A Novel Approximation of the Collatz Conjecture Through Discrete Fourier Transformations

By Michael Petrizzo

1. Introduction

a. Definition

The Collatz Conjecture, also known as the 3x+1 problem, is an infamously simplistic algorithm that remains unsolved to date. The conjecture is as follows: given any positive integer, if this is an even number, divide it by 2, else an odd number, triple it and add one. Do these series of operations result in a convergence or divergence? Surprisingly, this easily understandable conjecture is riddled with lengthy videos from popular Mathematicians such as Veritasium or Numberphile. I found myself discovering one of these videos, and becoming instantly intrigued, how could such an easy problem stump some of the world's greatest mathematicians? In fact, a video regarding the conjecture states it to be "The most dangerous problem in mathematics", due to the countless hours spent researching it. However, regardless of fierce warnings, this paper aims to explore approximations, end behaviors, and engage a relatively novel idea of utilizing machine learning optimization to model the stopping iterations of the conjecture at integer values. Ideally, these methods can create a generalized formula to better predict the relative maximum, minimum, and possibly end-behaviors of the conjecture.

Mathematically, the Collatz Conjecture is a piecewise function as follows,

$$f(n) = \left\{ \begin{array}{ll} \frac{n}{2} & if\ n\ mod\ 2{=}0, \\ 3n{+}1 & if\ n\ mod\ 2{=}1. \end{array} \right\}$$

Where the function is recursively inputted until the values converge to a 4→2→1→4 cycle (Or diverges!), each additional recursion increments a tracking value representing the stopping time of the initial number. The repeating 4,2,1 cycle is the only currently discovered cyclic

convergence of the conjecture. Interestingly, as a side note, the required length of a 'non-trivial' cycle - One which does not follow the 4,2,1 cycle- is ever-increasing as higher numbers are tested within the conjecture.

b. Function Visualizations

To demonstrate these values, hence the graph of a randomized integer, 10, and its respective recursion up until the stopping time.

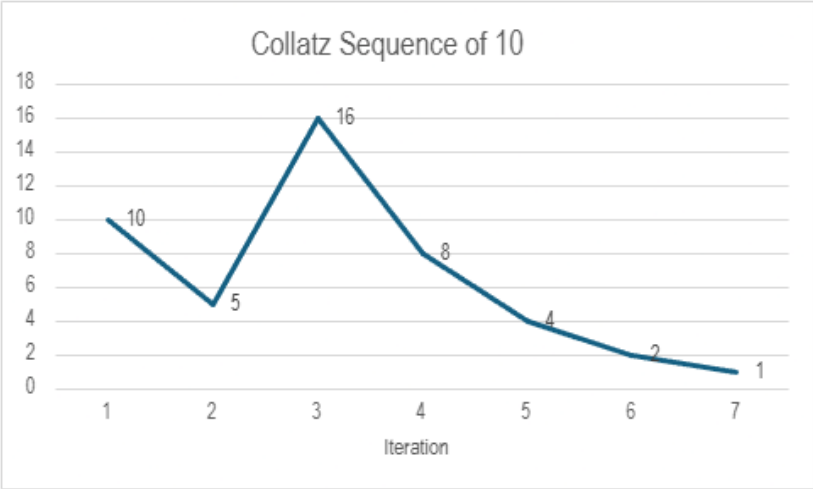

Figure 1: Collatz Sequence of Value 10

The stopping time of f(10) is hence 7, with each iteration representing unique values relative to the cycle. The relationship between starting values can be shown in a directed graph, where the stopping time is the total number of orbitals between the initial value and 1.

