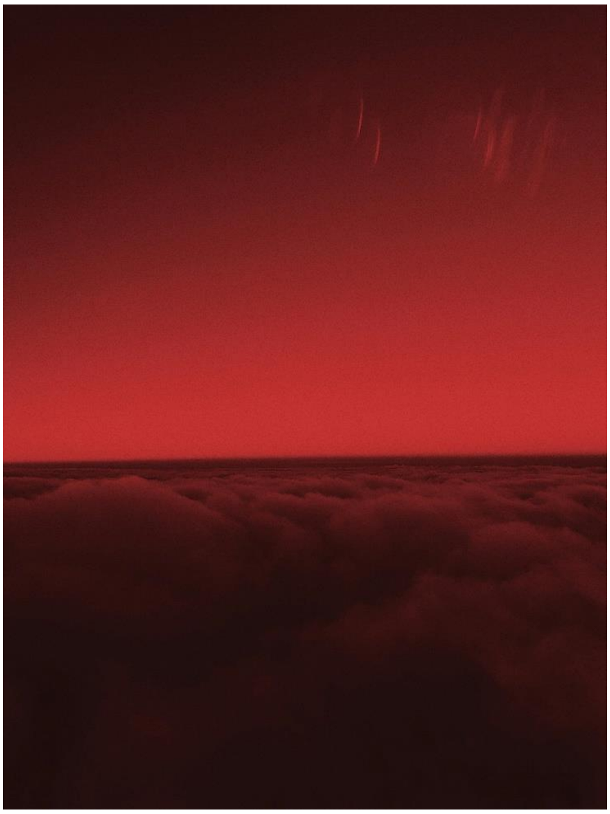


# THE NEGLIGIBILITY PRINCIPLE

Ayabonga S Mhlungu


## ABSTRACT

This paper introduces the Negligibility Principle, which simplifies summing large or infinite series by focusing on tolerance values. The principle asserts that for any given sum S or sum to infinity there is a value Sn where the terms are negligible and can be excluded without significantly affecting the overall sum. This method, similar to rounding 0.99 to 1, reduces computational complexity while maintaining accuracy. The Negligibility Principle has broad applications in fields requiring large-scale data analysis, offering a method to streamline calculations without compromising precision.

The Negligibility principle: There exist a value of $S_n$ that is negligible to the $S_\infty$ (or any given sum S) in relation to the tolerance value, the terms making up the sum are negligible.

## Terms and symbols

$\epsilon$: tolerance value

# $\epsilon = S_\emptyset$

$S_n$: sum of non-negligible terms

$S_\emptyset$: sum of negligible terms

n: Threshold term of negligible terms

#max of $S_n$

#n + 1 min value of $S_\emptyset$

# $a^|$ (the new a)

## Introduction

The negligible sum is the sum of terms that won't change the overall sum to infinity (or any given sum), just as 0.99 can be rounded as 1, the negligibility eliminates the terms that do not take a huge toil on the overall data, this is to have less terms but still get the close approximate value of the sum.

Although my mathematical skills are not very astonishing I can say with confidence that the tolerance value cannot be a fixed value for universal sequences rather it's a controlled variable to be able to discern how much can be ignored

The best way to do this is to find the percentage deviation or rather the tolerance value, the exact percentage deviation depends on how precise the approximation should be, but for the purpose of this research we will use 10% as the maximum deviation. With that we now know what gives us the tolerance value, it is the value we are willing to discard.

