Mathematica Eterna
ISSN: 1314-3344
OPEN ACCESS Freely available online
Review Article

# Stochastic Characterization and Numerical Analysis of Prime Gaps for Stationary Distribution Approximation

**George Gabriel***

*Department of Mathematics and Statistics, McGill University, Quebec, Canada*

## ABSTRACT

This paper presents a comprehensive stochastic framework for modeling the Probability Mass Function (PMF) of prime gaps, focusing on the existence and characterization of a stationary distribution. Through theoretical analysis, we establish that prime gap sequences can be effectively modelled as stochastic processes and provide a formal proof of the existence of a stationary distribution that governs their long-term behavior.

To support our theoretical findings, we conducted an extensive numerical analysis using the Canadian supercomputer "Béluga", computing prime gaps up to $10^{12}$ using the Sieve of Eratosthenes. The empirical results validate our models, particularly those incorporating arithmetic properties such as prime factorization and modulo 6 congruence, demonstrating that they capture the periodic and combinatorial nature of prime gap distributions more accurately than simpler models.

Our study reveals that approximations adopting a geometric PMF structure with piecewise components outperform non-piecewise models and align with the stationary distribution. Specifically, prime gaps of 2 and 4 exhibit a uniform distribution, comprising approximately 5% of all gaps, while larger gaps follow a geometric progression. This dual nature of prime gaps-uniform for smaller gaps and structured for larger ones-offers a novel perspective on their distribution. Our formal proof of the stationary distribution not only enhances the understanding of prime gap distributions but also contributes to the broader field of stochastic modeling in prime number theory.

**Keywords:** Multiple kernel learning; Multi-class classification; Kernel clustering

## INTRODUCTION

Prime gaps have been the subject of mathematical inquiry for centuries, with early contributions from Euclid to more recent advancements by mathematicians such as Hardy et al., Erdos [1-3]. Recent breakthroughs by Goldston et al., Zhang, have provided new impetus to this field of study [4,5]. The prime number theorem, which approximates the distribution of primes, suggests that primes become less frequent as numbers increase, yet does not fully explain the variability in the gaps between consecutive primes [6].

The distribution of prime gaps is conjectured to follow certain probabilistic models, including poisson-like distributions for larger gaps. The Hardy et al., conjectures, specifically the first conjecture, predict the density of prime pairs separated by a specific even gap, proposing that the occurrence of such pairs can be modelled by a non-trivial multiplicative function over the primes [2]. This study leveraged the computational resources of the Canadian supercomputer "Béluga" to perform large-scale calculations, generating up to $10^{12}$ prime gaps using the Sieve of Eratosthenes algorithm. This extensive dataset was crucial in validating the accuracy and robustness of the models developed in this research.

## MATERIALS AND METHODS

### Stochastic modeling of prime gaps using Markov chains

A discrete time Markov chain $\{X_n : n \in N\}$ is a stochastic process with states $s = \{j : j \in Z\}$ indexed by discrete time T such that:

$$P\left(X_n = j \mid X_0 = j_0, X_1 = j_1, \ldots, X_{n-1} = i\right) = P\left(X_n = j \mid X_n - 1 = i\right)$$

In the context of a Markov chain, the transition matrix P is a matrix that describes the probabilities of transitioning from one state to another. Each entry $P_{ij}$ in the transition matrix represents the probability of moving from state i to state j in one step.

The probability $P\ddot{u}X_n = j \; X_{n-1} = i$ is denoted by $P_{ij}$ and the matrix $P = (P_{ij})$ is called a one-step transition matrix. The matrix P is a stochastic matrix in the sense that:

**Correspondence to:** George Gabriel, Department of Mathematics and Statistics, McGill University, Quebec, Canada, E-mail: georgegabriel@gmail.com

**Received:** 20-Aug-2024, Manuscript No. ME-24-33569; **Editor assigned:** 23-Aug-2024, PreQC No. ME-24-33569 (PQ); **Reviewed:** 09-Sep-2024, QC No. ME-24-33569; **Revised:** 16-Sep-2024, Manuscript No. ME-24-33569 (R); **Published:** 23-Sep-2024, DOI: 10.35248/1314-3344.24.14.236

**Citation:** Gabriel G (2024). Stochastic Characterization and Numerical Analysis of Prime Gaps for Stationary Distribution Approximation. Math Eter. 14:236.

$$\sum_j P_{ij} = 1,\ j,i \in S$$

Required to use P as a model for the sequence of prime gaps $\{d_n\}_{n=1}^{\infty}$ where, $ü_n = \ _{n+1} - \ _n$ for $n§\in N_0$ and $P_n$ is the n-th prime number. Although primes are not random, their appearance in the number sequence is unpredictable and so, they are often modeled as random numbers. We remark that for small values of n, the model P is a poor representation of the behavior of prime gaps because the probability that the system is in state j at time n depends not only on the immediate past state i but also on the states before that down to the state at the 0-th step. For moderate to large values of n, this dependence on the other past states disappears.

The limiting distribution of a Markov chain is a probability distribution that remains unchanged 2 as the chain evolves over time. It represents the long-term behavior of the system. Define $\pi_i^{(n)} = P(X_n = i), i \in$ to be the marginal distribution of $X_n, n = 1,2,...,$ and let $\pi^{(n)}$ be the row vector:

$$\pi^{(n)} = \left(\pi_0^{(ü)}, \pi_1^{(\ )}, \pi_2^{(\ )}, ...\right)$$

Then, mathematically, a Markov chain is said to have a limiting distribution if for all i, j ∈ S:

$\exists \pi_j$ such that;

$$\pi_j = \lim_{n\to\infty} P_{ij}^{(n)} \wedge \pi_j \perp X_0 \wedge \sum_{j\in S} \pi_j = 1$$

i.e.,
$$\lim_{n\to\infty} P^{(n)} = \begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \pi_3 & \cdots \\ \pi_0 & \pi_1 & \pi_2 & \pi_3 & \cdots \\ \pi_0 & \pi_1 & \pi_2 & \pi_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The stationary distribution is a special type of limiting distribution that satisfies the equation:

$$\pi_j = \sum_{j\in S} \pi_j P_{ij} \quad \forall j \in S$$

Indicating that once the system reaches this distribution, it remains there indefinitely. The stationary distribution can be interpreted as the equilibrium state of the Markov chain.

The relationship between the transition matrix, limiting distribution, and stationary distribution is important for understanding the long-term behavior of the stochastic process. The existence of a limiting distribution implies the existence of a unique stationary distribution:

$$\exists \lim_{n\to\infty} P_{ij}^{(n)} \Rightarrow \exists! \pi$$

such that $\pi = \pi P$

**Proof:** Let S be the state space of a Markov chain with transition matrix P. Assume that there exists a limiting distribution, i.e., $\forall i, j \in S \lim_{n\to\infty} P_{ij}^{(n)} = \pi_j$ where, $\pi_j$ is independent of the initial state i and $\Sigma_{j\in S} \pi_j = 1$. We need to show that there exists a unique stationary distribution $\pi = (\pi_j)_{j\in S}$ such that $\pi P = \pi$. Since the limiting distribution exists, we have;

$$\lim_{n\to\infty} P_{ij}^{(n)} = \begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \cdots \\ \pi_0 & \pi_1 & \pi_2 & \cdots \\ \pi_0 & \pi_1 & \pi_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Let $\pi = (\pi_j)_{j\in S}$ be the row vector such that $\pi_j = \lim_{n\to\infty} P_{ij}^{(n)}$. We claim that π is a stationary distribution. To show this, consider the equation $\pi P = \pi$. Hence, we have:

$$(\pi P)_j = \sum_{i\in S} \pi_i P_{ij}$$

.