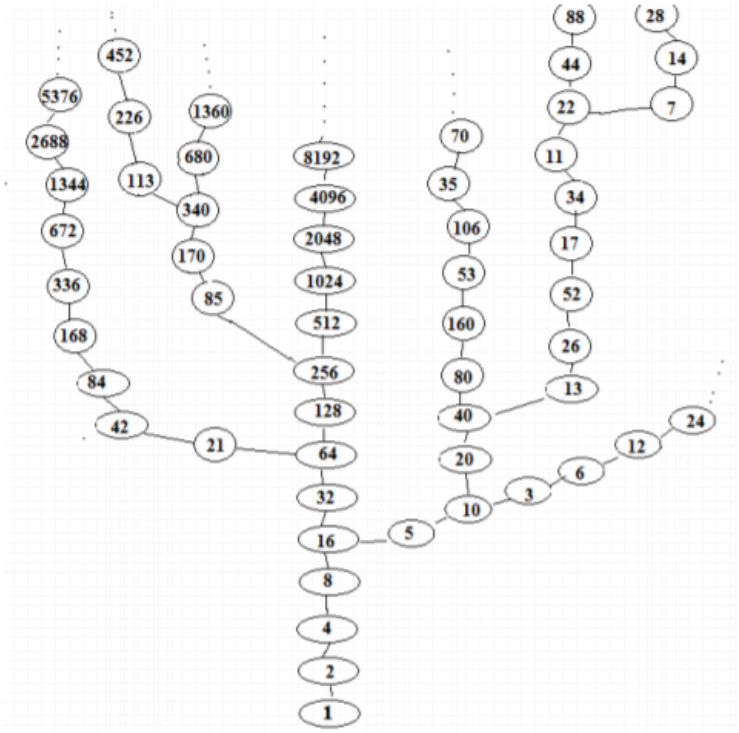

Figure 2: Collatz Conjecture Directed Graph -

https://ultimatetheorem.blogspot.com/2020/08/visualization-of-collatz-conjecture_10.html

The conjecture would hence be proven false with a value that does not converge into a known integer in the complete directed graph, or true with a definitive proof such that all values "branch" into each other. Crucially, a counter-example would prove this conjecture false, in which any integer has a diverging stopping time. A large value to counter a conjecture is also not unheard of, for example, the Polya Conjecture was disproved initially with a value of $1.845 * 10^{361}$.

c. Prior Research and Known Values

The known 4,2,1 cycle is not the only basis for current advances in the Conjecture. In fact, it has been developed since the 1990s with modern advancements in the 2000s. However, most of this research focused on discovering a proof to the conjecture, such as Terence Tao's work in 2019 discovering that *almost all* stopping times are shorter than any function which eventually diverges into infinity, no matter the growing speed. In this case, *almost all* is a fancy math term for stating a negligible amount of exceptions with respect to a logarithmic density. Furthermore, computational methods of brute-forced have been checked at least once for all numbers up to $\sim 2.36 * 10^{21}$ as of January 2025 (Roosendaal, 2025). This computational method holds multiple interesting aspects and defines additional valuable definitions that will be utilized throughout this paper.

N = The starting positive Integer

S_i = Series of N, where $S_0 = N$. (Etc. $S_0 = 4, S_1 = 2$)

$Mx(N) = \lim_{k \to \infty} Max(S_0, S_1, \ldots, S_k)$, The maximum value of S_i. (Etc. $Mx(5) = 16$, 5>16>8>4>2)

$G(N) = Min_k(S_K < N)$, The minimum number of iterations required before a sequence has a value below its starting value, called the Glide.

$D(N) = Min_k(S_k = 1)$, The minimum number of iterations required before a sequence has a value of 1.

$O(N)$ $E(N)$, The total odd and even numbers in the sequence respectfully.

$O(N) + E(N) = D(N)$

$S(N) = 5 * O(N) - 3 * E(N)$, The *Strength* of a sequence, which weighs the odd(growing) and even(shrinking) numbers to a growth of 5:3, or 1.67. Numbers with positive strength grow much more and often have higher delays.

$L(N) = [S(N)/8]\ where\ [*]\ denotes\ flooring$, The *Level* of a sequence, which scales the strength with advantages in representing similar numbers in distinct integers.

