DISTRIBUTIONS OF ULAM WORDS UP TO LENGTH 30

Paul Adutwum
Bates College, Lewiston, Maine
pauladutwum303@gmail.com

Hopper Clark
Bates College, Lewiston, Maine
hopperaclark@gmail.com

Ro Emerson
Bates College, Lewiston, Maine
remerson@bates.edu

Alexandra (Sasha) Sheydvasser
University of Massachusetts, Amherst, Massachusetts
asheydvasser@umass.edu

Arseniy (Senia) Sheydvasser
Department of Mathematics, Bates College, Lewiston, Maine
sheydvasser@gmail.com

Axelle Tougouma
Bates College, Lewiston, Maine
atougouma@bates.edu

Received: 10/16/24, Revised: 7/2/25, Accepted: 10/26/25, Published: 11/5/25

Abstract

We further explore the notion of Ulam words considered by Bade, Cui, Labelle, and Li, giving some lower bounds on how many there are of a given length. Gaps between words and words of special type also reveal remarkable structure. By substantially increasing the number of computed terms, we are also able to sharpen some of the conjectures made by Bade et al.

1. Introduction

In their 2020 paper, Bade, Cui, Labelle, and Li [1] introduced the notion of Ulam words, defined as follows. Consider the free semigroup $S[\{0,1\}]$ on two generators

DOI: 10.5281/zenodo.17535327

0 and 1. We say that 0 and 1 are *Ulam* and then define all other Ulam words inductively: a word $w \neq 0, 1$ is Ulam if and only if there exists exactly one pair of Ulam words $u_1 \neq u_2$ such that $w = u_1 \frown u_2$. (Here, $\frown$ denotes concatenation.) We shall denote the entire set of Ulam words as $\mathscr{U}$, and Ulam words of length n by $\mathscr{U}_n$. It is easy to check that:

$$\mathscr{U}_1 = \{0, 1\} \qquad \mathscr{U}_3 = \{001, 011, 100, 110\}$$
$$\mathscr{U}_2 = \{01, 10\} \qquad \mathscr{U}_4 = \{0001, 0010, 0100, 0111, 1000, 1011, 1101, 1110\}.$$

All Ulam words up to length 24 were computed in [1]; we were able to compute up to length 30. While this might appear as a small improvement at first glance, because the number of Ulam words of length n appears to (almost) double on each iteration, in reality, this represents nearly 60 times as much data.

It is an open question whether $|\mathscr{U}_n|$ (the size of $\mathscr{U}_n$) grows exponentially. The best lower bound that we can prove is linear, mainly using explicit constructions of words from [1].

Theorem 1. *For all $n \geqslant 6$, we have that $|\mathscr{U}_n| \geqslant 2n + 4$.*

However, we are able to demonstrate that there is a *subsequence* of Ulam words that grows exponentially, using a completely different argument.

Theorem 2. *There exists $1 < \alpha_0 \leqslant 2$ such that for all $1 < \alpha < \alpha_0$, we have that*

$$\limsup_{n\to\infty} \frac{|\mathscr{U}_n|}{\alpha^n} = \infty.$$

Concretely,

$$\alpha_0 = \left(\frac{101847671}{31}\right)^{1/5} \approx 1.648996$$

suffices.

We give proofs of both of these theorems in Section 3. Unfortunately, both of these results are still quite far from what is conjectured to hold. To wit, define the density

$$\rho(n) := \frac{|\mathscr{U}_n|}{2^n};$$

It was conjectured in [1] that $\rho(n) \to r$ for some $0 < r < 1$ (see Conjecture 3.10 in [1]); with our enlarged data set, we instead posit something a little stranger.

Conjecture 1. The density of Ulam words $\rho(n) = \Theta(n^{-3/10})$.

This conjecture is supported by the numerical evidence—see Figures 1 and 2 for an example—but it also has ties to another conjecture involving the average gap between Ulam words, which we shall describe below. In any case, observe that if

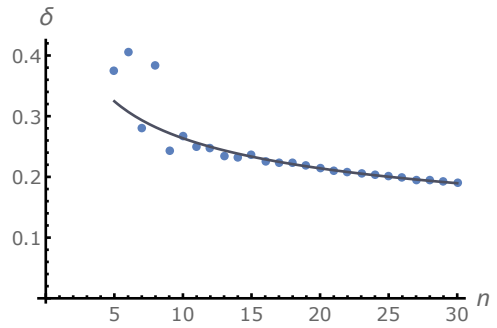

Figure 1: A plot of the densities $\rho(n)$ for $4 \leqslant n \leqslant 30$, together with a plot of $f(n) = 0.526n^{-3/10}$.

n	$\lvert\mathscr{U}_n\rvert$	n	$\lvert\mathscr{U}_n\rvert$	n	$\lvert\mathscr{U}_n\rvert$
13	1916	19	114300	25	6720784
14	3812	20	225166	26	13303332
15	7772	21	441724	27	26273948
16	14822	22	876238	28	52010642
17	29368	23	1717748	29	102933200
18	58478	24	3406884	30	203695342

Figure 2: The exact counts for $|\mathscr{U}_n|$ for $13 \leqslant n \leqslant 30$.

either conjecture is correct, the number of Ulam words grows only very slightly slower than 2^n.

This notion of Ulam words was built on the earlier notion of Ulam sets due to Kravitz and Steinerberger [8], which was itself a generalization of Ulam's eponymous integer sequence, also defined recursively [13]: the (classical) Ulam sequence begins with $1, 2$, and then every subsequent term is the next smallest integer that can be written as the sum of two distinct prior terms in exactly one way. Generalizations of Ulam's classic sequence have become an increasingly popular object of study: in 1972, Queneau did some preliminary work studying generalizations where the initial two terms of the integer sequence are varied [10]; in the 1990s, Cassaigne, Finch, Shmerl, and Spiegel determined some of the families of such sequences such that the consecutive differences are eventually periodic [2, 3, 4, 5, 11]; in 2017, [8] considered generalizing the Ulam condition for abelian groups; in 2020, [1] gave the aforementioned notion of Ulam words with some preliminary results; and in 2021, Sheydvasser showed that there is an analogous notion of Ulam sets for integer polynomials [12] by building off earlier work of Hinman, Kuca, Schlesinger, and Sheydvasser [6, 7].