Since $\lim_{n\to\infty} P_{ij}^{(n)} = \ _j$ and $\Sigma_{i\in S} \pi_j = 1$, we can substitute $\pi_j$ into the equation $\Sigma_{i\in S} \pi_j P_{ij} = \pi_j$.

Next, we show that the stationary distribution is unique. Suppose there exist two stationary distributions π and π′. Then, we have πP = π and π′P = π′. Consider the difference:

$$(\pi - \pi')P = \pi - \pi'.$$

Since both π and π′ are probability distributions, the sum of their components is 1:

$$\sum_{j\in S} (\pi_j - \pi_j') = 0$$

This implies that π = π′, proving the uniqueness of the stationary distribution of the Markov chain. Hence, the existence of a limiting distribution implies the existence of a unique stationary distribution.

Knowing this, notice how the heat maps indicate a pattern, suggesting the existence of a limiting distribution and, by consequence, the existence of the stationary distribution. These patterns are vital in understanding the stochastic behavior of prime gaps and their probabilistic distribution as shown in Figure 1.

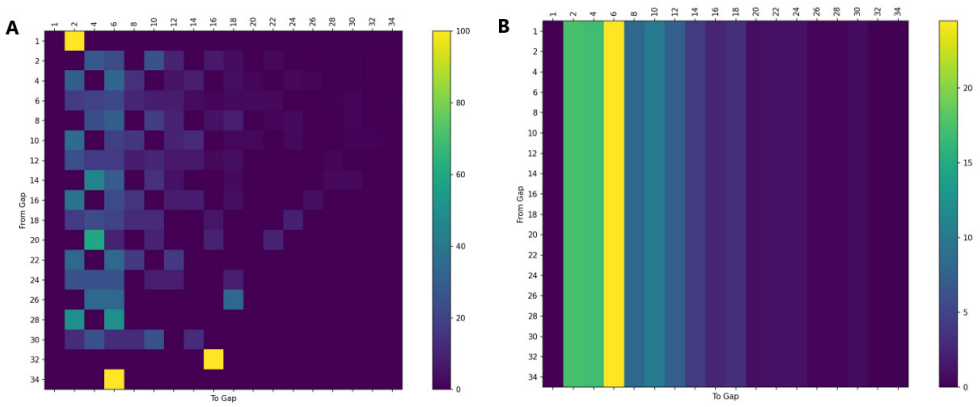


**Figure 1:** (A): Heat map of the transition matrix for the first 1000 prime gaps; (B): Heat map of the transition matrix[50] for the first 1000 prime gaps.

## Stationary distribution

**Lemmas:** These are intermediate propositions used to assist in the proof of a larger theorem or statement. They serve as stepping stones, simplifying complex arguments by breaking them down into more manageable parts, thereby enhancing clarity and rigor in mathematical reasoning [7].

***Lemma 0:*** Let $\{d_n\}_{n=1}^{\infty}; d_n = P_{n+1} - P_n, n \geq 0$ be the sequence of prime gaps. Then, $d_n \leq 6$ infinitely often. (This is also referred to as Tao's Lemma [8]).

**Proof:** Suppose, for the sake of contradiction, that $d_n \leq 6$ for all sufficiently large n. This implies that there exists an integer N such that for all $n \geq N$, $d_n \leq 6$. In other words, there exists a point beyond which every prime gap exceeds 6. Consider the sequence of primes $\{P_n\}$. If $d_n \leq 6$ for all $n \geq N$, this implies that the primes are becoming increasingly sparse as n increases. Specifically, the gaps between consecutive primes are at least 7. However, the prime number theorem tells us that the gaps between consecutive primes $P_n$ are asymptotically approximated by log $(P_n)$. While this allows for large gaps, it does not support the idea that all gaps beyond a certain point are strictly greater than 6. In fact, it is known that there are infinitely many pairs of primes (twinprimes) that are only 2 units apart. This directly contradicts the assumption that $d_n \leq 6$ for all sufficiently large n. Thus, our initial assumption must be false, and it follows that there must be infinitely many n for which $d_n \leq 6$.

***Lemma 1:*** Let $d_n$ be the most recent prime gap and $d_n \neq d_{n-k}$ for k = 1,2,...,(n−1). Then, $d_{n+k} \leq M$ where, $M = \max\{d_j\}$ for j=1,2,...,n+1.

Proof: Suppose, for the sake of contradiction, that $d_{n+1} > M$ for all n. This implies that $d_n < d_{n+1}$ for all n, leading to an unbounded sequence of increasing even numbers. This contradicts Tao's Lemma, which states that gaps less than or equal to 6 occur infinitely often. ⇒⇐ Therefore, our initial assumption must be false, and it follows that $d_{n+1} > M$.

Both lemmas ensure that each new prime gap will connect to previous states.

**Irreducibility:** A Markov chain $\{X_n\}_{n\in N}$ with state space S is said to be irreducible if for any two states i, j ∈ S, there exists a positive integer n such that the n-step transition probability $P_{ij}^n > 0$. Formally, this can be expressed as:

$$\forall i, j \in S, \exists n \in N \text{ such that } P_{ij}^n = P\left(X_n = j \mid X_0 = i\right) > 0$$

Irreducibility implies that it is possible to reach any state from any other state in a finite number of steps.

***Theorem 1:*** Let $\{X_n\}$ be the Markov chain of prime gaps. Then, all states of $\{X_n\}$ communicate with each other. Hence, $\{X_n\}$ is irreducible.

Proof: We need to show that $j_1$=2 communicates with all other states. Through enumeration, we find that $j_1$=2 communicates with {4,6}. Consider the one-step transition matrix P=($P_{ij}$). Let $\{j_k\}$ for k ≥ 4 be the state's such that $P_{2,jk}^{(m)} > 0$. For k=4, $j_4$=8 and $P_{4,8}$=0:

$$P_{2,8}^{(2)} + P_{2,4}P_{4,8} + P_{2,6}P_{6,8} + P_{2,4} \cdot 0 + P_{2,6} \cdot P_{6,8} = P_{2,6} \cdot P_{6,8} \neq 0$$

Thus, m=2. Therefore, it is possible to transition from 2 to 8 in 2 steps. Similarly, since $P_{8,4} \neq 0$ and $P_{ü} \neq 0$:

$$ü_{8,4}^{(2)} = {}_{8,4} \cdot {}_{4,2} \neq 0$$

This indicates that transitioning from 8 to 2 is possible *via* a gap of 4, hence 2↔8. By Lemma 1, 2 communicates with 2k for k ≤ 4. Assuming 2↔2k for all k, we must show that 2↔2 (k+1). By Lemma 1, 2 (k+1) is accessible from at least one smaller prime gap {2k,2 (k − 1),2(k−2)...,2}. Since 2 communicates with all smaller prime gaps, 2 (k+1) is accessible from 2. The prime gap 2k, prior to 2(k+1), communicates with 2, hence 2 is accessible from 2 (k+1). Thus, 2 communicates with 2 (k+1), i.e., 2 ↔ 2 (k+1).

**Aperiodicity:** A state i ∈ S of a Markov chain is said to be aperiodic if the greatest common divisor (gcd) of the

set of return times to i is 1. The chain is aperiodic if all states are aperiodic. Formally, for a state i ∈ S, let:

$$d(i) = gcd\left\{n \geq 1 \mid P_{ii}^n > 0\right\}.$$

The state i is aperiodic if d(i)=1. The entire Markov chain is aperiodic if: ∀i ∈ S, d(i)=1

***Theorem 2:*** The states $\{X_n\}$ are aperiodic.

Proof: Given that $P_{22} \neq 0$, it follows that di=1. Since all states communicate with 2, all states are aperiodic.

**Positive recurrence:** A state i ∈ S is positively recurrent if the expected return time to i is finite. Let $T_i$ be the first return time to state i. The state i is positively recurrent if:

$$E\left[T_i \mid X_0 = i\right] = \sum_{n=1}^{\infty} nP\left(T_i = n \mid X_0 = i\right) < \infty$$

The Markov chain is positively recurrent if all states are positively recurrent.