2. Fourier Analysis and Approximation

In attempting to discover underlying trends present within the conjecture, I first approximated the Strength and Level by utilizing a Fast Fourier Transform (FFT). The FFT is a tool utilized in signal processing, however, it is used to represent a complex function through the sum of simpler trigonometric functions. In this case, with discrete values, the more precise mathematical definition is the Discrete Fourier Transform through the FFT algorithm. But firstly, we must turn the Fourier Series (The sum of trig functions) into a form we could effectively utilize.

Fourier Series: $$f(x) = a_0 + \sum_{n=1}^{\infty}(a_n cos(n\omega x) + b_n sin(n\omega x))$$

a_0 =DC component/average value

a_n, b_n, n= magnitudes

ω=Period

To solve this, we must learn advanced calculus in the Maclaurin Series, whereas a function is represented by a series of its derivatives. By representing a function as a series of derivatives around a point (in this case, the origin), the Maclaurin Series is equivalent to the function as

each progressive derivative is equal. If you're interested, a non-origin based series is called a Taylor Series. Anyways, if two functions have 1st, 2nd, 3rd.. Kth, $\lim_{K \to \infty}$, they are equivalent because their rates of change are equal. The general formula for the Maclaurin series is as follows: $f(x) = f(0) + \frac{f'(0)x}{1!} + \frac{f''(0)x^2}{2!} \ldots \frac{f^{K'}(0)x^n}{n!}$ where ' denotes a derivative. By taking the derivative of the series, we can see the trends of cancellation. Further, this is applicable to all continuous functions, whereas we can discover an awfully interesting formula.

$e^x = 1 + \frac{e^0 x}{1!} + \frac{e^0 x^2}{2!} \ldots \frac{x^n}{n!}$

$cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} \ldots \frac{(-1)^n x^{2n}}{(2n)!}$ (cos(x)' = -sin(x))

$sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} \ldots \frac{(-1)^n x^{2n+1}}{(2n+1)!}$ (sin(x)' = cos(x))

Here are three series, $e^x, cos(x), and\ sin(x)$ as the Maclaurin series. If you don't believe me, I gladly suggest you attempt to create the series. The trig functions follow alternating series due to their derivatives looping in cycles of four, $(cos(x) >- sin(x) >- cos(x) > sin(x) > cos(x))$. And lastly, every other term cancels due to $sin(0)$. I'll write another variation of e^x, and you may see something very interesting.

$e^{ix} = 1 + ix + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} \ldots \frac{(ix)^m}{n!} = 1 + ix - \frac{x^2}{2!} - i\frac{x^3}{3!} + \frac{x^4}{4!} + i\frac{x^5}{5!} \ldots$

Which looks an awful lot like the trig series! In fact,

$isin(x) = ix - i\frac{x^3}{3!} + i\frac{x^5}{5!} \ldots$

And combining the terms, we can arrive at the most beautiful equation in all of mathematics.

Euler's Formula: $e^{xi} = cos(x) + isin(x)$

Where $cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2}$ $sin(\theta) = \frac{e^{i\theta} - e^{-i\theta}}{2i}$

(Proving this from Euler's formula is left as an exercise for the reader)

Now back to why we did this in the first place, hence again the Fourier Series:

$$f(x) = a_0 + \sum_{n=1}^{\infty}(a_n cos(n\omega x) + b_n sin(n\omega x))$$

We can now represent the trigonometry in terms of e,

$$cos(n\omega x) = \frac{e^{in\omega x} + e^{-in\omega x}}{2} \quad sin(n\omega x) = \frac{e^{in\omega x} - e^{-in\omega x}}{2i}$$

Substituting this into the Fourier series…

$$f(x) = a_0 + \sum_{n=1}^{\infty}[a_n \frac{e^{in\omega x} + e^{-in\omega x}}{2} + b_n \frac{e^{in\omega x} - e^{-in\omega x}}{2i}]$$

$$f(x) = a_0 + \sum_{n=1}^{\infty}[\frac{a_n}{2} e^{in\omega x} + \frac{a_n}{2} e^{-in\omega x} + \frac{b_n}{2i} e^{in\omega x} - \frac{b_n}{2i} e^{-in\omega x}]$$

$$f(x) = a_0 + \sum_{n=1}^{\infty}[(\frac{a_n}{2} + \frac{b_n}{2})e^{in\omega x} + (\frac{a_n}{2} - \frac{b_n}{2})e^{-in\omega x}]$$

$$c_n = \frac{a_n}{2} - \frac{b_n}{2i} \quad c_{-n} = \frac{a_n}{2} + \frac{b_n}{2i}$$

