The Collatz Problem: An Iterative Algorithm

International Baccalaureate Extended Essay, Mathematics
Submitted 3 March 1998 to Mr. Smyth, teacher-advisor

By Christopher D. J. Jones
Harry Ainlay High School
Edmonton, Alberta, Canada

Final Draft Copy,
Word count: 3052 words, plus equations

Table of Contents

Introduction

One of the major unsolved problems in modern mathematics is the Collatz Problem. It involves the behaviour under iteration of a particular algorithm, and predicts that, after sufficient iteration, the value will always reach 1, for any given initial value that is an integer. While the mathematics involved with the algorithm proper are simple -- multiplication, division, and addition, its solution seems to be impossibly complex. A more generalized form of the algorithm requires the use of modulus arithmetic, and appears to have ties to Ergodic Theory and Markov Chains [Matthews 1995].

It is this behaviour: its paradoxical mix of simplicity and intractability, that makes the Collatz Problem such an interesting and unique topic of study. This extended essay will be most concerned with presenting various aspects of the problem, including aspects related to graphs, and generalizations of the Collatz Problem.

Definitions

The Collatz Conjecture states that "starting from any positive integer n, repeated iteration of [the Collatz function] eventually produces the value 1." [Lagarias 1985] The function is extremely simple, yet the solution is apparently intractable:

> Where *k* is the number of iterations performed, and *n* is the positive integer used as the initial value in the iteration,
>
> $$T^k(n) = n_{k+1} = \begin{cases} 3n_k + 1 & \text{if } n_k \text{ is odd} \\ n_k / 2 & \text{if } n_k \text{ is even} \end{cases}$$

then, the Collatz conjecture asserts that for some *k*, $T^k(n) = 1$. In other words, if the algorithm is repeated enough (ie, *k*) times, the result will eventually be 1.

Some more useful terms may be defined from this: the path (or trajectory) is defined as the sequence of numbers given by the iteration, $\{n, T^1(n), T^2(n), ..., T^{k-1}(n), T^k(n)\}$ [Lagarias 1985)]; the height of the algorithm is defined as the number of iterations required such that the function reaches 1 [Sander 1971], or according to L. E. Garner, "the number of terms in the trajectory of *n* preceding the first occurrence of 1" [Garner 1985]; and the Collatz graph is defined as "the digraph on ù with edges $n \to T^k(n)$.... The figure is arranged so that integers of the same height are found on the same horizontal line." [Garner 1985]. Finally, the stopping time is defined as being *k* for which $T^k(n) < n$, or the number of iterations until the result is less than the starting value.

Table I: Hailstone Sequences for *n* from 1 to 50

1: 1	33: 33 100 50 [25]
2: 2 [1]	34: 34 [17]
3: 3 10 [5]	35: 35 106 53 160 80 [40]
4: 4 [2]	36: 36 [18]
5: 5 16 [8]	37: 37 112 56 [28]
6: 6 [3]	38: 38 [19]
7: 7 22 [11]	39: 39 118 59 178 89 268 134 67 201 604 302 151 454 227 682 341 1024 512 256 128 [32]
8: 8 [4]	40: 40 [20]
9: 9 28 [14]	41: 41 124 62 [31]
10: 10 [5]	42: 42 [21]
11: 11 34 [17]	43: 43 130 65 196 98 49 148 74 [37]
12: 12 [6]	44: 44 [22]
13: 13 40 [20]	45: 45 136 68 [34]
14: 14 [7]	46: 46 [23]
15: 15 46 [23]	47: 47 142 71 214 107 322 161 484 242 121 364 182 91 274 137 412 206 103 310 155 466 233 700 350 175 526 263 790 395 186 593 1780 890 445 1336 668 334 167 502 251 754 377 1132 566 283 850 425 1276 638 319 958 479 1438 719 2158 1079 3238 1619 4858 2429 7288 3644 1822 911 2734 1367 4102 2051 6154 3077 9232 4616 2308 1154 577 1732 866 433 1300 650 325 976 488 244 122 61 184 92 [46]
16: 16 [8]	48: 48 [24]
17: 17 52 [26]	49: 49 148 74 [37]
18: 18 [9]	50: 50 [25]
19: 19 58 [29]	
20: 20 [10]	
21: 21 64 [32]	
22: 22 [11]	
23: 23 70 [35]	
24: 24 [12]	
25: 25 76 [38]	
26: 26 [13]	
27: 27 82 [41]	
28: 28 [14]	
29: 29 88 [44]	
30: 30 [15]	
31: 31 94 [47]	
32: 32 [16]	

