Deriving a new formula for P_n based on empirical observations

Shreyansh Jaiswal
Atomic Energy Central School - 6 Mumbai, India

26 September 2024

Abstract

We discuss some simple ideas and facts about the following function

$$C(n) := \frac{P_n}{n \log n}$$

and it's implication for estimation of the nth prime number. We also present some rather non trivial empirical observations regarding this function, which are the base for a new function for the nth prime we propose:

$$P_n \approx \frac{(\log(n \log n) + \log(n \log n - n \log 2) - \log 2) \cdot n \log n}{\log(n \log n) + \log(\frac{n \log n - n \log 2}{2}) - \log(\log(n \log n) + \log(\frac{n \log n - n \log 2}{2}))}$$

Note - P_n and $p(n)$ are representations for the nth prime in this article. Any and all representations for log refer to the Natural logarithm.

1 Prime Number Theorem

The Prime Number Theorem (PNT) describes the asymptotic distribution of prime numbers. It provides a profound insight into how primes are distributed among the integers. Specifically, the PNT states that the number of primes less than or equal to n, denoted $\pi(n)$, is asymptotically equal to $\frac{n}{\log n}$ citeElementary Number Theory. In mathematical terms, this can be expressed as:

$$\lim_{n \to \infty} \frac{\pi(n)}{\frac{n}{\log n}} = 1$$

which can be written as,

$$\pi(n) \sim \frac{n}{\log n}$$

where the symbol $\sim$ means that the ratio of $\pi(n)$ to $\frac{n}{\log n}$ approaches 1 as n approaches infinity. This theorem was first conjectured by Gauss and Legendre in the late 18th century and was later proved independently by Hadamard and de la Vallée Poussin in 1896.[1]

The PNT also aids in estimating the nth prime number, $p(n)$. An important corollary of the PNT is that the nth prime can be approximated by:

$$p(n) \sim n \log n [1]$$

This approximation becomes more accurate as n increases.

The significance of the Prime Number Theorem lies not only in its ability to approximate the number of primes up to a given number but also in its implications for the overall understanding of number theory. The distribution of primes influences various areas of mathematics and has applications in fields such as cryptography, where large prime numbers are crucial.

2 Introduction

A well defined limit from PNT is that,

$$\lim_{x\to\infty} \frac{P_n}{n\log(n)} = 1$$

This trivially asserts that:

$$\lim_{x\to\infty} C(n) = 1$$

This function $C(n)$ is certainly not trivial, as its high precision values could help to estimate the primes with comparatively less extent of error.

A simple graph for this function can be shown:

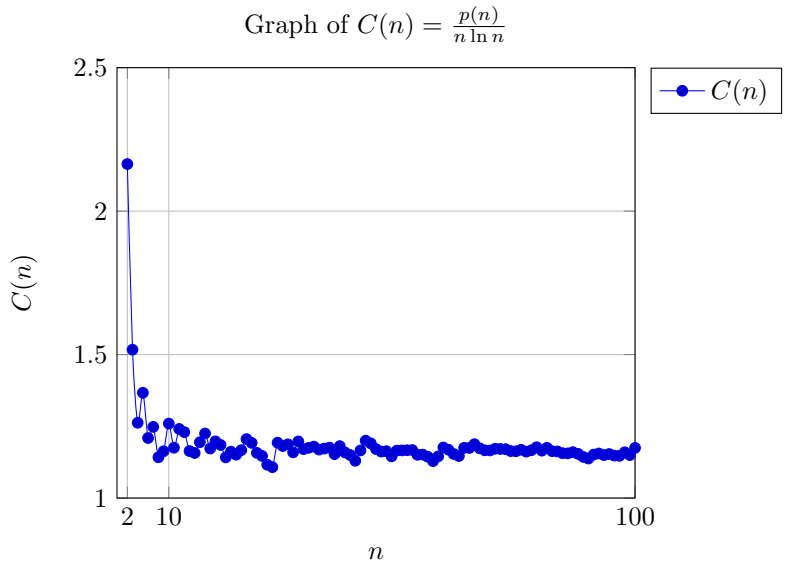

Figure 1: The function $C(n) = \frac{p(n)}{n \ln n}$ where $p(n)$ is the n-th prime, plotted till n = 100.

