Contents

While chatting about number theory problems, I'm often asked about the Collatz Conjecture and its difficulty. Many of my mathematical colleagues can tell me that the conjecture is really hard, but couldn't answer why (besides saying Terry Tao said it was hard). This paper summarizes a little project I've done in trying to understand why the Collatz Conjecture is still an unsolved problem. In particular, I would like to find a reformulation of the Collatz Conjecture that shows that the simple looking statement is not as innocent as it seems.

Part of the goal was to just explore as far as possible and see what I will find. In the process of exploring this question, I found many interesting mathematical results which I found to be interesting enough to compile into this paper.

A very interesting result is that there is probability distributions that can be put on a number that is recursively put into the Collatz function. To be more precise, let $n_0 \in \mathbb{N}_{>0}$, and let f be the Collatz function:

$$f(n) = \begin{cases} \frac{t}{2} & t \text{ is even} \\ 3t+1 & t \text{ is odd} \end{cases} \tag{1}$$

Let $n_k = f^{\circ k}(n_0)$ be the number we get after applying f to it k-times. Decompose n_k into it's sum of 2-powers, for example:

$$93 = 2^0 + 2^2 + 2^3 + 2^4 + 2^6$$

Take the difference in powers and put them all into an array (for example: $N_K = [2, 1, 1, 2]$). Call this array the *chain* representing n_k. Then I showed for an arbitrary position that is not too close to the end of the array (not too close to the tail), we have that:

$$P(N_{K,i} = 1) = 0.5 \qquad P(N_{K,i} = 2) = 0.25 \qquad P(N_{K,i} \geq 3) = 0.25 \tag{2}$$

The stability of this result is rather interesting: it applies to almost all position of the chain while iterating f, except when near the tail. The last digit is the most unstable, varying greatly in probability, but having a relatively stable average probability (as shall be outlined in greater detail in the paper). The further away from the last digit we are, the closer the probability is to the $0.5, 0.25, 0.25$ presented above. This effect is also extremely local: only the last 10 digits of a chain are usually over $> 1\%$ away from the probabilities given in equation (2), and this effect diminishes even more beyond 10 digits.

This complexity in the probability of the last digit turns out to be the key issue making the Collatz Conjecture so hard to solve: I showed that if the last digit as certain stable probability that I found heuristically, then we can provably show the probability of a number not reaching 1 in finite iterations is 0%. The fact of the mater is, it is rather hard to show this final result, and I outline my current best attempt at it in this paper.

1

Main Paper

Let $t \in \mathbb{N}$. Consider the following arithmetic progression:

$$f(n) = \begin{cases} \frac{t}{2} & t \text{ is even} \\ 3t+1 & t \text{ is odd} \end{cases} \tag{1.1}$$

Then the Collatz conjecture claims that there exists a k such that $f^{\circ k}(t) = 1$ for every t. With the goal of understanding this result better, I shall present a re-framing of the Collatz Conjecture in three steps:

1. To gain more information out of each number, I shall look at the array of the difference in exponents when a number is represented as a sum of 2-powers; these shall be called the *2-chain representation* of a natural number, or just *chain representation* for short. This is not the only way that we can gain more information from a natural number, in fact for the proof that will be presented in this paper, there are many equivalent formulations that can be used (for example, we don't need to take the difference, and we can keep track of information $\mod n^k$ for number $n \neq 2$), however this representation was chosen specifically as it will be illustrative on where the bottleneck in solving the Collatz Conjecture arises[1]

   After finding the chain representation, we shall defined *Collatz Arithmetic's*, which will be the arithmetical rules that will allow us to algorithmically find the next chain representation of a number without needing to reference the Collatz function, essentially translating the Collatz Conjecture into a problem about chains and Collatz Arithmetics.

2. We shall next define some functionals on chains that shall extract useful information. Two natural functional we may define are the *length* functional $m(-)$, and *size* functional $\sum(-)$, where the first measure the number of components of a chain while the other sums over all

[1] I have found after having already invented my "chain representation" of numbers that this can be reformulated in the language of the p-adics; however I think that the "Collatz" arithmetic's that will be presented feels easier to work with in the notation that will be presented

the components. We shall prove that the size functional shall have the following remarkable property: if N_k is the chain representation of $f^{\circ k}(n_0)$, and $F(N_k)$ is the chain representation of the next odd Collatz number, then:

$$\sum F(N_k) = \sum N_k + \{-1, 0, 1\} - 2\ell$$

where $\ell \in \mathbb{N}$ will be the number of 2's at the beginning of the chain. In other words, a chain will increase in size by at most 1, and has many opportunities to shrink. We shall show that the size only depends on the *first digit* and *last block* of the chain, where we shall define what is a block within this paper; it shall usually be the last few digits of a chain.

Thus, the Collatz Conjecture can be re-framed as asking to show that the size of a chain will always reach 0.

3. With the goal of showing that the size *any* the chain shrinks to 0 after applying a finite iteration of the Collatz function, we need to find a way to determine the value of the first digit and last few digits of a chain. This is where we shall introduce a probabilistic distribution on the digits of a chain. To understand where this distribution comes from, consider first any natural number $n \in \mathbb{N}_{>0}$. Representing it as a chain, we can show that the probability that one of the components is 1, 2, or ≥ 3 follows the probabilities given in equation 1.2, namely:

$$P(N_{0,i} = 1) = 0.5 \qquad P(N_{0,i} = 2) = 0.25 \qquad P(N_{0,i} \geq 3) = 0.25 \tag{1.2}$$

We shall show that these probabilities are stable after applying the Collatz function *away from the tail*.

Thus, the Collatz Conjecture can be re-framed to requiring the first digit and last block to have the right probabilities so that the size of any chain must eventually reach 0.

As stated in the abstract, the probability of the last block is difficult to ascertain and is in fact not very stable. Some empirical data was collected to see what would be the expected probability, and the results were strikingly simple. These finding shall be presented at the end of the paper as a new way of approaching the conjecture.

1.1 Chains and Collatz Arithmetic's

First, we must represent each natural number t as a *chain*. This chain will give way to *Collatz arithmetic's*, which will be important in computing various aspects of the probability problem. We may start with an easy simplification: if n is even, divide it until it's odd. Then if $t = 1$, we're done, so assume it's not the case. Represent n in the following way:

$$t = 1 + 2^{n_1}(1 + 2^{n_2}(1 + \cdots))$$

For example:

$$77 = 1 + 2^2(1 + 2(1 + 2^3))) = 1 + 2^2 + 2^3 + 2^6$$

Simplifying notation, we may write this as $\sim\ 2\ \sim\ 1\ \sim\ 3$ (representing the difference in power's of 2). This can be done for nay natural number. This will be called the *chain representation*

Definition 1: Chain Representation

Let $t \in \mathbb{N}_{>0}$. Then the *chain representation* of n is the string

$$t_0 \ \sim \ t_1 \ \sim \ t_2 \ \sim \ \cdots \ \sim \ t_k$$

where each t_i is the difference in powers in the binary representation. The string will be called a *chain*, and each number in the chain is called an *element* or *component* of the chain. The chain $\sim\ 0$ shall often be represented as $\sim$. If n is odd, then it will be represented as:

$$\sim \ t_1 \ \sim \ t_2 \ \sim \ \cdots \ \sim \ t_k$$

In the following, we will examine the behavior of a chain as it passes through f (see equation 1). As we shall see, the current representation only depends on what happens when n is odd, so define the function F to take in an odd number and return the next odd progression, for example if $n = 13$, then $F(13) = 5$. The collection of all progressions shall sometimes be of interest:

Definition 2: Collatz Sequence

Let t be a number which without loss of generality is odd. Then if N_0 is the chain representation of t, the sequence

$$(N_0, F(N_0), F(F(N_0)), ...)$$

is called the *Collatz sequence* with respect to t (or for t). Often, the shorthand $N_k = F^{\circ k}(N_0)$ will be used, and the sequence may be represented as $(N_{t,i})_i$ or just (N_i) is t is clear.

Note that the Collatz sequence is always infinite, however if the Collatz Conjecture is true this sequence shall at some point loop between

$$1 \quad \Rightarrow \quad 4 \quad \Rightarrow \quad 2 \quad \Rightarrow \quad 1 \quad \Rightarrow \quad \cdots$$

which in terms of chains would be represented $F(\sim) =\sim$. Hence given the Collatz Conjecture is true any sequence will eventually be:

$$(\cdots, \sim, \sim, \sim, \cdots)$$

1.1.1 Collatz Arithmetic's

What we shall next endeavor is to fully classify the behavior of chains when passing through F once. The behaviour shall be completely dependent on whether a component has value 1, 2, or $n \geq 3$. Starting with the case where a chain only consists of 1, after passing through F:

$$\sim 1 \ \sim \ 1 \ \sim \cdots \sim \ 1 \ \sim \ 1 \ \Rightarrow \ \color{red}{\sim \ 1} \sim 1 \ \sim \cdots \sim \ 1 \ \sim \ 1 \color{blue}{\sim \ 2} \ \equiv \sim \ 1 \sim \cdots \ \sim \ 1 \ \sim \ 1 \ \sim \ 2$$

We see that the first one gets "consumed", while a new power of two gets added at the end of the chain:

Lemma 1: Chain of 1

Let t be the odd number represented by a chain of ones: Let $t_i = 1$ for each i. Then:

$$F(\sim \ t_1 \ \sim \cdots \ \sim \ t_k) = \sim \ t_2 \ \sim \ \cdots \ \sim \ t_k \ \sim \ 2$$

***Proof* :**
In binary representation, this number is of the form:

$$1 + 2 + 2^2 + \cdots + 2^{t_k}$$

Tripling and adding one, we get:

$$1 + (1 + 2 + 2^2 + \cdots + 2^{t_k}) + (2 + 2^2 + 2^3 + \cdots + 2^{t_k+1}) = 2 + \cdots + 2^{t_k} + (2^{t_k+1} + 2^{t_k+1})$$

Dividing by 2 and transforming to the chain representation, we get:

$$\underbrace{\sim\ 1\ \sim \cdots \sim\ 1}_{(t_k-1)\text{ times}}\ \sim\ 2$$

as we sought to show.

If we had a single $\sim\ 1$, then we would get $F(\sim\ 1) = \sim\ 2$. Notice the first one always disappears. For reasons that shall be clear soon, we shall say that this one was *consumed* to start the *active propagation*. These terms shall be made rigorous soon, the vocabulary is introduced now to ease the reader into it. What if we have an arbitrary chain of only 2's? This is our ideal result:

Lemma 2: Chains Of 2

Let t be the number represented by a chain of two's: let $t_i = 2$ for each i. Then:

$$F(\sim\ t_1\ \sim \cdots\ \sim\ n_k) = \sim$$

that is t "collapses" to the 1

***Proof* :**
In binary representation, this number is of the form:

$$1 + 2^2 + 2^4 + \cdots + 2^{2t_k}$$

Tripling and adding one, we get:

$$1 + (1 + 2^2 + 2^4 + \cdots + 2^{2t_k}) + (2 + 2^3 + 2^5 + \cdots + 2^{2t_k+1}) = 2^{2t_k+2}$$

But this is a power of two, hence dividing away we get $F(t) = 1$, as we sought to show.

More generally, for any chain that starts with sub-chain of 2's, the then 2's will "disappear"

Lemma 3: Disappearing 2's

Let t be the number represented by the chain $\sim\ 2\ \sim \cdots \sim\ 2\ \sim\ C$ where C represents some arbitrary other chain. Then:

$$F(\sim\ 2\ \sim \cdots \sim\ 2\ \sim\ C) = F(C)$$

***Proof* :**
In binary representation, this number is of the form:

$$1 + 2^2 + 2^4 + \cdots + 2^{2t_k} + 2^{2n_k+1/n} + \cdots$$

where the next digit is either 1 more the power of the previous one, or a power $n \geq 3$ more than the previous one. Tripling and adding one, we get:

$$\begin{gathered} 1 + (1 + 2^2 + 2^4 + \cdots + 2^{2t_k} + 2^{2n_k+1/n} + \cdots) \\ +(2 + 2^3 + 2^5 + \cdots + 2^{2t_k+2} + 2^{2t_k+2/(n+1)} + \cdots) \\ = 2^{2t_k+2} + \cdots \end{gathered}$$

The beginning of the chain a power of two, hence dividing away we get $F(C)$, as we sought to show.