Based on [Criswell 1995-1998]

The table should be read as giving the path from a number to 1 for the $3x + 1$ algorithm. The use of [*n*] indicates that from that point forth, the path is identical to that of *n*, and the reader should use the path of *n* at that point. Ie, 50: {50, [25]} becomes {50, 25, 76, [38]} becomes {50, 25, 76, 38, [19]}, etc.

As an example, if the initial value of *n* is set to 5, the path given by repeated iteration of the algorithm is {5, 16, 8, 4, 2, 1}, with a height (*k* value) of 5. Similarly, if *n* is set to 85, its path is {85, 256, 128, 64, 32, 16,

4, 2, 1}. These can also be represented in a type of shorthand notation as {5, [16]} and {85, 256, 128, 64, 32, [16]}, respectively, where [*n*] simply represents the path from that *n* onwards -- in this case, [16] would be expanded as {16, 8, 4, 2, 1}. A listing of the paths for *n* from 1 to 50 is included in Table 1. [Based on Criswell 1995]

History of the Collatz Problem

Also known as the *3n + 1* problem, the Syracuse problem, Kakutani's problem, Hasse's algorithm, and Ulam's problem, the algorithm first came to prominence in the mathematical community when S. Ulam brought it to the attention of several mathematicians at Syracuse University in the 1950s, although L. Collatz was known to have been working with problems of this nature in the 1930s. It has since then reappeared often at universities and institutes around the world, most particularly at the University of Chicago. Scientific American first publicized the problem in 1984, but Popular Computing wrote several articles in 1973 and 1975, primarily referring to its iterative nature.

According to J. Lagarias, "The $3x + 1$ Conjecture is simple to state and apparently intractably hard to solve. It shares these properties with other iteration problems, for example, that of aliquot sequences and with celebrated Diophantine equations such as Fermat's last theorem. Paul Erdös commented concerning the intractability of the $3x + 1$ problem: 'Mathematics is not yet ready for such problems.'"[Lagarias 1985]. As a result of the great interest in the problem, both from the mathematical and the computer communities, the conjecture has been tested for numbers up to 10^{15} [Vardi 1991]. It has however, not yet been proven, nor is such a event considered likely in the near future. According to Weisstein [1998], Conway "proved that Collatz-type problems can be formally Undecideable" in 1972.

One of the more colourful names for the algorithm is the "Hailstone Algorithm", which produces "Hailstone Numbers". This appellation is due to the fact that the numbers, under iteration, have a tendency to vary wildly, not unlike the supposed altitude changes of hailstones in formation. Also, once the value reaches a power of 2, it drops extremely rapidly towards its termination, just as does a hailstone. This can especially be seen on a plot of the value against the number of iterations.
Another interesting fact is that some consecutive numbers (example, 340 and 341) have the same height, while other consecutive numbers (ie, 46 and 47) differ wildly. The first property can be explained by “twins” (see below), but the second is as yet inexplicable.

Properties of the Collatz Algorithm

The algorithm is deceptively simple, yet conceals a multitude of interesting properties related to its iterative nature. The most fundamental property at work in the algorithm is that it is deterministic -- once an *n* is chosen, the path is known, and the paths of any *n* that include any other known *n* share the path of that second *n*. Second, after every $(3n + 1)$ step, $T^k(n)$ is an even number. While this is not an especially profound observation, it will be important in dealing with the reverse Collatz algorithm.