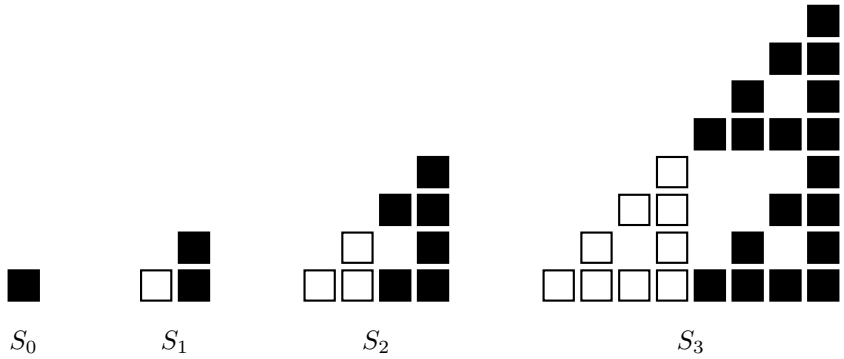

Figure 3: Visual of the first 4 steps of constructing the discrete Sierpiński triangle.

Earlier work around Ulam words has largely centered around giving simple criteria for when words of some special type are Ulam—for example, [1] showed that a word of the form 0^a10^b is Ulam if and only if $\binom{a+b}{a}$ is odd. Similarly, Mandelshtam [9] considered Ulam words of the form $0^a10^b10^c$ and demonstrated a connection to the Sierpiński gasket. We also prove a few such results, such as the following.

Theorem 3. *Consider the set of points $(x, y) \in \mathbb{Z}^2_{\geqslant 1}$ such that $1^y0^{x-y} \in \mathscr{U}$. This is the discrete Sierpiński triangle, union a point.*

We will discuss this construction more precisely in Section 4, but briefly, the discrete Sierpiński triangle is an approximation to the standard Sierpiński triangle. It can be constructed either iteratively (as in Figure 3) or by coloring Pascal's triangle by parity.

On the other hand, we also have a novel way of considering Ulam words by interpreting them as integers. Observe that there exists a natural map $\pi : S[\{0, 1\}] \to \mathbb{Z}_{\geqslant 0}$ via interpreting a word as the binary representation of an integer. In general, this map is not injective—for example, $\pi(0) = \pi(00) = \pi(000) = 0$. However, if we restrict it to words of a fixed length, then it is. In particular, the restrictions $\pi : \mathscr{U}_n \to \mathbb{Z} \cap [0, 2^{n-1}]$ are injective maps. This gives a natural ordering on $\mathscr{U}_n$ and allows us to ask questions about how Ulam words are distributed. For example, we might ask about the distribution of the gaps—differences between consecutive Ulam words, interpreted as integers.

Conjecture 2. Let $u_1 < u_2 < \ldots < u_{k_n}$ be the (ordered) elements of $\pi(\mathscr{U}_n)$. Define

$$\begin{aligned} p_n : \mathbb{Z}_{\geqslant 1} &\to [0, \infty) \\ g &\mapsto \frac{\left|\{i | u_{i+1} - u_i = g\}\right|}{k_n - 1}. \end{aligned}$$

This has a natural interpretation as a probability measure. As $n \to \infty$, the functions p_n converge pointwise to a probability measure $p : \mathbb{Z}_{\geqslant 1} \to [0, \infty)$. Furthermore, let $\mu_g(n)$ be the mean of the probability measure p_n. Then $\mu_g(n) = \Theta(n^{3/10})$—indeed, it may be that there is a constant $c \approx 1.9$ such that $\mu_g(n) = cn^{3/10} + o(1)$.

This conjecture is well-supported by our available data—see Section 5 for details, illustrations, and further odd properties of the apparent distribution. What is interesting about this statement about average gaps is that, if true, it immediately implies Conjecture 1.

Theorem 4. *As $n \to \infty$, we have that $\rho(n)^{-1} \asymp \mu_g(n)$. Consequently, Conjecture 2 implies Conjecture 1.*

This is salient, since our numerical evidence for Conjecture 2 is arguably much stronger than for Conjecture 1! Again, see Section 5 for details. Finally, in Section 6, we ask the question of how $\pi(\mathscr{U}_n)$ is distributed modulo N.

Conjecture 3. For any integer $N > 1$ and $a \in \mathbb{Z}/N\mathbb{Z}$, define the *relative density*

$$\rho_{a,N}(n) := \frac{\left|\{w \in \mathscr{U}_n | \pi(w) \equiv a \mod N\}\right|}{|\mathscr{U}_n|}.$$

Then $\lim_{n\to\infty} \rho_{a,N}(n) = 1/N$.

Remark 1. As we discuss in Section 6, while this conjecture is consistent with the available data, it is somewhat surprising. For one thing, $\rho_{5,6}(1) = \rho_{5,6}(2) = \rho_{5,6}(3) = 0$, and it takes some time before it appears to start to converge to $1/6$. For another, there is an apparent bias modulo 6 in the distribution of the gaps.

Our code and some of our data can be found on GitHub[1], but it is far from efficient—as was pointed out to us Tomás Oliveira e Silva, it is possible to use bitmaps to make these computations much faster; a good implementation should give a $O(2^n \log(n))$ running time. However, we leave this as material for future work.

2. Definitions and Visualizations

We start with some basic definitions and constructions. Given a word $w \in S[\{0, 1\}]$, we define its *complement* $\hat{w}$ to be the word with every instance of 0 replaced with a 1, and vice versa. We also define the *reverse* $\overline{w}$, which is the word obtained by reversing the order of the letters. It was shown in [1] that $w \in \mathscr{U}$ if and only if $\hat{w} \in \mathscr{U}$, if and only if $\overline{w} \in \mathscr{U}$.