This is rather special. The case of the first 2's of a chain disappearing will be called the *pre-active propagation.*

Let us now take a moment to analyze the case of a chain with an arbitrary number of $n \geq 3$, the behavior each n may differ slightly, and so we first tackle the case where we only have a single n:

Lemma 4: Single n

Let t be the number represented by the chain $\sim\ n$ for some $n \geq 3$. Then:

$$F(\sim\ n) = \sim\ (n-2)\ \sim\ 1$$

***Proof* :**
In binary representation, this number is of the form:

$$1 + 2^n$$

Tripling and adding one, we get:

$$1 + (1 + 2^n) + (2 + 2^{n+1}) = 2^2 + 2^n + 2^{n+1}$$

Dividing by 4, we get $1 + 2^{n-2} + 2^{n-1}$, which in chain representation is:

$$\sim\ (n-2)\ \sim\ 1$$

as we sought to show

This process adds a new number in our chain, namely a 1. We shall say that the n has *split.* Let us next consider a chain where each element is ≥ 3:

Lemma 5: Chain of n's

Let t be the number represented by a chain

$$\sim\ n_1\ \sim\ n_2\ \sim\cdots\sim\ n_k$$

where $n_i \geq 3$ for each i. Then:

$F(\sim n_1 \sim n_2\ \sim\cdots\sim n_k) = \sim (n_1{-}2)\ \sim\ 1\ \sim (n_2{-}1)\ \sim\ 1 \sim\ (n_3{-}1)\ \sim\cdots\sim\ (n_k{-}1)\ \sim\ 1$

Proof **:**
In binary representation, this number is of the form:

$$1 + 2^{n_1} + 2^{n_1+n_2} + \cdots$$

Tripling and adding one, we get:

$$1{+}(1{+}2^n_1{+}2^{n_1+n_2}{+}\cdots){+}(2{+}2^{n_1+1}{+}2^{n_1+n_2+1}{+}\cdots) = 2^2{+}2^{n_1}{+}2^{n_1+1}{+}2^{n_1+n_2}{+}{+}2^{n_1+n_2+1}{+}\cdots$$

Dividing by 4, we get $1 + 2^{n_1-2} + 2^{n_1-1} + 2^{n_1+n_2-2} + \cdots$, which in chain representation is:

$$\sim\ (n_1-2)\ \sim\ 1\ \sim\ (n_2-1)\ \sim\ 1 \sim\cdots$$

as we sought to show

The fact that the first n_1 behaves differently from the rest of the n_i's can be interpreted this way: when reading a chain from left-to-right to determine the next values of the chain, we always start in an *active state*, or an *active propagation.* If we encounter an $n \geq 3$ at the beginning, then n becomes $n-2$ and splits, and the active propagation becomes *inactive.* When in an inactive state, all $n \geq 3$ shall diminish by 1 to become $(n-1)$ as well as split (i.e. $\sim\ (n-1) \sim\ 1)$[2].

All the proceeding proofs shall follow a very similar pattern to the above, and so the details of each will now be left out (the reader is free to verify at least a few of the next lemmas as the intuition behind the computation of these chains by reading the chain from left to right will be taken for granted in all of sections after this one). Furthermore, notice that in each of these calculations where always calculated from left-to-right, that is, the calculations of the next component was done after determine the value and behaviour of the immediately previous component. We shall henceforth say that the process computing a chain and processing it's value by computing the values from left-to-right as the *propagation.*

Let us now mix n's and 2's. An interesting pattern to notice is that with the positioning of 2's in a chain. If it is in behind a power $n \geq 3$ (to the left), then the 2 "behaves" like a power greater than 2:

[2] In lemma 9, we shall see that the one that split can consume when an active propagation is triggered, hence it is possible that the 1 splits and gets immediately consumed

Lemma 6: Inactive 2

Let t be the number represented by the chain $\sim\ n\ \sim\ 2$ for $n \geq 3$. Then:

$$F(\sim\ n\ \sim\ 2) = \sim\ (n-2)\ \sim\ 1\ \sim\ (2-1)\ \sim\ 1 = \sim\ (n-2)\ \sim\ 1\ \sim\ 1\ \sim\ 1$$

***Proof* :**
This is a matter of computation

The idea behind this lemma is that if a 2 is "stuck", then the power's cannot group, and hence each 2 shall just be split. Notice that this lemma can be generalized for sequence of 2 so that:

$$F(\sim\ n\ \sim\ 2 \sim\ \cdots\ \sim\ 2) = \sim\ (n-2)\ \sim\ 1\ \sim \cdots \sim\ 1$$

given that the 2's are not at the beginning of the chain (in which case we apply lemma 3).

Next, let us mix 1's and 2's. This behaviour is in fact the most nuanced behaviour of any propagation. Let us point out the simplest case of there being a sequence of ones followed by a sequence of two's before getting into all the possible mixed cases:

Lemma 7: 2 In front of 1's

Let t be the number represented by the chain $\sim\ 1\ \sim\ \cdots \sim\ 1\ \sim\ \underbrace{2\ \sim \cdots \sim\ 2}_{k \text{ times}}$. Then:

$$F(\sim\ 1\ \sim\ \cdots \sim\ 1\ \sim\ 2\ \sim \cdots \sim\ 2) = \underbrace{\sim 1 \cdots \sim 1}_{\text{one less}}\ \sim\ (2k+2)$$

The idea is that the 2 that follow a 1 get *consumed*, and each following 2 shall also be consumed until the end is hit, at which point the value of the 2's will be *transmitted* at the end. The fact that we get $2k+2$ instead of $2k$ comes from the fact that the original 1 from the beginning of the chain must also get it's value transmitted, and that any active propagation will end by adding an extra value of 1 at the end of the active propagation.

For the rest of the possible cases of having components with values either 1 or 2, they are listed out bellow with a brief explanation of each. These explanations shall motivate the intuition that a chain can be computed by reading it left to right:

1. Consider first $\sim\ 1\ \sim\ 1\ \sim\ 2\ \sim\ 1\ \sim\ 1$. Then the first 1 is consumed to activate the propagation, then the 2nd one stays neutral, the first 2 gets consumed, then since this leaves a gap the next 1 gets modified and becomes a 3 (essentially, its value is transmitted at the next 1). The next one is unchanged, and a new 2 is added at the end, finishing the "active propagation". Ultimately, the change after applying F looks like:

$$\sim\ 1\ \sim\ 1 \sim\ 2 \sim\ 3\ \sim\ 1 \sim\ 2\ \Rightarrow \sim\ 1\ \sim\ 3\ \sim\ 1\ \sim\ 2$$

2. What if there is sequence of 2's followed by two ones. Then they shall all get absorbed into the one:

$$\sim\ 1\ \sim\ 1 \underbrace{\sim\ 2 \sim \cdots \sim\ 2}_{k \text{ times}} \sim\ (2k+1)\ \sim\ 1 \sim\ 2\ \Rightarrow \sim\ 1\ \sim\ (2k+1)\ \sim\ 1\ \sim\ 2$$

3. What if there is only one 1 following the 2's. Then that single one shall absorb the value of the 2's and a new 2 shall still occur:
$$\sim\ 1\ \sim\ 1 \underbrace{\sim\ 2 \sim \cdots \sim\ 2}_{k \text{ times}} \sim\ (2k+1) \sim\ 2\ \Rightarrow \sim\ 1\ \sim\ (2k+1)\ \sim\ 2$$

4. What if there is a single 1 in front. Then it is the catalyst for a propagation, or it starts the active propagation:
$$\sim\ 1 \underbrace{\sim\ 2 \sim \cdots \sim\ 2}_{k \text{ times}} \sim\ (2k+1) \sim\ 2\ \Rightarrow \sim\ (2k+1)\ \sim\ 2$$

5. What if there is no 1 following the 2's. Then we get the following behavior
$$\sim\ 1 \underbrace{\sim\ 2 \sim \cdots \sim\ 2}_{k \text{ times}} \sim\ (2k+2)\ \Rightarrow \sim\ (2k+2)$$
In particular, the extra value within the propagation is absorbed with the 2's. This is exactly the content of lemma 7.

6. Finally, the last case to explore is if a sequence of 2's starts the chain and is followed by a sequence of 1's. Since there is no leading 1, there is no term to absorb the values of these and so they disappear:
$$\underbrace{\sim\ 2 \sim \cdots \sim\ 2}_{k \text{ times}} \sim\ 1\ \sim\ 1 \sim\ 2\ \Rightarrow \sim\ 1\ \sim\ 2$$
This is simply lemma 1 and lemma 3 combined.

This gives a good idea for what happens when working exclusively with 1's and 2's. For example:
$$F(\sim\ 2\ \sim\ 2\ \sim\ 1\ \sim\ 2\ \sim\ 1\ \sim\ 2) =\sim\ 3\ \sim\ 4$$

Breaking down this example:

1. The first two 2's disappear as it is a pre-active propagation
2. first 1 activates the propagation hence gets consumed
3. The first 2 get absorbed into the next 1
4. Finally the propagation finishes, and as it is the end of the propagation it must transmit all values it has accumulated, which in this case is 4 as explained in lemma 7.

Lemma 8: Exclusively 1s And 2s

Let n be a number whose chain representation exclusively has 1's and 2's. Then after applying F:

1. Any two at the beginning is disappears from the chain
2. Any sequence of twos between ones will have its values absorbed and transmitted to the next one
3. Any sequence of two at the end (including a sequence of length zero) will "dump" it's value as well as add 2 to the total value

Let us now consider a chain with arbitrary values in each component. As we saw in lemma 5, how they transform through F seems to depend on their position: the first n_1 in the chain representation of lemma 5 diminished by 2 while all the other ones diminished by 1. Furthermore, by contrast to lemma 8, we saw that a 2 acts very differently when it's preceded by an $m \geq 3$; in fact it acts like a number ≥ 3 in the sense that it's value diminishes by 1 and splits just like the other numbers in lemma 5. Hence, the behavior of the 2 changes. This behavior is given a name:

Definition 3: Active and Inactive Elements

Let C be some arbitrary chain. Then if after applying F, a 2 is being removed from the chain (whether it's valued is transmitted or removed), this 2 is said to be an *active* 2. If the two is splitting into $-\ 1\ -\ 1$, then it said to be an inactive 2.

More generally, call all elements in a chain *inactive* if they are either an inactive 2 or ≥ 3, and call all elements in a chain *active* if it is either an active 2 or 1.

As a consequence of all the above lemma's, it is impossible that a 2 at the beginning of a chain be inactive; a 2 is only inactive it is preceded by another inactive 2 or by an element $n \geq 3$, and a 2 is active if it is at the beginning of a chain, preceded by an active 2, or a 1. This means that 1's are in some sense that "catalyst" to activate 2. These last few lemmas make this precise, and covers the case of mixing 1's and n's:

Lemma 9: Case 1: $\sim n \sim 1 \sim 1 \sim p$

Let t be a number represented by a chain $C \sim\ n\ \sim\ 1\ \sim\ 1\ \sim\ p\ \sim\ D$ for some chains C, D, inactive n, and $p \geq 3$. Then:

$$F(C\ \sim\ n\ \sim\ 1\ \sim\ 1\ \sim\ p\ \sim\ D) = F(C)\ \sim\ (n-1)\ \sim\ 2\ \sim\ 2\ \sim\ (p-2)\ \sim\ \cdots$$

where p and D do not affect the behavior earlier in the chain:

$$F(C\ \sim\ n\ \sim\ 1\ \sim\ 1) = F(C)\ \sim\ (n-1)\ \sim\ 2\ \sim\ 2$$

and if C is null:

$$F(\sim\ n\ \sim\ 1\ \sim\ 1) =\sim\ (n-2)\ \sim\ 2\ \sim\ 2$$

Proof **:**
this is a matter of computation

Lemma 10: Case 2: $\sim n \sim 1 \sim \cdots \sim 1 \sim p$

Let t be a number represented by a chain $C \sim\ n\ \sim\ 1\ \cdots\ \sim\ 1\ \sim\ p\ \sim\ D$ for some chains C, D, inactive n, and $p \geq 3$. Then:

$$F(C\ \sim\ n\ \sim\ 1\ \cdots\ \sim\ 1\ \sim\ p\ \sim\ D) = F(C) \sim (n{-}1)\ \sim\ 2\ \sim\ 1\ \sim\ \cdots\ \sim\ 1\ \sim\ 2\ \sim\ (p{-}2)\ \sim\ \cdots$$

where p and D does not affect the behavior earlier in the chain:

$$F(C\ \sim\ n\ \sim\ 1\ \cdots\ \sim\ 1) = F(C) \sim (n-1)\ \sim\ 2\ \sim\ 1\ \sim\ \cdots\ \sim\ 1\ \sim\ 2$$

and if C is null

$$F(\sim\ n\ \sim\ 1\ \cdots\ \sim\ 1) = (n-2)\ \sim\ 2\ \sim\ 1\ \sim\ \cdots\ \sim\ 1\ \sim\ 2$$

Proof **:**
this is a matter of computation

When proving the above lemma, notice that the n splits, and the added 1 gets consumed as part of initializing the active propagation. Furthermore, the 1 proceeding the n (to the right of it) also gets consumed. This is different from the case where a chain started with 1's, for in that case there is no extra 1 consumed from the split that gets transmitted to the next 1. This behavior is why when starting a propagation, it is in a *pre-active state* or *pre-active propagation*, and requires the consumption of a non-split one to become active. This is new behavior the reader should keep in mind while looking at the different subtleties that may occur.

