Divisions by Two in Collatz Sequences: A Data Science Approach

Christian Koch[1,a*], Eldar Sultanow[2,b] and Sean Cox[3,c]

[1]Technische Hochschule Nürnberg Georg Simon Ohm, Nuremberg, Germany

[2]Capgemini, Nuremberg, Germany

[3]RatPac-Dune Entertainment, Los Angeles, USA

[a]christian.koch@th-nuernberg.de, [b]eldar.sultanow@capgemini.com, [c]sean.cox@ratpacent.com

Keywords: Collatz Conjecture, Divisions by Two, Binary Representation, Data Science

Abstract. The Collatz conjecture is an unsolved number theory problem. We approach the question by examining the divisions by two that are performed within Collatz sequences. Aside from classical mathematical methods, we use techniques of data science. Based on the analysis of 10,000 sequences we show that the number of divisions by two lies within clear boundaries. Building on the results, we develop and prove an equation to calculate the maximum possible number of divisions by two for any given Collatz sequence. Whenever this maximum is reached, a sequence leads to the result one, as conjectured by Lothar Collatz. Furthermore, we show how many divisions by two are required for a cycle of a specific length. The findings are valuable for further investigations and could form the basis for a comprehensive proof of the conjecture.

Introduction

The Problem

The Collatz conjecture is a well-known number theory problem and is the subject of numerous publications.[1] Therefore, our description of the topic will be brief. The mathematician Lothar Collatz introduced a function $g : \mathbb{N} \to \mathbb{N}$ as follows:

$$g(x) = \begin{cases} 3x+1 & \text{if } x \equiv 1 (\text{mod } 2) \\ x/2 & \text{if } x \equiv 0 (\text{mod } 2) \end{cases} \tag{1}$$

The conjecture, as treated in this paper, claims that the above function leads to the final result one for every natural starting number, when applied recursively. A series of numbers involved in this process is called a Collatz sequence. With an aim to contribute to a proof of the conjecture, this paper analyses a central aspect of the problem: the divisions by two.[2]

Determining Odd Numbers

Sultanow, Koch and Cox demonstrated that odd numbers of Collatz sequences can be calculated with the following recursive equation:[3]

$$v_{n+1} = 3^n \cdot v_1 \cdot \prod_{i=1}^{n} \left(1 + \frac{1}{3v_i}\right) \cdot \prod_{i=1}^{n} 2^{-\alpha_i} \tag{2}$$

The variable v_1 denotes the first odd number of the sequence, that is, the starting value. The variable v_i symbolises the odd number that is the result of a particular iteration.[4] The exponent n stands for

[1] An overview is provided by Lagarias [1].

[2] Details on our scientific approach can be found in appendix "Scientific Approach".

[3] See Sultanow, Koch, and Cox [2, p. 10].

[4] For $n = 1$ this is the starting value v_1.

the count of odd numbers that are processed by the algorithm. In the further course of this paper we will call the parameter n the *length* of a sequence. The exponent α_i finally represents the number of divisions by two that are performed in a specific iteration. Accordingly, the sum of α_i is the count of divisions by two leading from the starting value v_1 to the outcome v_{n+1}.[5] Let us consider the example $v_1 = 13$ and $n = 2$. Applying equation 2 yields:[6]

$$v_{2+1} = 3^2 \cdot 13 \cdot \left(1 + \frac{1}{3 \cdot 13}\right) \cdot \left(1 + \frac{1}{3 \cdot 5}\right) \cdot 2^{-7} = 1$$

Starting with $v_1 = 33$ for $n = 3$ we obtain the result:

$$v_{3+1} = 3^3 \cdot 33 \cdot \left(1 + \frac{1}{3 \cdot 33}\right) \cdot \left(1 + \frac{1}{3 \cdot 25}\right) \cdot \left(1 + \frac{1}{3 \cdot 19}\right) \cdot 2^{-5} = 29$$

Improving readability, we denote the factor $\left(1 + \frac{1}{3 \cdot v_i}\right)$ with the variable β_i. In addition, we generalise the formula by replacing the factor three with the variable k. This will be useful for further analysis and leads us to the following generalised version of equation 2:

$$v_{n+1} = k^n \cdot v_1 \cdot \prod_{i=1}^{n} \left(1 + \frac{1}{k v_i}\right) \cdot \prod_{i=1}^{n} 2^{-\alpha_i} \tag{3}$$

$$v_{n+1} = k^n \cdot v_1 \cdot \prod_{i=1}^{n} \beta_i \cdot \prod_{i=1}^{n} 2^{-\alpha_i}$$

In order to correctly calculate odd numbers with formula 3, we must first define the halting conditions of the algorithm in the next section.

Halting Conditions

Being compliant with the Collatz conjecture, the algorithms 2 and 3 halt if at least one of the following conditions is fulfilled:

$$\begin{aligned} &1. \quad v_{n+1} = 1 \\ &2. \quad v_{n+1} \in \{v_1, v_2, v_3, \ldots, v_n\} \end{aligned} \tag{4}$$

When the first condition applies, the Collatz conjecture is true for a specific sequence. If the second condition is fulfilled, the sequence has led to a cycle. For every starting value, except $v_1 = 1$, the Collatz conjecture is therefore falsified.[7] Let us consider the example $k = 3$, $v_1 = 13$, and $n = 2$. Applying equation 3 yields:

$$v_{2+1} = 3^2 \cdot 13 \cdot \left(1 + \frac{1}{3 \cdot 13}\right) \cdot \left(1 + \frac{1}{3 \cdot 5}\right) \cdot 2^{-7} = 1$$

In the above example the algorithm halts after two iterations because the first condition is fulfilled. If we examine the case $v_1 = 1$, we realise that the algorithm finishes after the first iteration, since both halting conditions are true:

$$v_{1+1} = v_1 = 3^1 \cdot 1 \cdot \left(1 + \frac{1}{3 \cdot 1}\right) \cdot 2^{-2} = 1$$

The sequence stops in the example above due to the result being one. Apart from that, the sequence has led to a cycle.