[1] https://github.com/asheydva/Ulam-Words.git

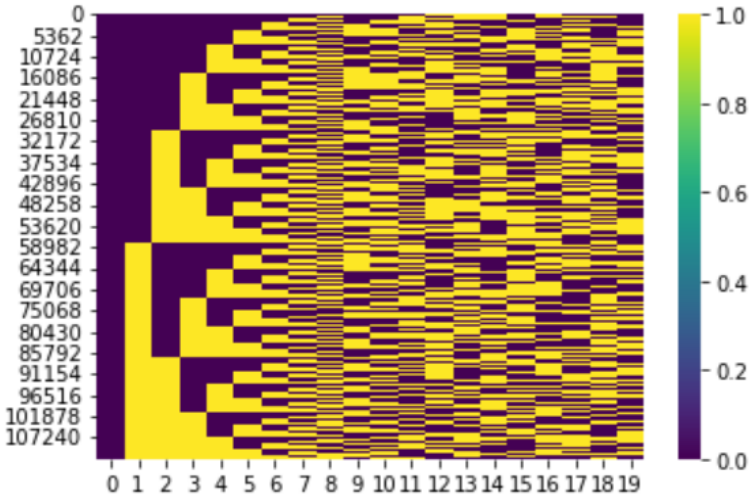

Figure 4: All words of length 20 beginning with a zero.

To better visualize the set $\mathscr{U}$, we made use of heat maps, which depict each Ulam word as a colored bar and stacks all of the words vertically—that is, for a given word, a 0 corresponds to a rectangle of one color, and a 1 corresponds to a rectangle of a second color. An example is provided in Figure 4. In general, we abridge such diagrams: we created figures only using all the Ulam words that started with zero, since Ulam words are closed under complements. Moreover, we impose the ordering discussed in the introduction, defining $w \leqslant w'$ if and only if $\pi(w) \leqslant \pi(w')$. Using this way of visualizing Ulam words allows us to easily see that there is both a clear binary tree structure that governs the existence of Ulam words, as well as a chaotic element to the set where the binary tree breaks down.

We can be more specific about our meaning regarding this breakdown: since Ulam words are preserved under the reverse map, this is equivalent to saying that for any n there exists $n > \ell_n > 0$ such that all possible subwords of length ℓ_n occur as the final ℓ_n characters of words in $\mathscr{U}_n$. In turn, that is equivalent to saying that the quotient map $\pi(\mathscr{U}_n) \to \mathbb{Z}/2^{\ell_n}\mathbb{Z}$ is surjective. Our observation is that ℓ_n appears to increase as a function of n, albeit not very quickly—see Figure 5. Assuming that Conjecture 3 is true, then it would follow immediately that $\ell_n \to \infty$ simply by considering the case where $N = 2^{\ell_n}$—indeed, the heat maps were the original impetus for our equidistribution conjectures. On the other hand, the "chaotic" latter half of the heat map is more of a mystery.

n	ℓ_n	n	ℓ_n	n	ℓ_n	n	ℓ_n	n	ℓ_n
1	1	7	4	13	5	19	9	25	11
2	1	8	4	14	5	20	9	26	11
3	1	9	4	15	6	21	9	27	12
4	1	10	4	16	7	22	10	28	12
5	3	11	4	17	7	23	10	29	13
6	2	12	5	18	8	24	11	30	13

Figure 5: Tables of n versus ℓ_n, where ℓ_n is the largest integer such that $\mathscr{U}_n \to \mathbb{Z}/2^{\ell_n}\mathbb{Z}$ is surjective.

3. Lower Bounds on Growth

Our goal in this section is to prove our lower bounds on $|\mathscr{U}_n|$; we begin with Theorem 1, for which we need some explicit examples of Ulam words. The first three are due to [1].

Theorem 5 ([1])**.** *There are $G(n-1)$ Ulam words of length n of the form 0^a10^b, where $G(n)$ is the n-th entry in Gould's sequence.*

Remark 2. Gould's sequence $G(n)$ is the number of odd entries in the n-th row of Pascal's triangle; equivalently, $G(n) = 2^{\#_1(n)}$, where $\#_1(n)$ is the number of non-zero bits in the binary representation of n.

Remark 3. Since $w \in \mathscr{U}$ if and only if $\overline{w} \in \mathscr{U}$, if and only if $\hat{w} \in \mathscr{U}$, we get analogous results with 1's replaced with 0's and the order of the letters reversed. This is true for all the results that we prove here.

Theorem 6 ([1])**.** *For any $a, b \in \mathbb{Z}_{\geqslant 0}$, the word $0^a1^20^b$ is in $\mathscr{U}$ if and only if the length of the word is odd (that is, $a + b \equiv 1 \pmod 2$).*

Theorem 7 ([1])**.** *For any $a, b \in \mathbb{Z}_{\geqslant 0}$ such that $a + b \geqslant 2$, the word 0^a1010^b is in $\mathscr{U}$ if and only if the length of the word is even (that is, $a + b \equiv 1 \mod 2$).*

Lemma 1. *For any $a, b \in \mathbb{Z}_{\geqslant 0}$ such that $a + b \geqslant 1$, the word $0^a1^40^b$ is in $\mathscr{U}$ if and only if $a + b \equiv 1 \pmod 4$.*

Proof. We will use proof by induction on the length of the word n, where the base cases $n = 5, 6, 7, 8$ can be verified directly. Assume the statement holds for all words of length strictly less than n, and consider the word $u = 0^k1^40^l$ of length n, where, since $n \geqslant 9$, at least one of k and l is at least 3. By applying the reverse map to switch k and l if necessary, we may assume that $k \geqslant 3$.
`Case 1:` $l = 0$. The only possible representations are $0^\frown0^{k-1}1^4$ and $0^k1^3{}^\frown1$. By the inductive hypothesis, the first is valid if and only if $n \equiv 2 \pmod 4$. By Lemma

3, the second is valid if and only if $n \equiv 1, 2 \pmod 4$. Thus, exactly one of these representations is valid if and only if $n \equiv 1 \pmod 4$.
Case 2: $l \geqslant 1$. There are five potential representations:

1. $0^\frown 0^{k-1}1^4 0^l$,
2. $0^k 1^\frown 1^3 0^l$,
3. $0^k 1^2 {}^\frown 1^2 0^l$,
4. $0^k 1^3 {}^\frown 10^l$, and
5. $0^k 1^4 0^{l-1} {}^\frown 0$.