Lemma 11: Case 3: $\sim n \sim 1 \sim p$

Let t be a number represented by a chain $C \sim\ n\ \sim\ 1\ \sim\ p\ \sim\ D$ for some chains C, D, inactive n, and $p \geq 3$. Then:

$$F(C\ \sim\ n\ \sim\ 1\ \sim\ p\ \sim\ D) = F(C)\ \sim\ (n-1)\ \sim\ 3\ \sim\ (p-2)\ \sim\ \cdots$$

where p and D does not affect the behavior earlier in the chain:

$$F(C\ \sim\ n\ \sim\ 1) = F(C)\ \sim\ (n-1)\ \sim\ 3$$

and if C is null

$$F(\sim\ n\ \sim\ 1) = \sim\ (n-2)\ \sim\ 3$$

Proof **:**
this is a matter of computation

Lemma 12: Case 4: $\sim 1 \sim n$

Let t be a number represented by a chain $\sim\ 1 \sim\ n\ \sim\ p\ \sim\ D$ for some chain D, $n \geq 3$, and inactive p Then:

$$F(\ \sim\ 1\ \sim\ n\ \sim\ p\ \sim\ D) =\ 2\ \sim\ (n-2)\ \sim\ (p-1)\ \sim\ \cdots$$

where p and D does not affect the behavior earlier in the chain:

$$F(\ \sim\ 1\ \sim\ n) =\ 2\ \sim\ (n-2)$$

Proof **:**
this is a matter of computation

Lemma 13: Case 5: $\sim 1 \sim \cdots \sim 1 \sim m$

Let t be a number represented by a chain $\sim\ 1 \sim\ \cdots\ \sim\ n\ \sim\ p\ \sim\ D$ for some chain D, $n \geq 3$ and inactive p. Then:

$$F(\underbrace{\sim\ 1\ \cdots\ \sim\ 1}_{k\text{ times}}\ \sim\ n\ \sim\ p\ \sim\ D) = \underbrace{\sim\ 1\ \cdots\ \sim\ 1}_{k-1\text{ times}}\ \sim\ 2\ \sim\ (n-2)\ \sim\ (p-1)\ \sim\ \cdots$$

where p and D does not affect the behavior earlier in the chain:

$$F(\underbrace{\sim\ 1\ \cdots\ \sim\ 1}_{k\text{ times}}\ \sim\ n) = \underbrace{\sim\ 1\ \cdots\ \sim\ 1}_{k-1\text{ times}}\ \sim\ 2\ \sim\ (n-2)$$

Proof **:**
this is a matter of computation

The 1's consistently diminish the value of an $n \geq 3$ by 2 in the same way an element n at the beginning of a chain gets it's value diminished by 2, and furthermore they are the key to activating 2's. With these lemmas, we shall make precise the terminology alluded to so far:

Definition 4: [Collatz Chain] Propagation: Pre-active, Active, and Inactive

Let C be a chain without leading 2's (due to lemma 3). Then the process to determine the value of $F(C)$ by reading the chain left-to-right to determine the next values is called reading the *propagation* of a chain. Each propagation starts at a *pre-active state* before the first number of considered. Then

- If the chain starts with an $n \geq 3$, then it comes an *inactive-propagation*, or goes into an *inactive state*.
- If the chain starts with a 1, then it becomes an *active-propagation*, or goes into an *active state* and the 1 gets consumed.

Furthermore:

1. When a chain is in an activate state, 2's are activated.
2. An active-propagation *terminates* when it hits an $n \geq 3$ or the end of the chain, at which point any 2's thereon are inactive. If a propagation terminates by hitting an $n \geq 3$, that n shall diminish its value by 2. All terminating propagation's shall end with an additional value of 2.
3. A propagation my reoccur if *catalyzed* by a 1 (consuming it in the process), at which point 2's thereon shall be activated.
4. A propagation catalyzed after an inactive element shall transfer the initial value of the 1 to the next available 1. If no next available 1 exists then the value is added to the end of the propagation.

Of all the possibilities we have covered, there is one combinations of 1's, 2's, and n's that we have yet to consider, namely when $\sim\ n\ \sim\ 1\ \sim\ 2\ \sim\ \cdots\ \sim\ 2\ \sim\ 1$. Covering this shall give the final possible arithmetical operation:

Lemma 14: Augmented 2

Let t be a number represented by a chain $C \sim n \sim 1 \sim 2 \sim \cdots \sim 2 \sim 1 \sim p$ where C is some chain, n is inactive and $p \geq 3$. Then:

$$F(C \sim n \sim 1 \underbrace{\sim 2 \sim \cdots \sim 2}_{k \text{ times}} \sim 1 \sim p) = F(C) \sim (n-1) \sim (2k+2) \sim 2 \sim (p-2)$$

where the C can as before be null and p has no effect on the earlier behavior. With color, this transformation can be represented as:

$$\sim (n-(1 \text{ or } 2)) \underbrace{\sim 1}_{\text{catalyst}} \underbrace{\sim 2 \sim \cdots \sim 2}_{\text{removed}} \underbrace{\sim (2k+2)}_{\text{absorbed 2's } +1 \text{ from } n} \underbrace{\sim 2}_{\text{propagation terminated}}$$

If the chain was of the form $n \sim 1 \sim 2 \sim \cdots \sim 2$ (or ending with a $p \geq 3$), then the propagation will look like:

$$\sim (n-(1 \text{ or } 2)) \underbrace{\sim 1}_{\text{catalyst}} \underbrace{\sim 2 \sim \cdots \sim 2}_{\text{removed}} \underbrace{\sim (2k+3)}_{\text{absorbed 2's, } +1 \text{ from } n \text{ } + \text{ } 2 \text{ termination}}$$

This leads us to the final state a propagation can be in: *augmented*.

Definition 5: [Collatz Chain] Propagation: Augmented

An *augmented propagation* or *augmented state* is a state triggered when an active propagation starts from an inactive element (an inactive 2 or $n \geq 3$). This state terminated either when:

1. The next 1 is encountered, in which case the 1 taken from the split n gets transmitted to it. It is possible that there is a sequence of 2's that preceded the one whose values also get transmitted
2. The next $n \geq 3$ is encountered. In which case, the augmented and active state both end (as shown in lemma 14).

Overall, a propagation can either be active or inactive, and an active propagation can further be augmented or not depending on its initialization. The only exception to these states is the pre-active state that the state before the propagation starts and the propagation remains in this state for each 2 at the beginning of the chain.

1.2 Quantifying chains

There are two natural functional that we can define on a chain, namely if $\mathcal{C}$ is the collection of all chains representing odd numbers, then:

$$\sum : \mathcal{C} \to \mathbb{N}, \ \sum(\sim n_1 \sim \cdots \sim n_k) = \sum_i n_i$$

$$m : \mathcal{C} \to \mathbb{N}, \ m(\sim n_1 \sim \cdots \sim n_k) = k$$

The second functional is rather complicated when iterating F, and we shall comment on it at the end of the paper. On the other hand, the first functional is remarkably simple when iterating F:

Theorem 1: Summing Properties

Let (C_i) be a Collatz sequence. Then:

1. If the chain starts with a 1 and ends with an active state, $\sum(C_{i+1}) = \sum(C_i) + 1$
2. If the chain starts with a 1 and ends with an inactive p, $\sum(C_{i+1}) = \sum(C_i)$
3. If the chain starts with an n and ends with an active state, $\sum(C_{i+1}) = \sum(C_i)$
4. If the chain starts with an $n \geq 3$ and ends with an inactive p, $\sum(C_{i+1}) = \sum(C_i) - 1$
5. If the chain starts with n number of 2's, then $\sum(C_{i+1}) = \sum(C_i) - 2n[\pm 1/0]$ where the additional values added or subtracted (or no value added) depends on the 4 above conditions.

***Proof* :**
The 5th case is a consequence of lemma 3 and cases (1)-(4), hence we focus on those.

Let us first decompose the chain into blocks, each block starting with either an $n \geq 3$, an inactive element, or if neither one nor the other are available than the first element of the chain. A block starts at this element, and continues onto the next inactive element or $n \geq 3$. Examples of blocks in a chain are:

- The chain $\sim\ 1 \sim\ 1 \sim\ 1$ is composed of a single block
- The chain $\sim\ 1 \sim\ 2 \sim\ 1$ is composed of a single block
- The chain $\sim\ n \sim\ 1 \sim\ 2 \sim\ 1$ is composed of a single block
- The chain $\sim\ 1\ \sim\ n \sim\ 1 \sim\ 2 \sim\ 1$ is composed of two blocks, $[\sim\ 1]$ and $[\sim\ n \sim\ 1 \sim\ 2 \sim\ 1]$
- The chain $\sim\ 2\ \sim\ 2 \sim\ 1\ \sim\ n \sim\ 1 \sim\ 2 \sim\ 1\ \sim\ n\ \sim\ p$ is composed of 4 blocks, $[\sim\ 2\ \sim\ 2\ \sim\ 1]$, $[\sim\ n \sim\ 1 \sim\ 2 \sim\ 1]$, $[\sim\ n]$, and $[\sim\ p]$

Starting off with the simplest case of a chain of 1, by lemma 1 we get that

$$\sum(C_{i+1}) = \sum(C_i) + 1$$

If the chain started with an $n \geq 3$, then we see that after applying F we get that n becomes $n-2$, the $\sim\ 1$ that would be generated by n is absorbed into the propagation, while an extra value of 1 is added at the end of the propagation, balancing out to:

$$\sum(C_{i+1}) = \sum(C_i)$$

Similarly, if C_i was a chain of 1's that ended with an $n \geq 3$, then the extra value introduced by the propagation would be countered by the hit on n since $\sim\ n$ would become $\sim\ (n-2)\ \sim\ 1$ and hence:

$$\sum(C_{i+1}) = \sum(C_i)$$

Finally, if the starts and with an n with 1's in the middle, it never had the opportunity to add the extra 1 from the initial propagation. After applying F, The first block would have the same sum, while the last block would diminish in value, giving:

$$\sum(C_{i+1}) = \sum(C_i) - 1$$

Now, if we introduced active 2's in the chain, their values will be moved around in the chain, and hence not affecting the overall sum. Elements $n \geq 3$ that are not hit and are not contributing their 1 to initialize a propagation have their sum unchanged between propagations, for example:

$$\sim\ n\ \sim \qquad \Rightarrow \qquad \sim\ (n-1)\ \sim\ 1\ \sim$$

If the one is contributed to the chain propagation, but the element is not hit, then we include the entire propagation in the block, and its total sum goes up by 1, however the next block will diminish its sum by 1 For example:

$$\begin{gathered}\sim\ 1 \sim\ 1 \sim\ 4 \sim\ 1 \sim\ 1 \sim\ 4 \sim\ 4 \sim\ 1 \sim\ 1 \\ \Rightarrow \\ \sim\ 2\ \sim\ 2\ \sim\ 2\ \sim\ 2\ \sim\ 2\ \sim\ 2\ \sim\ 1\ \sim\ 3\ \sim\ 2\ \sim\ 2\end{gathered}$$

Hence, every propagation adds a value, but if it hits another block that block diminishes in value, and if a block has no propagation and is not hit by a propagation its value is unchanged.

Finally, we may determine the effects of all the propagations simply by looking at the first and last element of the chain. If the first element is an $n \geq 3$, then there is no initial propagation, only a partial propagation, starting off the sum of C_{i+1} with a -1. Then every propagation would add a value of 1, but as long as it's followed up by another block the value will be countered. Hence, if the chain ends with an inactive element,

$$\sum(C_{i+1}) = \sum(C_i) - 1$$

If the chain ends with a 1 or active 2, then the value added by the last propagation cancels out the minus 1 value from the beginning, hence:

$$\sum(C_{i+1}) = \sum(C_i)$$

If the chain starts with a 1, then we get the same process except there is an additional value at to contend with, giving us the other two possibilities, completing the proof.

Thus, there is a surprising amount of rigidity over the size of a chain, it may only grow by at most 1, and it shrinks by 1 unless there is a sequence of two in-front of the chain that drastically diminish the size of a chain. The Collatz Conjecture can now be rephrased as:

Collatz Conjecture rephrased Given any [odd] $n \in \mathbb{N}_{>0}$, if C is the chain representation, then it's size (with respect to Σ) is eventually 0.

As theorem 1 tells us, this size depends on the first and last digit and their states. These digits dynamically change as F iterates. If there were some regularity to these digits behaviour under F, we may be able to draw some interesting conclusion. Fortunately, if we turn to probability, we shall see that these digits do behave in a stable manner when considering the right probability distribution.

1.3 Distribution of Most Digits

Given an arbitrary chain $C \in \mathcal{C}$ before applying F, the digits within a chain are distributed rather nicely, that is given an odd $t \in \mathbb{N}_{>0}$ represented as a chain $\sim\ t_1\ \sim\ t_2\ \sim\ \cdots \sim\ t_k$:

$$P(t_i = 1) = 0.5 \qquad P(t_i = 2) = 0.25 \qquad P(t_i \geq 3) = 0.25$$

This shall be proven shortly. The fact shouldn't be a surprise as chain representation comes from counting in binary. What is more surprising is that this distribution remains the same for *most* digits while applying F! The goal of this section is to show just that. The digits for which this does not apply to is the *tail* of the chain, more particularity, it in general doesn't apply for about the last 5 digits of an arbitrary chain iterating through F, and these digits shall require special consideration.