---

[5] For a glossary of notations see section "Glossary of Notations" in the appendix.

[6] The result of the first iteration v_{1+1} equals five.

[7] This statement refers to the Collatz conjecture in its original form $3v + 1$.

Boundaries of α_i

We know that in every iteration of the equations 2 and 3 at least one division by two is performed. This follows from the constraints of the Collatz problem. Consequently, we can define the minimum of α_i with the following condition:

$$1 \leq \alpha_i$$

The maximum can be specified in a similarly easy way. According to the halting conditions, defined in the previous section, a Collatz sequence finishes when $v_{n+1} = 1$. The maximum of α_i, hereinafter called $\hat{\alpha}_i$, can hence be defined as:

$$\begin{aligned} 2^{\hat{\alpha}_i} &= k \cdot v_i + 1 \\ \hat{\alpha}_i &= \log_2 k + \log_2 v_i + \log_2 \beta_i \end{aligned} \tag{5}$$

The formula above builds on the fact that the expression $2^{\hat{\alpha}_i}$ must equal the next even number $k \cdot v_i + 1$ in order to lead to $v_{n+1} = 1$. Being greater, the result v_{n+1} would be less than one. The second step inverses the exponentiation of $\hat{\alpha}_i$ by taking the binary logarithm. Appropriately, we replace the operation *plus one* by β_i. For a better understanding of the above term, let us consider the example $k = 3$ and $v_1 = 5$. In this case equation 5 results in:

$$\alpha_1 = \hat{\alpha}_1 = 4 = \log_2 3 + \log_2 5 + \log_2 \left(1 + \frac{1}{3 \cdot 5}\right)$$

Whenever a sequence reaches the maximum $\hat{\alpha}_i$, it finishes with one, thus verifying the Collatz conjecture. If we could prove that every odd number finally leads to this maximum for $k = 3$, the Collatz problem would be solved. Summarising, we can define the following boundaries for α_i:

$$1 \leq \alpha_i \leq \log_2 k + \log_2 v_i + \log_2 \beta_i \tag{6}$$

Before we continue, we validate theorem 6 empirically. We will do so at various points in this paper to avoid obvious errors in our mathematical reasoning. The basis for the validation is a sample of $10,000$ Collatz sequences. The data set comprises information about sequences for the odd starting numbers $v_1 \in \{1, 3, 5, \ldots, 3999\}$ and $k \in \{1, 3, 5, 7, 9\}$. Since we do not know that all generated sequences halt, we limited the number of iterations per sequence to $n = 100$. For further details on the data set, see section "Data Set" in the appendix.

Unsurprisingly, we found that all values of α_i in the sample are compliant with theorem 6.[8] In the next section we move on to more sophisticated considerations and study the properties of $\prod_1^n 2^{\alpha_i}$.

Analysing α

Boundaries of α

In equations 2 and 3, the expression $\prod_{i=1}^n 2^{\alpha_i}$ represents the divisions by two performed by the algorithms. The number of divisions by two can be determined with the following formula and will be symbolised by α:

$$\alpha = \sum_{i=1}^{n} \alpha_i$$

[8] Source: Own empirical analysis, see appendix "Data Set" for details.

Based on theorem 6 we can define the minimum of α as follows:

$$n \leq \alpha$$

Since we carry out at least one division by two in every iteration of formulas 2 and 3, the minimum of α equals the sequence's length. The maximum value is harder to determine. In the first step we derive it empirically from the data set mentioned in the previous section. Based on the observed data we formulate the hypothesis that the maximum of α can be calculated with the following equation:

$$\begin{aligned} \hat{\alpha} &= \lfloor n \cdot \log_2 k + \log_2 v_1 \rfloor + 1 \\ \alpha &\leq \hat{\alpha} \end{aligned} \tag{7}$$

The hypothesis holds for all Collatz sequences in the empirical data set.[9] If a Collatz sequence reaches the above stated maximum, it finishes with one, as conjectured by Lothar Collatz.[10] Let us, for example, consider the case where $v_1 = 13$, $n = 2$ and $k = 3$. Applying theorem 7 and formula 3 leads to:

$$\begin{aligned} \hat{\alpha} &= \lfloor 2 \cdot \log_2 3 + \log_2 13 \rfloor + 1 = 7 \\ v_{2+1} &= 3^2 \cdot 13 \cdot \left(1 + \tfrac{1}{3 \cdot 13}\right) \cdot \left(1 + \tfrac{1}{3 \cdot 5}\right) \cdot 2^{-7} = 1 \end{aligned}$$

The empirical validation supports our hypothesis, but does not prove it for all Collatz sequences. Throughout the next sections we will formulate a comprehensive proof of theorem 7 step by step.