Lagarias [1985] noted that there are three possibilities for the behaviour of a particular trajectory: it can either be convergent, such that some $T^k(n) = 1$; non-trivially cyclic, such that the trajectory becomes periodic and no $T^k(n) = 1$; or divergent, such that the limit of $T^k(n)$ is infinity. In order that the function become convergent or non-trivially cyclic, $T^k(n)$ must be less than n, for some value of *k*. Obviously, this is due to the fact that only if the $T^k(n) > n$ for all *k* can the function become divergent. Thus, for $n > 1$, to have $T^k(n) = 1$ requires that first, $T^k(n) < n$. This also allows the simplification of path statements, by permitting the use of [x] to represent the path of *x*, within the trajectory of another number.

Then, is the algorithm always convergent or cyclic? Lagarias [1985], while not offering a proof, provides a probabilistic argument in favour of this: "Pick an odd integer n_o at random. Iterate *T* until another odd integer, n_1, occurs. Then, ½ of the time, $n_1 = (3n_0 + 1) / 2$; 1/4 of the time, $n_1 = (3n_0 + 1) / 4$; 1/8 of the time $n_1 = (3n_0 + 1) / 8$; and so on." Hayes [1984] also indicates that, on average, the value of *n* decreases by approximately 3/4 per iteration. It would thus seem likely that the function decreases for most, if not all values of *n*.

Terras [1976, 1979], showed that "If *A* is a set of positive integers then the density of *A* is defined in terms of the counting function μ to be":

$$o(A) \;'\; \lim\ (— \ \mu) \qquad \{n \ \#\ m \ \textit{and}\ n \ \text{Ʊ}\ A\}$$

Thus, the integers S_k such that *n* has a stopping time ˜ *k*, because $\mu = 1$ in the case that the stopping time of *n* ˜ *k*, provided *n* ˜ *x* and the stopping time ˜ k, have the limiting asymptotic density

$$ó(k) \;'\; \lim \quad (\frac{1}{m})$$

This limit approaches 1 as *k* approaches infinity, and so “almost all integers have a finite stopping time” [Weisstein 1998], and thus satisfy the conjecture. From this result, the minimum length of any possible cycle which does not obey the conjecture can be determined. Lagarias demonstrated that no nontrivial cycles shorter than 275,000 iterations exist. [Lagarias 1985]

The Reverse Collatz Algorithm

While the "forwards" algorithm is deterministic, the "reverse" algorithm, in which the sequence proceeds from 1 to *n*, is not necessarily so. This is due to the possibility that a number x could be generated by either portion of the algorithm, depending on its parent. While this behaviour is relatively infrequent, it creates a very interesting behaviour: "forking" or "branching", which can be observed in table 2: "Paths of Hailstone Iteration" [From Criswell 1997].

Since a fork can only occur when the values given by $(3n + 1)$ and $n / 2$ are equal, all forks must follow a pattern of $6a + 4$, where a 0 I and a $ 0. Evidently, the first fork ($n = 4$) is a trivial case, due to the infinite loop which begins at that point. The following fork values (10, 16, 22, etc) are all non-trivial. This creates a type of "tree", on which it is possible to see that there are some "branches" which seem to be "fork-rich", and others which appear to be "fork-poor". A good example of a fork-rich branch is that proceeding {1, 2, 4, 8, 2*k}*, which branches at every second level, while a fork-poor branch is simply one that starts with a multiple of 3, as it will never fork, and simply becomes (3 * 2 ^ *I*, where *I* is an arbitrary number). [Criswell 1995]

Upon closer inspection of the chart, an interesting oddity becomes apparent: there are "twins", consecutive numbers at the same level (having the same height), which are on different branches. These twins are invariably 3 steps above a fork point, and are a consequence of the following:

Even	*Odd*
$\frac{(2)(2)n\&1}{3}$	$\frac{(n\&1)(2)(2)}{3}$
$\frac{4n\&1}{3}$	$\frac{4n\&4}{3}$
$\frac{4n}{3}\&\frac{1}{3}$	$\frac{4n}{3}\&\frac{4}{3}$
$\frac{4n}{3}\&\frac{1}{3}$	$\frac{4n}{3}\&\frac{1}{3}\&1$

But, *(4n - 1 / 3)* and *(4n - 4 / 3)* have a difference of (3 / 3), or 1, and thus are consecutive. The number produced by following the "even" branch from any fork will always be one higher than that from the "odd" branch. Similarly, one level further removed produces a pair of consecutive even numbers (as a result of the multiplication by two), and so on. Garner [1985] believes that pairs of consecutive integers of the same height occur infinitely often and in infinitely many patterns, and there is no evidence to contradict this belief. Between 75 and 95, there are 6 pairs of numbers with the same heights and 98, 99, 100, 101, and 102 all have a height of 18.