***Theorem 3:*** All states of $\{X_n\}$ are positive recurrent.

Proof: For a fixed number of states S = $\{j_1, j_2, ..., j_m\}$, Theorem 1 and Theorem 2, show that $\{X_n\}$ is positive recurrent.

We use induction on the number of states m. Let $S = \{j_k\}_{k=1}^{\infty}$ and assume the statement is true for k=m. Suppose $P_{2,jm+1} = 0$. By Theorem 1, there exists a state $j_l$ with 1 ≤ l ≤ m such that $P_{2,jl} \neq 0$ and $P_{jl,jm+1} \neq 0$, indicating that $j_{m+1}$ communicates with 2. Hence, $j_{m+1}$ is positive recurrent.

**Ergodicity:** A Markov chain is said to be Ergodic if it is irreducible, aperiodic, and positively recurrent.

***Theorem 4:*** The Markov chain of prime gaps is ergodic.

Proof: To establish that the Markov chain $\{X_n\}$ of prime gaps is ergodic, we need to verify that it satisfies the conditions of irreducibility, aperiodicity, and positive recurrence.

From Theorem 1, we have shown that all states of $\{X_n\}$ communicate with each other. Hence, the Markov chain is irreducible.

From Theorem 2, it is established that the states $\{X_n\}$ are aperiodic since $P_{22} \neq 0$ and all states communicate with 2. Thus, the chain is aperiodic.

From Theorem 3, all states of $\{X_n\}$ are shown to be positively recurrent. Since $\{X_n\}$ is irreducible, aperiodic, and positively recurrent, it follows that the Markov chain of prime gaps is ergodic. Consequently, the Markov chain $\{X_n\}$ has a unique stationary

distribution $\pi = (\pi_i)_{i \in S}$ satisfying $\pi P = \pi$ and $\sum_{i \in S} \pi_i = 1$.

Furthermore, the distribution of states converges to the stationary distribution, i.e.,

$$\lim_{n \to \infty} \| \mu_n - \pi \| = 0$$

Where $\mu_n$ is the state distribution at time n and $\| \cdot \|$ denotes an appropriate norm such as the total variation norm.

Since the Markov chain of prime gaps is irreducible, aperiodic, and positive recurrent, it is ergodic. Therefore, the Markov chain has a stationary distribution $\pi$ such that, $\pi \quad = \pi$.

## Modular pattern of prime gaps

**Visual pattern:** In this section, we identify a modular pattern in prime gaps. Specifically, we categorize prime gaps based on their congruence modulo 6 to simplyfy the pattern.

The following categorization reveals a significant structure in the distribution of prime gaps. The histogram illustrates the distribution of prime gaps modulo 6, with each category of bins represented in different colors for clarity. This pattern not only supports our theoretical findings but also highlights the regularity and structure inherent in the distribution of prime numbers as shown in Figure 2.

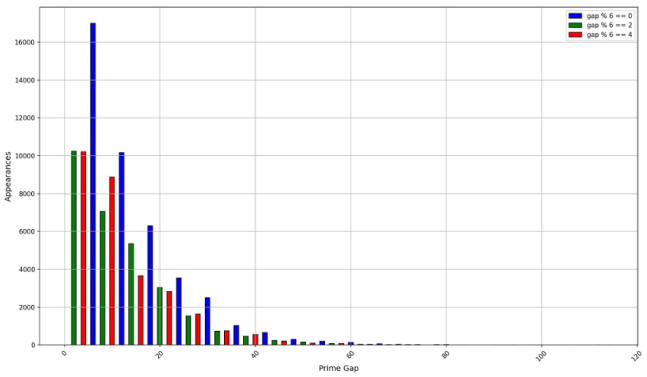


**Figure 2:** Histogram of the first 100,000 prime gaps categorized by their congruence modulo 6. **Note:** The colors indicate different categories: Prime gap ≡ 0 mod 6 (blue); Prime gap ≡ 2 mod 6 (green); Prime gap ≡ 4 mod 6 (red).

**Arithmeticity of prime numbers:**

***Theorem 5:*** Every prime number p>3 can be written in the form 6k ± 1 for some k ∈ Z.

Proof: Given any integer p>3, we know by the division algorithm that there exist unique integers q ≥ 0 and r ∈ {−1,0,1,2,3,4} such that: $p = 6q + r$. Suppose that p is prime. Then observe that $r \notin \{0,2,4\}$, since otherwise p would be even (contradicting the fact that $p \neq 2$) $\Rightarrow\Leftarrow$. Likewise, observe that $r \neq 3$, since otherwise $p = 6q + 3$ would be divisible by 3 (contradicting the fact that $p \neq 3$) $\Rightarrow\Leftarrow$ So, $r = \pm 1$, as desired.

**Prime gaps modulo 6:**

***Theorem 6:*** Approximately 50% of prime gaps are congruent to 0 mod 6 and approximately 25% of prime gaps are congruent to 2 mod 6.

Proof: Given that we proved that prime gaps have a stationary distribution, this implies that the probabilities of prime gaps falling into different congruence classes modulo 6 stabilize over time. Therefore, the long-term proportions of prime gaps that are 0 mod 6, 2 mod 6, and 4 mod 6 will remain constant. And since the primes are equally likely to be 6k+1 or 6k-1, and the stationary distribution reflects these equal probabilities, it follows that:

$$\lim_{n \to \infty} P(\textit{prime gap} \equiv 0 \, mod \ 6) = \frac{1}{2} \wedge \lim_{n \to \infty} P(\textit{prime gap} \equiv 2 \equiv 4 \, mod \ 6) = \frac{1}{4}$$

## Prime factorization of prime gaps

**Prime factorization of even numbers:**

***Theorem 7:*** Every prime gaps greater than 1 can be factorized in at least one of the following six ways: $2^a \cdot q$, $2^a \cdot 3^b \cdot q$, $2^a \cdot 5^b \cdot q$, $2^a \cdot 7^b \cdot q$, $2^a \cdot 11^b \cdot q$, and $2^a \cdot 3^b \cdot 5^c \cdot q$

Where, a, b, and c are non-negative integers, and q is an odd integer.

**Proof:** Let $g_n = p_{n+1} - P_n$ represent the prime gap between two consecutive primes $P_n$ and $p_{n+1}$, where, $g_n > 1$. According to the Fundamental Theorem of Arithmetic, every integer $g_n$ greater than 1 can be factored into primes.

- Factorization into powers of 2: Any integer can be factored into powers of 2. If $g_n$ is divisible by 2, this is captured by 2a.
- Divisibility by small primes: If $g_n$ is divisible by 3, 5, 7, or 11, it can be captured by the inclusion of $3^b$, $5^b$, $7^b$, or $11^b$ in the factorization.
- Odd integer q: Any remaining factor that is not divisible by 2, 3, 5, 7, or 11 is captured by q, which must be an odd integer.

Thus, every prime gap greater than 1 can be factored into one of the six forms by choosing the appropriate powers of 2, 3, 5, 7, 11, and an odd q, completing the proof.

## Prime factorization pattern in prime gaps

Using Theorem 5 and Theorem 6, we will patternize the prime gaps by determining the percentage of prime gaps that fall into each category based on their prime factorization.

***Theorem 8:***

- Approximately 19.96% of prime gaps can be factorized as $2^a \cdot q$
- Approximately 33.33% of prime gaps can can be factorized as $2^a \cdot 3^b \cdot q$
- Approximately 20% of prime gaps can be factorized as $2^a \cdot 5^b \cdot q$
- Approximately 14.29% of prime gaps can be factorized as $2^a \cdot 7^b \cdot q$
- Approximately 9.09% of prime gaps can be factorized as $2^a \cdot 11^b \cdot q$
- Approximately 3.33% of prime gaps can be factorized as $2^a \cdot 3^b \cdot 5^c \cdot q$

Where, a, b, and c are non-negative integers, and q is an odd integer.