Proving $\hat{\alpha}$ for $k = 1$

First, we examine the case $k = 1$, where theorem 7 can be simplified as follows:

$$\hat{\alpha} = \lfloor n \cdot \log_2 1 + \log_2 v_1 \rfloor + 1 = \lfloor \log_2 v_1 \rfloor + 1 \tag{8}$$

In order to prove theorem 7, we have to demonstrate that the number of divisions by two, α, is less than or equal to the maximum $\hat{\alpha}$. This can be achieved by analysing the binary representation of Collatz numbers.[11] Let us consider the case $v_1 = 25$ and $k = 1$ in the decimal system. Applying equation 3 leads to the sequence shown in the following table.

n	variable	decimal	log 2	binary	binary length	α_i	α	operation
1	v_1	25	4.64	11001_2	5			$+1$
	$v_1 + 1$	26	4.70	11010_2	5	1	1	$\cdot 2^{-1}$
2	v_2	13	3.70	1101_2	4			$+1$
	$v_2 + 1$	14	3.81	1110_2	4	1	2	$\cdot 2^{-1}$
3	v_3	7	2.81	111_2	3			$+1$
	$v_3 + 1$	8	3.00	1000_2	4	3	5	$\cdot 2^{-3}$
4	v_4	1	1.00	1_2	1			

Table 1: Binary representation of a Collatz sequence for $k = 1$

The sequence presented in table 1 starts with the decimal number $v_1 = 25$ at $n = 1$. Subsequently it comprises the odd numbers $v_2 = 13$, $v_3 = 7$ and finally $v_4 = 1$. In the binary system the sequence starts accordingly with $v_1 = 11001_2$. The binary length of the starting number $len(v_1)$ equals five.[12]

[9] Source: Own empirical analysis, see appendix "Data Set" for details.

[10] The parameter n, representing the length of a sequence, cannot be predicted for a specific k and v_1 with the formula.

[11] To avoid confusion between decimal and binary numbers, we will label binary numbers with a subscripted 2.

[12] With binary length we mean the count of digits of a binary number.

This observation is crucial for our proof. For clarification, it is important to note that the length of a binary number can be calculated with the following equation:[13]

$$len(v_i) = \lfloor \log_2 v_i \rfloor + 1 \tag{9}$$

For example, consider the case $v_i = 13$ in decimal, rendered in binary that means $v_i = 1101_2$. Here, equation 9 leads to the following result:

$$len(13) = len(1101_2) = \lfloor \log_2 13 \rfloor + 1 = 4$$

The comparison of equation 9 with formula 8 makes it clear that they are identical. This raises the question why the maximum number of divisions by two of a Collatz sequence corresponds to the binary length of v_1.[14] To answer this, we take a closer look at the mechanics of a Collatz sequence in the binary system.

We start with $v_1 = 11001_2$ in the above example. Adding one, we obtain the even number $v_1 + 1 = 11010_2$. The binary length of v_1 equals the binary length of $v_1 + 1$, which is five. Due to the trailing zero we immediately realise that $v_1 + 1$ is even. A division by two can be performed in the binary system by deleting the trailing zero. The result is $v_2 = 1101_2$. Adding one again, leads to the next even number $v_2 + 1 = 1110_2$. Deleting the trailing zero once more, results in $v_3 = 111_2$.

Up to this point we have performed two divisions by two. The parameter α therefore equals two. The case $v_3 = 111_2$ is very important for our proof. Adding one to $v_3 = 111_2$, leads to an overflow of the binary number. As a result, we obtain the even number $v_3 + 1 = 1000_2$, which is a power of two and equals 2^3 in decimal. Knowing that every power of two in a Collatz sequence directly leads to the terminal value $v_{n+1} = 1$, we can deduce that the sequence ends after the third iteration.

The binary length $len(v_3) = 3$ increases to $len(v_3 + 1) = 4$ in the final step. This situation only occurs once in a Collatz sequence for $k = 1$. Whenever adding one to a number v_n causes an overflow of its binary representation, the result $v_n + 1$ will be a power of two. The binary length will, in this scenario, increase from $len(v_n)$ to $len(v_n) + 1$. The sequence will consequently halt. For all other cases the following condition applies:[15]

$$len(v_n) = len(v_n + 1) > len(v_{n+1})$$

Only the final iteration increases the length of the binary number. In any other case the binary length decreases from v_n to v_{n+1}.

Let us now reflect what this implies for the maximum $\hat{\alpha}$. We know that the binary length of the starting value v_1 can be calculated with equation 9. In order to reach the final result $v_{n+1} = 1$, starting at v_1, we have to perform the following number of divisions by two:

$$\alpha = \hat{\alpha} = len(v_1) + 1 - 1 = \lfloor \log_2 v_1 \rfloor + 1 \tag{10}$$

The equation builds on the binary length of the starting value $len(v_1)$. We add one to respect the binary overflow in the final iteration. Furthermore, we subtract the binary length of the final result $v_{n+1} = len(v_{n+1}) = 1$. No value of α can possibly exceed this maximum, since $\hat{\alpha}$ directly leads to the terminal value $v_{n+1} = 1$, halting the sequence.

The above equation thus proves theorem 7 for $k = 1$. In the next section we will explain why this argumentation is in principle valid for all k.

---

[13] See Sedgewick and Wayne [3, p. 185].
[14] The statement is only true for $k = 1$.
[15] The statement is only true for $k = 1$.

