Computational and Empirical Mathematics Summer 2019

Project Leader: James Schmidt
IGL Scholars: Ethan Ashbrook, Collin Jung, Akash Pardeshi

1 Introduction

The main focus of this project was to better understand prime numbers. We did this by first examining the Prime Number Theorem (PNT).

Definition 1.1. *The Prime Number Theorem states that the prime counting function* $\pi(x)$ *asymptotically approaches* $\frac{x}{\log x}$

Since the PNT makes an asymptotic statement, it required a rigorous understanding of limits, which we accomplished to an extent by studying the $\varepsilon - \delta$ definition of limits. However, we realized that it would be difficult to fully understand this definition of limits with enough time to study the primes, so we explored other ways of understanding the PNT. Another way we tried to understand the PNT was to visualize it. The standard visualization of the PNT showed us important details of primes, such as the fact that there are infinitely many of them and that they become less frequent, since the density of the primes is asymptotic to $\frac{1}{\log x}$. Most importantly, it also showed us that the ratio of $\pi(x)$ to $\frac{x}{\log x}$ converges very slowly.

Through our research on the Prime Number Theorem, we expanded our perspective on the topic and decided we wanted to understand primes from another perspective. We realized that the PNT primarily told us about the "density" of primes among the natural numbers, but less about the "distribution", namely *where* primes are located in relation to each other. We sought to investigate where these primes were located in more detail, leading us to using a probabilistic method of analysis.

2 Probabilistic Analysis

The tools of analytic number theory gave rise to the PNT. However, the PNT only makes a statement about the density about the primes. Our second goal is to find the primes more precisely. As such, we examine the prime gaps:

Definition 2.1. *Prime gap: the* n*th prime gap is the difference between the* $n+1$*st and* n*th primes. Denote* $\gamma(n)$ *to be the* n*th prime gap:* $\gamma(n) = p_{n+1} - p_n$.

To perform statistical analysis on γ, we require a data set. Hence, let $\mathcal{D}_n = \{\gamma(1), \gamma(2), \ldots \gamma(n)\}$ be a data set. Natural parameters of $\mathcal{D}_n$ to examine are its mean and variance. As a single number such as a mean or variance is not very illustrative of the nature of γ, we define mean and variance *functions* α and β.

Definition 2.2. *Let* $\alpha(n) = \frac{1}{n}\sum_{i=1}^{n}\gamma(i)$ *represent the cumulative mean of the first* n *prime gaps. Similarly, let* $\beta(n) = \frac{1}{n}\sum_{i=1}^{n}(\alpha(n)-\gamma(i))^2$ *represent the cumulative variance of the first* n *prime gaps.*

As seen in figures 2 and 3, there is striking regularity in α and β. To see if the trend was unusual, we applied the functions to other data sets. For instance, figure 4 shows the result when α is applied to data from the stock market. No obvious trend emerged.

Given the regularity in the mean and variance plots for the prime gap data, we performed least-squares regression on α and β with the following model:

$$a\log(1+n)^c + b.$$

The results are shown in figures 5 and 6.

Note that while the regression model does model the data very well and the residuals seem to approach 0, regression was performed on a finite data set. However, we make a conjecture based on this evidence.

Conjecture 2.1. $\alpha(n) \sim a_1\log(1+n)^{c_1} + b_1$ *and* $\beta(n) \sim a_2\log(1+n)^{c_2} + b_2$.

While we cannot conclude that α and β are logarithmic, we can make a statement about the coefficients of regression if we suppose they are.

Theorem 2.1. *If functions* $f(n)$ *and* $an+b$ *are asymptotic (with* $a, b > 0$*), then the coefficients of regression on the points* $(i, f(i))$ *for* $i = 1, 2, \ldots, n$ *approach the true parameters as* $n \to \infty$:

$$f \sim an+b \to a = \lim_{n\to\infty} a_n \text{ and } b = \lim_{n\to\infty} b_n$$

where a_n *and* b_n *are the coefficients of regression on the first* n *points.*

As a direct consequence of 2.1, we know that by performing regression on α with more and more points, the coefficients of regression a_n, b_n, c_n will converge to the true parameters.