Let us first establish a fact about the length of a chain. Given some $N \in \mathbb{N}$, we may ask what is the chance to have a chain of length k, $1 \leq k \leq N$. This turns out to be normally distributed, as can be shown with a counting argument. For example, here are the distributions of the length of chains up to length 21 (i.e. the first ≈1'000'000 odd numbers):

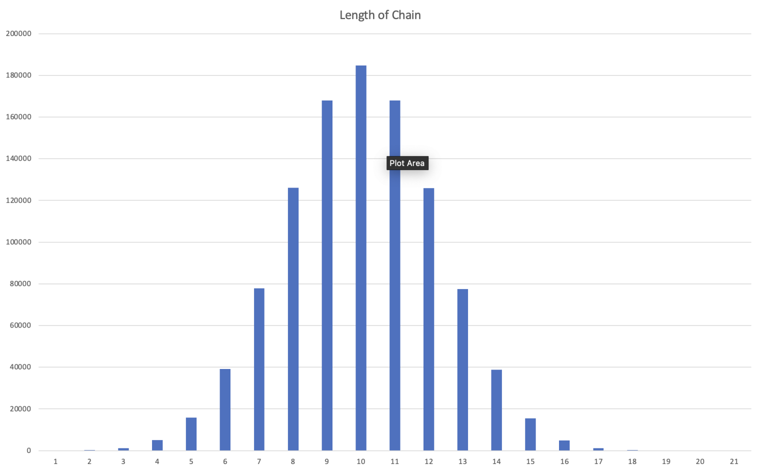

Figure 1.1: Length Distribution

It is a counting and convergence argument to see that as $N \to \infty$, this distribution remains normal, and hence when grabbing a random chain, we may always assume we are taking the chain from a set $\mathcal{C}_N$ for N large enough so that the probability that we have a number at the position we want (ex. the kth position) is ∼100%. In particular, as $N \to \infty$, we may assume that the probability of each digit of the chain is essentially independent of length consideration (i.e. pick N large enough so that $P(m(C) \leq k) \approx 0$ where $m(C)$ is the length of $C \in C_N$).

With the length of chain taken care of, let us calculate the probabilities of having either a 1, 2, or

$n \geq 3$ at each position of a chain. An arbitrary chain can be represented as (for large enough N):

$$\sim \begin{bmatrix} 1 \\ 2 \\ n \end{bmatrix} \sim \begin{bmatrix} 1 \\ 2 \\ n \end{bmatrix} \sim \cdots$$

Lemma 15: Random Chain Digit Distribution

Let $C \in \mathcal{C}_N$ be an arbitrary non-trivial chain of length $\leq N$ for some arbitrary large N. Consider the functionals $\pi_i(C) = n_i$ which returns the ith digit of the chain. Then as $N \to \infty$:

$$P(\pi_i(C) = 1) = 50\% \qquad P(\pi_i(C) = 2) = 25\% \qquad P(\pi_i(C) \geq 3) = 25\%$$

Proof **:**
Let C be an arbitrary chain representing the number t. Recall that a chain represents a number of the form:

$$1 + 2^{n_1}(1 + 2^{n_2}(1 + \cdots))$$

Then the probability question becomes: after subtracting 1, what is the probability of being able to divide n_1 times for some $n_1 \in \mathbb{N}$? This is in fact a uniform distribution argument; given any 2^k, half of the numbers are divisible by 2, a quarter divisible by 4, an eighth by 8, and so forth. Hence, there is a 50% chance $n_1 = 1$, a 25% chance that $n_1 = 2$, a 12.5% chance that $n_1 = 3$, and so froth. Summing up all the possibilities for $n_1 \geq 3$ gives us 25% chance. Hence:

$$P(\pi_1(C) = 1) = 50\% \qquad P(\pi_1(C) = 2) = 25\% \qquad P(\pi_1(C) \geq 3) = 25\%$$

Next, for π_2, we would be concerned whether there is a second digit, however as we have N large enough to not need to worry about the length, we may without consequence assume the required digits needed. Then the argument repeats itself verbatim and can be continued inductively, giving us:

$$P(\pi_i(C) = 1) = 50\% \qquad P(\pi_i(C) = 2) = 25\% \qquad P(\pi_i(C) \geq 3) = 25\%$$

as we sought to show.

Corollary 1: First Digit $n \geq 3$ after F

Let $C \in \mathcal{C}_N$ be an arbitrary non-trivial chain of length $\leq N$ for some arbitrary large N. Assume $\pi_1(C) \geq 3$ Then as $N \to \infty$:

$$P(\pi_1(F(C)) = 1) = 50\% \qquad P(\pi_1(F(C)) = 2) = 25\% \qquad P(\pi_1(F(C)) \geq 3) = 25\%$$

Proof **:**
The first $\sim n \sim$ will become $\sim (n-2) \sim$, and may be 1 2, or ≥ 3. We need to calculate: given an $n \geq 3$, what is the probability of landing on a 1, 2, or ≥ 3. This can be figured out by finding the probability of n having value 3, 4, or $n \geq 5$, conditional on $n \geq 3$. By lemma 15, we know the distribution of all possible digits, namely every time we increase a number by 1, we half its possibility. Given that $n \geq 3$, we may sum up all the possibilities to get:

$$3 : 50\%, \qquad 4 : 25\%, \qquad \geq 5 : 25\%$$

as we sought to show.

Let us now consider the harder case of the behavior throughout the entire chain while taking F. We shall start with the special case of *infinitely long chains*. Naturally, such a chain does not represent a natural number[3]. Notice that Collatz Arithmetics is well-defined on infinitely long chain's as we may simply propagate indefinitely (in a similar way that arithmetic's is well-defined on the 2-adics). This result shall give us enough information to control the length of the chain, which shall then let us conclude the result for arbitrarily long chains up to the tails of chains:

Theorem 2: Distribution - Infinitely Long

Let C_0 be an infinitely long chain, and take the Collatz sequence $(C_0, C_1, ...)$. Consider the functionals $\pi_j(C_i) = n_{i,j}$ which returns the jth component of the chain C_i. Then

$$P(\pi_j(C_i) = 1) = 50\% \qquad P(\pi_j(C_i) = 2) = 25\% \qquad P(\pi_j(C_i) \geq 3) = 25\%$$

Proof **:**
Let $C_0 \in \mathcal{C}_\infty$, and take the Collatz sequence $(C_0, C_1, C_2, ...)$. Since C_0 has yet to have F applied to it, then by lemma 15 for all j

$$P(\pi_j(C_0) = 1) = 50\% \qquad P(\pi_j(C_0) = 2) = 25\% \qquad P(\pi_j(C_0) \geq 3) = 25\%$$

Let us now do induction on i and say that for all j

$$P(\pi_j(C_i) = 1) = 50\% \qquad P(\pi_j(C_i) = 2) = 25\% \qquad P(\pi_j(C_i) \geq 3) = 25\%$$

Let us now calculate the probabilities for the positions of C_{i+1}. What we shall do is manually calculate the first few positions of a general chain (i.e. the "head" of a general chain), and then show that any arbitrary element in the middle of the chain will have probability 0.5/0.25/0.25.

Starting with the case $\pi_1(C_i) = 1$, we know that by the rules of a propagation that the first one is "consumed", ostensibly it disappears. To calculate the probabilities of $\pi_1(C_{i+1})$, we need at least 3 digits for precise results:

%	25	12.5	12.5	12.5	6.25	6.25	12.5	6.25	6.25
C_i	$\sim 1 \sim 1 \sim 1 \sim$	$\sim 1 \sim 2 \sim 1 \sim$	$\sim 1 \sim n \sim 1 \sim$	$\sim 1 \sim 1 \sim 2 \sim$	$\sim 1 \sim 2 \sim 2 \sim$	$\sim 1 \sim n \sim n \sim$	$\sim 1 \sim 1 \sim n \sim$	$\sim 1 \sim 2 \sim n \sim$	$\sim 1 \sim n \sim n' \sim$
C_{i+1}	$\sim 1 \sim 1 \sim$	$\sim 3 \sim$	$\sim 2 \sim m \sim$	$\sim 1 \sim n \sim$	$\sim n \sim$	$\sim 2 \sim n \sim 1 \sim$	$\sim 1 \sim 2 \sim m \sim$	$\sim 4 \sim m \sim$	$\sim 2 \sim m \sim 1 \sim m' \sim$

Summing up, we get:

$$\begin{aligned} P(\pi_1(C_{i+1}) = 1 \,|\, \pi_1(C_i) = 1) = 50\% \\ P(\pi_1(C_{i+1}) = 2 \,|\, \pi_1(C_i) = 1) = 25\% \\ P(\pi_1(C_{i+1}) \geq 3 \,|\, \pi_1(C_i) = 1) = 25\% \end{aligned}$$

Let's next consider the probabilities when $\pi_1(C_i) \geq 3$. The first $\sim n \sim$ will become $\sim (n-2) \sim$, and may be 1 2, or ≥ 3. We need to calculate: given an $n \geq 3$, what is the probability of landing on a 1, 2, or ≥ 3. This can be figured out by finding the probability of n having value 3, 4, or $n \geq 5$, conditional on $n \geq 3$, which by corollary 1:

$$3 : 50\%, \qquad 4 : 25\%, \qquad \geq 5 : 25\%$$

[3] It represents a 2-adic number

Thus, we get:

$$P(\pi_1(C_1) = 1 \,|\, \pi_1(C_0) \geq 3) = 50\%$$
$$P(\pi_1(C_1) = 2 \,|\, \pi_1(C_0) \geq 3) = 25\%$$
$$P(\pi_1(C_1) \geq 3 \,|\, \pi_1(C_0) \geq 3) = 25\%$$

For the probabilities if $\pi_1(C_i) = 2$, we know that by the rules of a propagation that the 2 will disappear. If there is any other 2 to the right of it, then that 2 will disappear as well. Hence, the possible first values of C_i are:

$$\begin{array}{lr} [\sim 2 \sim] \sim 1 \sim 1 \sim & 25\% \\ [\sim 2 \sim] \sim n \sim 1 \sim & 12.5\% \\ [\sim 2 \sim] \sim 1 \sim 2 \sim & 12.5\% \\ [\sim 2 \sim] \sim n \sim 2 \sim & 6.25\% \\ [\sim 2 \sim] \sim 1 \sim n \sim & 12.5\% \\ [\sim 2 \sim] \sim n \sim n' \sim & 6.25\% \end{array}$$

then given $m = n - 2$, the resulting value after the propagation is:

$$\begin{array}{l} \color{blue}{\sim 1 \sim} \\ \sim m \sim \\ \color{green}{\sim m \sim} \\ \sim m \sim \\ \color{red}{\sim 2 \sim} \\ \sim m \sim 1 \sim m' \sim \end{array}$$

The blue marks chains where the 2nd digit is 1, the red where the 2nd digit is 2, the green where the 2nd digit has a chance of being greater than or equal to 3, and black is one of these three. Visually, we may see these as:

$$\begin{array}{r} 25\% :\sim 1 \sim \\ 12.5\% :\sim 2 \sim \\ 12.5\% :\sim n \sim \\ 25\% : \{1, 2, n\} \end{array}$$

To determine the last 25%, we use our earlier calculations to redistribute the probabilities. Hence, appropriately re-distributing the probabilities we get:

$$P(\pi_1(C_1) = 1 \,|\, \pi_1(C_0) = 2) = 50\%$$
$$P(\pi_1(C_1) = 2 \,|\, \pi_1(C_0) = 2) = 25\%$$
$$P(\pi_1(C_1) \geq 3 \,|\, \pi_1(C_0) = 2) = 25\%$$

Though we have done this calculation manually, we could have also combined the previous two cases to

Combining all these results, we get:

$$P(\pi_1(C_1) = 1) = 50\%$$
$$P(\pi_1(C_1) = 2) = 25\%$$
$$P(\pi_1(C_1) \geq 3) = 25\%$$

We shall now show that the probabilities $P(\pi_i(C_{i+1}) = 1)$, $P(\pi_i(C_{i+1}) = 2)$, and $P(\pi_i(C_{i+1}) \geq 3)$ is exactly the same as for C_i, that is we keep the same type of randomness throughout the propagation!