Proving $\hat{\alpha}$ for $k > 1$

Let us now examine the case $k = 3$, which is most interesting because it relates to the original Collatz conjecture. The first question we need to address is whether or not the principles discussed in the previous paragraph are transferable to this form of the problem. To find an answer, we analyse a sequence, starting with $v_1 = 17$ and $k = 3$. The results are displayed in the following table.

n	variable	decimal	log 2	binary	binary length	α_i	α	operation
1	v_1	17	4.09	10001_2	5			$\cdot 3$
	$3v_1$	51	5.67	110011_2	6			$+1$
	$3v_1 + 1$	52	5.70	110100_2	6	2	2	$\cdot 2^{-2}$
2	v_2	13	3.70	1101_2	4			$\cdot 3$
	$3v_2$	39	5.29	100111_2	6			$+1$
	$3v_2 + 1$	40	5.32	101000_2	6	3	5	$\cdot 2^{-3}$
3	v_3	5	2.32	101_2	3			$\cdot 3$
	$3v_3$	15	3.91	1111_2	4			$+1$
	$3v_3 + 1$	16	4.00	10000_2	5	4	9	$\cdot 2^{-4}$
4	v_4	1	1.00	1_2	1			

Table 2: Binary representation of a Collatz sequence for $k = 3$

The example presented in table 2 reveals that in comparison to the previous case $k = 1$, the algorithm performs an additional operation, which is the multiplication with three. This operation leads to a growth of the binary length when comparing v_n to $3v_n$. The result of the operation can be calculated as follows:

$$len(3v_n) = \lfloor \log_2 3 + \log_2 v_n \rfloor + 1$$

In determining the maximum $\hat{\alpha}$ for $k = 3$, we have to take the additional binary growth into account. With regard to the operation $+1$ we can utilise the same arguments as in the previous section. Whenever adding one leads to an overflow in the binary representation of $3v_n$, the result will be a power of two, halting the sequence. The length of $(3v_n + 1)$ will, in this case, increase by one in contrast to $3v_n$. This can happen only once in a Collatz sequence, since the resulting power of two will lead to a termination.

In order to prove our hypothesis, we have to adjust equation 8 by considering the additional binary growth that is caused by the multiplications with three. Therefore, we obtain the following formula:

$$\alpha = \hat{\alpha} = \lfloor n \cdot \log_2 3 + \log_2 v_1 \rfloor + 1 \tag{11}$$

The above term proves theorem 7 for the case $k = 3$. A closer look makes clear that it is not only valid for $k = 3$, but for all k. The appendix outlines an alternative approach to verification of theorem 7. In conclusion, we can define the following boundaries for the number of divisions by two in a Collatz sequence:

$$n \leq \alpha \leq \hat{\alpha} \tag{12}$$

If one could establish that every sequence finally leads to $\hat{\alpha}$, that means to a binary overflow of $3v_n + 1$, the Collatz problem would be solved. In the following we will discuss the consequences of our findings for the occurrence of cycles and further confirm our line of reasoning.

Occurence of Cycles

Definition

A promising possibility to falsify the Collatz conjecture in its original form is a cycle. We have found such a counterexample if the following halting condition from section ”Introduction” is fulfilled:

$$v_{n+1} \in \{v_1, v_2, v_3, \ldots, v_n\}$$

The single known cycle for $k = 3$ is the trivial one starting with $v_1 = 1$:

$$v_1 = 1 = v_{1+1} = 3 \cdot 1 \cdot \left(1 + \frac{1}{3 \cdot 1}\right) \cdot 2^{-2}$$

The Collatz conjecture claims that the above example is the only possibility of a cycle for $k = 3$. Based on equation 3 we derive the following condition for the occurrence of a cycle within a Collatz sequence:[16]

$$2^{\alpha} = k^n \cdot \prod_{i=1}^{n} \beta_i \tag{13}$$

For the convenience of the reader, the expression $\prod_{i=1}^{n} \beta_i$ will be referred to as β subsequently. Showing that equation 13 is true for $k = 3$ would partially prove the Collatz conjecture. Yet there would still remain the possibility of an eternally growing sequence. This makes theorem 7 particularly interesting.

A major difficulty in analysing cycles in Collatz sequences is that there seems to be just one example. This is, however, not true for our generalised form of the problem. Let us consider the case $k = 5$ and $v_1 = 13$. Applying formula 3 leads to a cycle of the length $n = 3$:

$$13 = 5^3 \cdot 13 \cdot \left(1 + \frac{1}{5 \cdot 13}\right) \cdot \left(1 + \frac{1}{5 \cdot 33}\right) \cdot \left(1 + \frac{1}{5 \cdot 83}\right) \cdot 2^{-7}$$

Setting $k = 5$ and $v_1 = 13$ in equation 13, we obtain the following result after three iterations:

$$128 = 5^3 \cdot \left(1 + \frac{1}{5 \cdot 13}\right) \cdot \left(1 + \frac{1}{5 \cdot 33}\right) \cdot \left(1 + \frac{1}{5 \cdot 83}\right)$$

To determine the number of divisions by two, which can lead to a cycle, we need to investigate the parameter β more thoroughly.