3 Approximations For Primes

From the definition of $\gamma(n)$ (definition 2.1), we know that $p_{n+1} = p_n + \gamma(n)$. Since $\alpha(n)$ is the expected gap i.e. $\gamma(n)$ (if we assume all gaps are equally likely), we have $p_{n+1} \approx p_n + \alpha(n)$. Hence, we take $\hat{p}_{n+1} = p_n + \alpha(n)$ to be an approximation of the nth prime.

As seen in figure 7 the predictions for p_n are accurate. The histogram of the errors $|p_n - \hat{p}_n|$ shows that the errors are concentrated on the smaller side and the magnitude of the errors themselves are small.

		moment				
		1	2	3	4	5
n	10^3	0.9754	1.8371	2.5144	3.1186	3.8459
	10^5	1.0001	2.0620	3.2025	4.2306	4.9009
	10^7	0.9473	1.9795	2.9967	3.9405	4.8066

Table 1: For different values of n, we calculate the coefficient c when fitting the first five moments.

A crucial consequence of theorem 2.1 is that by performing regression once on a very large number of points, we obtain very accurate approximations for a, b, c. After such a computation, we may use $a_n \log(1+n)^{c_n} + b_n$ in place of α to speed up computation while retaining the same level of accuracy.

4 Future Work

The research we have conducted gives rise to many new questions and ideas to be followed.

The first includes proving (or disproving) that α does actually follow a logarithmic model. If it does, then subsequent work may establish statements regarding bounds on the error of our predictions and studying the implications on computation times for finding primes.

We made interesting observation we made when fitting the model $a \log(1+n)^c + b$ to α and β. It seemed that the coefficient c was approaching 1 as n grew larger for α, and approached 2 when fitting β. Note that the first moment of a random variable is simply the mean, and the second moment is related to the variance ($\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$.) From this, we have the following conjecture:

Conjecture 4.1. *Define a* k*th moment function* $\mu_k(n) = \frac{1}{n}\sum_{i=1}^{n} \gamma(i)^k$*, where* n *is the number of data points. Then the coefficient* c *when performing regression with the model* $a \log(1+n)^c + b$ *on* $\mu_k(n)$ *will approach* k *as* $n \to \infty$.

We have already computed approximations for c on samples of n. 1 shows our current results.

Another area of work includes formalizing the notion of weighted least-squares regression for asymptotics. It may be the case that a weighted regression may improve the approximations for p_n for particular choices of weight functions. Formalizing the implications of weighted regression, including choices for weight functions, may be studied in the future.

A Proof of Theorem 2.1

Recall theorem 2.1: "If functions $f(n)$ and $an + b$ are asymptotic (with $a, b > 0$), then the coefficients of regression on the points $(i, f(i))$ for $i = 1, 2, \ldots, n$ approach the true parameters as $n \to \infty$:

$$f \sim an + b \to a = \lim_{n\to\infty} a_n \text{ and } b = \lim_{n\to\infty} b_n$$

where a_n and b_n are the coefficients of regression on the first n points."

Proof. In order to prove a statement about the coefficients of least-squares regression, we must first derive the coefficients.

Lemma A.1. *In linear least-squares regression, the coefficients* a *and* b *of* $\widehat{y} = ax + b$ *that minimize the sum of the squared residuals are*

$$a = \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) - n\sum_{i=1}^{n} x_i y_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - n\sum_{i=1}^{n} x_i^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

where n *is the number of points.*

Proof. Define $L(a, b)$ to be the function representing the sum of the squared residuals: $\sum(y_i - \widehat{y})^2$. The coefficients from least-squares regression are those that minimize L. In order to find the coefficients that minimize L, we set the partial derivatives of L with respect to a and b to 0 and solve for a and b..

$$\frac{\partial}{\partial a}L = \sum \frac{\partial}{\partial a}(y_i - ax_i - b)^2 = \sum 2(y_i - ax_i - b)\cdot -x_i \tag{1}$$

$$\frac{\partial}{\partial b}L = \sum \frac{\partial}{\partial b}(y_i - ax_i - b)^2 = \sum 2(y_i - ax_i - b)\cdot -1 \tag{2}$$