We shall split up the case-work depending on the value of $\pi_1(C_i)$. Let's consider $\pi_1(C_i) \geq 3$ as it is the simplest. Then as always, $\pi_2(C_{i+1})$ on the first 3 digits of C_i. Let's split these into 9 cases representing the possibilities of the beginning of C_i:

$\sim n \sim 1 \sim 1 \sim$	25%
$\sim n \sim 2 \sim 1 \sim$	12.5%
$\sim n \sim n' \sim 1 \sim$	12.5%
$\sim n \sim 1 \sim 2 \sim$	12.5%
$\sim n \sim 2 \sim 2 \sim$	6.125%
$\sim n \sim n' \sim 2 \sim$	6.125%
$\sim n \sim 1 \sim n' \sim$	12.5%
$\sim n \sim 2 \sim n' \sim$	6.125%
$\sim n \sim n' \sim n'' \sim$	6.125%

Then given $m = (n - 2)$ and m', m'' is the resulting value after the propagation:

$$\sim m \sim 2 \sim$$
$$\sim m \sim 1 \sim 1 \sim$$
$$\sim m \sim 1 \sim m' \sim$$
$$\sim m \sim m' \sim$$
$$\sim m \sim 1 \sim 1 \sim 1 \sim 1 \sim$$
$$\sim m \sim 1 \sim m' \sim 1 \sim 1 \sim$$
$$\sim m \sim 3 \sim m' \sim$$
$$\sim m \sim 1 \sim 1 \sim 1 \sim m' \sim$$
$$\sim m \sim 1 \sim m' \sim 1 \sim m'' \sim$$

The blue marks chains where the 2nd digit is 1, the red where the 2nd digit is 2, the green where the 2nd digit has a chance of being greater than or equal to 3. Summing all the probabilities, we get:

$$P(\pi_2(C_1) = 1 \,|\, \pi_1(C_0) \geq 3) = 50\%$$
$$P(\pi_2(C_1) = 2 \,|\, \pi_1(C_0) \geq 3) = 25\%$$
$$P(\pi_2(C_1) \geq 3 \,|\, \pi_1(C_0) \geq 3) = 25\%$$

Let us next do the case of $\pi_2(C_0) = 1$. This case requires a good amount more work, which can be summarized with these image:

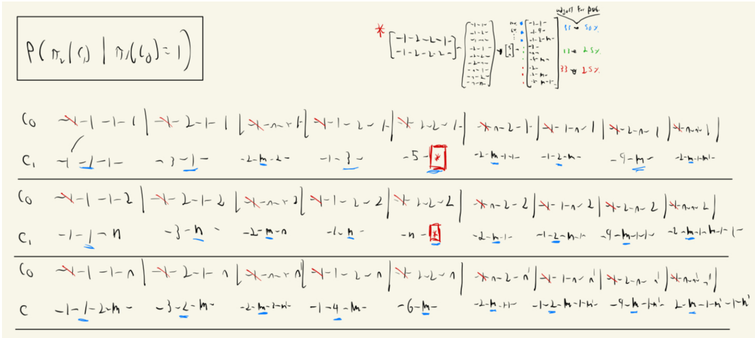

Figure 1.2: Calculations for possible 2nd digit given a length of 4

The probability calculations are given by this table:

probabilities counted up in binary						2ND DIGIT	1	2	n	m	X	
1	1	1	1	0.125	111	1	0.125					
1	0.5	1	1	0.0625	211	1	0.0625					
1	0.5	1	1	0.0625	n11	m				0.0625		
1	1	0.5	1	0.0625	121	3			0.0625			
1	0.5	0.5	1	0.03125	221	X					0.03125	
1	0.5	0.5	1	0.03125	n21	m				0.03125		
1	1	0.5	1	0.0625	1n1	2		0.0625				
1	0.5	0.5	1	0.03125	2n1	m				0.03125		
1	0.5	0.5	1	0.03125	nn1	m				0.03125		
1	1	1	0.5	0.0625	112	1	0.0625					
1	0.5	1	0.5	0.03125	212	n			0.03125			
1	0.5	1	0.5	0.03125	n12	m				0.03125		
1	1	0.5	0.5	0.03125	122	n			0.03125			
1	0.5	0.5	0.5	0.015625	222	X					0.015625	
1	0.5	0.5	0.5	0.015625	n22	m				0.015625		
1	1	0.5	0.5	0.03125	1n2	2		0.03125				
1	0.5	0.5	0.5	0.015625	2n2	m				0.015625		
1	0.5	0.5	0.5	0.015625	nn2	m				0.015625		
1	1	1	0.5	0.0625	11n	1	0.0625					
1	0.5	1	0.5	0.03125	21n	2		0.03125				
1	0.5	1	0.5	0.03125	n1n	m				0.03125		
1	1	0.5	0.5	0.03125	12n	4			0.03125			
1	0.5	0.5	0.5	0.015625	22n	m				0.015625		
1	0.5	0.5	0.5	0.015625	n2n	m				0.015625		
1	1	0.5	0.5	0.03125	1nn	2		0.03125				
1	0.5	0.5	0.5	0.015625	2nn	m				0.015625		
1	0.5	0.5	0.5	0.015625	nnn	m				0.015625		FINAL TALLY
				1			0.3125	0.15625	0.15625	0.328125	0.046875	1
						accounting for M	0.4765625	0.2382813	0.2382813			
						Accounting for X	0.5	0.25	0.25			

Figure 1.3: calculations for all probabilities

Which we see gives us:

$$P(\pi_2(C_{i+1}) = 1 \,|\, \pi_1(C_i) = 1) = 50\%$$
$$P(\pi_2(C_{i+1}) = 2 \,|\, \pi_1(C_i) = 1) = 25\%$$
$$P(\pi_2(C_{i+1}) \geq 3 \,|\, \pi_1(C_i) = 1) = 25\%$$

Finally, for $\pi_1(C_i) = 2$, this is a combination of the other two cases since by lemma 3, the 2's will disappear and reduces to the case of 1 or n that we have just covered, which gives us:

$$P(\pi_2(C_{i+1}) = 1 \,|\, \pi_1(C_i) = 2) = 50\%$$
$$P(\pi_2(C_{i+1}) = 2 \,|\, \pi_1(C_i) = 2) = 25\%$$
$$P(\pi_2(C_{i+1}) \geq 3 \,|\, \pi_1(C_i) = 2) = 25\%$$

Combining all the results gives:

$$P(\pi_2(C_{i+1}) = 1) = 50\%$$
$$P(\pi_2(C_{i+1}) = 2) = 25\%$$
$$P(\pi_2(C_{i+1}) \geq 3) = 25\%$$

We may repeat this argument for the head digits for as long as we like, the details of which are similar and repetitive, and hence will be taken for granted for now. What is important is that the probabilities of the components at the head are the same for any state that we may be in, which will be important for the next part of the proof.

Now, with probabilities of the digits of head of the chain after applying F are calculated for each possible stated, let us now consider being in the middle of the chain. To do so, we shall introduce some new vocabulary to keep track of the behaviour of digits. In particular, the digits in the chain may have multiple different behaviour depending on what digits precede them. It is thus practical to have more granular control over which "actions" are happening on each digit:

1. $\alpha^n_{n-(1/2)}$: This represents the value of an inactive element with value n becoming $n-1$ or $n-2$ after applying F, for example given the chain $\sim\ n_a\ \sim\ n_b$ with $n_a, n_b \geq 3$, $F(\sim n_a \sim n_b)$ will create $(n_a - 2)$ and $(n_b - 1)$ which will have actions $\alpha^{n_a}_{n_a-2}$ and $\alpha^{n_b}_{n_b-1}$
2. β: this represents 1's which did not chain their values. In $F(\sim\ 1\ \sim\ 1)$, the 2nd one would have action β as it remains unchanged
3. $\gamma^{1^*}_2$: represents a 1 that changed values in the case of $\sim\ i\ \sim\ 1\ \sim\ 1$ where i represents an inactive element.
4. $\gamma^{2,k[*]}_{2k+1/2}$: represents a 1 that changed values. For example in $F(\sim 5 \sim 1 \sim 2 \sim 1) =\sim 3 \sim 4 \sim 2$, the 3rd digit, 2nd two, went from a 1 to a 2 (do the proof of lemma 9 if more motivation is needed). The star, *, in the superscript represents a 1 that has had an augmented propagation effect its value.
5. δ_1: represents a new element that comes if the 1 created by an $n \geq 3$ is not consumed. For example, $F(\sim n_a \sim n_b)$ at least one 1 since n_a is followed by $n_b \geq 3$.
6. δ_{2k+2}: represents creating a new value at the end of a propagation that is not augment
7. δ^*_{2k+3} represents creating a new value at the end of a propagation that *is* augmented.
8. $\epsilon^{1,*}$: represents the one that is consumed to start a new propagation that is *not* the first digit of the chain. This explains why it is augmented.
9. $\epsilon^{2,k}$: represents consumed 2 that are not augmented.
10. $\epsilon^{2,k*}$: represents consumed 2 that *are* augmented.
11. ϵ^1: represents the first consumed 1
12. $\epsilon^{2,\varphi}$ Represents consumed two at the beginning of the chain. These are special as they completely disappear from the chain, and hence there is the φ symbol to represent this.

As we are working in the middle of the chain, it is not necessary to consider the last two possibilities as they only occur at the beginning. Using these, we may completely determine the value of a digit t_i in a chain given the digits and that precede it as well as which action is happening on those digits.

What we shall do now is calculate the following probability: given a digit t_i in a chain that is either 1, 2, or $n \geq 3$, what is the probability that the next digit after applying F will be 1, 2, n, or be part of a propagation. Being part of a propagation means that the digit had an ϵ-action, meaning it was consumed. Those shall be considered separately. This shall be a local analysis, in particular we are determining the probability of an individual digit given the behaviour of what is happening around the digit (in particular what's happening before the digit). This shall almost be enough to calculate the distribution of the digits, after having calculated the local behaviour we shall turn to a global perspective to finish off the calculations.

Starting with $t_i = 1$, we get the following table:

-1-	necessary circumstances	probabilities	result
alpha^n_{n-1/2}	(i2/n)-1	0.33333333	collapse
beta	(1/a2)-1-1-1	0.16666667	1
gamma^{1*}_2	-(n/i2)-1-1-1	0.08333333	1
gamma^{2,k*}_{2k+2}	-(n/i2)-1-2-[...]-2-1-1	0.02777778	1
gamma^{2,k}_{2k+1}	-1-1-2-[...]-2-1-1	0.04166667	1
^	-(a2)-1-2-[...]-2-1-1	0.01388889	1
epsilon^{1,*}	-n-1-1 and -n-2-[...]-2-1-1	0.16666667	2
epsilon^{2,k}	-1-1-2-[...]-2-1	0.08333333	n
^	(a2)-1-2-[...]-2-1	0.02777778	n
e^{2,k*}	-(n/i2)-1-2-[...]-2-1	0.05555556	n
	total	1	
P (digit's next value | digit is 1)			

Figure 1.4: P(digit's next value $\mid$ digit is 1)

Most of the calculations are a continuation of what was done before. The new circumstance to consider are the chains that have $2 \sim [...] \sim 2$. The [...] represents all possible chains composed of 2's (ex, $\sim$, $\sim 2 \sim$, $\sim 2 \sim 2 \sim$, and so forth). This is because we are considering *all* possible values before the chain. If we had infinitely many digits that precede t_i (to the left of it), then the sum of all possible chains in [...] would be (1/3). In practice this sum is finite, and we would break down the calculations depending on our choice of depth. Note, that the longer the chain of twos, the less likely it is to appear, namely for each extra ~ 2 added the probability of the chain diminishes by 1/4. In fact, within "[...]", the chances of a chain being length ≥ 5 is $< 0.1\%$. Furthermore, these probabilities would be re-distributed to other cases that balance out the probabilities (something that could be verified inductively). Hence, for an arbitrary position in an arbitrarily long chain, the calculations presented in figure 1.4 is an accurate representation of the breakdown of the probabilities. Note that in the first row, the 1 has action ϵ^1 applied to it and hence is consumed. We shall see how to deal with consumed digits soon (in the figures, this is labeled as "collapse").