Proof: Prime gaps can be seen as even numbers, so they can be analysed by considering their divisibility by small primes (since every prime gap g>1 is even). The proportion of numbers divisible by 2, 3, 5, 7, and 11 among all even numbers can be approximated using the following reasoning:

Prime gaps divisible by 2a · q where q is not divisible by 3, 5, 7, or 11:

$$\text{P (Category 1)} = \prod_{p\in\{3,5,7,11\}}\left(1-\frac{1}{p}\right)\approx\frac{2}{3}.\frac{4}{5}.\frac{6}{7}.\frac{10}{11}\approx 0.1996$$

- Prime gaps divisible by $2^a \cdot 3^b \cdot q$ where q is not divisible by 5, 7, or 11:

$$\text{P (Category 2)} = \frac{1}{3}.\prod_{p\in\{5,7,11\}}\left(1-\frac{1}{p}\right)\approx\frac{1}{3}.\frac{4}{5}.\frac{6}{7}.\frac{10}{11}\approx 0.3333$$

- Prime gaps divisible by $2^a \cdot 5^b \cdot q$ where q is not divisible by 3, 7, or 11:

$$\text{P (Category 3)} = \frac{1}{5}.\prod_{p\in\{3,7,11\}}\left(1-\frac{1}{p}\right)\approx\frac{1}{5}.\frac{2}{3}.\frac{6}{7}.\frac{10}{11}\approx 0.2000$$

- Prime gaps divisible by $2^a \cdot 7^b \cdot q$ where q is not divisible by 3, 5, or 11:

$$\text{P (Category 4)} = \frac{1}{7}.\prod_{p\in\{3,5,11\}}\left(1-\frac{1}{p}\right)\approx\frac{1}{7}.\frac{2}{3}.\frac{4}{5}.\frac{10}{11}\approx 0.1429$$

- Prime gaps divisible by $2^a \cdot 11^b \cdot q$ where q is not divisible by 3, 5, or 7:

$$\text{P (Category 5)} = \frac{1}{11}.\prod_{p\in\{3,5,7\}}\left(1-\frac{1}{p}\right)\approx\frac{1}{11}.\frac{2}{3}.\frac{4}{5}.\frac{6}{7}\approx 0.0909$$

- Prime gaps divisible by $2^a \cdot 3^b \cdot 5^c \cdot q$ where q is not divisible by 7 or 11:

$$\text{P (Category 6)} = \frac{1}{3}.\frac{1}{11}.\prod_{p\in\{7,11\}}\left(1-\frac{1}{p}\right)\approx\frac{1}{3}.\frac{1}{5}.\frac{6}{7}.\frac{10}{11}\approx 0.0333$$

However, it is important to note that these events are not disjoint, i.e. $60 \equiv 0 \pmod{2,6,10,14,22,30}$. Therefore, the probability of the union of all given states is:

$$P\left(A_1 \vee A_2 \vee A_3 \vee A_4 \vee A_5 \vee A_6\right) =$$

$$\sum_i p(A_i) - \sum_{i<j} p(A_i \wedge A_j) + \sum_{i<j<k} p(A_i \wedge A_j \wedge A_k) - \cdots + (-1)^{n+1} P\left(A_1 \wedge A_2 \wedge \cdots \wedge A_6\right)$$

## Approximation of expectation

In this section, we approximate the expectation of prime gaps using a sample of 1 billion prime gaps.

By applying logarithmic polynomial fitting to the data, we estimated the formulas for the expectation and variance of a given set of n consecutive prime gaps.

**Logarithmic polynomial and fitting:** A logarithmic polynomial is a polynomial function in terms of the natural logarithm of a variable. It is generally expressed in the form:

$$P(\ln(n)) = a_0 + a_1 \ln(n) + a_2 \ln(n))^2 + \cdots + a_k \ln(n))^k$$

Where, $a_0, a_1, \ldots, a_k$ are coefficients and k is the degree of the polynomial.

Logarithmic polynomial fitting is a statistical method used to approximate a set of data points by finding the coefficients $a_0, a_1, \ldots, a_k$ that best fit the data in a least-squares sense. The goal is to minimize the sum of the squared differences between the observed values and the values predicted by the polynomial model. Formally, given a set of data points $(n_i, y_i)$ for i=1,2,...,m, the fitting process involves solving the following optimization problem:

$$\min_{a_0, a_1, \ldots, a_k} \sum_{i=1}^{m} (y_i - P(\ln(n_i)))^2$$

Where,

$$P(\ln(n_i)) = a_0 + a_1 \ln(n_i) + a_2 \ln(n_i))^2 + \cdots + a_k \ln(n_i))^k$$

Knowing this, for a given sample of $n=10^{12}$ prime gaps, we approximate

$$E(n) \approx 1.08556674 \cdot \ln(n) + 0.4495518$$

This formula indicates a linear relationship between the expectation of prime gaps and the natural logarithm of n, adjusted by a constant term.

## Probabilistic mass function of prime gaps

**Structure of the PMF of prime gaps:** Cramer's conjecture states that the average gap between consecutive primes near a large number $n$ is approximately $\ln(n)$ [9]. In other words, the gap between the $n^{th}$ prime, pn, and the $(n+1)^{th}$ prime, pn+1, is bounded by:

$$g_n = p_{n+1} - p_n = O(\ln^2(p_n))$$

This approximation is also supported by the prime number theorem, which states that the number of primes less than or equal to $x$ is approximately:

$$\pi(x) \approx \frac{x}{\ln(x)}$$

Consequently, from these two statements, the density of primes around x is given by:

$$g(x) \approx \frac{x}{\pi(x)} \approx \frac{x}{\frac{x}{\ln(x)}} = \ln(x)$$

Hence, the average gap length between consecutive primes near $x$ is then:

$$g(x) \approx \frac{x}{\pi(x)} \approx \frac{x}{\frac{x}{\ln(x)}} = \ln(x)$$

However, in this research study, we are not basing our calculations on a number $x \in R$; instead, we are basing our calculations on $n \in N$, representing the set of the first n consecutive prime gaps. Thus, we need to adapt Cramer's conjecture and the prime number theorem on n: the set of the first n consecutive prime gaps.

Given primes $p_1, p_2, \ldots, p_n$, the average prime gap length is the total gap divided by n−1, the number of gaps.

Let $P_n$ be the n-th prime. The prime number theorem approximates $P_n$ as $p_n \approx n \cdot \ln(n)$. Plus, the total gap length between the first n primes is approximately $p_n - 2$, since the first prime is 2. So, for large n, the average gap $g_n$ between the first n primes is then:

$$g_n \approx \frac{pn-2}{n-1} \approx \frac{n.\ln(n)-2}{n-1} \approx \ln(n)$$

This result aligns with Cramer's conjecture and the Prime Number Theorem, indicating that the average gap between consecutive primes increases logarithmically with the number of primes considered.