We arrive at this, the Fourier Transform.

$$f(x) = a_0 + \sum_{-\infty}^{\infty} c_n e^{in\omega x}$$

The range is adjusted from -∞ to +∞ due to the cyclic nature of the trig functions and hence c_{-n} is omitted. Also, fun fact, $e^{in\omega x}$ is oftentimes called the twiddle factor because it rotates the function.

This is also commonly written as

$$f(x) = a_0 + \sum_{-\infty}^{\infty} c_n e^{in\frac{2\pi}{T}x}$$, where $\omega = \frac{2\pi}{T}$

However… this function still cannot be used because it utilizes a continuous interval, and our S(N) and L(N) are discrete values, aka Integers.

To fix this, the function must be converted discretely, more specifically, the twiddle factor will be altered into a regular interval.

$x_n = f(\frac{kT}{N}), k = A\ Regular\ Interval,\ ie\colon\ 0, 1, \ldots N - 1$

Further, the Summation range is reestablished in interval from 0 to N-1. Because of this, the continuous coefficients are mirrored once more and c_{-n} is used rather than c_n.

Adjusting the twiddle factor: $$e^{\frac{i2\pi nx}{T}} \longrightarrow e^{\frac{i2\pi n(\frac{kT}{N})}{T}} = e^{\frac{i2\pi kn}{N}}$$

$$DFT(x) = a_0 + \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N}$$

And finally at the Discrete Fourier Transform:

Then finally, in the context of S(N) (or identically for L(N))

$$F(k) = DC + \sum_{n=0}^{N-1} S(n)e^{-i2\pi kn/N}$$

a daunting formula derived from the coefficient analog of a Fourier series(citation). The FFT, which is actually a set of algorithms, is usually defined through the Cooley-Tukey algorithm which utilizes recursive computation of divisional aspects of the DFT, such as

$$F(k) = \sum_{n=0}^{N/2-1} S(2n)e^{-i2\pi k2n/N} + \sum_{n=0}^{N/2-1} S(2n+1)e^{-i2\pi k(2n+1)/N}$$

, with the most vital part of this denoting the n/2-1, splitting the DFT into two parts, odd(2n) and even(2n+1) indices. Then the twiddle term can be factored out, and each Summation is represented as Even and Odd

$$F(k) = EF(k) + e^{-i2\pi k/N}OF(k)$$

And the next period of the wave,

$$F(k + N/2) = EF(k) - e^{-i2\pi k/N}OF(k)$$

This effectively splits the DFT into two parts by period, and the speed is saved by reusing the results of multiple DFT cycles computationally, into $O(nlog(n))$ complexity. The magnitude of each Fourier Component, indicating the strength of the sinusoidal wave in representing trends, is calculated through Pythagorean formula substitution,

$$|F(k)| = \sqrt{Real(F(k))^2 + Imaginary(F(k))^2}$$

combining both the Real and Imaginary coefficients. The resulting magnitudes result in a symmetrical image due to the positive and negative elements of a sinusoidal wave, meaning half

of the graph, left or right, is indicative. Thankfully, I never will have to code the entirety of this algorithm, as most scientific libraries include it. In much of code, algorithms like these are abstracted, only returning the output, but learning the internal formulas greatly provides reasoning to a process. The series utilized in this case is the strength values from 1 to an upper limit, where strength was calculated by counting the odd and even counts through the S(N). While this method is not the most efficient, upper bounds can be reached to an extent.