We set equations (1) and (2) equal to 0 and solve for a and b.

$$\sum_{i=1}^{n}(y_i - ax_i - b)\cdot x_i = 0 \tag{3}$$

$$\sum_{i=1}^{n}(y_i - ax_i - b) = 0. \tag{4}$$

We solve for b in terms of a in equation (4):

$$\sum y_i - a\sum x_i - bn = 0$$

$$b = \frac{\sum y_i - a\sum x_i}{n}.$$

We solve for a by substituting $\frac{\sum y_i - a\sum x_i}{n}$ for b in equation (3):

$$\begin{aligned}
\sum x_i^2 + \frac{\sum y_i - a\sum x_i}{n}\sum x_i &= \sum x_i y_i \\
a\sum x_i^2 + \frac{\sum x_i \sum y_i - a\left(\sum x_i\right)^2}{n} &= \sum x_i y_i \\
an\sum x_i^2 + \sum x_i \sum y_i - a\left(\sum x_i\right)^2 &= n\sum x_i y_i \\
an\sum x_i^2 + \sum x_i \sum y_i - a\left(\sum x_i\right)^2 &= n\sum x_i y_i \\
a\left(\left(\sum x_i\right)^2 - n\sum x_i^2\right) &= \sum x_i \sum y_i - n\sum x_i y_i \\
a &= \frac{\sum x_i \sum y_i - n\sum x_i y_i}{\left(\sum x_i\right)^2 - n\sum x_i^2}.
\end{aligned}$$

Thus the coefficients that minimize the sum of squared residuals are as desired.

□

Lemma A.2. *If* $f(n) \sim an + b$ *(with* $a, b > 0$*) then* $\sum_{i=1}^{n} f(i) \sim \sum_{i=1}^{n} an + b$.

Proof. The Stolz-Cesaro theorem states for infinite sequences of real numbers $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty} > 0$ with $\sum b_n = \infty$, if $\lim_{n\to\infty} \frac{a_n}{b_n} = L$, for a finite value of L, then $\lim_{n\to\infty} \frac{\sum a_n}{\sum b_n} = L$.

Suppose $f(n) \sim an + b$. Let $\{a_n\}$ to be the sequence defined by $f(i)$, and let $\{b_n\}$ to be the sequence defined by $ai + b$ for $i = 1, 2, \ldots, n$. The summation of b_n must diverge, since $a > 0$, by assumption. We have $f(n) \sim an + b$ which means $\lim_{n\to\infty} \frac{f(n)}{an+b} = 1$, by the definition of an asymptotic. Since 1 is finite and $a, b > 0$, we may apply the Stolz-Cesaro theorem to get $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(i)}{\sum_{i=1}^{n} (ai+b)} = 1$, meaning $\sum_{i=1}^{n} f(i) \sim \sum_{i=1}^{n} (ai + b)$, by the definition of an asymptotic. Hence the lemma. □

Lemma A.3. *If* $f(n) \sim an + b$ *(with* $a > 0$*), then* $\sum_{i=1}^{n} i(ai + b) \sim \sum_{i=1}^{n} if(i)$.

Proof. Similar to the proof of lemma A.2, we use the Stolz-Cesaro theorem.

Suppose $f(n) \sim an + b$. Let $\{a_n\}$ to be the sequence defined by $if(i)$, and let $\{b_n\}$ to be the sequence defined by $i(ai + b)$ for $i = 1, 2, \ldots, n$. The summation of b_n must diverge, since $a > 0$, by assumption. We have $f(n) \sim an + b$ which means $\lim_{n\to\infty} \frac{nf(n)}{n(an+b)} = \frac{f(n)}{an+b} = 1$, by the definition of an asymptotic. Since 1 is finite and $a, b > 0$, we may apply the Stolz-Cesaro theorem to get

$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} if(i)}{\sum_{i=1}^{n} i(ai+b)} = 1$, meaning $\sum_{i=1}^{n} if(i) \sim \sum_{i=1}^{n} i(ai+b)$, by the definition of an asymptotic. Hence the lemma. □