Summing the probabilities we get:

	prob's
1	0.33333333
2	0.16666667
n	0.16666667
collapse	0.33333333

Figure 1.5: Summing probabilities from figure 1.4

Let us next do the same thing but taking the case of $t_i = 2$. Following the same procedure where we look at the possible digits before t_i and consider what possible actions may occur, we get:

-2-	necessary circumstances	prob	result
alpha^n_{n-(1/2), 1 or nothing}	(inactive 2/n)-2	0.33333333	1
beta	-(active 2/-1)-1-1-2	0.16666667	collapse
gamma^{1,*}_2	-(inactive 2/n)-1-1-2	0.08333333	collapse
gamma^{2,k}_{2k+1}	-(active 2/-1)-1-2[...]-2-1-2	0.05555556	collapse
gamma^{2,k*}_{2k+2}	-(inactive 2/n)-1-2[...]-2-1-2	0.02777778	collapse
epsilon^{1,*}	-(inactive 2/n)-1-2	0.16666667	collapse
epsilon^{2,k}	-(active 2/1)-1-2[...]2-2	0.11111111	collapse
epsilon^{2,k^*}	-(inactive 2/n)-1-2[...]-2-2	0.05555556	collapse
	total:	1	
P (digit's next value | digit is 2)			

Figure 1.6: P(digit's next value | digit is 2)

Summing the probabilities we get:

1	0.33333333
2	0
n	0
collapse	0.66666667

Figure 1.7: Summing probabilities from figure 1.6

The final case is where $t_i \geq 3$. In this case, we get:

-n-	concrete	prob	result
alpha^n_{n-(1/2), 1 or nothing}	(inactive 2/n)-n	0.33333333	(n-1)
delta_2	-(1/active 2)-1-n	0.33333333	(n-2)
delta_{2k+2}	(-1/active 2)-1-2[...]-2-n	0.11111111	(n-2)
delta^*_{2k+3}	-(n/inactive 2)-1[-2-...-2]-n	0.22222222	(n-2)
	total:	1	

Figure 1.8: P(digit's next value | digit is ≥ 3)

Building off of the proof of corollary 1 it can be showed that:

$$P(n = 3 | n \geq 3) = 0.5 \qquad P(n = 4 | n \geq 3) = 0.25 \qquad P(n \geq 5 | n \geq 3) = 0.25$$

From this, we may calculate the following:

(inactive 2/n)-n	0.33333333	(n-1)
2	0.16666667	
3	0.08333333	
n \ge 3	0.08333333	0.33333333
...		
-(1/active 2)-1-n	0.33333333	(n-2)
1	0.16666667	
2	0.08333333	
n \ge 3	0.08333333	0.33333333
...		
(-1/active 2)-1-2[...]-2-n	0.11111111	(n-2)
1	0.05555556	
2	0.02777778	
n \ge 3	0.02777778	0.11111111
...		
-(n/inactive 2)-1[-2-...-2]-n	0.22222222	(n-2)
1	0.11111111	
2	0.05555556	
n \ge 3	0.05555556	0.22222222

Figure 1.9: breaking down the probabilities in figure 1.8

Summing, this gives the result:

RESULTS	
1	0.33333333
2	0.33333333
n	0.33333333

Figure 1.10: Summing probabilities from figure 1.9

Putting all of this together and correctly weighting the probabilities for each case, we get that the probability of the next digit is:

-1- -> -1-	0.3333333	0.1666667
-1- -> -2-	0.1666667	0.0833333
-1- -> -n-	0.1666667	0.0833333
-1- -> collapse	0.3333333	0.1666667
-2- -> -1-	0.3333333	0.0833333
-2- -> -2-	0	0
-2- -> -n-	0	0
-2- -> collapse	0.6666667	0.1666667
-n- -> -1-	0.3333333	0.0833333
-n- -> -2-	0.3333333	0.0833333
-n- -> -n-	0.3333333	0.0833333

Figure 1.11: adjusting probabilities of previous calculations

P(next = 1) =	0.3333333
P(next = 2) =	0.1666667
P(next \ge 3) =	0.1666667
P(next=col) =	0.3333333

Figure 1.12: Taking adjusted probabilities from figure 1.11 and summing

This gives us the local behaviour of the chain. Notice that the local behaviour does not take into account δ_1 (i.e. consumed digits), namely because this is an element that does not come from previous elements, and furthermore it does not allow us to "resolve" propagations. To solve these problems, we require taking a global perspective.

Let us first calculate the probability of δ_1. This is the probability that we have an inactive element followed by a 2 or $n \geq$, The probability of an inactive element is 1/3 (the probability of $n \geq$, plus the probability of $\sim n \sim 2$, plus $\sim n \sim 2 \sim 2$, and so forth). The probability of an n or 2 is 0.5, and hence

$$P(\sim i \sim (n/2)) = 1/6$$

This means that if we were to pick an arbitrary element in an arbitrarily long chain, the chances that it is a 1 coming from a δ_1 is 1/6. Next, let us calculate the probability of there being a propagation that results in a δ_k $(k \geq 2)$, since the results from a γ_b^a are already factored into the calculations for the next digit (for ex. look at the row for $\sim 2[...]2 \sim 1$).

$$i \sim 1 \sim n, i \sim 1 \sim 2 \sim n, ... \qquad \text{or} \qquad a \sim 1 \sim n,\ a \sim 1 \sim 2 \sim 2 \sim n, ...$$

where i represents an inactive component, and a represents an active component. A quick calculations shows that the probability of i is 1/3 and a is $2/3^a$. Calculating we get that:

$$P(\text{consumed}) = 1/6$$

Finally, we may look at our above table to determine that the value after an active propagation terminates is:

$$P(\text{value} = 1|\text{propagation}) = 0$$
$$P(\text{value} = 2|\text{propagation}) = 0.5$$
$$P(\text{value} \geq 3|\text{propagation}) = 0.5$$

This leaves us with the chances that we have the same digit being 2/3 which are distributed 0.5/0.25/0.25. Putting together all the possibilities, we get:

same digit	collapse	delta_1	
0.6666667	0.16666667	0.1666667	THIS!
0.5	0	1	0.5
0.25	0.5	0	0.25
0.25	0.5	0	0.25

Finally, to show that we get back the same probability distribution for each $n \geq 3$ (which is necessary for the computations of the next iteration), we require showing that:

$$P(n = 3|n \geq 3) = 0.5 \qquad P(n = 4|n \geq 3) = 0.25 \qquad P(n = 5|n \geq 3) = 0.125 \cdots$$

Here are some of the calculations for these probabilities:

	beta	gamma*	delta*	gamma*	delta*	gamma*	delta*	gamma*	delta*	gamma*	delta*
(i2/n)=ん	-あ-1-1	-ん-1-1	-ん-1-n	-ん-1-2-1	-ん-1-2-n	-ん-1-2-2-1	-ん-1-2-2-n	-ん-1-2-2-2-1	-ん-1-2-2-2-n	-ん-1-2-2-2-2-n	...
	1	2	3	4	5	6	7	8	9	10	11
	0.333333333	0.166666667	0.083333333	0.041666667	0.020833333	0.010416667	0.005208333	0.002604167	0.001302083	0.000651042	0.000325521
		0.166666667	0.083333333	0.041666667	0.020833333	0.010416667	0.005208333	0.002604167	0.001302083	0.000651042	0.000325521
	1	2	3	4	5	6	7	8	9	10	11
(a2/1)=あ	-あ-1-1	-あ-1-n	-あ-1-2-1	-あ-1-2-n	-あ-1-2-2-1	-あ-1-2-2-n	-あ-1-2-2-2-1	-あ-1-2-2-2-n	-あ-1-2-2-2-2-1	-あ-1-2-2-2-2-n	...
		delta	gamma	delta	gamma	delta	gamma	delta	gamma	delta	gamma
sum is:	1	2	3	4	5	6	7	8	9	10	11
total:	0.333333333	0.333333333	0.166666667	0.083333333	0.041666667	0.020833333	0.010416667	0.005208333	0.002604167	0.001302083	0.000651042
gamma/delta (<-*)		0.166666667	0.083333333	0.041666667	0.020833333	0.010416667	0.005208333	0.002604167	0.001302083	0.000651042	0.000325521
adjusted for guaranteed num change:	0	0.5	0.25	0.125	0.0625	0.03125	0.015625	0.0078125	0.00390625	0.001953125	0.000976563
¬2 g/d	0	0.25	0.125	0.0625	0.03125	0.015625	0.0078125	0.00390625	0.001953125	0.000976563	0.000488281

Figure 1.13: Part of the calculations

Summing up, we get that for all j:

$$P(\pi_j(C_{i+1}) = 1) = 50\%$$
$$P(\pi_j(C_{i+1}) = 2) = 25\%$$
$$P(\pi_j(C_{i+1}) \geq 3) = 25\%$$

which completes the inductive argument, and hence the proof.

[a] For i, consider $P(n) + P(\sim n \sim 2) + P(n \sim 2 \sim 2)$... and for a consider $P(1) + P(\sim 1 \sim 2) + P(1 \sim 2 \sim 2) + ...$

Corollary 2: Average Length of Chain

Let $C \in \mathcal{C}$ be an arbitrary chain of finite length. Then as $N \to \infty$, the average size decrease goes to

$$-0.2$$

under the assumption that the distribution at the tail has a consistent effect on the length when iterating the Collatz function.

Proof **:**
If C were infinite, then we already get this result from when we were calculating the probability of an element being consumed for a collapse and along with the probability of the new element caused by the termination of a propagation, namely this comes from the calculations of the ϵ's and the δ's, which comes out to an average probability -0.2 new elements are added to a chain.

For the finite case, for arbitrarily large N, we can conclude from the numbers figure 1.13 that the distributions at the tails has any effect with probability $> 1\%$ for the first 10 digits, meaning if we assume the effect of the length of the entire chain from the tail is consistent, as N get's large we see that its effect are negligible so that the average length from C_i to C_{i+1} is

$$-0.2$$

completing the proof.

Corollary 3: Normalization of Chain

Let $C_0 \in \mathcal{C}_N$ be an arbitrary chain of length $\leq N$, and consider its Collatz sequence $(C_0, C_1, ...)$. Then as $n \to \infty$. Take the sets $D_{i,j}$ which keep track of how many 1's, 2's and $n \geq 3$'s there are at position j of the chain at the ith position of the Collatz Sequence. If C_i has no position j, skip this position when counting. For example, it is possible that $D_{2,3} = \{1 : 2,\ 2 : 1,\ n : 0\}$.

Then as $N \to \infty$, the ratio between these three quantities approach

$$0.5 \qquad 0.25 \qquad 0.25$$

This is called the "normalization" as it says that given enough iterations, any chain must start having the expected probabilities at each component. For example, if you consider the chain $(\sim\ 1) \times 10'000$ it will either stay the same size or grow for $10'000$ iterations before eventually reaching a "normal" state.

Proof **:**
If C_0 was infinite, this result is immediate from the above. For the finite case, we use corollary 3 to conclude that we will have enough position to calculate the probabilities.

We can in fact do one better: we can calculate the speed of at which these ratios converges to $0.5, 0.25, 0.25$. However, I have some studying to do and so I'll leave this as an open problem.

From these results, we can finally conclude the following:

Theorem 3: Distribution a.e. Digit

Let $C \in \mathcal{C}_N$ be an arbitrary non-trivial chain of length $\leq N$ for some arbitrary large N with sequence $(C_0, C_1, ...)$. Consider the functionals $\pi_j(C_i) = n_{i,j}$ which returns the jth digit of the chain C_i. Then as $N \to \infty$:

$$P(\pi_j(C_i) = 1) = 50\% \qquad P(\pi_j(C_i) = 2) = 25\% \qquad P(\pi_j(C_i) \geq 3) = 25\%$$

for all but the last few digits. In other words, given any propagation, on average, the distribution of 1's 2's and n's in a chain is the same as picking an arbitrary chain; F does not alter the probability as it is iterated over.

***Proof* :**
Let N be some arbitrarily large number so that the chain is as long as we need it to be (a.e. 100%). Then theorem 2 gives the result for almost all positions, and corollary 5 and corollary 3 tells us that there is in general and eventually no sudden change is length of the chain, allowing for the arguments to follow.

With this, we have the distribution of most digit. The problem lies with the distribution of the last few digits which exhibit different behavior. Indeed, if π_{-1} returns the last digit, for an arbitrary chain C, we have in general that:

$$P(\pi_{-1}(F(C)) = 1) = 1/3 \qquad P(\pi_{-1}(F(C)) = 2) = 1/3 \qquad P(\pi_{-1}(F(C)) \geq 3) = 1/3$$

Which breaks the nice pattern that we've established. Furthermore, we may heuristically see that the probability at the end of the chain doesn't behave as well. Here is a screenshot of the graph of the probability of the last digit of 5'000 random chains (of length at most 1000) as we iterate through the Collatz function:

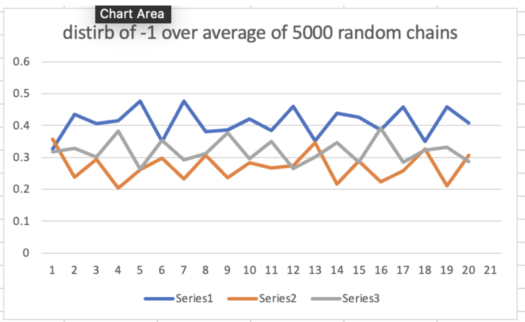

Figure 1.14: Series1 $P(\pi_{-1} = 1)$, Series2 is $P(\pi_{-1} = 2)$, Series3 is $P(\pi_{-1} \geq 3)$

This is rather unfortunate given how nice the above theorems turned out to be. However, if we take the average of these probabilities as we iterate the Collatz function, we shall start converging to some stable values, with probabilities that are in fact very simple (as we'll show in the next two sections). Part of what makes the calculation of the tail so hard is that the arithmetic's presented so far is good at computing what happens from left-to-right, but not right-to-left. The only place were we can "sort of" read from right-to-left is when finding the probability of $n \geq$ in figure 1.13, however this is not enough information to understand the behaviour at the tail. With how nice the numbers line up at the end, I suspect that there is another way of representing natural numbers from which the complexity at the tail is represented by something more manageable, however answering that question seems to be the same a solving the conjecture.