Now based on this, we can proceed. Knowing that n represents the number of prime gaps, the simplest model for the gap $g(n) = p_{n+1} - p_n$ is that it follows approximately an exponential distribution of parameter $\frac{1}{\ln(n)}$ Following from that, the probability that $p_{n+1} + 2k'$ (for $k' \in N$) is not prime is about:

$$1 - \frac{2}{\ln(n+2k')} \approx 1 - \frac{2}{\ln(n)} \Rightarrow P(g_n = 2k) \approx \left(1 - \frac{2}{\ln(n)}\right)^{k-1} \cdot \frac{2}{\ln(n)}$$

## Best approximation of the PMF of prime gaps

**Models to be tested:** Based on all of these results, we will analysed compare different models of approximated PMFs for $P(g_n = 2k)$ to see which approximation fits better the real PMF of Prime Gaps:

**Model 1 and 2:**

$$P_{Model}(g_n = 2k) \approx \left(1 - \frac{2}{\mathrm{E}(n)}\right)^{k-1} \cdot \frac{2}{\mathrm{E}(n)}$$

With:

-Model 1: $E(n) \approx \ln(n)$

-Model 2: $E(n) \approx 1.08556674 \cdot \ln(n) + 0.4495518$

**Model 3 and 4:**

$$P_{Model}(g_n = 2k) = \begin{cases} \left(1 - \frac{2}{E(n)}\right)^{k-1} \cdot \frac{2}{E(n)} .0.5 \ \ if\ 2k \equiv 0 (mod\ 6) \\ \left(1 - \frac{2}{E(n)}\right)^{k-1} \cdot \frac{2}{E(n)} .0.25 \ \ if\ 2k \equiv 2 (mod\ 6) \\ \left(1 - \frac{2}{E(n)}\right)^{k-1} \cdot \frac{2}{E(n)} .0.25 \ \ if\ 2k \equiv 4 (mod\ 6) \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases}$$

**With:**

-Model 3: $E(n) \approx \ln(n)$

-Model 4: $E(n) \approx 1.08556674 \cdot \ln(n) + 0.4495518$

**Model 5 and 6:**

$$P_{Model}(g_n = 2k) = \left(1 - \frac{2}{E(n)}\right)^{k-1} \cdot \frac{2}{E(n)} \cdot P, E(n) = \ln(n)$$

With:

-Model 5: $E(n) \approx \ln(n)$

-Model 6: $E(n) \approx 1.08556674 \cdot \ln(n) + 0.4495518$

$\neg(A_1 \uplus A_2 \uplus A_3 \uplus A_4 \uplus A_5 \uplus A_6)$

$$P(A_1) = \begin{cases} 0.1996 \ if\ \exists a \in N, q \in N_{odd}\ such\ that\ g = 2^a \cdot q, \\ 0 \qquad otherwise \end{cases}$$

$$P = P(A_1 \vee A_2 \vee A_3 \vee A_4 \vee A_5 \vee A_6)$$

$$P(A_2) = \begin{cases} 0.3333 \ if\ \exists a,b \in N, q \in N_{odd}\ such\ that\ g = 2^a \cdot 3^b \cdot q, \\ 0 \qquad otherwise \end{cases}$$

$$= \sum_{i=1}^{6} P(A_i) - \sum_{1 \le i < j \le 6} P(A_i \wedge A_j)$$

$$P(A_3) = \begin{cases} 0.2000 \ if\ \exists a,b \in N, q \in N_{odd}\ such\ that\ g = 2^a \cdot 5^b \cdot q \\ 0 \qquad otherwise \end{cases},$$

$$+ \sum_{1 \le i < j < k \le 6} P(A_i \wedge A_j \wedge A_k)$$

$$P(A_4) = \begin{cases} 0.1429 \ if\ \exists a,b \in N, q \in N_{odd}\ such\ that\ g = 2^a \cdot 7^b \cdot q \\ 0 \qquad otherwise \end{cases},$$

$$- \sum_{1 \le i < j < k < l \le 6} P(A_i \wedge A_j \wedge A_k \wedge A_l)$$

$$P(A_5) = \begin{cases} 0.0909 \ if\ \exists a,b \in N, q \in N_{odd}\ such\ that\ g = 2^a \cdot 11^b \cdot q, \\ 0 \qquad otherwise \end{cases}$$

$$+ \sum_{1 \le i < j < k < l < m \le 6} P(A_i \wedge A_j \wedge A_k \wedge A_l \wedge A_m)$$

$$P(A_6) = \begin{cases} 0.0333 \ if\ \exists a,b,c \in N, q \in N_{odd}\ such\ that\ g = 2^a \cdot 3^b \cdot 5^c \cdot q, \\ 0 \qquad otherwise \end{cases}$$

$$- P(A_1 \wedge A_2 \wedge A_3 \wedge A_4 \wedge A_5 \wedge A_6)$$

### Statistical tests

We will compare these 6 different models to determine which best approximates the true Probability Mass Function (PMF) of prime gaps. The following statistical tests will be used on sample sizes of n = $10^3$, $10^6$, $10^9$, and $10^{12}$ prime gaps to assess the impact of sample size on model performance:

- Chi-Square test: Assesses the difference between observed and expected frequencies:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Kolmogorov-Smirnov test: Compares the empirical and reference cumulative distribution functions:

$$D_n = \sup_x \left| F_n(x) - F(x) \right|$$

- Anderson-Darling (AD) test: Measures how well data fits a specified distribution, improving sensitivity at distribution tails:

$$A^2 = -n - S$$

With, S representing a modified sum of differences between empirical and theoretical distributions. However, another common form for a finite sample is:

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x),$$

- Mean Squared Error (MSE): Quantifies the average squared difference between observed and predicted values;

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2$$

- Kullback-Leibler Divergence (KL Divergence): Measures divergence between true and predicted probability distributions;

$$D_{KL}(P \| Q) = \sum_{i}^{\infty} p(i) \log\left( \frac{p(i)}{q(i)} \right)$$

Notice that Non-parametric statistics, such as the K-S and Anderson-Darling tests, are particularly useful when no specific parametric distribution is assumed, allowing for a more flexible evaluation of model fit. These tests, combined with parametric approaches like the Chi-Square Test, provide a robust framework to evaluate how well the models to empirical data and determine which best approximates the prime gap distribution.

## RESULTS

Please note that the graphs below have been adapted up to k=21 to enhance readability and make the results easier to interpret.

**Sample Size:** $n = 10^3$ as shown in Table 1.

**Sample size:** $n = 10^6$ as shown in Table 2.

**Sample size:** $n = 10^9$ as shown in Tables 3.

**Sample size:** $n = 10^{12}$ as shown in Figure 3 and Table 4.

**Table 1:** Combined test results with best results highlighted, $n = 10^3$.

| Model | Chi-square statistic | p-value (Chi-square) | KS statistic | p-value (KS) | AD statistic | MSE value | KL divergencevalue |
|---|---|---|---|---|---|---|---|
| Model 1 | χ2 =0.1734 | 0.9999 | 0.1176 | 0.9999 | 1.0069 | 0.00155 | 0.0805 |
| Model 2 | χ2 =0.1333 | 0.9999 | 0.1176 | 0.9999 | 0.9714 | 0.001094 | 0.0624 |
| Model 3 | χ2 =0.0574 | 1 | 0.1176 | 0.9999 | 1.1548 | 0.000343 | 0.0264 |
| Model 4 | χ2 =0.0406 | 1 | 0.1176 | 0.9999 | 1.0204 | 0.000193 | 0.0197 |
| Model 5 | χ2 =0.0192 | 1 | 0.0588 | 1 | 1.245 | 0.000125 | 0.0098 |
| Model 6 | χ2 =0.0296 | 1 | 0.1765 | 0.9631 | 0.9341 | 0.000135 | 0.016 |

**Note:** K-S: Kolmogorov-Smirnov Test; AD: Anderson-Darling Test; MSE: Mean Squared Error; KL: Kullback-Leibler Divergence