Observe that by the inductive hypothesis, representations (1) and (5) are valid if and only if $n \equiv 2 \pmod 4$, which is to say that $k + l \equiv 2 \pmod 4$. By Theorem 8 and Lemma 3, representation (2) is valid if and only if $l \equiv 0, 3 \pmod 4$; similarly, representation (4) is valid if and only if $k \equiv 0, 3 \pmod 4$. Finally, by Theorem 6, representation (3) is valid if and only if $k \equiv l \equiv 1 \pmod 2$. This allows us to count the number of valid representations in terms of the congruence classes of k and l modulo 4, as seen in Figure 6. In particular, there is a unique representation if and only if $n \equiv 1 \pmod 4$.

$k\backslash l$	0	1	2	3
0	2	1	3	2
1	1	3	0	2
2	3	0	0	1
3	2	2	1	5

Figure 6: Table of number of representations for $0^k 1^4 0^l$ for values of $n = k + l$ modulo 4.

□

With this, we are ready to give a proof of the general linear bound.

Proof of Theorem 1. We consider three cases.
Case 1: n is even. By Theorem 7, we know that $0^a 1010^{n-a-3} \in \mathscr{U}_n$ for all $0 \leqslant a \leqslant n-3$—this yields $n-2$ Ulam words. By Theorem 5, we also know that there are $G(n-1)$ Ulam words of length n of the form $0^a 10^b$. Note that these two sets of Ulam words do not intersect (they have different numbers of ones), and since $n-1$ is odd, $G(n-1) \geqslant 2^2 = 4$. In total, this yields $n+2$ Ulam words.

Note that $0^a\widehat{1010^{n-a-3}} = 1^a0101^{n-a-3} = 0^{a_1}10^{b_1}$ if and only if $n = 3$, so the reverses of the constructed Ulam words are also distinct Ulam words. Therefore, we have at least $2n + 4$ Ulam words in this case.

`Case 2:` $n \equiv 3 \pmod 4$. By Theorem 6, we know that $0^a1^20^{n-a-2} \in \mathscr{U}_n$ for all $0 \leqslant a \leqslant n-2$—this yields $n-1$ Ulam words. Since $n-1 \equiv 2 \pmod 4$, $G(n-1) \geqslant 2^2 = 4$, and so we can again use Theorem 5 to conclude that there are at least 4 words 0^a10^b of the right length. In total, this yields $n+3$ Ulam words.

Note that $0^a\widehat{1^20^{n-a-2}} = 1^a0^21^{n-a-2} = 0^{a_1}10^{b_1}$ if and only if $a = 0$ and $a_1 = 2$. Therefore, the reverses of our two families of constructed Ulam words intersect, but only in two places; therefore, we have $2(n+3) - 2 = 2n + 4$ Ulam words.

`Case 3:` $n \equiv 1 \pmod 4$. As in the previous case, we have $n-1$ words of the form $0^a1^20^{n-a-2}$, but it is possible that $G(n-1) = 2$, so we have to argue differently: specifically, we use Lemma 1 to conclude that $0^a1^40^{n-4-a} \in \mathscr{U}$ for all $0 \leqslant a \leqslant n-4$, which yields another $n-3$ Ulam words, for a total of at least $2n-4$.

Observe that $0^a\widehat{1^20^{n-a-2}} \neq 1^a0^21^{n-a-2} = 0^{a_1}1^40^{n-4-a_1}$ ever, so we may simply double our count of Ulam words. In total, we have $4n-8$, which is at least $2n+4$ if $n \geqslant 6$.

In each case, we have identified at least $2n + 4$ distinct Ulam words. □

Next, we tackle the exponential bound, which we approach in a completely different fashion using the following lemma.

Lemma 2. *For any $n \in \mathbb{Z}_{\geqslant 1}$,*

$$|\mathscr{U}_n|^2 \leqslant |\mathscr{U}_n| + |\mathscr{U}_{n+1}| + \ldots + |\mathscr{U}_{2n}|.$$

Proof. Consider the set

$$X := \left\{(w_1, w_2) \in \mathscr{U}_n^2 \middle| w_1 \neq w_2\right\}.$$

For any $(w_1, w_2) \in X$, either $w_1{}^\frown w_2 \in \mathscr{U}_{2n}$ or there exists $v_1 \in \mathscr{U}_k$, $v_2 \in \mathscr{U}_{2n-k}$ such that $w_1{}^\frown w_2 = v_1{}^\frown v_2$, where $k \in [1, n-1] \cup [n+1, 2n-1]$; of course, if $k \in [1, n-1]$, then $2n - k \in [n+1, 2n-1]$, and so we may conclude that

$$|X| \leqslant |\mathscr{U}_{n+1}| + \ldots + |\mathscr{U}_{2n}|.$$

On the other hand,

$$|X| = |\mathscr{U}_n|^2 - |\mathscr{U}_n|.$$

□

As a consequence of Lemma 2, we get the following very weak lower bound: for any $n \in \mathbb{Z}_{\geqslant 1}$,

$$\max_{n \leqslant i \leqslant 2n} |\mathscr{U}_i| \geqslant \frac{|\mathscr{U}_n|^2}{n+1}. \tag{1}$$

This is sufficient for our purposes.

Proof of Theorem 2. Choose any $n_1 \in \mathbb{Z}_{\geqslant 6}$ and let

$$\alpha_0 := \left(\frac{|\mathscr{U}_{n_1}|}{2(n_1+1)}\right)^{1/n_1}.$$

Then $|\mathscr{U}_{n_1}| = 2(n_1+1)\alpha_0^{n_1}$. By Theorem 1, we know that $|\mathscr{U}_{n_1}| > 2n_1+2$, hence $1 < \alpha_0 < 2$. Ergo, for any $1 < \alpha < \alpha_0$, $|\mathscr{U}_{n_1}| = 2C(n_1+1)\alpha^{n_1}$ for some $C > 1$. Recursively define a sequence $n_1, n_2, \ldots$ such that n_i is the unique integer $n_{i-1} \leqslant n_i \leqslant 2n_{i-1}$ such that $|\mathscr{U}_{n_i}|$ is as large as possible. We will prove by induction that $|\mathscr{U}_{n_k}| \geqslant 2C^{2^k}(n_k+1)\alpha^{n_k}$. The base case $k = 1$ is clear; for the induction step, we can apply our bound (Equation 1) to see that

$$\begin{aligned}
\mathscr{U}_{n_{k+1}} &= \max_{n_k \leqslant i \leqslant 2n_k} |\mathscr{U}_i| \geqslant \frac{|\mathscr{U}_{n_k}^2|}{n_k+1} \\
&> \frac{4C^{2^{k+1}}(n_k+1)^2}{n_k+1}\alpha^{2n_k} \\
&> 2C^{2^{k+1}}(2n_k+1)\alpha^{2n_k} \\
&\geqslant 2C^{2^{k+1}}(n_{k+1}+1)\alpha^{n_{k+1}}.
\end{aligned}$$