## The Tolerance value ($\epsilon$)

The tolerance value is a fascinating variable that helps us to see how much data…in this case the sum we can lose as stated in the introduction, the $\epsilon$ takes 2 forms these are:

**The percentage form:**

$\epsilon$ = x% of $s_\infty$ (where x ≤ 10)

This can be used to determine if the $\epsilon$ does follow the 10% rule or to find the $\epsilon$ using the given sum

**The difference form**:

$\epsilon = S_\infty - Sn$

This form enables us to do so much with the tolerance value and gives us 2 more equations.

$$\epsilon = |S_\infty - S_n|$$

$$\epsilon = |\frac{a}{1-r} - \frac{a(1-r^n)}{1-r}|$$

$$\epsilon = |\frac{a-a(1-r^n)}{1-r}|$$

$$\epsilon = \frac{a[1-(1-r^n)]}{1-r}$$

$$\epsilon = \frac{a(1-1+r^n)}{1-r}$$

$\epsilon = \frac{ar^n}{1-r}$ ...① this simplifies $\epsilon$ to this simple equation

We can make n the subject of the formula:

$$\in = \frac{ar^n}{1-r}$$

$$r^n = \frac{\in(1-r)}{a}$$

$$log_r(r^n) = log_r[\frac{\epsilon(1-r)}{a}]$$

$$n = log_r[\frac{\epsilon(1-r)}{a}] \text{ ...②}$$

This now gives us tools we can use to reduce the series to terms we need and discard other terms

## The “n” and “a|”

The n + 1 is the threshold for the negligible terms, the sum of those terms is equal to the sum of the tolerance value:

$$S_{\emptyset} = \epsilon = \frac{a^{|}(1 - r^n)}{(1 - r)}$$

The $a^{|}$ is the term n + 1 and the $a^{|}$ takes the form:

$$a^{|} = ar^{(n+1)-1}$$

$$a^{|} = ar^n$$

This equation finds us the minimum point for our $S_{\emptyset}$

This can enable us to work these later if we want to verify our $\epsilon$ and n

## The new problem

The $a^{|}$ if we are observing it closely matches with the numerator of the $\epsilon$ equation:

$$a^{|} = ar^{n}$$

$$\epsilon = \frac{ar^n}{1 - r}$$

Therefore it can be re arranged as:

$$\epsilon = \frac{a^{|}}{1 - r}$$

$$a^{|} = \epsilon\,(1 - r)$$

To illustrate the problem with this equation let's have an example:

We will use a series: $1 + \frac{1}{2} + \frac{1}{4} + \cdots$

$a = 1$

$r = \frac{1}{2}$

$S_\infty = \frac{a}{1-r}$

$S_\infty = \frac{1}{1-\frac{1}{2}}$

$S_\infty = 2$

Using the percentage form of ϵ:

$\in = \text{x\% of s}\infty$

$\in = \text{5\% of 2}$

$\in = 0{,}1$

We can find n with the tolerance value:

$\text{n} = log_r(\frac{\in(1-r)}{a})$

$\text{n} = log_{\frac{1}{2}}(\frac{0{,}1(1-\frac{1}{2})}{1})$

$n = 4{,}32192 \ldots$

$n \approx 4$

So this tells us that we can reduce the series to 4 terms to get a value close to the sum to infinity but the n is a float value so then the n itself is approximated this leads to this problem:

$a^{|} = \in (1-r)$

$a^{|} = 0{,}1\left(1-\frac{1}{2}\right)$

$a^{|} = \frac{1}{20}$

But:

$a^{|} = ar^n$

$a^{|} = 1(\frac{1}{2})^4$

$a^{|} = \frac{1}{16}$

The a$^{|}$ is inconsistent, the reason is the ϵ which brings us to the ϵ authenticity

# The ϵ authenticity

The ϵ authenticity will help solve the $a^{|}$ inconsistency problem, how?

The equation for n gives us a float value most of the time so the n is not a true reflection of the ϵ

But we can work in reverse to find the actual ϵ, this will be a true reflection of the n

This method will enable us to solve the inconsistency of the $a^{|}$ but challenges the Authenticity of ϵ

The ϵ chosen using the percentage of the sum to infinity will differ to the actual ϵ to illustrate let's use our previous example:

$$\in= \frac{ar^n}{1-r}$$

$$\in= \frac{1(\frac{1}{2})^4}{1-\frac{1}{2}}$$

$$\in= 0.125$$

The actual ϵ (which can be denoted by $\epsilon_{a)}$ will give us a true picture but will deviate from the intended ϵ

In this case it was a 0,025 deviation which generally doesn't change the non-negligible sum.