From this figure, it is obvious that, the function converges pretty quickly to 1, even for very small values of n

3 Understanding C(n) as a number theoretic function

By the fact that, for $C(n)$,

$$C : \mathbb{N} \to \mathbb{R}$$

, it is clear that it satisfies the definition of any number theoretic function. From this, we can use Apostol's definition for derivative of arithmetical functions [2], which is given as:

Definition: For any arithmetical function f, we define its derivative f' to be the arithmetical function given

by the equation

$$f'(n) = f(n)\log n \quad \text{for } n \geq 1.$$

This definition, aims to capture the intricate behaviour of the growth of many number theoretic functions, which is often tied to logarithmic terms. From this definition, we have,

$$C'(n) = \frac{P_n}{n} \to \log(n)$$

$$C'(n) \sim \log(n)$$

This derivative presents us a simple but important fact, that, the "growth" of $C(n)$ before it converges to 1, is mainly governed by a logarithmic factor.

4 Using empirical analysis to derive an asymptotic function for $C(n)$ inorder to derive a new formula for the nth prime

In this section, we present our formula for the function $C(n)$, which arose from empirical calculations, particularly from trying to understand the function using the primes themselves.

The fascinating result that we get from empirical data is that:

$$C(n) \approx \frac{\log(P_n \cdot P_{n/2})}{\log(P_n \cdot P_{n/2}) - \log(\log(P_n \cdot P_{n/2}))}$$

$$C(n) \approx \frac{\log(P_n) + \log(P_{n/2})}{\log(P_n) + \log(P_{n/2}) - \log(\log(P_n) + \log(P_{n/2}))}$$

Note - P_n is the nth prime, and $P_{n/2}$ is the (n/2)th prime.
Using this relation, we can rather quickly derive ourselves a general formula for $C(n)$ using the facts that, by the PNT:

$$P_n \approx n\log(n)$$

$$P_{n/2} \approx \frac{n}{2}\log(\frac{n}{2})$$

First, we can define:

$$a = \log(P_n) + \log(P_{n/2})$$

We simplify,

$$a = \log(n\log(n)) + \log(\frac{n}{2}\log(\frac{n}{2}))$$

$$a = \log(n\log(n)) + \log(\frac{n}{2}(\log(n) - \log(2))$$

$$a = \log(n\log(n)) + \log(\frac{n\log(n)}{2} - \frac{n\log(2)}{2})$$

$$a = \log(n\log(n)) + \log(\frac{n\log(n) - n\log(2)}{2})$$

$$a = (\log(n\log(n)) + \log(n\log(n) - n\log(2)) - \log(2))$$

It is obvious, that:

$$C(n) \approx \frac{a}{a - \log(a)}$$

Hence, combining,

$$C(n) \approx \frac{(\log(n\log n) + \log(n\log n - n\log 2) - \log(2)}{\log(n\log n) + \log(\frac{n\log n - n\log 2}{2}) - \log(\log(n\log n) + \log(\frac{n\log n - n\log 2}{2}))}$$

From the definition of $C(n)$, we can understand that:

$$C(n) \cdot n\log(n) = P_n$$

but as we are taking approximations for $C(n)$,

$$C(n) \cdot n\log(n) \approx P_n$$

Hence, we can finally arrive:

$$P_n \sim \frac{(\log(n\log n) + \log(n\log n - n\log 2) - \log 2) \cdot n\log n}{\log(n\log n) + \log(\frac{n\log n - n\log 2}{2}) - \log(\log(n\log n) + \log(\frac{n\log n - n\log 2}{2}))}$$

5 Validating our formula using asymptotic analysis

In this section, we work to validate our formula. Particularly, we prove, that for our approximation of $C(n)$:

$$\lim_{x\to\infty} C(n) = 1$$

5.1 Proof for the limit

Step 1: Analyzing the Numerator

The numerator is:

$$\log(n\log n) + \log(n\log n - n\log 2) - \log 2.$$

For large n, we have the following approximations: $\log(n\log n)$ grows without bound. $\log(n\log n - n\log 2) \approx \log(n\log n)$ because $n\log n$ dominates $n\log 2$ as $n \to \infty$.