We first show that $a = \lim a_n$, where a_n is the coefficient of regression on the points $(i, f(i))$, for $i = 1, 2, \ldots, n$. Note that $x_i = i$ for all i. By lemma A.1 we have

$$a_n = \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} f(x_i)\right) - n\left(\sum_{i=1}^{n} x_i \cdot f(x_i)\right)}{\left(\sum_{i=1}^{n} x_i\right)^2 - n\sum_{i=1}^{n} x_i^2} \tag{5}$$

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{\left(\sum_{i=1}^{n} i\right)\left(\sum_{i=1}^{n} f(i)\right) - n\left(\sum_{i=1}^{n} if(i)\right)}{\left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2}. \tag{6}$$

By the linearity of the limit, we can rewrite equation (6) as

$$= \frac{\lim_{n\to\infty}\left(\sum_{i=1}^{n} i\right)\left(\sum_{i=1}^{n} f(i)\right) - \lim_{n\to\infty} n\left(\sum_{i=1}^{n} if(i)\right)}{\lim_{n\to\infty}\left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2}. \tag{7}$$

We may multiply by 1 in the numerator without changing the equality.

$$= \frac{\lim_{n\to\infty}\left(\sum_{i=1}^{n} i\right)\left(\sum_{i=1}^{n} f(i)\right)\cdot\frac{\sum_{i=1}^{n} ai+b}{\sum_{i=1}^{n} ai+b} - \lim_{n\to\infty} n\left(\sum_{i=1}^{n} if(i)\right)\cdot\frac{\sum_{i=1}^{n} i(ai+b)}{\sum_{i=1}^{n} i(ai+b)}}{\lim_{n\to\infty}\left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2} \tag{8}$$

$$= \frac{\lim_{n\to\infty}\left(\sum_{i=1}^{n} i\right)\left(\sum_{i=1}^{n} ai+b)\right)\cdot\frac{\sum_{i=1}^{n} f(i)}{\sum_{i=1}^{n} ai+b} - \lim_{n\to\infty} n\left(\sum_{i=1}^{n} i(ai+b)\right)\cdot\frac{\sum_{i=1}^{n} if(i)}{\sum_{i=1}^{n} i(ai+b)}}{\lim_{n\to\infty}\left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2} \tag{9}$$

By lemmas A.2 and A.3, we know $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(i)}{\sum^{n}(ai+b)} = 1$ and $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} if(i)}{\sum_{i=1}^{n} i(ai+b)} = 1$. Therefore,

$$= \frac{\lim_{n\to\infty} \left(\sum_{i=1}^{n} i\right)\left(\sum_{i=1}^{n} ai+b)\right) - \lim_{n\to\infty} n\left(\sum_{i=1}^{n} i(ai+b)\right)}{\lim_{n\to\infty} \left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2}. \tag{10}$$

We pull the limits out and simplify.

$$= \lim_{n\to\infty} \frac{\left(\sum_{i=1}^{n} i\right)\left(\sum_{i=1}^{n} ai + \sum_{i=1}^{n} b\right) - n\left(\sum_{i=1}^{n} ai^2 + bi\right)}{\left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2} \tag{11}$$

$$= \lim_{n\to\infty} \frac{a\left(\sum_{i=1}^{n} i\right)^2 + bn\left(\sum_{i=1}^{n} i\right) - na\left(\sum_{i=1}^{n} i^2\right) - bn\sum_{i=1}^{n} i}{\left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2} \tag{12}$$

$$= \lim_{n\to\infty} \frac{a\left(\sum_{i=1}^{n} i\right)^2 - na\left(\sum_{i=1}^{n} i^2\right)}{\left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2} \tag{13}$$

$$= \lim_{n\to\infty} a\frac{\left(\sum_{i=1}^{n} i\right)^2 - n\left(\sum_{i=1}^{n} i^2\right)}{\left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2} \tag{14}$$

$$= a. \tag{15}$$

In a similar way, we evaluate $\lim_{n\to\infty} b_n$. Again, note that $x_i = i$, for all i. From lemma A.1 we have

$$b_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{a_n}{n}\sum_{i=1}^{n} x_i \tag{16}$$

$$\lim_{n\to\infty} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(i) - \frac{a_n}{n}\sum_{i=1}^{n} i. \tag{17}$$