Table II: Hailstone Trajectories for Selected Numbers

```
0  1  2  3  4  5  6   7   8   9   10   11   12   13    14    15    16                Depth

                 -  3 * 2 ^ I
                 |
                 |                                -    7
                 |                                |
                 |                        -   11 -   22 -   44
                 |                        |
                 |                        |         -   45 * 2 ^ I
                 |                        |         |
                 |                 -  17 -  34 -   68 -  136 -  272
                 |                 |
                 |                 |        -   69 * 2 ^ I
                 |                 |        |
                 |                 |        |         -  277
                 |                 |        |         |
                 |       -  13 -  26 -  52 - 104 - 208 -  416 -  832 - 1664
                 |       |
                 |       |                       -   23 -   46
                 |       |                       |
                 |       |                -  35 -   70 -  140 -  280
                 |       |                |
                 |       |                |         -  141 * 2 ^ I
                 |       |                |         |
                 |       |        -  53 - 106 - 212 -  424 -  848 - 1696
                 |       |        |
                 |       |        |        - 213 * 2 ^ I
                 |       |        |        |
                 |       |        |        |         -  853 - 1706
                 |       |        |        |         |
            -  5 - 10 -  20 -  40 -  80 - 160 - 320 -  640 - 1280 -  2560 -  5120 - 10240
            |
            |      - 21 * 2 ^ I
            |      |
            |      |                        -   75 * 2 ^ I
            |      |                        |
            |      |                        |         -  301
            |      |                        |         |
            |      |                 - 113 - 226 -  452 -  904 - 1808
            |      |                 |
            |      |                 |        -  453 * 2 ^ I
            |      |                 |        |
            |      |                 |        |         - 1813
            |      |                 |        |         |
            |      |       - 85 - 170 - 340 - 680 - 1360 -  2720 -  5440 - 10880
            |      |       |
            |      |       |                       -  151 -  302
            |      |       |                       |
            |      |       |                - 227 -  454 -  908 - 1816
            |      |       |                |
            |      |       |                |         -  909 * 2 ^ I
            |      |       |                |         |
            |      |       |        - 341 - 682 - 1364 -  2728 -  5456 - 10912
            |      |       |        |
            |      |       |        |        - 1365 * 2 ^ I
            |      |       |        |        |
            |      |       |        |        |         -  5461 - 10922
            |      |       |        |        |         |
1 - 2 - 4 - 8 - 16 - 32 - 64 - 128 - 256 - 512 - 1024 - 2048 - 4096 - 8192 - 16384 - 32768 - 65536
```

This is a Collatz Graph, showing (at the bottom) the "base" of the tree. One could pick any number shown on the tree, and follow its path to 1 graphically, reading from right to left, or reading from left to right, run "up" the tree from 1 to any given number on the tree. Branches are indicated by "T-intersections", such as that at 32, leading to 64 and $21*2^i$.

From [Criswell 1995-1998]

Parity Vectors

In the midst of all these branches, is there a way to concisely represent the path taken, both in the forwards and backwards algorithms? Since there are only two branches available at any time, it is possible to, arbitrarily assigning the binary value 0 to the "even" part of the algorithm, and 1 to the "odd" part, represent the path not as a sequence of numbers, but rather as a sequence of binary values representing the unique path to and from a number. Garner [1985] defines this to be a "Parity Vector", with the following definition:

> "V(n) of *n* of length *k* is found by reducing the first *k* terms of the trajectory of *n* (mod 2). That is, $V(n) = \{v_0, v_1, \ldots v_{k-1}\}$, with $v = 0$ if $T^k(n)$ is even, and $v = 1$ if $T^k(n)$ is odd."