Analysing β

The starting point of our analysis of β is theorem 7. The formula can be used to calculate the maximum possible divisions by two of a Collatz sequence:

$$\hat{\alpha} = \lfloor n \cdot \log_2 k + \log_2 v_1 \rfloor + 1$$

In section "Analysing $\boldsymbol{\alpha}$" we showed that $\hat{\alpha}$ relates to the binary length of the starting value v_1. Furthermore, the equation accounts for the binary growth, caused by the n-fold multiplication with k as well as the final overflow, triggered by the operation $+1$. If a sequence reaches $\hat{\alpha}$, it halts at the terminal value $v_{n+1} = 1$. In order to learn more about the parameter β, we take a look at the relation between theorem 7 and equation 3. We examine the situation in which formula 3 leads to the final result one. Consequently, we set $v_{n+1} = 1$ and $\alpha = \hat{\alpha}$:

$$\begin{aligned}
&1 = k^n \cdot v_1 \cdot \textstyle\prod_{i=1}^{n} \beta_i \cdot 2^{-\hat{\alpha}} \\
&1 = k^n \cdot v_1 \cdot \beta \cdot 2^{-\hat{\alpha}} \\
&2^{\hat{\alpha}} = k^n \cdot v_1 \cdot \beta \\
&\hat{\alpha} = n \cdot \log_2 k + \log_2 v1 + \log_2 \beta \\
&\lfloor n \cdot \log_2 k + \log_2 v_1 \rfloor + 1 = n \cdot \log_2 k + \log_2 v_1 + \log_2 \beta \\
&\log_2 \beta = -n \cdot \log_2 k - \log_2 v_1 + \lfloor n \cdot \log_2 k + \log_2 v_1 \rfloor + 1
\end{aligned} \tag{14}$$

[16] See Sultanow, Koch, and Cox [2, p. 11].

For a better understanding of the above term, let us examine two examples. We begin with the border case where $k = 1$ and $v_1 = 1$. Here, equation 14 leads to:

$$\log_2 \beta = 1 = -n \cdot \log_2 1 - \log_2 1 + \lfloor n \cdot \log_2 1 + \log_2 1 \rfloor + 1$$
$$\beta = 2$$

Moreover, we study the example where $k = 5$, $v_1 = 19$ and $n = 2$. Equation 14 in this case results in:

$$\log_2 \beta = 0.1082 = -2 \cdot \log_2 5 - \log_2 19 + \lfloor 2 \cdot \log_2 5 + \log_2 19 \rfloor + 1$$
$$\beta = 1.0780$$

Based on equation 14 and the fact that β must always be greater than one, we define the following boundaries of β:

$$0 < \log_2 \beta \leq 1$$
$$1 < \beta \leq 2 \tag{15}$$

The limits formulated by theorem 15 are confirmed through a validation with our empirical data set.[17] Figure 1 shows the maximum β for different values of k in the sample. The diagram as well depicts the corresponding starting number v_1, which leads to this maximum.