Thus, the numerator simplifies as follows for large n:

$$\log(n\log n) + \log(n\log n - n\log 2) - \log 2 \approx 2\log(n\log n) - \log 2.$$

This means the numerator behaves like $2\log(n\log n)$ as $n \to \infty$.

Step 2: Analyzing the Denominator

The denominator is:

$$\log(n\log n) + \log\left(\frac{n\log n - n\log 2}{2}\right) - \log\left(\log(n\log n) + \log\left(\frac{n\log n - n\log 2}{2}\right)\right).$$

For large n:

$\log(n\log n)$ is large, and we have already approximated the second term $\log\left(\frac{n\log n - n\log 2}{2}\right) \approx \log(n\log n) - \log 2$. The last term involves $\log\left(\log(n\log n) + \log\left(\frac{n\log n - n\log 2}{2}\right)\right)$, which simplifies for large n to $\log(2\log(n\log n))$.

Thus, the denominator behaves like:

$$2\log(n\log n) - \log 2 - \log(2\log(n\log n)) \approx 2\log(n\log n) - \log 2 - \log 2 - \log(\log(n\log n)).$$

For large n, the $\log(\log(n \log n))$ term grows much slower than the $\log(n \log n)$ term, so the denominator can be approximated by $2 \log(n \log n)$ for large n.

Step 3: Taking the Limit

Now we can rewrite $C(n)$ for large n:

$$C(n) \approx \frac{2 \log(n \log n) - \log 2}{2 \log(n \log n) - \log 2 - \log(\log(n \log n))}.$$

As $n \to \infty$, the term $\log(\log(n \log n))$ becomes negligible compared to $\log(n \log n)$, so both the numerator and denominator behave asymptotically like $2 \log(n \log n)$. Thus, the ratio tends to 1:

$$\lim_{n \to \infty} C(n) = 1.$$

Conclusion

We have shown that:

$$\lim_{n \to \infty} C(n) = 1.$$

Having proven this limit, it is obvious that:

$$\lim_{n \to \infty} P_n = n \log n.$$

5.2 Validation through bounds for P_n

Through empirical analysis, we can show that, for $n > 150$:

$$n \log n < \frac{(\log(n \log n) + \log(n \log n - n \log 2) - \log 2) \cdot n \log n}{\log(n \log n) + \log(\frac{n \log n - n \log 2}{2}) - \log(\log(n \log n) + \log(\frac{n \log n - n \log 2}{2}))} < n(\log n + \log \log n)$$

These bounds, can be obtained from M Cipolla's asymptotic formula for the nth prime.[3]

6 Graphical analysis of P_n

In this section, we present some graphs for comparing two functions:

$$P_n \approx \frac{(\log(n \log n) + \log(n \log n - n \log 2) - \log 2) \cdot n \log n}{\log(n \log n) + \log(\frac{n \log n - n \log 2}{2}) - \log(\log(n \log n) + \log(\frac{n \log n - n \log 2}{2}))}$$

and

$$P_n \sim n \log n$$

One thing to note that, all the empirical calculations that are referenced in this paper, were carried out using mpmath library in python.

Graph 1:

This graph below is a graph showing the values of n on the x axis and the values for absolute errors produced by both equations. The yellow graph represents $P_n \sim n \log n$'s error. The range of primes tested were from $n = 1$ till $n = 10^6$.