Since $C^{2^k} \to \infty$, we can conclude that for all $c > 0$, there exist infinitely many n such that $|\mathscr{U}_n| \geqslant c\alpha^n$. Therefore,

$$\limsup_n \frac{|\mathscr{U}_n|}{\alpha^n} = \infty,$$

as was claimed. All that remains is to find a value of α_0 that works. In our case, we used $n_1 = 30$, which gives the α_0 in the statement of the theorem. □

As can be seen from the proof of this theorem, if it is true that $|\mathscr{U}_n| = \Theta(n^{-3/10}2^n)$ as we conjecture, then it will be possible to improve the constant α_0 arbitrarily close to 2 simply by computing $|\mathscr{U}_n|$ for larger and larger n. Unfortunately, this rapidly becomes impractical and in any case will never suffice to prove the theorem with $\alpha_0 = 2$, as we suspect must be true.

4. Patterns in Ulam Words

Various results regarding Ulam words containing certain patterns were proven in [1]. We offer a similar collection of results. We begin with a couple of intermediate results that allow us to prove Theorem 3.

Theorem 8 ([1])**.** *A word of the form 0^a10^b is Ulam if and only if*

$$\binom{a+b}{a} \equiv 1 \mod 2.$$

Lemma 3. *For any $n \geqslant 3$, the word $1^3 0^{n-3}$ is in $\mathscr{U}$ if and only if $n \equiv 0 \pmod 4$ or $n \equiv 1 \pmod 4$.*

Proof. We will use proof by induction, where the base cases $n = 3, 4, 5, 6$ can all be verified by direct computation. Assume the statement holds for all words of length strictly less than n, and consider the word $u = 1^3 0^{n-3}$. The only two possible representations for u are

1. $1^\frown 1^2 0^{n-3}$ and
2. $1^3 0^{n-4\frown} 0$,

since neither 1^a nor 0^a are Ulam words for any $a > 1$. By Theorem 6, $1^2 0^{n-3} \in \mathscr{U}$ if and only if $n \equiv 0 \pmod 2$; thus, representation (1) is valid if and only if $n \equiv 0, 2 \pmod 4$. On the other hand, by the inductive hypothesis, representation (2) is valid if and only if $n \equiv 1, 2 \pmod 4$. Ergo, exactly one of the representations is valid if and only if $n \equiv 0, 1 \pmod 4$. □

The proof of Lemma 3 illustrates how the modular length restrictions for $1^{a+1}0^b$ can easily be found using the length restrictions for $1^a 0^b$, which, in turn, means we could generate countless additional theorems, providing length restrictions for $1^4 0^b$, $1^5 0^b$, and so forth. However, this would quickly prove tedious. Instead, we will demonstrate the unifying pattern between all words of the form $1^a 0^b$. Recursively define

$$\begin{aligned} S_0 &:= (2, 1) \\ S_{n+1} &:= S_n \cup (S_n + (2^n, 0)) \cup (S_n + (2^n, 2^n)) \\ \mathcal{S} &:= \bigcup_{n=0}^{\infty} S_n. \end{aligned}$$

The set $\mathcal{S}$ is sometimes referred to as the discrete Sierpiński triangle. The reason for this is that if one considers a suitable limit of the sets $2^{-n}\mathcal{S}$, the result is the usual Sierpiński triangle. Remarkably, this is exactly the correct construction to determine whether a word $1^a 0^b$ is Ulam or not.

Theorem 9. *Let $\mathcal{S}$ be defined as above. Then*

$$\left\{(x, y) \in \mathbb{Z}^2_{\geqslant 1} \middle| 1^y 0^{x-y} \in \mathscr{U}\right\} = (1, 1) \cup \mathcal{S}.$$

Remark 4. Observe that this is Theorem 3, but stated more precisely.

Proof of Theorem 9. It is easy to check that the only elements in both sets with $x \leqslant 2$ are $(1, 1)$ and $(2, 1)$. Our goal is to show that the iterative process for producing points in $\mathcal{S}$ with larger x values is the same as for the first set. To do so,

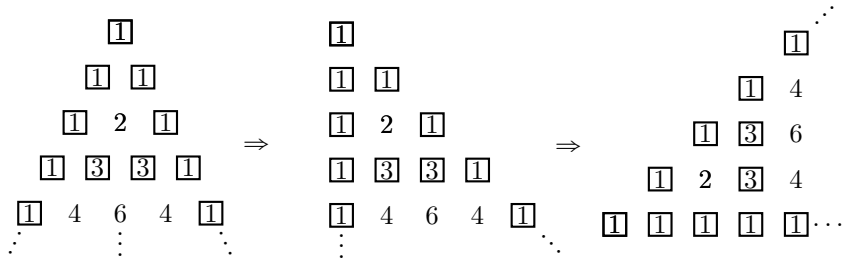

Figure 7: Visual of steps to create modified Pascal's triangle.

we will use a modified version of Pascal's triangle, achieved by aligning all entries to the left and then rotating the resulting image counterclockwise 90 degrees, as illustrated in Figure 7. In the modified version, if we let $p(x, y)$ be the entry in row y and column x, then

$$p(x, y) = p(x - 1, y) + p(x - 1, y - 1).$$

We make two additional modifications:

1. we align Pascal's triangle such that the bottom left entry occurs at $(2, 1)$ and
2. we shade all odd entries.

On the one hand, this is well-known to be the discrete Sierpiński triangle $\mathcal{S}$. On the other hand, because only odd entries are shaded, an entry (x, y) is shaded if and only if exactly one of $(x - 1, y)$ or $(x - 1, y - 1)$ is shaded.