**Table 2:** Combined test results with best results highlighted, $n = 10^6$.

| Model | Chi-square statistic | p-value (chi-square) | KS statistic | p-value (KS) | AD statistic | MSE value | KL divergence value |
|---|---|---|---|---|---|---|---|
| Model 1 | $\chi^2$ = 0.1359 | 1 | 0.1 | 0.9999 | 1.1084 | 0.000533 | 0.0642 |
| Model 2 | $\chi^2$ = 0.1159 | 1 | 0.15 | 0.9831 | 0.9211 | 0.000441 | 0.055 |
| Model 3 | $\chi^2$ = 0.0309 | 1 | 0.15 | 0.9831 | 0.9921 | 0.000105 | 0.0149 |
| Model 4 | $\chi^2$ = 0.0188 | 1 | 0.15 | 0.9831 | 1.0874 | 0.000057 | 0.0089 |
| Model 5 | $\chi^2$ = 0.0098 | 1 | 0.1 | 0.9999 | 1.1088 | 0.000025 | 0.0048 |
| Model 6 | $\chi^2$ = 0.0085 | 1 | 0.1 | 0.9999 | 1.1568 | 0.000021 | 0.0043 |

**Note:** K-S: Kolmogorov-Smirnov Test; AD: Anderson-Darling Test; MSE: Mean Squared Error; KL: Kullback-Leibler Divergence

**Table 3:** Combined test results with best results highlighted, $n = 10^9$.

| Model | Chi-square statistic | p-value | KS statistic | p-value (KS) | AD statistic | MSE value | KL divergence value |
|---|---|---|---|---|---|---|---|
| (chi-square) | KS statistic | p-value (KS) | AD statistic | MSE value | KL divergence value | 0.000533 | 0.0642 |
| Model 1 | $\chi^2$ = 0.1271 | 1.0000 | 0.1000 | 0.9999 | 1.0613 | 0.000387 | 0.0595 |
| Model 2 | $\chi^2$ = 0.1156 | 1.0000 | 0.1500 | 0.9831 | 0.8666 | 0.000346 | 0.0543 |
| Model 3 | $\chi^2$ = 0.0241 | 1.0000 | 0.1500 | 0.9831 | 0.9509 | 0.000070 | 0.0116 |
| Model 4 | $\chi^2$ = 0.0165 | 1.0000 | 0.1000 | 0.9999 | 1.1158 | 0.000044 | 0.0079 |
| Model 5 | $\chi^2$ = 0.0098 | 1.0000 | 0.1000 | 0.9999 | 1.0666 | 0.000021 | 0.0048 |
| Model 6 | $\chi^2$ = 0.0083 | 1.0000 | 0.1000 | 0.9999 | 1.1350 | 0.000019 | 0.0041 |

**Note:** K-S: Kolmogorov-Smirnov Test; AD: Anderson-Darling Test; MSE: Mean Squared Error; KL: Kullback-Leibler Divergence

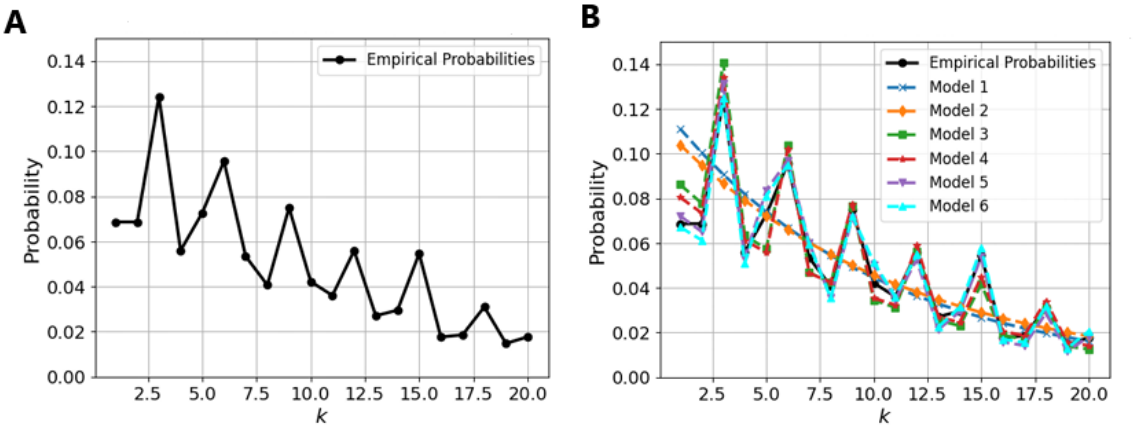


**Figure 3:** (A): Empirical probabilities of prime gaps (sample size: n=$10^{12}$ prime gaps; (B): Comparison of empirical and model predictions of prime gaps (sample size: n=$10^{12}$ prime gaps)

**Table 4:** Combined test results with best results highlighted,

| Model | Chi-square statistic | p-value | KS statistic | p-value (KS) | AD statistic | MSE value | KL divergence value |
|---|---|---|---|---|---|---|---|
| (chi-square) | KS statistic | p-value (KS) | AD statistic | MSE value | KL divergence value | 0.000533 | 0.0642 |
| Model 1 | $\chi^2$ = 0.1134 | 1.0000 | 0.2000 | 0.8319 | 0.7446 | 0.000329 | 0.0533 |
| Model 2 | $\chi^2$ = 0.1105 | 1.0000 | 0.2000 | 0.8319 | 0.6624 | 0.000314 | 0.0521 |
| Model 3 | $\chi^2$ = 0.0152 | 1.0000 | 0.1000 | 0.9999 | 1.1562 | 0.000039 | 0.0073 |
| Model 4 | $\chi^2$ = 0.0145 | 1.0000 | 0.1000 | 0.9999 | 1.1841 | 0.000034 | 0.0070 |
| Model 5 | $\chi^2$ = 0.0130 | 1.0000 | 0.1000 | 0.9999 | 1.2085 | 0.000033 | 0.0065 |
| Model 6 | $\chi^2$ =0.0090 | 1.0000 | 0.1000 | 0.9999 | 1.1658 | 0.000021 | 0.0045 |

**Note:** K-S: Kolmogorov-Smirnov Test; AD: Anderson-Darling Test; MSE: Mean Squared Error; KL: Kullback-Leibler Divergence

## DISCUSSION

### Test specific observations

The models under consideration can be categorized into three distinct groups based on their approach to the prime gaps:

- **Category 1:** Basic models (Models 1 and 2)-These models do not incorporate any specific categorization of prime gaps, relying solely on a fundamental approximation of E(n).
- **Category 2:** Modulo 6 congruence models (Models 3, 4,)-These models classify prime gaps based on their congruence modulo 6, capturing the periodic properties inherent in prime distributions.
- **Category 3:** Prime factorization models (Models 5 and 6)-These models categorize prime gaps based on their prime factorization, leveraging the deeper arithmetic structure of the gaps.

While all three categories exhibit commendable performance, effectively approximating the PMF of prime gaps to a significant extent, notable distinctions emerge across different statistical tests. For all tests, excluding Anderson-Darling (AD) test, the category 3 models, which incorporate prime factorization, consistently demonstrate the highest accuracy, followed by category 2 models that consider modulo 6 congruence. The basic models in category 1 rank the lowest. This trend underscores the superiority of models that account for more intricate arithmetic structures.

Interestingly, the AD test results deviate from the aforementioned trends. Category 1 models attain the highest ranking in this test. This anomalous result can be attributed to the AD test's sensitivity to the tails of the distribution. The basic models' simplicity and broader assumptions may lead to a better fit for the extreme values of the prime gap distribution, explaining their superior performance in this specific test.

**Approximation of the expectation of prime gaps:** The expectation of prime gaps, $E(n)$, is an essential parameter in accurately modeling the distribution of gaps between consecutive prime numbers. It is generally approximated by the natural logarithm of n, i.e., $E(n) \approx \ln(n)$. This approximation is widely accepted in number theory due to its basis in the prime number theorem, which states that the average gap between consecutive primes near a large number n is approximately $\ln(n)$. However, our analysis shows that more refined approximations of $E(n)$ can significantly improve the accuracy of models, $E(n) \approx 1.08556674 \cdot \ln(n) + 0.4495518$

In summary, while $E(n) \approx \ln(n)$ is a widely known and useful approximation, incorporating refined estimates of $E(n)$ can lead to more accurate models of the distribution of prime gaps. These refinements account for empirical observations and theoretical insights that extend beyond the basic logarithmic approximation.

