Divisions by Two in Collatz Sequences

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 a Collatz sequence. Besides 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 formulate and prove several theorems on the occurrence of cycles and the termination of Collatz sequences. The findings are useful 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 we treat it 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 will be called a Collatz sequence. With the 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 showed that odd numbers of Collatz sequences can be calculated with the following recursive equation[3]:

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

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 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 is performed in a specific iteration.

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

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

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

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

Accordingly, the sum of α_i is the count of divisions by two that leads from the starting value v_1 to the outcome v_{n+1}. Let us consider the example $v_1 = 13$ and $n = 2$. Applying equation **??** leads to

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

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 the subsequent chapters. 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:

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

In order to correctly calculate odd numbers with equation 3, we have to define the halting conditions of the algorithm in the next chapter.

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.[5] Let us consider the example $k = 3$, $v_1 = 13$, and $n = 2$. Applying equation 3 leads to:

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

In the above example the algorithm halts after two iterations, since 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:

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

The sequence stops in the example above, because the result is one. Furthermore, the sequence has led to a cycle.

[5] 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 likewise easy way. According to the halting conditions, defined in the previous chapter, 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 theorem 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 theorem 2 results in:

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

If a sequence reaches the maximum $\hat{\alpha}_i$, it finishes with one, 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 interval $v_1 \in [1 \ldots 3999]$ 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".

Not surprisingly, we found that all values of α_i in the sample are compliant with theorem 6.[6] In the next chapter we move on to more sophisticated considerations and study the properties of $\prod_1^n 2^{\alpha_i}$.

Analyzing 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:

$$\alpha = \sum_{i=1}^{n} 2^{\alpha_i} \tag{7}$$

[6] Source: Own empirical analysis, see section "Data Set" for details.

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

$$\alpha \geq n$$

Since we carry out at least one division by two in every iteration of theorem 2 and theorem 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 chapter. Based on the empirical data, we formulate the hypothesis that the maximum of α can be calculated with the following equation:

$$\begin{aligned} \hat{\alpha} &= \lfloor n \cdot log_2 k + log_2 v_1 \rfloor + 1 \\ \alpha &\leq \hat{\alpha} \end{aligned} \tag{8}$$

The hypothesis holds for all Collatz sequences in the empirical data set.[7] If a Collatz sequence reaches the above formulated maximum, it finishes with one, as conjectured by Lothar Collatz. Let us for example consider the case $v_1 = 13$, $n = 2$ and $k = 3$. Applying theorems 8 and 3 leads to:

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

Throughout the next sections we will formulate a proof of the hypothesis step by step.

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

First, we examine the case $k = 1$, where theorem 8 can be simplified as follows:

$$\hat{\alpha} = \lfloor n \cdot log_2 1 + log_2 v_1 \rfloor + 1 = \lfloor log_2 v_1 \rfloor + 1 \tag{9}$$

In order to prove theorem 8, we have to show that the number of divisions by two, α is less or equal than the maximum $\hat{\alpha}$. This can be achieved by analysing the binary representation of Collatz numbers.[8] Let us consider the case $v_1 = 25$ and $k = 1$ in the decimal system. Applying theorem 3 leads to the sequence shown in the following table .

n	variable	decimal	log2	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 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.[9] This observation is crucial for our proof.

[7] Source: Own empirical analysis, see section "Data Set" for details.

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

[9] With binary length we mean the count of the digits of a number expressed in the decimal.

For understanding, it is important to note that the length of a binary number can be calculated with the following equation[10]:

$$len(v_i) = \lfloor log_2 v_i \rfloor + 1 \tag{10}$$

For example, consider the case $v_i = 13$ in decimal, that means $v_i = 1101_2$ in binary. Here, the equation 10 leads to the following result:

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

The comparison of equation 10 with theorem 12 makes 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?[11] 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 tell 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[12]:

$$len(v_n) = len(v_n + 1) > len(v_{n+1}) \tag{11}$$

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 theorem 10. 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:

$$\alpha = \hat{\alpha} = len(v_1) + 1 - 1 = \lfloor log_2 v_1 \rfloor + 1 \tag{12}$$

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[13].

[10] See Sedgewick and Wayne [3, p. 185].

[11] The statement is only true for $k = 1$.

[12] The statement is only true for $k = 1$.

[13] The following notebook can be used to validate the proof experimentally:
`https://github.com/c4ristian/collatz/blob/master/notebooks/binary.ipynb`

The above equation thus proves theorem 8 for $k = 1$. In the next chapter we will explain why this argumentation is in principle valid for all k.

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. Are the principles discussed in the previous chapter 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 shown in the following Table .

n	variable	decimal	log2	binary	binary length	α_i	α	operation
1	v_1	17	4.09	10001_2	5			$\cdot 3$
	$3 \cdot v_1$	51	5.67	110011_2	6			$+1$
	$3 \cdot v_1 + 1$	52	5.70	110100_2	6	2	2	$\cdot 2^{-2}$
2	v_2	13	3.70	1101_2	4			$\cdot 3$
	$3 \cdot v_2$	39	5.29	100111_2	6			$+1$
	$3 \cdot v_2 + 1$	40	5.32	101000_2	6	3	5	$\cdot 2^{-3}$
3	v_3	5	2.32	101_2	3			$\cdot 3$
	$3 \cdot v_3$	15	3.91	1111_2	4			$+1$
	$3 \cdot v_3 + 1$	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 makes clear, 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:

$$len(3 \cdot v_n) = \lfloor log_2 3 + log_2 v_n \rfloor + 1 \tag{13}$$

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 argue in the same way as in the previous chapter. 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 resultant power of two will lead to a termination.

In order to prove our hypothesis, we have to adjust theorem 12 by considering the additional binary growth that is caused by the multiplications with three. Thereby we obtain the following formula:

$$\alpha = \hat{\alpha} = \lfloor n \cdot log_2 3 + log_2 v_1 \rfloor + 1 \tag{14}$$

The above term proves theorem 8 for the case $k = 3$. A closer look makes clear that it is not only valid for $k = 3$, but for all k. In conclusion, we can define the following boundaries for the number of divisions by two in a Collatz sequence:

$$n \leq \alpha \leq \hat{\alpha} \tag{15}$$

If one shows 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 next chapter we will discuss the consequences of our findings for the occurrence of cycles, further confirming our argumentation.

Occurences of Cycles

Definition
tbd

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 und the underlying operations. Theorem 4.4 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 5 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.

Data Set

This empirical data set was used to derive and validate theorems 6 and 8. The sample was generated with a Python script and comprises information about sequences for the interval $v_1 \in [1 \dots 3999]$ and $k \in \{1, 3, 5, 7, 9\}$.[14] 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. Both starting values, $v_1 = 13$ and $v_1 = 53$, i.e. result in the number five.

Cycle Finder

This Python script was used to validate theorem 4.4.[15] The program performs a linear search for the intervals $v_1 \in [1 \dots 9999]$ and $k \in [1 \dots 999]$. 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 classical maths alone, we moreover used techniques of data science. We combined the two fields in different ways. Firstly, we analysed sequences and related features empirically, 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.

[14] `https://github.com/c4ristian/collatz/blob/master/run_alpha_export.py`
[15] `https://github.com/c4ristian/collatz/blob/master/run_cycle_finder.py`

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