Now, observe that there are only two potential representations for $1^x 0^{x-y}$, those being

1. $1^y 0^{x-y-1} \frown 0$ and
2. $1 \frown 1^{y-1} 0^{x-y}$.

Thus, (x, y) is in the desired set if and only if exactly one of $(x-1, y)$, $(x-1, y-1)$ is in the set. □

Remark 5. This is not the first time that Pascal's triangle and the Sierpiński triangle have shown up in the study of Ulam words. As mentioned previously, there was already an analogous pattern for words $0^x 10^y 1^z$ [9]; even before that, [1] proved that the number of words of length n with one 1 is the n-th term in Gould's sequence—that is, the number of elements in the $(n-1)$-th row of Pascal's triangle.

There are many consequences of this result. First, it means that determining the set of Ulam words of the form 1^a0^b is quite simple: it is a straightforward iterative procedure. Counting the number of elements of such form is also simple. For example, consider the following simple corollary.

Corollary 1. *Fix $n \in \mathbb{Z}_{\geqslant 2}$; let L_n be the set of $a \in \mathbb{Z}_{\geqslant 1}$ such that $0^a1^{n-a} \in \mathscr{U}$. Then $|L_n| = G(n)$.*

Proof. The k-th term of Gould's sequence is the number of odd entries in the k-th row of Pascal's triangle. After rotating and shifting, the k-th row corresponds to L_{k+1}. □

There are certainly more symmetries lurking within $\mathcal{S}$ that would lead to more theorems and patterns. In particular, we believe that it might be possible to demonstrate the following.

Conjecture 4. Let u be a word of length n of the form $0^a1^{2^k}0^b$ for $a, b, k \in \mathbb{Z}_{\geqslant 1}$. The word u is in $\mathscr{U}$ if and only if $n \equiv 1 \mod 2^k$.

We close by giving one more result, which is analogous to Theorem 7.

Theorem 10. *For any $a, b \in \mathbb{Z}$ such that $a + b \geqslant 1$, the word $0^a101010^b$ is in $\mathscr{U}$ if and only if $a, b \in 2\mathbb{Z}$ and one is zero.*

Proof. We induct on $a + b$. The base cases $a + b = 1, 2$ are easily established by pure computation. First, assume that both $a, b \neq 0$. Then there are six possible representations:

1. $0^\frown0^{a-1}101010^b$,
2. $0^a1^\frown01010^b$,
3. $0^a10^\frown1010^b$,
4. $0^a101^\frown010^b$,
5. $0^a1010^\frown10^b$, and
6. $0^a101010^{b-1}{}^\frown0$.

Applying the inductive hypothesis and Theorem 7, we find that in all cases either none of these are valid representations, or multiple are simultaneously. This leaves the case where either $a = 0$ or $b = 0$—without loss of generality, we assume that $a = 0$, since we can always apply the reverse map if needed. We have three possible representations:

1. $1^\frown01010^b$,
2. $10^\frown1010^b$, and
3. $101010^{b-1}{}^\frown0$.

If b is odd, then (2) and (3) are valid representations due to Theorem 7; therefore, this word is not Ulam. If b is even, then (1) is a valid representation, but (2) and (3) are not; therefore, this word is Ulam. □

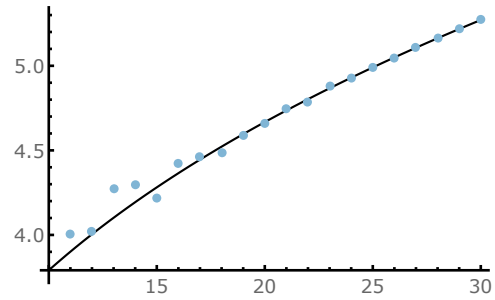

Figure 8: A plot of $\mu_g(n)$ for $11 \leqslant n \leqslant 30$ and $f(n) = 1.9n^{3/10}$.

5. Density and the Distribution of the Gaps

We begin by proving that there is a relationship between the size of the average gap between consecutive Ulam words of length n, and the density of Ulam words.

Proof of Theorem 4. Let $u_1 < u_2 < \ldots < u_{k_n}$ be the (ordered) elements of $\pi(\mathscr{U}_n)$—that is, the integers that are images of the Ulam words of length n. The gaps between them are the consecutive differences $g_1 = u_2 - u_1$, $g_2 = u_3 - u_2$, and so on. It is easy to see that $u_1 = \pi(0000\ldots01) = 1$ and $u_{k_n} = \pi(1111\ldots10) = 2^n - 2$—this follows from Theorem 5, for example. Ergo, the average gap between Ulam words of length n is

$$\begin{aligned}\mu_g(n) := \frac{g_1 + g_2 + \ldots + g_{k_n-1}}{k_n - 1} = \frac{u_{k_n} - u_1}{k_n - 1} = \frac{2^n - 3}{k_n - 1} \\ = \frac{1}{\rho(n)} \frac{k_n}{k_n - 1} - \frac{3}{k_n + 1},\end{aligned}$$

from which we conclude that $\rho(n)^{-1} \asymp \mu_g(n)$. (We already know from Theorem 1 that $k_n \to \infty$ as $n \to \infty$.) In particular, if Conjecture 2 is true and $\mu_g(n) \asymp n^{3/10}$, then $\rho(n) \asymp n^{-3/10}$, which is Conjecture 1. □

Notice in particular that for $\rho(n)$ to converge to a non-zero constant, it must be that $\mu_g(n)$ is bounded. But this does not appear to be the case, as evidenced by Figures 8 and 9. Instead, as near as we can tell, the order of growth seems to be around $n^{3/10}$. With this in mind, it is perhaps worthwhile to examine the distribution of the gaps more closely.

It is obvious that any gap is at least 1 (in fact, it is an easy exercise to show that a gap of 1 is always attained); on the other hand, numerical evidence suggests that the maximal gap grows exponentially—specifically, it is $O(r^n)$ for some constant $r \approx 1.35$ (see Figure 10). This is a little strange, in that one would not immediately guess this looking at the probability distributions, as in Figure 11.

n	Relative Error	n	Relative Error	n	Relative Error
13	4.08765%	19	0.196991%	25	0.0447467%
14	2.43579%	20	0.222403%	26	0.0981692%
15	1.54247%	21	0.24375%	27	0.028047%
16	1.27957%	22	0.332677%	28	0.0353147%
17	0.402629%	23	0.33529%	29	0.0365254%
18	0.87545%	24	0.104678%	30	0.00663021%

Figure 9: The relative error between the actual $\mu_g(n)$ and the estimate $1.9n^{3/10}$.