For example, the path vector of 34 can be determined as follows:

> Path(34) = {34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1}
> V(34) = {0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0}

This vector sequence uniquely identifies 34 in terms of the Collatz Problem's two choices -- either dividing by 2, or multiplying by 3, and adding 1. To reach 34 from 1, one must multiply by 2 four times, subtract one and divide by 3, etc....

Evidently, all even numbers will commence with a 0, and odd ones with a 1, because an odd number cannot be created by a multiplication by two of an integer (which a Parity Vector value of 0 requires). Also, because there is only one way to reach a *k* of more than 4 (multiplying by 2 four times), all numbers having a *k* of greater than 4 will have a terminating 0000 group.

Further, no two 1s can be consecutive, as the algorithm ensures the creation of even numbers from odd numbers. [Criswell 1995] It is intuitively obvious that there will, on the whole, be more 0s than 1s in parity vectors, due to the fact that there are more multiples of two (and thus more opportunities for a 0 to occur) than multiples of three (required for a 1 to occur). Indeed, one-half of all natural numbers are multiples of two, while one-third are multiples of three.

Finally, "twin pairs" have their first three vector parity bits reversed: 001 in the "even" twin, and 100 in the "odd" twin. This is simply due to the pattern which is found in the twin pairs, where the “odd” one is multiplied by two at the “fork” (the 3rd bit), while the “even” twin is not multiplied by two (thus giving the 1 as the 3rd bit).

The Generalized Collatz Problem

In order to generalize the Collatz problem, so as to examine it for broader patterns, it is helpful to redefine the $3x+1$ algorithm so as to divide by 2, regardless of whether the number is even or odd, such that:

$$T(x) = \begin{cases} (3x + 1) / 2 & \text{if } x \text{ odd} \\ x / 2 & \text{if } x \text{ even} \end{cases}$$

This then allows the division to regulated by one variable, as will be seen below. One consequence of this change (first suggested by Terras [1976, 1979]) is that the paths may be altered slightly, but the rest of the behaviour of the algorithm remains the same.

$T^k(x) =$	{ $m_0 x + r_0$ / 0 mod d { $m_1 x + r_1$ / 1 mod d { [and so forth to I] { $m_i x + r_i$ / I mod d

where d is a positive integer greater than or equal to 2, m_0,... m_{d-1} are nonzero integers, and mod d is the modulus function, to the base d (ie, "clock arithmetic" such that 23 mod 12 = 11). Also, r_i Θ Z must satisfy r_i / I m_i (mod d). Thus,

$$T(x) \quad ' \quad \frac{m_i \; x \& r_i}{d}$$

when x / I (mod d) is a generalized Collatz mapping. [Weisstein 1998] This opens the door to similar, but not identical functions. Do they all necessarily follow the same behaviour? Hayes [1984], noted that the $3n$-1 algorithm is equivalent to $3n+1$.

For example, let d be 3, m_0 be 1, r_0 be 4, m_1 be 2, r_1 be 7, m_2 be 17, and r_2 be 1. (These numbers were chosen randomly by the author of this extended essay, to clarify how the generalized Collatz mapping works, and have no known special characteristics) Then, the Collatz algorithm for those values is,

$$T(x) = \begin{cases} (x - 4) / 3 & \text{if x is divisible by 3, or is 0} \\ (2x - 7) / 3 & \text{if x has a remainder of 1 after division by 3} \\ (17x - 1) / 3 & \text{if x has a remainder of 2 after division by 3} \end{cases}$$

For this particular algorithm, then, the path of 5 would be {5, 28, (49 / 3), ...}. Evidently, the restriction that the results be integers does not apply in this case, but there is no readily apparent reason why the function cannot operate with the real numbers, instead of only with the integers, except simplicity.

According to Weisstein, the following conjectures were proposed by Matthews and Watts in 1984:

> "* If $[m_0 \ldots m_{d-1}] < d^d$ then all trajectories T(n) for *n* Θ Z eventually cycle.
> * If $[m_0 \ldots m_{d-1}] > d^d$, then almost all trajectories for *n* Θ Z are divergent, except for an exceptional set of Integers *n* satisfying # {*n* Θ *S* | -X# *n* # X} = o(X).
> * The number of cycles is finite."