1.4 Finding the Tail: Consequence

Calculating the tail distribution is crucial. If we can figure out what the eventual probability of an arbitrary chain having an active or inactive component as the last digit, we can invoke theorem 1 to find the eventual average difference between each iteration of the Collatz Function. Ideally, we would need this average to be *negative*. To do this, we need to know how often the last digit is active/inactive. If we look at heuristic data, we will see that the final digit ought to have:

$$P(\text{last digit active}) = 0.6 \qquad P(\text{last digit inactive}) = 0.4$$

If we take this as correct, we can use theorem 1 to get that average decrease in size of a chain ought to be -0.4. This can again be heuristically verified. Furthermore, corollary 2 would be proved. We can verify heuristically that by computing many different Collatz sequences that start at the same length:

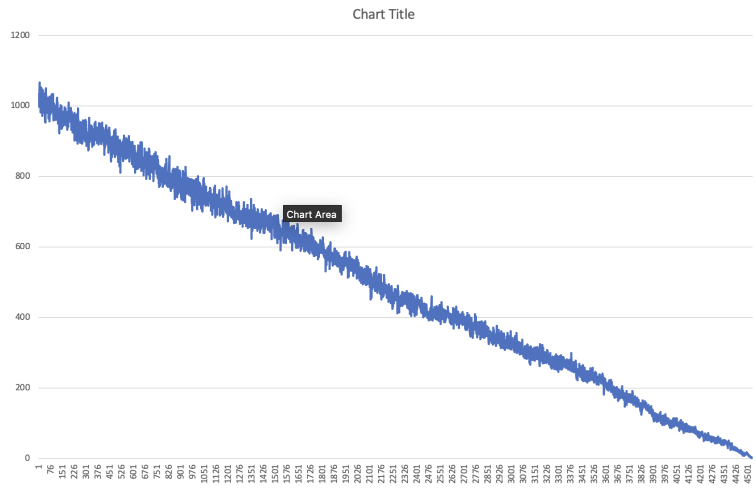

Figure 1.15: Length of random 1000 long chain during propagation

Showing one of these result would be enough to solve the Collatz Conjecture, however these numbers are based on results by running millions of random different chains and gathering data, which that is not a proof. For that, we would need to calculate the tail.

1.5 Tail Distribution Ideas

This is where the true complexity of the Collatz Conjecture lies. I have done many attempts, with the following steps being the closest to getting the desired 0.6/0.4 probability:

1. Given some natural number N, take all tails of length N.
2. Extend these tails (i.e. increase their length) so that after applying the Collatz function, the length remains N.
3. After applying N, compute what is the probability of achieving each tail
4. Update the distribution for the $n \geq 3$ numbers.
5. Calculate the probability of the last digit being active/inactive.
6. Now iterate this process. If all goes well, this process should converge.

So far, I wrote a program that almost does this; it does everything except update the distribution for $n \geq 3$ digits. With the assumption that n has the same distribution, we converge to:

$$\text{active: } 0.5687851787616299 \qquad \text{inactive: } 0.43121482123837$$

That is, it is off by a little more than 3% for each percentage. It would be very exciting if updating the distribution of the $n \geq 3$ would resolve this issue and get the desired output (which I have not done yet due to school work piling up). If this were to be the case, or if another method finds the 0.6/0.4 probabilities, we can have the following theorem:

Theorem 4: Tail Active/Inactive Distribution

For an arbitrary chain C and Collatz Sequence $(C, F(C), ..., F^{\circ k}(C))$, the probability of the last digit being active or inactive converges to:

$$P(\text{last digit active}) = 0.6 \qquad P(\text{last digit inactive}) = 0.4$$

which converges in less than 100 iterations.

Proof **:**
The best guess at what would be a proof of this result is the program in the Appendix.

An immediate corollary we shall get form this theorem is the following:

Corollary 4: Tail Distribution

For an arbitrary chain C and Collatz Sequence $(C, F(C), ..., F^{\circ k}(C))$, the probability of the tail digits converges to:

$$\begin{array}{lll}
P(\pi_{-1} = 1) = 41.5\% & P(\pi_{-1} = 2) = 26.3\% & P(\pi_{-1} \geq 3) = 32.2\% \\
P(\pi_{-2} = 1) \approx 47.6\% & P(\pi_{-2} = 2) \approx 25.5\% & P(\pi_{-2} \geq 3) \approx 26.8\% \\
P(\pi_{-3} = 1) \approx 49.3\% & P(\pi_{-3} = 2) \approx 25.2\% & P(\pi_{-3} \geq 3) \approx 25.5\% \\
P(\pi_{-4} = 1) \approx 49.7\% & P(\pi_{-4} = 2) \approx 25.1\% & P(\pi_{-4} \geq 3) \approx 52.2\% \\
P(\pi_{-5} = 1) \approx 49.9\% & P(\pi_{-5} = 2) \approx 25.04\% & P(\pi_{-5} \geq 3) \approx 25.03\%
\end{array}$$

and the position form 6-10 on rapidly on 0.5/0.25/0.25 This converges in less than 100 iterations.

Proof **:**
This would be the conclusion from the above theorem since we would be calculating the tails probabilities and can keep track of the tails digit distributions. The results shown above are what was found empirically. What was found from the program I wrote to try and prove these numbers are off by $\sim$ 2-4% per digit.

Using corollary 4 with theorem 2, we can properly conclude corollary 3. Combining the above corollary with theorem 4, we have the following corollary:

Corollary 5: Average Size Change

Let C be an arbitrary chain and $(C, F(C), ..., F^{\circ k}(C), ...)$ be the Collatz sequence associated to it. Then eventually, the average size shall change by:

$$-0.4$$

Proof **:**
By using Theorem 1 and theorem 4, can calculate the expected value of all chains starting with $1/n$ and ending with a/i, in addition to adding all possible $\sim$ 2 in-front; say each collection of chains that produce the same difference is collected in the set $\mathcal{T}$. Then if $\Delta_T = \sum(F(T)) - \sum(T))$ we must compute:

$$\sum_{T \in \mathcal{T}} P(T) \cdot \Delta_T$$

Delta_T	chain	prob	expected value
	(1/n)-(a/i)	0.75	0.2
-1	-2-1-..-a	0.075	-0.075
-2	-2-1-..-i	0.05	-0.1
-2	-2-n-..-a	0.0375	-0.075
-3	-2-n-..-i	0.025	-0.075
-3	-2-2-1-..-a	0.01875	-0.05625
-4	-2-2-1-..-i	0.0125	-0.05
-4	-2-2-n-..-a	0.009375	-0.0375
-5	-2-2-n-..-i	0.00625	-0.03125
-5	-2-2-2-1-..-a	0.0046875	-0.0234375
-6	-2-2-2-1-..-i	0.003125	-0.01875
-6	-2-2-2-n-..-a	0.00234375	-0.0140625
-7	-2-2-2-n-..-i	0.0015625	-0.0109375
-7	-2x4-1-..-a	0.001171875	-0.008203125
-8	-2x4-1-..-i	0.00078125	-0.00625
-8	-2x4-n-..-a	0.000585938	-0.0046875
-9	-2x4-n-..-i	0.000390625	-0.003515625

Figure 1.16: First few values of the calculation

The formal calculations shall be written down here later, for now it is import to know that after iterating through the first 60 blocks, we have an expected value of -0.399999999985, in other words it is converging to -0.4:

-26		2.23517E-09	-5.81145E-08
-27		1.49012E-09	-4.02331E-08
-27		1.11759E-09	-3.01749E-08
-28		7.45058E-10	-2.08616E-08
-28		5.58794E-10	-1.56462E-08
-29		3.72529E-10	-1.08033E-08
-29		2.79397E-10	-8.10251E-09
-30		1.86265E-10	-5.58794E-09
			-0.399999985

Figure 1.17: after 60 iterations

which is as we sought to show.

1.6 Conclusion

Given the assumptions and consequence thereof laid out in 1.5, we would be able to conclude the conjecture by the following reasoning:

Theorem 5: [Conjectured] Collatz Conjecture Proof

Let $t \in \mathbb{N}$ and f the Collatz Function:

$$f(t) = \begin{cases} \frac{t}{2} & t \text{ is even} \\ 3t + 1 & t \text{ is odd} \end{cases}$$

Then there exists a k such that

$$f^{\circ k}(t) = 1$$

Proof **:**
Let us now assume we have a Collatz sequence for which there does not exist a $K \in \mathbb{N}$ such that $f^{\circ K}(n) = 1$, that is the chain never "reaches" 1[a]. This also means that the chain must "normalize", as it grows for long enough that theorem 3 and corollary 4 to dictate the probabilities of the digits. Then as each chain has a finite sum and never reaches zero, the average difference must be ≥ 0. However, by corollary 5, the average sum is -0.4, a contradiction.

[a]Possible behavior may be looping, growing, chaotically moves around

A

Tail Distribution Program

There are many functions that go into calculation the tail distributions. Here is the main function that makes references to a lot of helper functions. The docstring of the function describes the desired behaviour:

```
def BlockedTailDistribution(iterations: int, length: int) -> dict:
    """Running this function will return an approxmimation of the last digit
    probabilities given the number of iterations and the length of the tails
    being considered. The larger they are, the more accurate the result ought
    to be.

    [INITIALIZATION]
    1. Create:
        a) all tails of length N (say, the list is called Tails). Include their
        probabilities
            ex. {[1, 1, 1]: 0.125, [1, 1, 2]: 0.0625, ..., [300, 300, 300]: 0.015625}
        b) Initialize TailDigitProbs with std digit distrib.
            ex : {-1: {1: 0.5, 2: 0.25, 300: 02.5}, ...}
        c) Initialize the LastDigitStateProb:
            ex: {'a': 2/3, 'i': 1/3}
        c) initialize TailProbs
            ex. {(1, 1, 1): 0.125, (1, 1, 2): 0.0625, ...}
    2. Filter all tails in Tails with Active twos into ActiveTwoTails.
    3. Now extend to take into account all possible actions:
        a) what is left over from removing ActiveTwoTails, extended normally
        b) for active two tails, extend far enough so that after applying F, it
           will be of length at least N
    4. Combine the two lists as a dict. Say the dict is called ExtendedTails. It is
       a dict bc we will be keeping track of the probabilities of each ExtendedTails
       ex. {(1, 1, 1, 1): #, (1, 2, 1, 1, 1): #, ..., [300, 300, 300]: #}
    5. Take F(ExtendedTails) and make it a dict, called FofTails (F of
        [Extended]Tails). The value of the dictionary will be the F(extended_tail):
        ex. {(1, 1, 1, 1): [1,1,1,2], (1,2,1,1): [3, 1, 2], ..., (300, 300, 300):
    [298, 1, 299, 1, 299, 1]}

    [ITERATION STEPS]
```

```
    Each iteration step represents the # of F's applied before returning the last
    digit prob.

    6. Update TailDigitProbs using FofTails (this is useful for the return
       value, but not the calculations). At the same time, update lastDigitStatus.
    7. Update Tails (be sure to tally appropriately). Using this, update
    ExtendedTails (this is key for the
       calculations of the next iteration)

    [FINAL]
    8. return the LastDigitStateProb dictionary.

    Note that currently, the progrma does not return the TailDigitProbs, however it
    keeps track of it bc I may want to look at these values in the future or
    while debugging

    Args:
        iterations: As space and time are finite, decide how many times the
        progrma iterates.
        length: As memory is finite, calculate the probility for the last
        "length" digitis. The rest will be assumed to always have
        0.5/0.25/0.25. This assumption is valied for large enough "length"

    Returns:
        The probability of a/i (active/inactive) of the last digits.
    """
    # Initialize the data structures keeping track of the data.

    tailDigitProbs: dict[int, dict[str | int, Fraction]] = {}
    for pos in range(1, length + 1):
        tailDigitProbs[-pos] = {
            1: Fraction(1, 2),
            2: Fraction(1, 4),
            "n": Fraction(1, 4),
        }
    lastAIDistrib: dict[str, Fraction] = {'a': Fraction(2,3), 'i': Fraction(1,3)}

    # Only once will tails probability be given by the specific digits. The rest
    # of the time it will be given by tailProb calculations.
    tails = initializeTailProbs(createTails(length))

    # print(fOfExtendedTails)

    # loop ITERATION time:
    for i in range(iterations):
        print(f'ITERATION # : {i+1}')
        # calculate probability of each exented tail using tails
        # NOTE: a current inefficiency: The extTails are being re-calculated,
        # when only the percentages need to be updated. This will be fixed later
        # when needed, rn it just slows down the program for large inputs.
        extendedTails = initializeExtendedTails(tails, tailDigitProbs, length)

        # save those probabilitie in fOfTails
        fOfExtendedTails = nextTails(extendedTails)
        #            extTail         Prob            F(extTail)
        # format: (1, 1, 1, 1): (Fraction(1, 16), (1, 1, 1, 2)),

        # calculate prob of each position
        tailDigitProbs = updateDigitPositionProb(fOfExtendedTails, length)