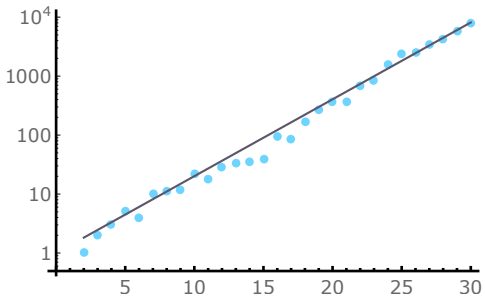

Figure 10: A logarithmic plot of the maximal gap between words of length n for $2 \leqslant n \leqslant 30$, and $f(n) = 1.35^n$.

Indeed, there are some curious details to this apparent distribution. The first is that there is a clear bias against gaps that are congruent to either 2 or 4 modulo 6. This is very odd, considering that we conjecture that Ulam words are equidistributed modulo 6 in the limit (See Section 6 for more about this). The second is that just as the average gap appears to grow without bound, so does the standard deviation. However, the standard deviations are quite small, of order around $n/3$—here is a table of the last few we were able to compute:

n	Standard Deviation	n	Standard Deviation
21	6.09043461	26	8.41026105
22	6.57391412	27	8.83842107
23	6.95198536	28	9.34566047
24	7.48652894	29	9.94055302
25	7.95451379	30	10.5007497

This means that the distribution is very tightly clustered toward the smaller side. However, it has extreme outliers: the maximal gap between words of length 30 is 8030, which is more than 764 standard deviations away from the mean! Somehow, this should be typical: as we noted already, the size of the maximal gap appears to grow exponentially, but the same is not true of either the average gap or the standard deviation.

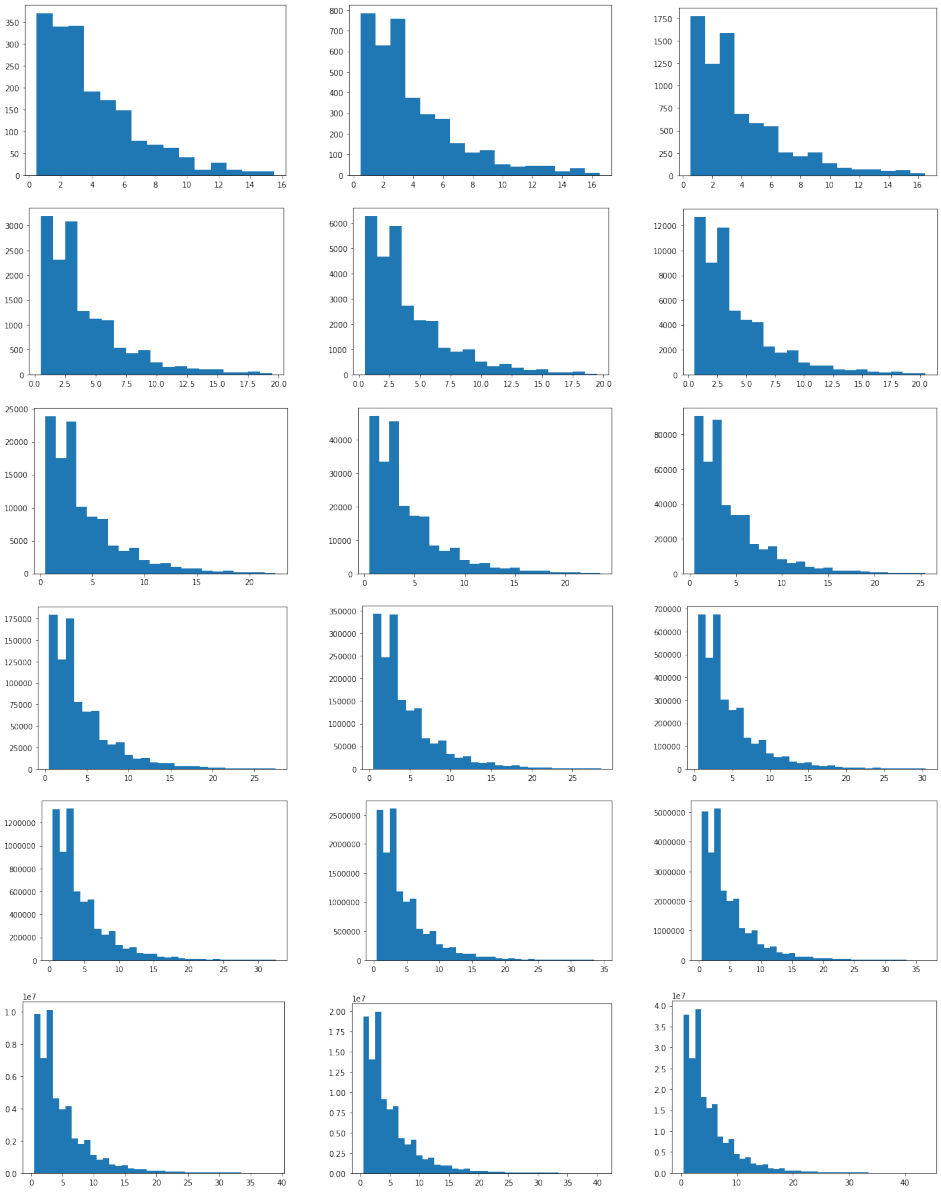

Figure 11: From left to right, top to bottom: bar graphs showing the frequency of gaps of various sizes between consecutive words in $\mathscr{U}_n$ for $n = 13, \ldots, 30$, shown out to 4 standard deviations.

6. Modular Distribution

Let us now consider the relative density of Ulam words. As we mentioned earlier, the set of Ulam words is preserved under the complement map. This forces a symmetry on congruence classes.