By the linearity of the limit, we have

$$= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(i) - \lim_{n\to\infty} a_n \cdot \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} i. \tag{18}$$

By (15), we know $\lim_{n\to\infty} a_n = a$. Therefore,

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(i) - \lim_{n\to\infty} \frac{a}{n} \sum_{i=1}^{n} i. \tag{19}$$

We may multiply by 1 without changing the equality.

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(i) \cdot \frac{\sum_{i=1}^{n} ai + b}{\sum_{i=1}^{n} ai + b} - \lim_{n\to\infty} \frac{a}{n} \sum_{i=1}^{n} i \tag{20}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (ai + b) \cdot \frac{\sum_{i=1}^{n} f(i)}{\sum_{i=1}^{n} ai + b} - \lim_{n\to\infty} \frac{a}{n} \sum_{i=1}^{n} i \tag{21}$$

We pull out the limit and simplify.

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (ax_i + b) - \frac{a}{n} \sum_{i=1}^{n} x_i \tag{22}$$

$$= \lim_{n\to\infty} \frac{1}{n} \left(a \sum_{i=1}^{n} x_i + bn \right) - \frac{a}{n} \sum_{i=1}^{n} x_i \tag{23}$$

$$= \lim_{n\to\infty} \frac{a}{n} \left(\sum_{i=1}^{n} x_i \right) + b - \frac{a}{n} \sum_{i=1}^{n} x_i \tag{24}$$

$$= b. \tag{25}$$

Therefore, $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$.

□

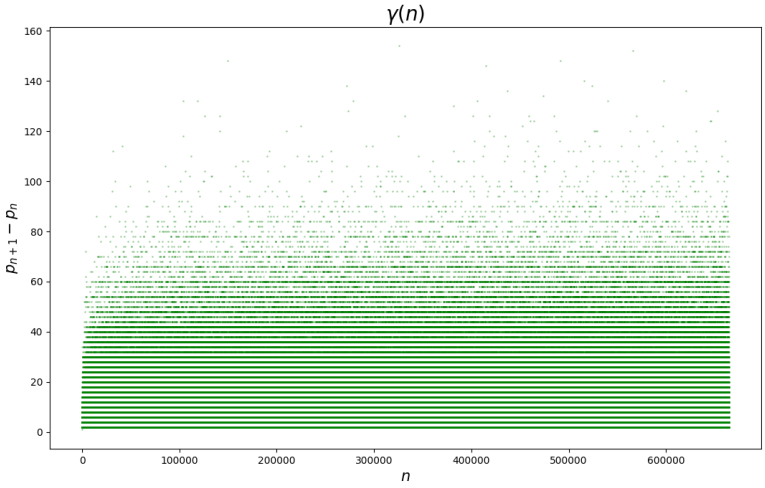

Figure 1: The first ten million prime gaps are plotted.

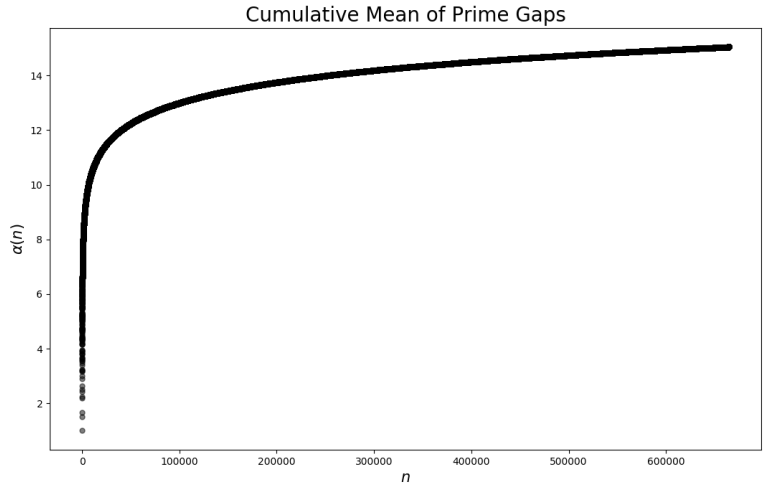

Figure 2: $\alpha(n)$ is plotted up to $n = 10^7$. A clear, roughly logarithmic curve appears.