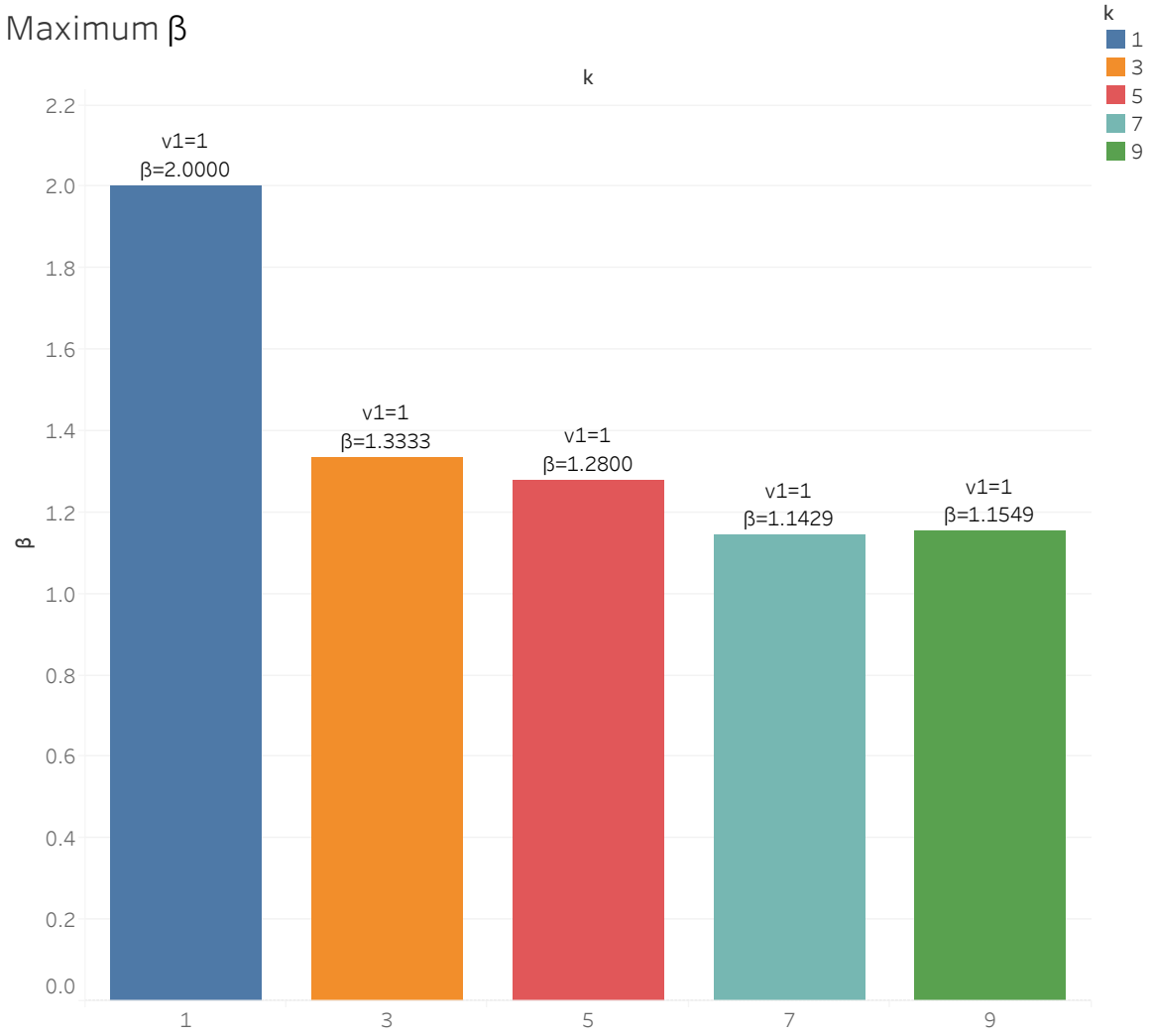

Fig. 1: Maximum β for different k

[17] Source: Own empirical analysis, see appendix "Data Set" for details.

As we can see from figure 1, the maximum for $k = 1$ equals 2. The limit for the other k is beneath. For example, the maximum for $k = 3$ equals $1.\overline{3} = \frac{4}{3}$. The diagram reveals that the maximum β for every k is reached for the starting number $v_1 = 1$. A proof for this finding will be provided in a future article. In the next section we will discuss the implications of theorem 15 on the occurrence of cycles.

Analysing $\bar{\alpha}$

How many divisions by two can lead to a cycle within a Collatz sequence? We can derive an equation for this number, subsequently called $\bar{\alpha}$, on the basis of formula 3 and theorem 15. Therefore, we examine the case in which equation 3 leads to a cycle by setting $v_{n+1} = v_1$:

$$\begin{aligned} v_1 &= k^n \cdot v_1 \cdot \beta \cdot 2^{-\bar{\alpha}} \\ 2^{\bar{\alpha}} &= k^n \cdot \beta \\ \bar{\alpha} &= n \cdot \log_2 k + \log_2 \beta \\ \bar{\alpha} &= \lfloor n \cdot \log_2 k \rfloor + 1 \end{aligned} \tag{16}$$

The last transformation above is applied, since $\bar{\alpha}$ is a whole number.[18] Now that it is clear that $1 < \beta \leq 2$, we truncate the fractional part of $(n \cdot \log_2 k)$ and add one to the result. In a Collatz sequence a cycle can only occur if the number of divisions by two equals $\bar{\alpha}$. Conversely, this does not imply that reaching $\bar{\alpha}$ inevitably leads to a cycle. The following example demonstrates this. Let us consider the case where $k = 3$, $v_1 = 83$ and $n = 3$. Here, theorem 16 and formula 3 yield the following result:

$$\bar{\alpha} = 5 = \lfloor 3 \cdot \log_2 3 \rfloor + 1$$
$$71 = 3^3 \cdot 83 \cdot \left(1 + \tfrac{1}{3 \cdot 83}\right) \cdot \left(1 + \tfrac{1}{3 \cdot 125}\right) \cdot \left(1 + \tfrac{1}{3 \cdot 47}\right) \cdot 2^{-5}$$

Before we continue, we will empirically validate theorem 16. Our tool is a linear search performed by a Python script. For details on the program see section "Cycle Finder" in the appendix. With the script we searched and evaluated cycles in Collatz sequences for the odd starting numbers $v_1 \in \{1, 3, 5, \ldots, 9999\}$ and $k \in \{1, 3, 5, \ldots, 999\}$. In order to restrict the runtime of the program we limited the length of the investigated cycles to $n = 100$. The results of our empirical validation are shown in the following table.

k	$v_1 \,..\, v_n$	n	α	$\bar{\alpha}$
1	(1)	1	1	1
3	(1)	1	2	2
5	(1, 3)	2	5	5
5	(13, 33, 83)	3	7	7
5	(27, 17, 43)	3	7	7
7	(1)	1	3	3
15	(1)	1	4	4
31	(1)	1	5	5
63	(1)	1	6	6
127	(1)	1	7	7
181	(27, 611)	2	15	15
181	(35, 99)	2	15	15
255	(1)	1	8	8
511	(1)	1	9	9

Table 3: Cycles in Collatz sequences

[18] Accordingly, the sum of the mantissas of $n \cdot \log_2 k + \log_2 \beta$ must be one.

As one can see in table 3, we found several cycles for our generalised form of the Collatz problem. All of which comply with theorem 16.[19]

Binary Growth

As we have emphasised at several points in this paper, theorem 7 builds on the binary length of the starting value $len(v_1)$. Furthermore, it accounts for the maximum binary growth, henceforth denoted with $\hat{\Lambda}$. We define binary growth as the total number of digits by which the binary length of v_1 increases in a sequence.[20] In order to reach the final result $v_{n+1} = len(v_{n+1}) = 1$, we have to substract $\hat{\alpha}$ from the sum of the binary length of v_1 and the binary growth:

$$\begin{aligned} 1 &= len(v_1) + \hat{\Lambda} - \hat{\alpha} \\ \hat{\Lambda} &= \hat{\alpha} + 1 - len(v_1) \\ \hat{\Lambda} &= \lfloor n \cdot \log_2 k + \log_2 v_1 \rfloor + 1 + 1 - \lfloor \log_2 v_1 \rfloor - 1 \\ \hat{\Lambda} &= \lfloor n \cdot \log_2 k + \log_2 v_1 \rfloor + 1 - \lfloor \log_2 v_1 \rfloor \\ \lfloor n \cdot \log_2 k \rfloor + 1 &\leq \hat{\Lambda} \leq \lfloor n \cdot \log_2 k \rfloor + 2 \end{aligned} \tag{17}$$