Let's solve the problem:

$$a^{|} =\in (1-r)$$

$$a^{|} = 0{,}125\left(1-\frac{1}{2}\right)$$

$$a^{|} = \frac{1}{16}$$

The $a^{|}$ is consistent if we use the actual ϵ.

# The deviation of $\epsilon_a$

As we have explored on the previous subheading the solution to the inconsistency will impact the data discarded either giving us more accurate data to our intended toleration or giving us less accurate data, this will have an impact on the sum. Although I haven't found the solution of the deviation of $\epsilon_a$ we can track how much accurate we are to our intended data and the actual data

Using our previous example:

$$d = \frac{\epsilon_{a-\epsilon}}{\epsilon} \times 100$$

$$d = \frac{0{,}125 - 0{,}1}{0{,}1} \times 100$$

$$d = 25\%$$

The $\epsilon$ deviated from the actual $\epsilon$ by 25%

We can also use this method to track the deviation

$$x = \left(\frac{\epsilon}{S_\infty} \times 100\right)\%\ )$$

$$x = \left(\frac{0.125}{2} \times 100\right)\%\ )$$

$$x = 6{,}25\%$$

Therefore the actual percentage of the data that we will lose is 6.25% of the sum rather than the 5% we intended at the start

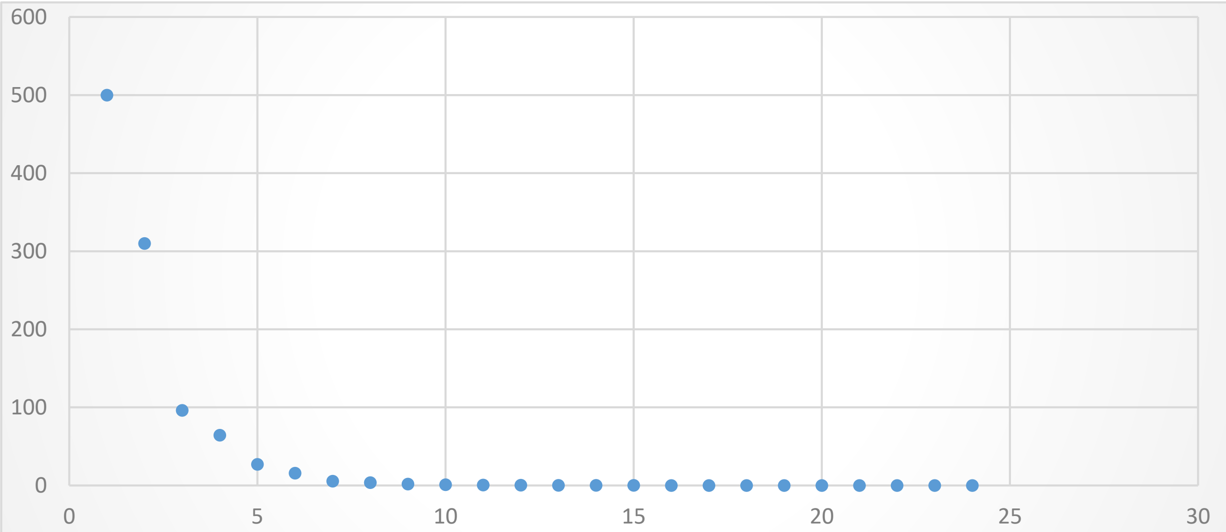
To avoid the deviation is to choose how much data (terms) you want but that wouldn't allow you to control exactly how much data (sum) you'll be losing, that method would be substandard because you might lose important data (sum). So for now the deviation is here to stay.