Figure 3:

Values beyond this bound result in large processing time exceeding an hour. However, generally, as the range increases, the curve ultimately becomes smoother with peaks remaining consistent at generally similar regions. The large peak at 0, known as the DC component (the average of all the data) is not important - or so I thought - having removed the DC component, this peak at 0 is still present, demonstrating the potential for a large periodic behavior. The presence of additional peaks however infers additional potential repetition between numbers, as the strength of a specific sinusoidal component is high, being a cyclic function. Furthermore, the presence of the peaks appears to be from a source that decreases in cyclic frequency, resulting in numerous “V” shapes, where there are multiple values that appear to be multiples in strength compared to each other. Lastly, the bottom “line” of continual plots indicates background noise or non-dominant strengths.

Investigating further and more precisely determining these frequencies, I sought to analyze multiple high-strength frequencies and detect if they appear periodically. For example, with a peak, detect its frequency and see if values at multiples of this are similar in strength to each other. I computed the Mean Absolute Error $= \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$ to reflect this

similarity. Defining this broadly, find various (~100) potential frequencies, comparing 50 frequency multiples.

Table 1: Frequencies and MAE - Strength

Fundamental Frequency	MSE
1.00E-06	8.28E-06
2.20E-05	0.000535
2.60E-05	0.000637
0.031250	0.397188
0.062500	1.156875
0.093750	1.941250
0.187500	4.320000
0.203125	4.718125
0.250000	5.910625
0.312500	7.503125

Table 2: Frequencies and MAE - Level

Fundamental Frequency	MSE
1.00E-06	5.70E-06
1.10E-05	0.000252
1.50E-05	0.000354
2.10E-05	0.000507
2.90E-05	0.000710
0.062500	1.156875
0.093750	1.941250
0.187500	4.320000
0.203125	4.718125
0.250000	5.910625

Interestingly, the dominant frequencies between Strength and Level are slightly different, not drastically, but have notable differences. Furthermore, the frequency near 0 is shown with a very small MSE, possibly representing a very large oscillating trend. There are further fundamental frequencies beyond these low values which provide additional possible indicators

but suffer from a larger relative MSE. The addition of a larger sequence would assist in either adding additional frequencies or providing a more solid MSE.

3. Conclusion

While not exhaustive, this paper has provided additional insight into the underlying connections within the Collatz Conjecture. By applying a novel approach—utilizing a Fast Fourier Transform-assisted Discrete Fourier Transform—I uncovered potential periodic structures that may have implications in advanced number theory. This exploration demonstrated the effectiveness of sinusoidal modeling in analyzing the Strength and Level of the Collatz Conjecture, further validated by Mean Squared Error (MSE) as an accuracy metric.

A key limitation in any approximation technique for this conjecture is sequence bounding, as computational constraints restrict the upper limits of analysis. While optimization in C++ could extend these boundaries, the overall trends remained consistent across different data ranges. Though approximating behaviors over the entire sequence is not definitive, this study suggests promising directions for further exploration.

As a next step, I plan to implement a Neural Network to analyze deeper relationships within the conjecture. By leveraging the FFT-identified strengths, this network aims to develop a more structured method for predicting positive Strength values—those most closely associated with extended glides. Such advancements could provide a more systematic approach to identifying rare cases and refining our understanding of the conjecture’s underlying patterns.

4. Bibliography

Lagarias, Jeffrey C. (1985). The 3x+1 Problem and its Generalizations. *The American Mathematical Monthly*. https://doi.org/10.1080/00029890.1985.11971528

Numberphile. UNCRACKABLE? The Collatz Conjecture - Numberphile. (2016). Retrieved from https://www.youtube.com/watch?v=5mFpVDpKX70

Roosendall, E. (2025). On the 3x+1 Problem. Retrieved from http://www.ericr.nl/wondrous/index.html#status

Tao, T. (2022). Almost all orbits of the Collatz map attain almost bounded values. *Forum of Mathematics, Pi*. https://doi.org/10.1017/fmp.2022.8

Veritasium. The Simplest Math Problem No One Can Solve - Collatz Conjecture. (2021). Retrieved from https://www.youtube.com/watch?v=094y1Z2wpJg.