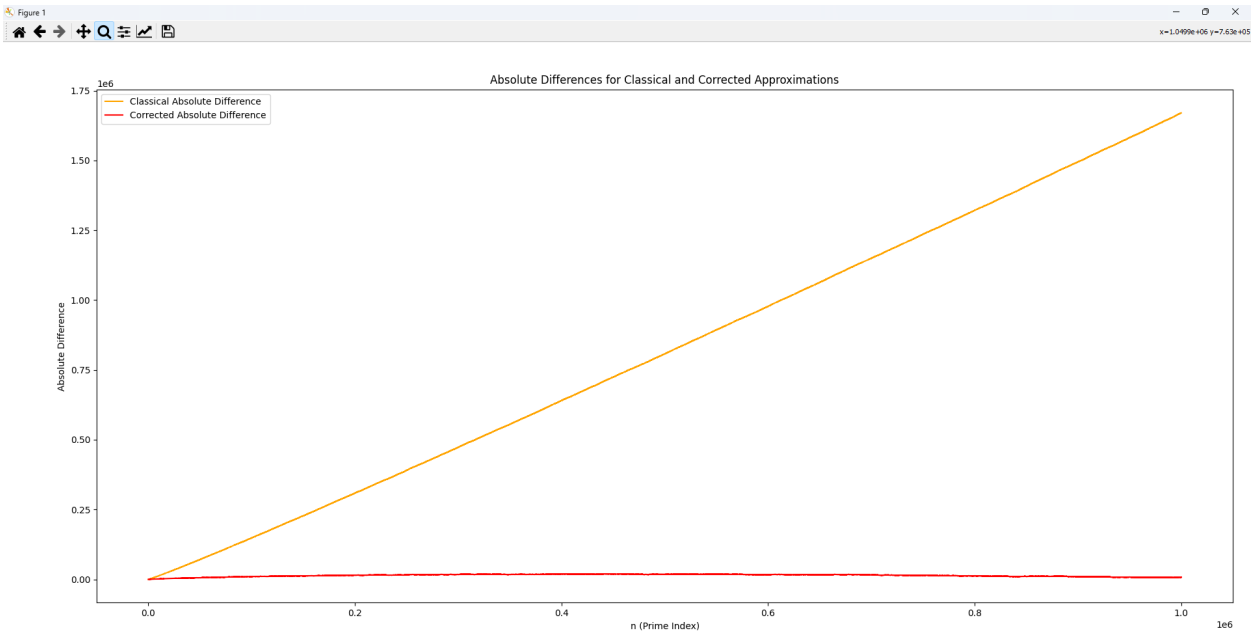

Figure 2: Comparison of Prime Estimation Formulas

Graph 2:

This second graph below represents the error for our function for estimating the nth prime. The semicirclic pattern of the error, seems to be something non trivial on its own, however, this "observation" doesn't affect the overall estimations, as can be seen from graph 1.