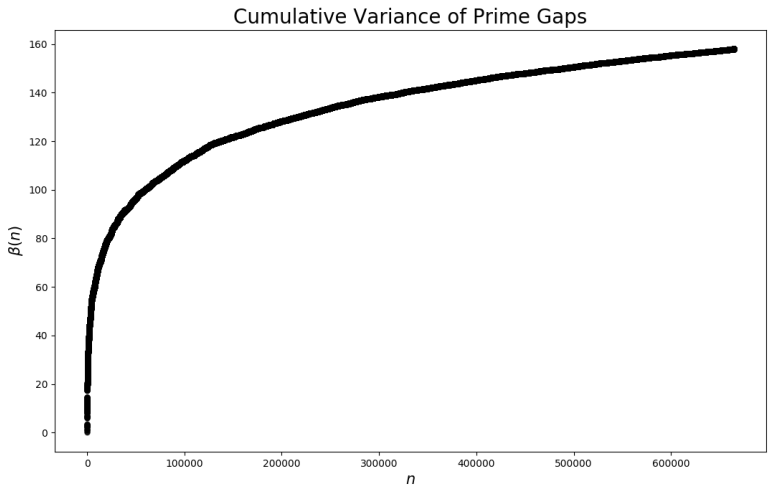

Figure 3: $\beta(n)$ is plotted up to $n = 10^7$. Again, a clear, roughly logarithmic curve appears.