## The Practical application

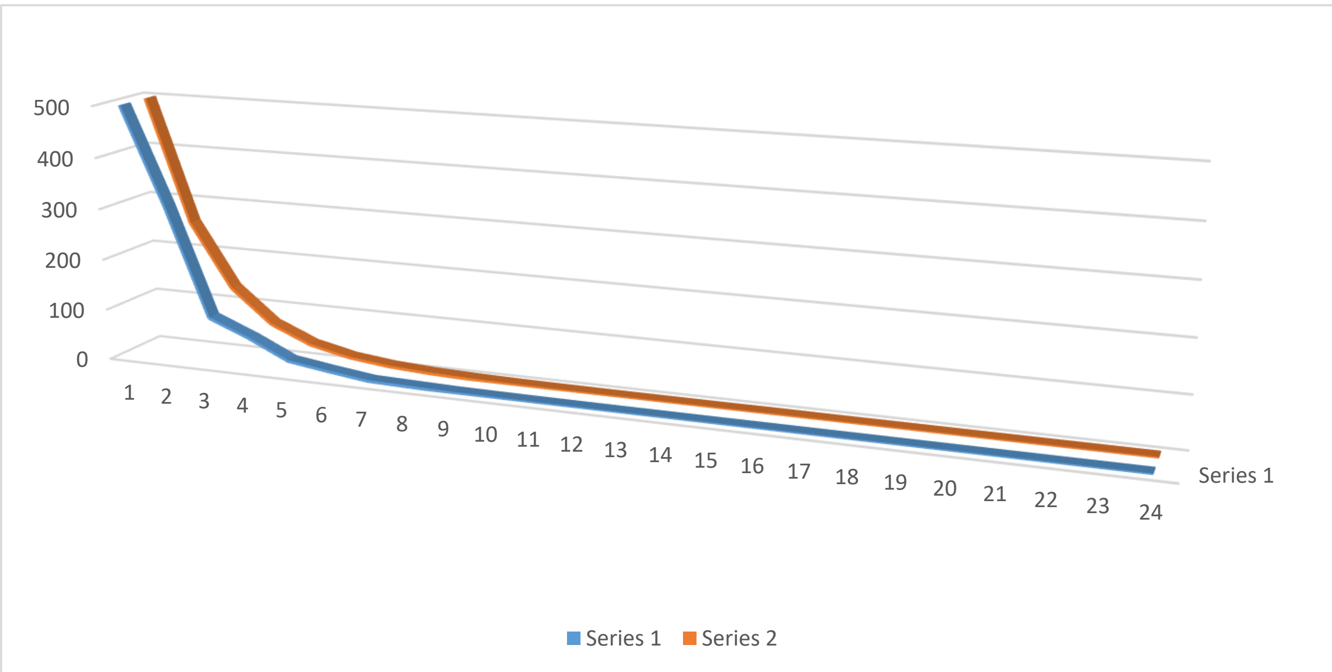
Although the principle is abstract let's try to apply it into a real life scenario,

Let's say we have a chemical that decays effectiveness in the rates between (0.4; 0.6) in hourly increments, and we are given 24 data sets for each hour, but we want to work with less data. We are given that we can ignore anything less that 1% of the chemicals combined effectiveness (for multiple hours)

We will use the scatter plot of that graph to represent the data




This is an erratic trend but follows a decay trend, we can get an average ratio and get a more smooth function



Given the new smooth function we can use its average ratio to apply the negligibility principle

**The average ratio is 0.518 and the starting point is 500 therefore:**

$a = 500$

$r = 0{,}518$

**Let's find the tolerance value using the percentage form:**

We will discard 1% of the sum.