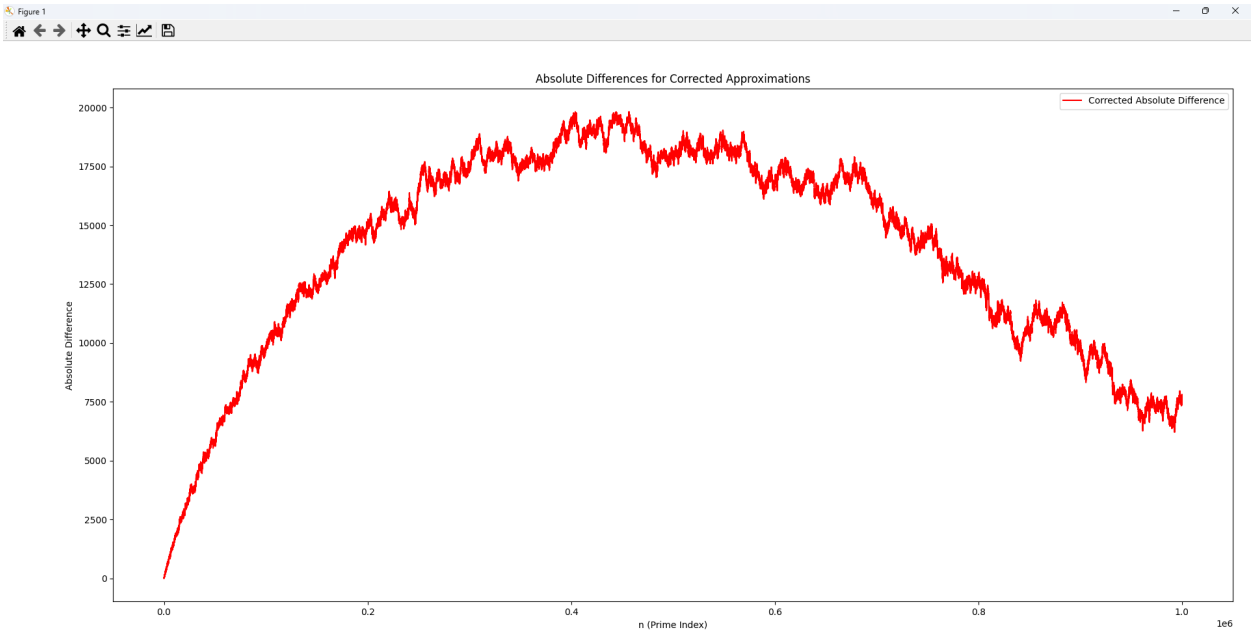

Figure 3: Absolute Difference vs nth Prime for Corrected Approximations

Graph 3 and 4:

These graphs gives us the understanding for the frequency and magnitude or errors, through a histogram chart. We understand that, our function tends to produce high frequency but low magnitude errors. The PNT's function displays a much more uniform distribution of errors, ie. The errors are roughly equal in frequency, but keep on increasing.

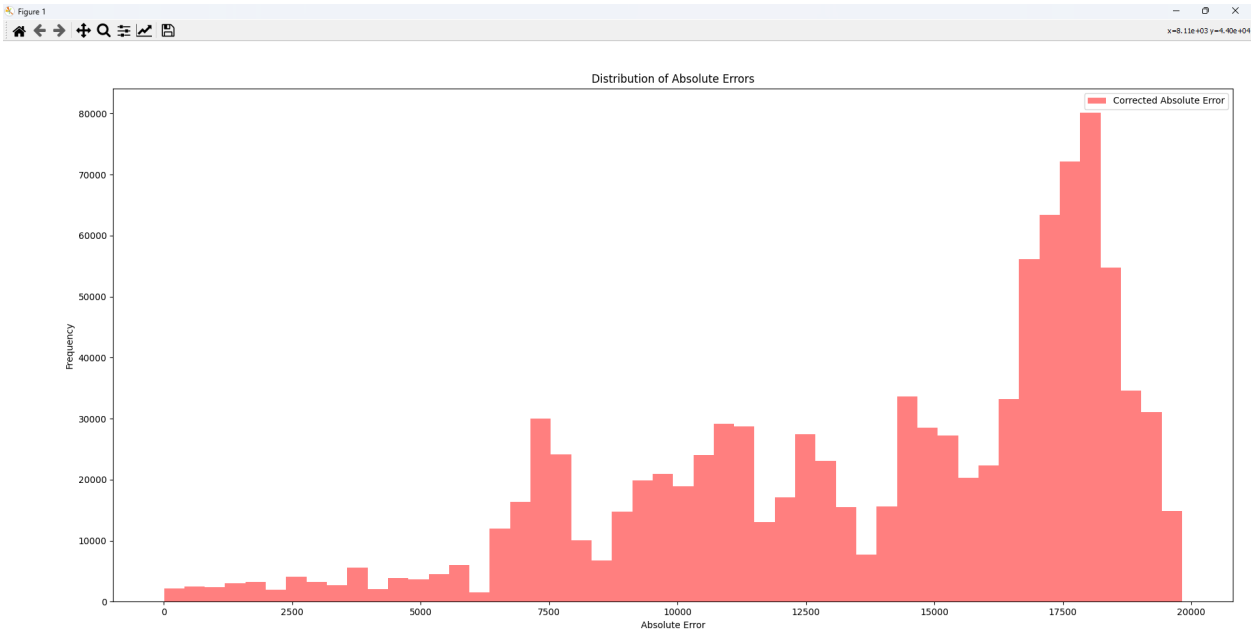

Figure 4: Histogram Analysis of Frequency vs Magnitude of Errors for New Equation