[Weisstein 1998, in reference to Matthews, K. R. and Watts, A. M. "A Generalization of Hasse's Generalization of the Syracuse Algorithm", Acta Arithmetica vol 43, pp. 167-175. (1984).]

If Matthews' and Watts' conjectures are also correct, then the example given above of a non-$3x+1$ Collatz algorithm should eventually cycle (as all of m_0, m_1, and m_2 are less than 3^3). Since all of the possible algorithmic rules (depending on the divisibility of the number by three) involve a division by three, it seems likely that the numbers iterated through that algorithm would, on average, decrease, and approach zero. Of course, this is not a proof, but it is similar in nature to Lagarias' argument for the validity of the (original) Collatz Conjecture.

Conclusion

The question of whether or not the conjecture is, in fact, correct, remains unresolved, and, barring breakthrough advances in the field, is, in the consensus opinion of experts on the conjecture, likely to remain unsolved for quite some time. It is, in a sense, a testament to the beauty and power of mathematics that, from such a simple set of instructions, an intricate and complex relationship between numbers can be formed. From a simple question: if numbers are run through a given algorithm repeatedly, will they all end up at 0?", comes a complex set of relationships between modulus arithmetic, graph theory, and other disciplines. The discovery of "twins" in the reverse algorithm shows how, without prior warning, mathematics has the ability to continually surprise and delight its practitioners in creating structure that was not apparent until examined in a new light. It is, perhaps, this outgrowth of its logical foundations that is the key to mathematics' complexity and richness.

"Hailstone" numbers and the Collatz Problem do not seem to have any practical application: they have no known new fundamental properties, but the experience the mathematical community has gained in attempting to prove or disprove the conjecture may be useful in suggesting possibilities for other iterative problems. This type of problem does not seem to be vulnerable to solution by algebraic methods, and requires "brute-force" solution by the means of trial and error and computational solution.

Interesting possibilities for further research in this field include cryptological applications: one could use the property of the family of algorithms that it is much easier to go "forwards" than "backwards" as a building-block of an encryption mechanism. This clearly has "real-world" uses, but the existence of parity vectors could create a possible method of circumventing such a mechanism.

Bibliography

Bohm, C. & G. Sontacchi, On the existence of cycles of given length in integer sequences like $x_{n+1} = x_n / 2$ if x_n even, and $x_{n+1} = 3x_n + 1$ otherwise.
Atti della Accademia Nazionale dei Lincei Rendiconti, vol. 64, pp 260-264. (1978)

Criswell, Evans A. Collatz Conjecture.
INTERNET http://mickey.cs.uah.edu/~criswell/collatz.html (1995 - 1997)

Garner, L. E. On heights in the Collatz 3n + 1 problem.
Discete Math, vol. 55, pp. 57-64. (1985)

Hayes, Brian. Computer Recreations: on the ups and downs of hailstone numbers.
Scientific American, January 1984. (1984)

Lagarias, J. The set of rational cycles for the 3x + 1 problem.
Acta Arithmetica, vol. 46, pp. 33-53. (1973)

Lagarias, J. The 3x + 1 problem and its generalizations.
American Math Monthly, vol. 92, pp 3-23. (1985)

Matthews, K. R. The Generalized 3x + 1 Mapping.
INTERNET: http://www.maths.uq.oz.au/~krm/survey.dvi (1995)

Sander, J. On the (3N + 1)-conjecture.
Acta Arithmetica, vol. 55, pp. 241-248. (1990)

Terras, R. A stopping-time problem on the positive integers.
Acta Arithmetica, vol. 30, pp. 241-252. (1976)

Terras, R. On the existence of a density.
Acta Arithmetica, vol. 35, pp. 101-102. (1979)

Weistein, Eric W. Eric's Treasure Trove of Mathematics.
INTERNET http://www.astro.virginia.edu/~eww6n/math (1998)

Wirsching, G. A Markov chain underlying the backwards Syracuse algorithm.
Rev. Roum. Math pures appl., vol. 39, pp. 915-926. (1994)