**Geometric structure:** All the models considered in this study follow a geometric Probability Mass Function (PMF) structure. This choice is based on the assumption that the distribution of prime gaps can be effectively approximated using a geometric distribution. The general form of the geometric PMF used in these models is:

$$P(g_n = 2k) \approx \left(1 - \frac{2}{E(n)}\right)^{k-1} \cdot \frac{2}{E(n)}$$

Where, $E(n)$ represents the expected value, which varies across different models. This structure implies that the probability of observing a gap of size 2k between consecutive primes decreases exponentially with k, consistent with the properties of geometric distributions.

The geometric PMF structure captures the idea that larger prime gaps are less likely than smaller ones, a characteristic feature of prime gaps. In other words, the geometric PMF inherently incorporates an exponential decay, reflecting the decreasing likelihood of larger prime gaps, which aligns well with empirical observations.

**Piecewise approximations:** The comparative analysis reveals that piecewise approximations, which categorize prime gaps based on modulo 6 congruence or their prime factorization, consistently outperform non piecewise approximations. This observation suggests that the PMF of prime gaps exhibits a combinatorial nature influenced by specific sub-properties intrinsic to the prime numbers themselves:

$$P_{\mathrm{mod}6}(g_n = 2k) = \begin{cases} \left(1 - \frac{2}{E_0(n)}\right)^{k-1} \cdot \frac{2}{E_0(n)} \cdot 0.5 & if\ 2k \equiv 0 (mod 6) \\ \left(1 - \frac{2}{E_2(n)}\right)^{k-1} \cdot \frac{2}{E_2(n)} \cdot 0.25 & if\ 2k \equiv 2 (mod 6) \\ \left(1 - \frac{2}{E_4(n)}\right)^{k-1} \cdot \frac{2}{E_4(n)} \cdot 0.25 & if\ 2k \equiv 4 (mod 6) \\ 0 & otherwise \end{cases}$$

and

$$P_{PrimeFactorization}(gn = 2k) = \left(1 - \frac{2}{E(n)}\right)^{k-1} \cdot \frac{2}{E(n)} . P$$

Where,

$\neg(A_1 \uplus A_2 \uplus A_3 \uplus A_4 \uplus A_5 \uplus A_6)$

$$P = P(A_1 \vee A_2 \vee A_3 \vee A_4 \vee A_5 \vee A_6)$$

$$P(A_1) = \begin{cases} 0.1996 \text{ if } \exists a \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$P(A_2) = \begin{cases} 0.3333 \text{ if } \exists a,b \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 3^b \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$P(A_3) = \begin{cases} 0.2000 \text{ if } \exists a,b \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 5^b \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$P(A_4) = \begin{cases} 0.1429 \text{ if } \exists a,b \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 7^b \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$P(A_5) = \begin{cases} 0.0909 \text{ if } \exists a, b \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 11^b \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$P(A_6) = \begin{cases} 0..0333 \text{ if } \exists a,b,c \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 3^b \cdot 5^c \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

The superior performance of piecewise models highlights the necessity of incorporating de tailed arithmetic properties and combinatorial classifications when approximating the PMF of prime gaps. This approach not only provides a more accurate fit to empirical data but also aligns with the intrinsic mathematical complexities of prime distributions.

**Best approximation:** The overall best approximation for the PMF of prime gaps is provided by Model 6. This model leverages a sophisticated combinatorial approach, integrating the arithmetic proper ties of prime numbers and utilizing both prime factorization and refined expectation.

- **Model 6:**

$$P_{Model6}(g_n = 2k) = \left(1 - \frac{2}{E(n)}\right)^{k-1} \cdot \frac{2}{E(n)} \cdot p$$

for

$$E(n) \approx 1.08556674 \cdot \ln(n) + 0.4495518$$

¬(A1⊎ A2 ⊎ A3 ⊎ A4 ⊎ A5 ⊎ A6)

$$P(A_1) = \begin{cases} 0.1996 \text{ if } \exists a \in N, q \in N_{odd} \text{ such that } g = 2a \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$P = P(A_1 \vee A_2 \vee A_3 \vee A_4 \vee A_5 \vee A_6)$$

$$P(A_2) = \begin{cases} 0.3333 \text{ if } \exists a,b \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 3^b \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$= \sum_{i=1}^{6} P(A_i) - \sum_{1 \le i < j \le 6} P(A_i \wedge A_j)$$

$$P(A_3) = \begin{cases} 0.2000 \text{ if } \exists a,b \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 5^b \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$+ \sum_{1 \le i < j < k \le 6} P(A_i \wedge A_j \wedge A_k)$$

$$P(A_4) = \begin{cases} 0.1429 \text{ if } \exists a,b \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 7^b \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$- \sum_{1 \le i < j < k < l \le 6} P(A_i \wedge A_j \wedge A_k \wedge A_l)$$

$$P(A_5) = \begin{cases} 0.0909 \text{ if } \exists a,b \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 11^b \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$+ \sum_{1 \le i < j < k < l < m \le 6} P(A_i \wedge A_j \wedge A_k \wedge A_l \wedge A_m)$$

$$P(A_6) = \begin{cases} 0.0333 \text{ if } \exists a,b,c \in N, q \in N_{odd} \text{ such that } g = 2^a \cdot 3^b \cdot 5^c \cdot q, \\ 0 \quad \text{otherwise} \end{cases}$$

$$-P(A_1 \wedge A_2 \wedge A_3 \wedge A_4 \wedge A_5 \wedge A_6)$$

The superior performance of Model 6 implies that the best approximation for the PMF of prime gaps must be a combinatorial PMF that leverages the arithmetic properties of prime numbers. Specifically, it suggests that modular behaviors of the gaps, as well as the prime factorization of even numbers, are important in accurately modeling prime gaps. The intricate structure captured by model 6, which combines these properties, provides a more precise fit for the empirical distribution of prime gaps as shown in Figure 4.

Moreover, the combinatorial nature of this approximation can also influence $E(n)$, indicating that the formula for

the average gap length must be more complex than just $E(n) \approx \ln(n)$. The refined function

$E(n) \approx 1.08556674 \cdot \ln(n) + 0.4495518$ used in model 6 underscores this complexity, suggesting that a more sophisticated understanding of prime gaps involves a detailed analysis of their arithmetic properties and combinatorial classifications.

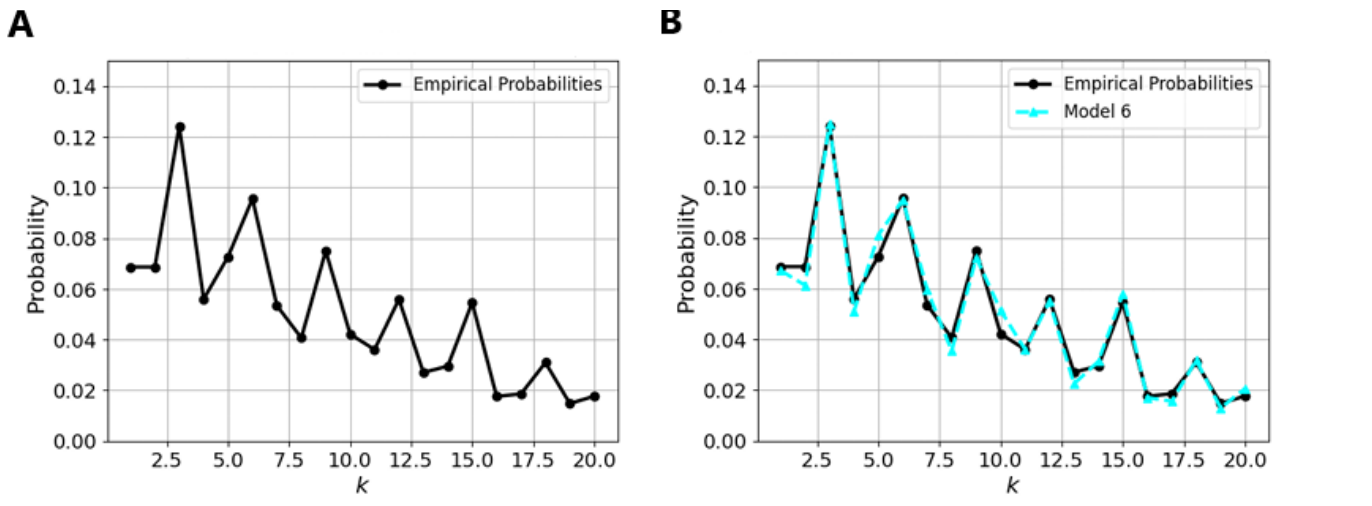


**Figure 4:** (A): Empirical probabilities of prime gaps (sample size: n=$10^{12}$ prime gaps); (B) Comparison of empirical and model 6 (sample size: n = $10^{12}$ prime gaps)