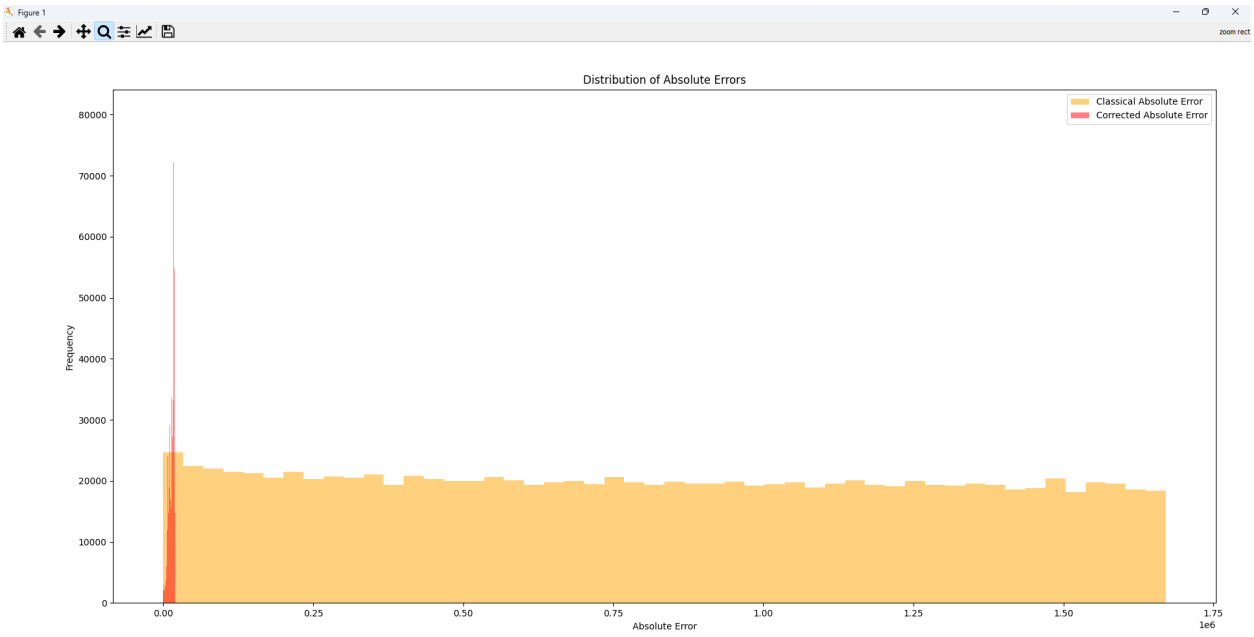

Figure 5: Comparison of Histogram Analysis of our equation with PNT's analysis

7 Applications

This section deals with the potential applications for the work.

1. This work demonstrates the usefulness of using empirical data inorder to pick up certain hidden connections. Also, this work helps to uncover some new approaches and ideas, which could help bring out some connections in this field.

2. Primes are often used in pseudorandom number generators. Accurate estimates of prime distributions can enhance the quality and randomness of generated numbers.

3. Many cryptographic algorithms, such as RSA, rely on the difficulty of factoring large prime numbers. A more accurate estimation of the nth prime can help in generating large prime numbers efficiently, which is crucial for secure key generation.

8 Conclusion

In this paper, we have derived a new formula for the n-th prime number P_n based on empirical observations of the function $C(n) = \frac{P_n}{n \log n}$. Our analysis demonstrates that the proposed formula offers a more refined estimate for P_n compared to traditional approaches. This work highlights the potential of utilizing empirical data in the development of mathematical models for prime number estimation. Future research could explore further refinements and applications of this formula in number theory.

9 References

References

[1] Elementary Number Theory - David M Burton

[2] Introduction to Analytical Number theory - Tom M Apostol

[3] "New Estimates for the nth Prime Number" - Christian Axler
Published in Journal of Integer Sequences, May 23 2019
https://cs.uwaterloo.ca/journals/JIS/VOL22/Axler/axler17.pdf