In the final step the above equation is condensed by subtracting the starting value v_1. As a result, we obtain a range for $\hat{\Lambda}$. The reason is a possible overflow which can be instigated by the expression $n \cdot \log_2 k + \log_2 v_1$. Let us examine two examples to illustrate this. Starting with the case $k = 3$, $v_1 = 13$ and $n = 2$ we find that the result is equal to the lower limit of $\hat{\Lambda}$:

$$\hat{\Lambda} = \lfloor 2 \cdot \log_2 3 + \log_2 13 \rfloor + 1 - \lfloor \log_2 13 \rfloor = \lfloor 2 \cdot \log_2 3 \rfloor + 1 = 4$$

Setting $v_1 = 7$, $k = 3$ and $n = 5$ leads to the upper limit of the variable:

$$\hat{\Lambda} = \lfloor 5 \cdot \log_2 3 + \log_2 7 \rfloor + 1 - \lfloor \log_2 7 \rfloor = \lfloor 5 \cdot \log_2 3 \rfloor + 2 = 9$$

The parameter $\hat{\Lambda}$ represents the maximum binary growth of a Collatz sequence. In other words, the binary growth of a sequence cannot exceed $\hat{\Lambda}$, even if we would not perform any divisions by two. Examining formula 17, it is not surprising that we find the following relation to theorem 16:

$$\bar{\alpha} = \lfloor n \cdot \log_2 k \rfloor + 1 \leq \hat{\Lambda}$$

As we know, a cycle occurs in a Collatz sequence when the condition $v_1 = v_{n+1}$ is fulfilled. The binary length of the starting number v_1, must therefore grow exactly as much as it is reduced by the divisions by two. Thus, for a cycle to occur, the number of divisions by two has to be equal to the binary growth.

One might argue that this reasoning is erroneous since a sequence does not necessarily reach the maximum binary growth. We build on formula 3 to show that our arguments are valid. By setting $v_{n+1} = v_1$ we examine the case where the growth of the binary length of a sequence is neutralised by the divisions by two:

$$\begin{aligned} v_1 &= k^n \cdot v_1 \cdot \beta \cdot 2^{-\alpha} \\ 2^\alpha &= k^n \cdot \beta \\ \alpha &= n \cdot \log_2 k + log_2 \beta \end{aligned}$$

Knowing that $1 < \beta \leq 2$, we derive the following limits for the binary growth of a cycle, subsequently called $\bar{\Lambda}$:

$$n \cdot \log_2 k < \bar{\Lambda} \leq \lfloor n \cdot \log_2 k \rfloor + 1$$

The binary growth of every Collatz sequence that leads to a cycle must lie within these boundaries. Due to the fact that $\bar{\alpha}$ is a whole number, it is obvious that it must equal the maximum on the right side of the expression. For all other cases a cycle is impossible.

---

[19] Source: Own empirical analysis, see appendix "Cycle Finder" for details.

[20] This means that $\hat{\Lambda}$ does not account for the divisions by two that reduce the binary length of v_1.

Summary

In our paper we have shed light on a central aspect of the Collatz conjecture: the divisions by two. We analysed the problem in its original form $3v + 1$ as well as in the generalised variant $kv + 1$. Based on mathematical reasoning and empirical studies we derived and proved theorems on the occurrence of cycles and the termination of sequences. Our reasoning primarily builds on the binary representation of Collatz numbers and the underlying operations. Theorem 16 determines the number of divisions by two that can lead to a cycle. The theorem is based on the simple truth that a cycle can only occur if the binary growth of a sequence is exactly neutralised by the divisions by two. Theorem 7 determines the maximum number of divisions by two that can be performed in a sequence. If one could show that every starting number finally leads to this maximum, the Collatz problem would be solved. We are convinced that a profound study of the binary mechanics of Collatz sequences will lead to this proof.

Appendix

Data Set

This empirical data set was used to derive and validate theorems 6, 7 and 15. The sample was generated with a Python script and comprises information about sequences for the odd starting numbers $v_1 \in \{1, 3, 5, \ldots, 3999\}$ and $k \in \{1, 3, 5, 7, 9\}$.[21] Since we do not know that all generated sequences halt, we limited the number of iterations per sequence to $n = 100$. In total, the sample contains 651,159 Collatz numbers, which are not necessarily distinct. This is due to the fact that different starting numbers can lead to the same subsequent values. For example, both starting values, $v_1 = 13$ and $v_1 = 53$, result in the number five.

Cycle Finder

This Python script was used to validate theorem 16.[22] The program performs a linear search for the odd starting numbers $v_1 \in \{1, 3, 5, \ldots, 9999\}$ and $k \in \{1, 3, 5, \ldots, 999\}$. To restrict the runtime of the script, we limited the length of the investigated cycles to $n = 100$. Furthermore, the results are not persisted. In order to reproduce our findings, the program must be executed again.

Scientific Approach

The contents published in this paper have been achieved with an interdisciplinary approach. Not surprising, we applied classic mathematical theory and reasoning. Since we are convinced that the Collatz problem cannot be solved with traditional maths alone, we additionally used techniques of data science. We combined the two fields in different ways. On one hand, we analysed sequences and related features empirically, in order to derive new formulas and theorems. On the other hand, we used data science to validate our proofs. As suggested by Karl Popper, we tried to falsify them with counterexamples. In the course of our work, we have learned that the combination of the two fields leads to a very efficient working mode. This might be the topic of another paper.

