_i divided by $(N-M)$, the standard deviation to the average divided by $(N-M)$, and Kuzmin's predicted value of $\phi(k_i)$ respectively. The averages and standard deviation are divided by the number of digits simply so that we can compare probabilities with Kuzmin's result conveniently.

n=3

k_i	*average*	*stdev*	*Kuzmin*	$av - stdev$	$av + stdev$
1	0.415100	0.001956	0.415037	0.413144	0.417056
2	0.169844	0.001646	0.169925	0.168198	0.171490
3	0.093078	0.001296	0.093109	0.091782	0.094374
4	0.058879	0.001017	0.058893	0.057862	0.059896
5	0.040667	0.000881	0.040641	0.039786	0.041548

n=4

k_i	*average*	*stdev*	*Kuzmin*	$av - stdev$	$av + stdev$
1	0.415001	0.002099	0.415037	0.412902	0.417100
2	0.169945	0.001697	0.169925	0.168248	0.171642
3	0.093135	0.001322	0.093109	0.091813	0.094457
4	0.058912	0.001059	0.058893	0.057853	0.059971
5	0.040639	0.000890	0.040641	0.039749	0.041529

n=5

k_i	*average*	*stdev*	*Kuzmin*	$av - stdev$	$av + stdev$
1	0.414976	0.002066	0.415037	0.412910	0.417042
2	0.170000	0.001691	0.169925	0.168309	0.171691
3	0.092987	0.001301	0.093109	0.091686	0.094288
4	0.058908	0.001057	0.058893	0.057851	0.059965
5	0.040623	0.000894	0.040641	0.039729	0.041517

n=103

k_i	*average*	*stdev*	*Kuzmin*	$av - stdev$	$av + stdev$
1	0.415021	0.002136	0.415037	0.412885	0.417157
2	0.169920	0.001620	0.169925	0.168300	0.171540
3	0.093101	0.001321	0.093109	0.091780	0.094422
4	0.058887	0.001069	0.058893	0.057818	0.059956
5	0.040699	0.000886	0.040641	0.039813	0.041585

As we can see from the data, all average probabilities fall within the standard deviation from Kuzmin's predictions. This data is strong support for the conclusion that numbers x, of the form

$$x^n - P = 0 \tag{4.7}$$

for $n \geq 3$ are "general" enough, in the sense that the properties of their continued fraction expansions have few enough restrictions, that the set of these numbers can be treated like the continuum with respect to Kuzmin's Theorem. The main point of ambiguity is that Kuzmin's Theorem holds up to sets of measure 0, but what is the structure of sets of measure 0, if they do in fact exist, that Kuzmin's Theorem also holds for? Through this investigation, we have narrowed down our possibilities by confirming that Kuzmin's Theorem appears to hold for at least algebraic numbers of order $n \geq 3$.

References

[Ki] A.Y.Khinchin, *Continued Fractions*, Third Edition, The University of Chicago Press, Chicago 1964.

[HSMT] S.J.Miller, R.Takloo-Bighash, H.Helfgott, F.Spinu, *Junior Research Seminar: Diophantine Analysis and Approximations*, Princeton University, Princeton, 2002