Theorem 11. *If $w \in \mathscr{U}_n$ then $\pi^{-1}\left(2^{n+1}-1-\pi(w)\right) \in \mathscr{U}_n$. Consequently, for any positive integer N and $a \in \mathbb{Z}/N\mathbb{Z}$,*

$$\rho_{a,N}(n) = \rho_{2^{n+1}-1-a,N}(a).$$

Proof. Given $w \in \mathscr{U}_n$, write $x = \pi(w) = a_n 2^n + \ldots + a_0$ in binary. Then

$$\pi\left(\hat{w}\right) = (1-a_n)2^n + \ldots + (1-a_0) = 2^{n+1} - 1 - x.$$

But $\hat{w} \in \mathscr{U}_n$. Now, observe that if $\pi(w) \equiv a \mod N$, then $2^{n+1}-1-\pi(w) \equiv 2^{n+1}-1-a \mod N$, which forces the equality of the relative densities. □

For $N = 2, 3$, this is particularly simple.

Corollary 2. *For any positive integer n, we have that $\rho_{0,2}(n) = \rho_{1,2}(n)$. Furthermore, for any $a \in \mathbb{Z}/3\mathbb{Z}$,*

$$\rho_{a,3}(n) = \begin{cases} \rho_{1-a,3}(n) & \text{if } n \equiv 0 \mod 2 \\ \rho_{-a,3}(n) & \text{if } n \equiv 1 \mod 2. \end{cases}$$

Proof. Observe that $2^{n+1}-1-x \equiv x+1 \mod 2$, from which $\rho_{0,2}(n) = \rho_{1,2}(n)$ follows immediately. For the second part, observe that

$$2^{n+1}-1-x \mod 3 \equiv \begin{cases} 1-x & \text{if } n \equiv 0 \mod 2 \\ -x & \text{if } n \equiv 1 \mod 2. \end{cases}$$

□

While there must always exist for any n two congruence classes $a, b \in \mathbb{Z}/3\mathbb{Z}$ such that $\rho_{a,3}(n) = \rho_{b,3}(n)$, there is no reason why the last congruence class c should be roughly equal. Indeed, for $n \leqslant 5$, we see that $\rho_{c,3}(n) = 0$. However, for larger n, it does appear to be the case that $\rho_{c,3}(n) \to \rho_{a,3}(n) = \rho_{b,3}(n)$; as we will illustrate presently. To help measure the extent to which words are equidistributing modulo N, we define the modular discrepancy.

Definition 1. For any positive integers n, N, the *modular discrepancy* is

$$d_N(n) := \max_{a,b \in \mathbb{Z}/N\mathbb{Z}} \left|\rho_{a,N}(n) - \rho_{b,N}(n)\right|.$$

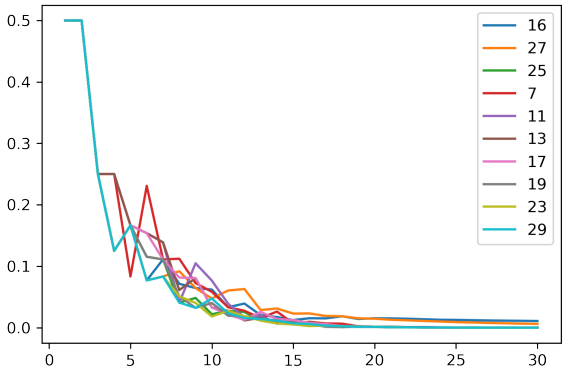

Figure 12: A plot of the modular discrepancies for prime power moduli $p^k < 30$.

Trivially, saying that Ulam words equidistribute modulo N is equivalent to saying that $d_N(n) \to 0$ as $n \to \infty$. Moreover, by appealing to the Chinese remainder theorem, proving that $d_N(n) \to 0$ for all N is reducible to proving that $d_{p^k}(n) \to 0$ for all prime powers p^k. To investigate Conjecture 3, we computed $d_{p^k}(n)$ for all prime powers $p^k < 30$—as near as we can tell, $d_{p^k}(n)$ decays exponentially as a function of n (see Figure 12).

Acknowledgements. Our collaboration was funded by four Bates College grants, all awarded by the Dean of Faculty's office: a STEM Faculty-Student Summer Research Grant and three Summer Research Fellowships. We would also like to thank Tomás Oliveira e Silva for confirming our computations of $|\mathscr{U}_n|$ and giving many helpful suggestions for improving the exposition.

References

[1] T. Bade, K. Cui, A. Labelle, and D. Li, Ulam sets in new settings, preprint, `arXiv:2008.02762`.

[2] J. Cassaigne and S. R. Finch, A class of 1-additive sequences and quadratic recurrences, *Exp. Math.* **4** (1995), 49–60.

[3] S. R. Finch, Conjectures about s-additive sequences, *Fibonacci Quart.* **29** (1991), 209–214.

[4] S. R. Finch, On the regularity of certain 1-additive sequences, *J. Combin. Theory Ser. A* **60** (1992), 123–130.

[5] S. R. Finch, Patterns in 1-additive sequences, *Exp. Math.* **1** (1992), 57–63.

[6] J. Hinman, B. Kuca, A. Schlesinger, and A. Sheydvasser, Rigidity of Ulam sets and sequences, *Involve* **12** (2019), 521–539.

[7] J. Hinman, B. Kuca, A. Schlesinger, and A. Sheydvasser, The unreasonable rigidity of Ulam sequences, *J. Number Theory* **194** (2019), 409–425.

[8] N. Kravitz and S. Steinerberger, Ulam sequences and Ulam sets, *Integers* **18** (2018), A80.

[9] A. Mandelshtam, On fractal patterns in Ulam words, preprint, `arXiv: 2211.14229`.

[10] R. Queneau, Sur les suites s-additives, *J. Combin. Theory Ser. A* **12** (1972), 31–71.

[11] J. Schmerl and E. Spiegel, The regularity of some 1-additive sequences, *J. Combin. Theory Ser. A* **66** (1994), 172–175.

[12] A. Sheydvasser, The Ulam sequence of linear integer polynomials, *J. Integer Seq.* **24** (2021).

[13] S. Ulam, Combinatorial analysis in infinite sets and some physical theories, *SIAM Rev.* **6** (1964), 343–355.