$\in = 1\%\ of\ 1035{,}9$

$\in = 10{,}359$

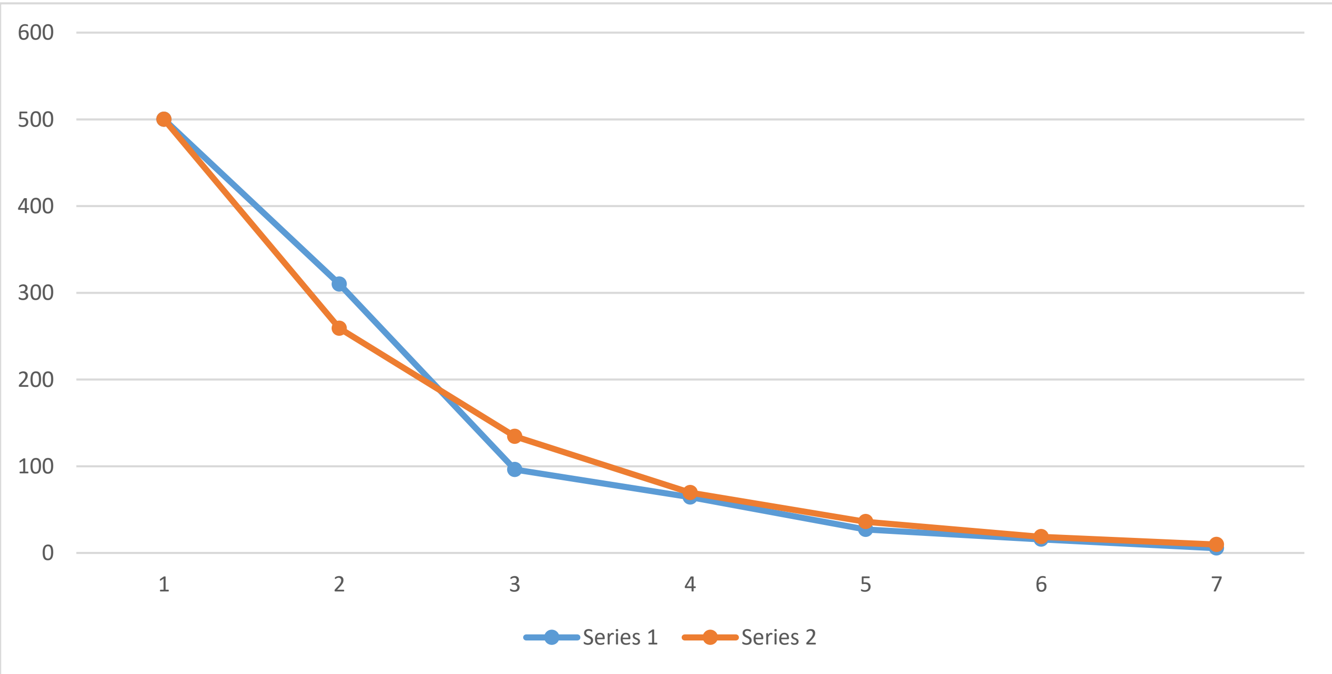
**Let's find n:**

$n = log_r[\frac{\epsilon(1-r)}{a}]$

$n = log_{(0{,}518)}[\frac{10{,}359(1-0{,}518)}{500}]$

$n \approx 7$

The 1st seven hours are only going to be considered which gives us this



With only seven data point we can be able to see the general trend that is significant.

Let’s consider the actual ϵ:

$$\epsilon = \frac{ar^n}{(1-r)}$$

$$\epsilon = \frac{(500)(0{,}518)^7}{(1-0{,}518)}$$

$$\epsilon = 10{,}38$$

This doesn’t deviate as much to the intended tolerance value.

To conclude the application: The combined effectiveness from hours 8 to 24 are negligible, discarding 17 data points, that’s 70.8% data points discarded in exchange for only 1% of the sum. That’s astonishing

## Conclusion

The principle is simple but can yield results. Though I can formally state that the principle is sealed I would not say it is perfect, rather I challenge those who will read this paper to carefully try to pinpoint any mistakes or inconsistencies I’ve made. The principle displays a potential in data science when developed further it can yield good results, although limited to decay functions which have a common ratio this is a good start to the vast world of approximation.

The paper covered the most basic parts of the principle, why we need it, what it does, and its limitations. We also looked at a practical way to use it on pre-existing data to make informed decisions. This can be a breakthrough in the field of data science when the limitations are overridden and develop more features to the principle.

I completed this work at the age of 17 just after completing the 12$^{th}$ grade so my mathematical skills and knowledge are not yet mature enough, so I plead with anyone who has greater knowledge than me to use this as a foundation to improve the principle. The paper is the base of something far greater, and the principle is subject to evolve as our mathematical understanding refines.

For inquires email: ayabongamhlungu126@gmail.com