[21] `https://github.com/c4ristian/collatz/blob/v1.3/run_alpha_export.py`
[22] `https://github.com/c4ristian/collatz/blob/v1.3/run_cycle_finder.py`

Alternative Verification of $\hat{\alpha}$

We may alternatively verify theorem 7 for $k = 3$ with the following equation. Formula 18 builds on the so-called Engel expansion and calculates the odd number v_{n+1} for a Collatz sequence in which we divide by two only once per iteration:[23]

$$v_{n+1} = \frac{3^n \cdot (v_1 + 1) - 2^n}{2^n} \tag{18}$$

The above term represents the (hypothetical) case in which a sequence rises to its maximum for a specific starting value v_1. For the Collatz conjecture this is the worst-case scenario because the equation never leads to the result one due to the steady increase. Let us consider the example $v_1 = 7$ and $n = 1$. Applying equation 18 yields:

$$v_{1+1} = \frac{3^1 \cdot (7 + 1) - 2^1}{2^1} = 11$$

Setting $v_1 = 31$ and $n = 3$ results in:

$$v_{3+1} = \frac{3^3 \cdot (31 + 1) - 2^3}{2^3} = 107$$

We use formula 18 to verify theorem 7 by proving that dividing $\hat{\alpha}$ times by two will lead to a number $v_{n+2} < 2$ for every worst-case sequence.[24] For this purpose, we extend formula 18 by an additional step and divide by $2^{\hat{\alpha}-n}$ in the final iteration:

$$v_{n+2} = \left[\left(\frac{3^n \cdot (v_1 + 1) - 2^n}{2^n}\right) \cdot 3 + 1\right] \cdot 2^{-(\hat{\alpha}-n)}$$

We assume that the sequence leads to a result $v_{n+2} < 2$ and thus verify theorem 7:

$$\begin{aligned}
2 \;&>\; \left[\left(\tfrac{3^n \cdot (v_1+1) - 2^n}{2^n}\right) \cdot 3 + 1\right] \cdot 2^{-(\hat{\alpha}-n)} \\
2^{\hat{\alpha}-n+1} \;&>\; \left[\left(\tfrac{3^n \cdot (v_1+1) - 2^n}{2^n}\right) \cdot 3 + 1\right] \\
2^{\hat{\alpha}-n+1} - 1 \;&>\; \tfrac{3^{n+1} \cdot (v_1+1) - 3 \cdot 2^n}{2^n} \\
2^{\hat{\alpha}+1} - 2^n \;&>\; 3^{n+1} \cdot (v_1 + 1) - 3 \cdot 2^n \\
2^{\hat{\alpha}+1} - 2^n + 3 \cdot 2^n \;&>\; 3^{n+1} \cdot (v_1 + 1) \\
2^{\hat{\alpha}+1} + 2^{n+1} \;&>\; 3^{n+1} \cdot (v_1 + 1) \\
2^{\lfloor (n+1) \cdot \log_2 3 + \log_2 v_1 \rfloor + 2} + 2^{n+1} \;&>\; 3^{n+1} \cdot (v_1 + 1)
\end{aligned} \tag{19}$$

When resolving $\hat{\alpha}$ in the last transformation above we have to use the factor $n + 1$, since we have extended our worst-case sequence by an additional iteration. Even though it is not obvious at the first glance, the above condition is true for any given n and v_1. Inequality 19 therefore verifies theorem 7 for $k = 3$, at least for all worst-case sequences.[25] We will elaborate on this approach more deeply in a future article.

[23] See Laarhoven [4, p. 11]

[24] The case is hypothetical. Not every worst-case sequence actually ends with $v_{n+2} = 1$. Consequently, we have to define the condition $v_{n+2} < 2$ in order to cover all possible sequences.

[25] Verifying theorem 7 for all k requires a generalised version of equation 18.

Glossary of Notations

Notation	Description
v_1	First odd number of a Collatz sequence, also referred to as starting value
v_i	Odd number that is the result of a particular iteration. In the first iteration this is the starting value v_1
k	Factor that is multiplied with odd numbers; equals three for the original Collatz conjecture
n	Count of odd numbers in a sequence, also referred to as the sequence's length
β_i	Symbolises the term $\left(1 + \frac{1}{k \cdot v_i}\right)$
β	Product of all β_i
α_i	Represents the number of divisions by two that are performed in a specific iteration
α	Number of divisions by two that leads from the starting value v_1 to the result v_{n+1}
$\hat{\alpha}_i$	Maximum possible number of divisions by two in a specific iteration
$\hat{\alpha}$	Maximum possible number of divisions by two in a Collatz sequence
$\bar{\alpha}$	Number of divisions by two that is required for a cycle
$\hat{\Lambda}$	.md Maximum binary growth of a Collatz sequence
$\bar{\Lambda}$	Binary growth that is required for a cycle

References

[1] J. C. Lagarias: The Ultimate Challenge: The 3x+1 Problem. American Mathematical Society, 2010, ISBN 978-0821849408

[2] E. Sultanow, C. Koch and S. Cox: Collatz Sequences in the Light of Graph Theory (Fourth Version). University of Potsdam, 2020, DOI https://doi.org/10.25932/publishup-44325

[3] R. Sedgewick and K. Wayne: Algorithms (Fourth Edition). Addison-Wesley Professional, 2011, ISBN 978-0321573513

[4] T.M.M. Laarhoven: The 3n + 1 conjecture. Eindhoven University of Technology, July 2009.