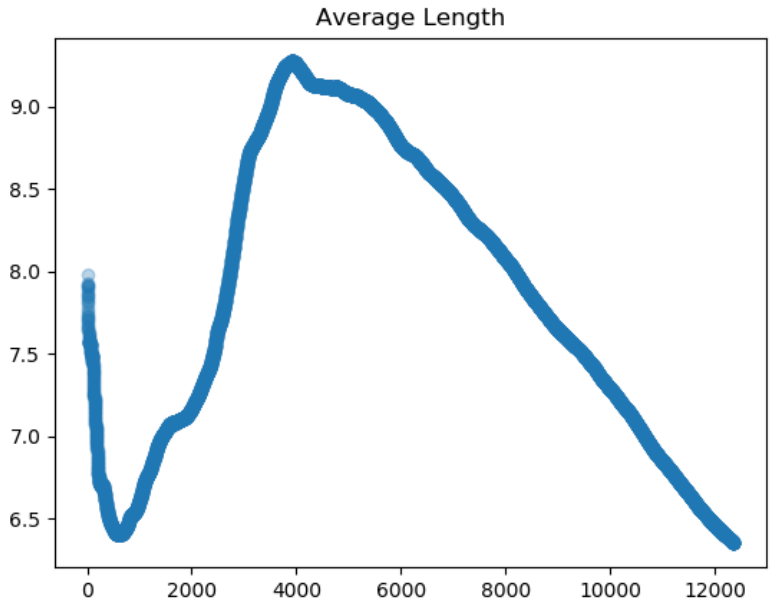

Figure 4: Average value of TNX stock prices

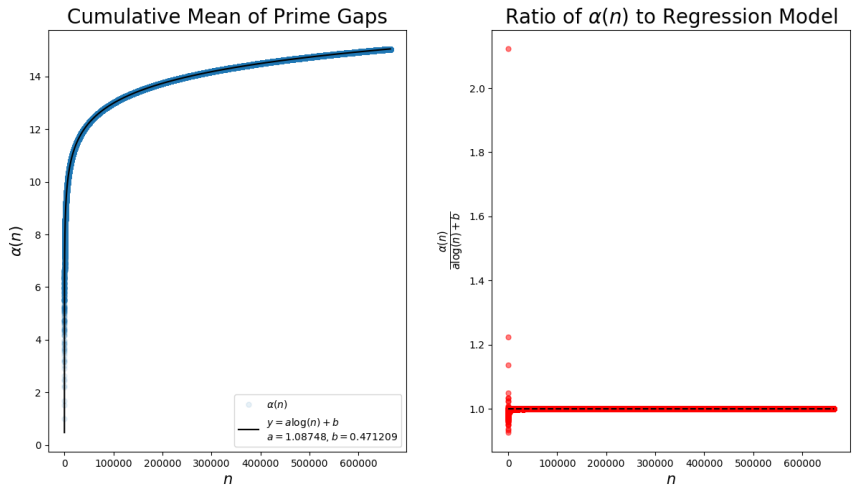

Figure 5: On the left, $\alpha(n)$ is plotted, along with the regression model obtained from ten million data points. On the right, their ratio is plotted and seems to converge to 1.

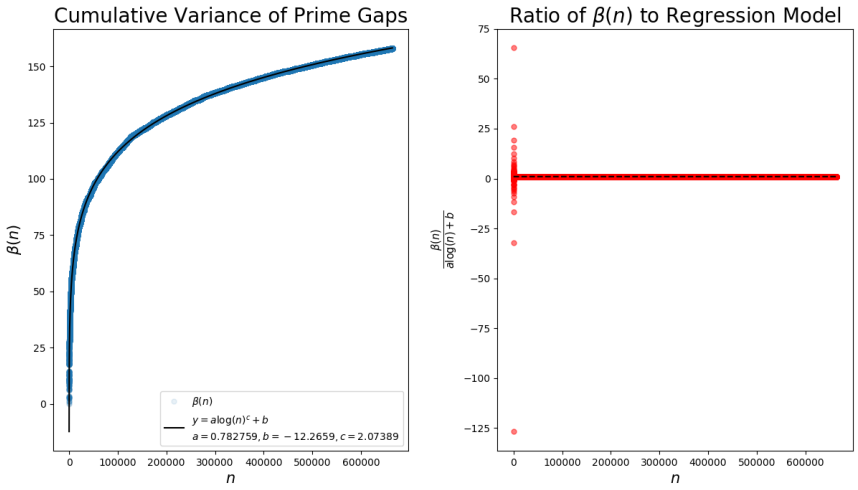

Figure 6: On the left, $\beta(n)$ is plotted, along with the regression model obtained from ten million data points. On the right, their ratio is plotted and seems to converge to 1.

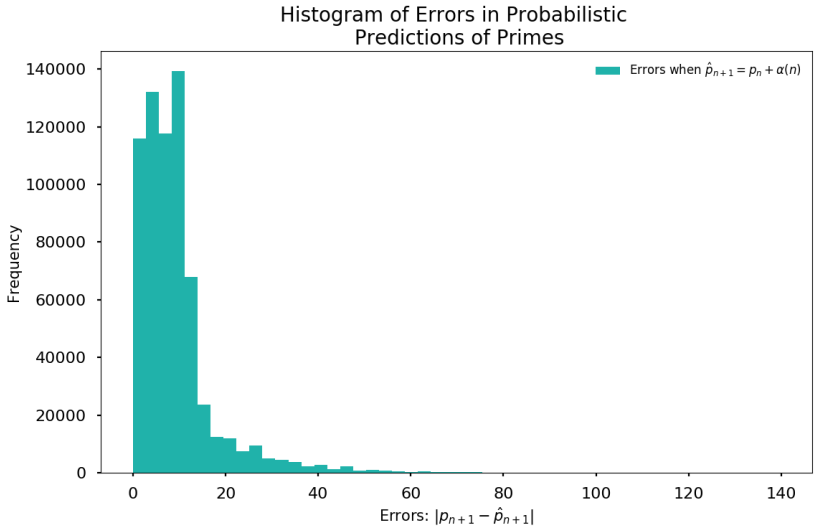

Figure 7: A Histogram of the first ten million errors $|\mathrm{p_n} - \hat{\mathrm{p}}_\mathrm{n}|$ is shown. Note the magnitude of the errors is small, and the errors are skewed towards the smaller side.