```

```
        # TEST: debugging
        for pos, value in tailDigitProbs.items():
            print('current position: ' + str(pos))
            totalProb = Fraction(0,1)
            for theType, prob in value.items():
                totalProb += prob
                print(str(theType) + ' : ' + str(float(prob)))
            print(f"totalProb: {totalProb}")
            print()

        # making an excel friendly printout
        # value = tailDigitProbs[-1]
        # print(f"{float(value[1])}, {float(value[2])}, {float(value['n'])}")


        # calc the prog of each tail
        tails = nextTailProb(fOfExtendedTails, length)
        lastAIDistrib = updatelastAIDistrib(fOfExtendedTails)
        print('last distrib:' +  str(float(lastAIDistrib['a'])) + ' , ' + str(float(
    lastAIDistrib['i'])))

        print('-------------------------')

        # TEST: debugging
        # for tail, probs in tails.items():
        #     print(f"{tail}, prob: {probs}")
        # print()


    # return digit probabilities
    return tailDigitProbs
```

Here is the output of :

`BlockedTailDistribution(50, 7)`

```

ITERATION # : 1
current position: -1

1 : 0.3333333333333333
2 : 0.3333333333333333
n : 0.3333333333333333

current position: -2

1 : 0.5694444444444444
2 : 0.2152777777777778
n : 0.2152777777777778

current position: -3

1 : 0.5185185185185185
2 : 0.24074074074074073
n : 0.24074074074074073

current position: -4

1 : 0.558641975308642
2 : 0.22067901234567902
```

```
n : 0.22067901234567902

current position: -5

1 : 0.5282921810699589
2 : 0.23585390946502058
n : 0.23585390946502058

current position: -6

1 : 0.5517832647462277
2 : 0.22410836762688616
n : 0.22410836762688616

current position: -7

1 : 0.5338363054412437
2 : 0.23308184727937814
n : 0.23308184727937814

--------------------------

last a/i  distrib:0.5300030073838305 , 0.4699969926161694

--------------------------

ITERATION # : 2
current position: -1

1 : 0.4391547309027778
2 : 0.2433810763888889
n : 0.3174641927083333

current position: -2

1 : 0.6100452564380787
2 : 0.20376643428096064
n : 0.18618830928096064

current position: -3

1 : 0.5239951051311729
2 : 0.26242555217978397
n : 0.21357934268904322

current position: -4

1 : 0.5882962701742541
2 : 0.215839024924447
n : 0.19586470490129887

current position: -5

1 : 0.545651827002422
2 : 0.24055811668783222
n : 0.2137900563097458

current position: -6

1 : 0.5729340255451246
```

```
2 : 0.22064923433391917
n : 0.2064167401209562

current position: -7

1 : 0.5455822565991274
2 : 0.2367871212824836
n : 0.21763062211838896

--------------------------

last a/i  distrib:0.5888249833778971 , 0.41117501662210293

--------------------------

ITERATION # : 3
current position: -1

1 : 0.3671488192836934
2 : 0.32211291152263377
n : 0.31073826919367287

current position: -2

1 : 0.6067816128619578
2 : 0.20152026308580354
n : 0.19169812405223874

current position: -3

1 : 0.5033766313617394
2 : 0.26194154491258853
n : 0.23468182372567203

current position: -4

1 : 0.5959863043518677
2 : 0.21227204903160776
n : 0.19174164661652449

current position: -5

1 : 0.5474744539700139
2 : 0.24091830543001413
n : 0.21160724059997194

current position: -6

1 : 0.5694298007979722
2 : 0.22128428297780756
n : 0.20928591622422021

current position: -7

1 : 0.5453421379395408
2 : 0.23569230665528615
n : 0.21896555540517307

--------------------------
```

```
last a/i  distrib:0.5705320016979236 , 0.4294679983020763

--------------------------

ITERATION # : 4
current position: -1

1 : 0.37764388707112545
2 : 0.2559975639571596
n : 0.3663585489717149

current position: -2

1 : 0.5914134394546104
2 : 0.20801421934521575
n : 0.2005723412001738

current position: -3

1 : 0.49051487216528067
2 : 0.2723889575062913
n : 0.23709617032842803

current position: -4

1 : 0.5978900909399921
2 : 0.21114353958763074
n : 0.1909663694723771

current position: -5

1 : 0.542673468461553
2 : 0.24592568302269688
n : 0.2114008485157502

current position: -6

1 : 0.5661950860049343
2 : 0.2221442801167614
n : 0.21166063387830433

current position: -7

1 : 0.5440382131308459
2 : 0.23579598620117617
n : 0.22016580066797795

--------------------------

last a/i  distrib:0.5390792739558125 , 0.4609207260441876

--------------------------

ITERATION # : 5
current position: -1

1 : 0.4301568325870183
2 : 0.270672234505298
n : 0.2991709329076837

```

```
current position: -2

1 : 0.5940244528095208
2 : 0.2056204842932277
n : 0.20035506289725152

current position: -3

1 : 0.46776859506259894
2 : 0.2823718743166058
n : 0.24985953062079527

current position: -4

1 : 0.5976267951000072
2 : 0.20822019444546103
n : 0.1941530104545318

current position: -5

1 : 0.5426985463265667
2 : 0.2418997327264257
n : 0.21540172094700766

current position: -6

1 : 0.5680214317313089
2 : 0.2217665768035711
n : 0.21021199146511999

current position: -7

1 : 0.5448828717657515
2 : 0.23556201547027106
n : 0.21955511276397752

--------------------------

last a/i  distrib:0.6004941575019017 , 0.3995058424980983

--------------------------

ITERATION # : 6
current position: -1

1 : 0.3624661029538874
2 : 0.30261167819028467
n : 0.3349222188558279

current position: -2

1 : 0.5940097321338363
2 : 0.20847889139547884
n : 0.19751137647068479

current position: -3

1 : 0.4868215742045899
2 : 0.2635774306345537
n : 0.2496009951608564
```

```
current position: -4

1 : 0.5962346452425306
2 : 0.21067223512913397
n : 0.19309311962833547

current position: -5

1 : 0.5446479443911582
2 : 0.2411513435919012
n : 0.21420071201694066

current position: -6

1 : 0.5668020302647171
2 : 0.2226692527243512
n : 0.21052871701093173

current position: -7

1 : 0.5439480413763997
2 : 0.23632877168798047
n : 0.21972318693561976

--------------------------

last a/i  distrib:0.5530175661126723 , 0.4469824338873277

--------------------------

ITERATION # : 7
current position: -1

1 : 0.40853096296824837
2 : 0.2580272828589134
n : 0.3334417541728382

current position: -2

1 : 0.5977355635189312
2 : 0.20526667981759472
n : 0.1969977566634741

current position: -3

1 : 0.48318407521060747
2 : 0.27622498497027764
n : 0.24059093981911486

current position: -4

1 : 0.5983744540087899
2 : 0.20770558330389569
n : 0.19391996268731448

current position: -5

1 : 0.5441122699122931
2 : 0.24269140144246606
```

```
n : 0.21319632864524082

current position: -6

1 : 0.5678931163564992
2 : 0.22136480306311293
n : 0.21074208058038787

current position: -7

1 : 0.543423480903079
2 : 0.23620307897607265
n : 0.2203734401208484

--------------------------

last a/i  distrib:0.5707662645639857 , 0.4292337354360143

--------------------------

ITERATION # : 8
current position: -1

1 : 0.3934938346024613
2 : 0.2916210612554627
n : 0.314885104142076

current position: -2

1 : 0.5948070914596424
2 : 0.20787491960878027
n : 0.19731798893157737

current position: -3

1 : 0.4819972176594717
2 : 0.27205630336563935
n : 0.24594647897488897

current position: -4

1 : 0.5977488086201077
2 : 0.20854339253756604
n : 0.19370779884232625

current position: -5

1 : 0.5448183197030112
2 : 0.24155526172995148
n : 0.21362641856703732

current position: -6

1 : 0.5674266658713385
2 : 0.2223120085507094
n : 0.21026132557795216

current position: -7

1 : 0.5441725240424239
```

```
2 : 0.23608105335537044
n : 0.21974642260220564

--------------------------

last a/i  distrib:0.5768535893339889 , 0.42314641066601105

--------------------------

ITERATION # : 9
current position: -1

1 : 0.38208539222409715
2 : 0.27659015713264323
n : 0.34132445064325956

current position: -2

1 : 0.5940528308486999
2 : 0.20865657285135367
n : 0.19729059629994652

current position: -3

1 : 0.4878447001554972
2 : 0.2695999841279894
n : 0.24255531571651343

current position: -4

1 : 0.5976043313193933
2 : 0.20918986773802387
n : 0.1932058009425829

current position: -5

1 : 0.5438017457209309
2 : 0.2430149338994269
n : 0.2131833203796422

current position: -6

1 : 0.56691677808222
2 : 0.22206618257532593
n : 0.21101703934245414

current position: -7

1 : 0.5436631743912321
2 : 0.23610191420847493
n : 0.22023491140029294

last a/i  distrib:0.5561271059055728 , 0.44387289409442726

.

.

.
```

```
.

.

--------------------------

last distrib:0.5687990043348553 , 0.4312009956651447

--------------------------

ITERATION # : 47
current position: -1

1 : 0.3933811159237998
2 : 0.27887324496759003
n : 0.3277456391086102
totalProb: 1

current position: -2

1 : 0.5949047729067221
2 : 0.2077227052901531
n : 0.19737252180312487
totalProb: 1

current position: -3

1 : 0.483200947672088
2 : 0.27259812976693104
n : 0.24420092256098097
totalProb: 1

current position: -4

1 : 0.5979931865627105
2 : 0.20862254492856905
n : 0.19338426850872045
totalProb: 1

current position: -5

1 : 0.544282171117185
2 : 0.2422926980109296
n : 0.21342513087188542
totalProb: 1

current position: -6

1 : 0.5672652554066429
2 : 0.22203540139574177
n : 0.21069934319761532
totalProb: 1

current position: -7

1 : 0.5438844307841912
2 : 0.23608164054122247
n : 0.22003392867458632

--------------------------
```

```

last distrib:0.5687852565688525 , 0.4312147434311475

--------------------------

ITERATION # : 48
current position: -1

1 : 0.393394511056407
2 : 0.2788577723303876
n : 0.3277477166132054
totalProb: 1

current position: -2

1 : 0.5949056914267313
2 : 0.20772179239652983
n : 0.1973725161767388
totalProb: 1

current position: -3

1 : 0.4831995781439204
2 : 0.27260222553941926
n : 0.24419819631666034
totalProb: 1

current position: -4

1 : 0.5979936496626285
2 : 0.20862196314970993
n : 0.19338438718766157
totalProb: 1

current position: -5

1 : 0.5442817017897226
2 : 0.24229330337753124
n : 0.21342499483274618
totalProb: 1

current position: -6

1 : 0.5672654852111614
2 : 0.2220349857918922
n : 0.21069952899694638
totalProb: 1

current position: -7

1 : 0.5438842058238458
2 : 0.23608161715058287
n : 0.22003417702557132
totalProb: 1

--------------------------

last distrib:0.5687888805923247 , 0.43121111940767526

--------------------------
```

```

ITERATION # : 49
current position: -1

1 : 0.39339268011718637
2 : 0.27886824084330764
n : 0.327739079039506
totalProb: 1

current position: -2

1 : 0.5949053439704781
2 : 0.20772204516099343
n : 0.1973726108685285
totalProb: 1

current position: -3

1 : 0.4831980690580662
2 : 0.2726015702102815
n : 0.24420036073165233
totalProb: 1

current position: -4

1 : 0.5979935058886033
2 : 0.20862205206205708
n : 0.19338444204933952
totalProb: 1

current position: -5

1 : 0.5442821187605815
2 : 0.24229265014912693
n : 0.21342523109029155
totalProb: 1

current position: -6

1 : 0.5672654866923511
2 : 0.2220352257565879
n : 0.21069928755106102
totalProb: 1

current position: -7

1 : 0.5438844786221726
2 : 0.23608158945325539
n : 0.220033931924572
totalProb: 1

-------------------------

last distrib:0.5687936116360708 , 0.4312063883639292

-------------------------

ITERATION # : 50

current position: -1
```

```

1 : 0.3933859220133807
2 : 0.27886549082455886
n : 0.32774858716206046
totalProb: 1

current position: -2

1 : 0.5949051609782107
2 : 0.2077223440029836
n : 0.19737249501880574
totalProb: 1

current position: -3

1 : 0.48320090356614087
2 : 0.2725996244829115
n : 0.24419947195094766
totalProb: 1

current position: -4

1 : 0.5979933739394176
2 : 0.2086223337136595
n : 0.19338429234692284
totalProb: 1

current position: -5

1 : 0.5442819096951946
2 : 0.24229306563586886
n : 0.21342502466893648
totalProb: 1

current position: -6

1 : 0.56726532918622
2 : 0.22203520398526644
n : 0.21069946682851362
totalProb: 1

current position: -7

1 : 0.5438842862468833
2 : 0.23608164029505907
n : 0.22003407345805756
totalProb: 1

--------------------------

last a/i distrib:0.5687851787616299 , 0.43121482123837

--------------------------
```