## Results-based conjecture

**Hypotheses:** Based on the empirical results obtained and the comprehensive investigation conducted, we hypothesize that the Probability Mass Function (PMF) of prime gaps exhibits a uniform distribution for gaps of size 2 and 4, while for larger gaps, it follows a more complex structure. Specifically, we conjecture that the PMF of prime gaps from 6 onward is best described by an infinite combination of geometric distributions, each corresponding to a distinct prime factorization of even numbers. This framework suggests a deep connection between the arithmetic properties of integers and the distribution of prime gaps, providing a novel perspective on their underlying structure.

**Conjectures:**

***Conjecture 0:*** The percentage of prime gaps of length 2 and length 4 among all prime gaps is approximately 5% for both gaps respectively. I.e. for π being the stationary distribution of prime gaps, then $\pi_2 = \pi_4$

≈ 0.05.

Proof: First we will prove that as n approaches infinity, the percentage of prime gaps of lengths 2 and 4 asymptotically converges to the same value. We'll combine the provided model for prime gaps, the given probabilities $P(A_i)$, and theorems related to the distribution of prime gaps.

Theorem 4, indicates that the Markov chain made of prime gaps has a stationary distribution, implying that the proportion of different prime gaps stabilizes as n increases. And based on the outstanding performance of model 6 in approximating the probabilistic distribution of prime gaps, we will use it to compute the probabilities of specific

prime gaps:

- **Model 6:**

$$P_{Model\,6}\left(g_n = 2k\right) = \left(1 - \frac{2}{E(n)}\right)^{k-1} \cdot \frac{2}{E(n)} \cdot p$$

for

$$E(n) \approx 1.08556674 \cdot \ln(n) + 0.4495518$$

$\neg(A_1 \uplus A_2 \uplus A_3 \uplus A_4 \uplus A_5 \uplus A_6)$

$$P(A_1) = \begin{cases} 0.1996 \text{ if } \exists a \in N,\ q \in N_{odd} \text{ such that } g = 2a \cdot q, \\ 0 \qquad \text{otherwise} \end{cases}$$

$$P = P\left(A_1 \vee A_2 \vee A_3 \vee A_4 \vee A_5 \vee A_6\right)$$

$$P(A_2) = \begin{cases} 0.3333 \text{ if } \exists a,b \in N,\ q \in N_{odd} \text{ such that } g = 2^a \cdot 3^b \cdot q, \\ 0 \qquad \text{otherwise} \end{cases}$$

$$= \sum_{i=1}^{6} P(A_i) - \sum_{1 \le i < j \le 6} P\left(A_i \wedge A_j\right)$$

$$P(A_3) = \begin{cases} 0.2000 \text{ if } \exists a,b \in N,\ q \in N_{odd} \text{ such that } g = 2^a \cdot 5^b \cdot q, \\ 0 \qquad \text{otherwise} \end{cases}$$

$$+ \sum_{1 \le i < j < k \le 6} P\left(A_i \wedge A_j \wedge A_k\right)$$

$$P(A_4) = \begin{cases} 0.1429 \text{ if } \exists a,b \in N,\ q \in N_{odd} \text{ such that } g = 2^a \cdot 7^b \cdot q, \\ 0 \qquad \text{otherwise} \end{cases}$$

$$- \sum_{1 \le i < j < k < l \le 6} P\left(A_i \wedge A_j \wedge A_k \wedge A_l\right)$$

$$P(A_5) = \begin{cases} 0.0909 \text{ if } \exists a,b \in N,\ q \in N_{odd} \text{ such that } g = 2^a \cdot 11^b \cdot q, \\ 0 \qquad \text{otherwise} \end{cases}$$

$$+ \sum_{1 \le i < j < k < l < m \le 6} P\left(A_i \wedge A_j \wedge A_k \wedge A_l \wedge A_m\right)$$

$$P(A_6) = \begin{cases} 0.0333 \text{ if } \exists a,b,c \in N,\ q \in N_{odd} \text{ such that } g = 2^a \cdot 3^b \cdot 5^c \cdot q, \\ 0 \qquad \text{otherwise} \end{cases}$$

$$-P\left(A_1 \wedge A_2 \wedge A_3 \wedge A_4 \wedge A_5 \wedge A_6\right)$$

Using the approximated PMF and $E(n)$,the ratio of the probabilities of gaps 4 to 2 is:

$$\frac{P_{Model\,6}(g_n = 4)}{P_{Model\,6}(g_n = 2)} = \frac{\frac{2}{E(n)} \cdot P}{\left(1 - \frac{2}{E(n)}\right) \cdot \frac{2}{E(n)} \cdot P} = \left(1 - \frac{2}{E(n)}\right)$$

As $n \to \infty$:

$$\frac{2}{E(n)} \to 0 \Rightarrow \left(1 - \frac{2}{E(n)}\right) \to 1 \Rightarrow \frac{P_{Model\,6}(g_n = 4)}{P_{Model\,6}(g_n = 2)} \to 1$$

Given the existence of a stationary distribution of the Markov chain made of prime gaps, the probabilities for gaps

of lengths 2 and 4 converge to the same asymptotic value. Let the prime counting function π be our stationary distribution:

$$\exists \pi_{pn+1-pn} \wedge \lim_{n \to \infty} P_{Model7}\left(g_n = 2\right) = \lim_{n \to \infty} P_{Model7}\left(g_n = 4\right) \to \pi_2 = \pi_4$$

The twin prime conjecture states that there are infinitely many pairs of primes $(p, p+2)$where the gap between the primes is 2. Although this conjecture has not yet been proven, substantial evidence supports its validity.

Empirically, it is observed that prime gaps of length 2 are quite frequent among smaller primes. However, as numbers get larger, the distribution of prime gaps becomes more varied. Despite this, twin primes continue to appear even among large numbers, though less frequently.

Let $\pi_2(x)$ be the number of twin primes less than or equal to x, and $\pi(x)$ be the total number of primes less than

or equal to $x$. The proportion of twin primes relative to all primes can be estimated using:

$$Proportion\ of\ twin\ primes = \lim_{x \to \infty} \frac{\pi 2(x)}{\pi(x)}$$

The Hardy-Little wood conjecture for the density of twin primes suggests that

$$\pi_2(x) \sim 2C_2 \int_2^x \frac{dt}{(\log t)^2}$$

Where, $C_2$ is the twin prime constant, approximately 0.660161.

Using the prime number theorem:

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

While an exact percentage is difficult to calculate without exhaustive data, we can use empirical and asymptotic

results to estimate the percentage of prime gaps that are length 2.

According to these estimates, the proportion of twin primes (gaps of 2) tends to:

$$\frac{\pi_2(x)}{\pi(x)} \approx \frac{2C_2\int_2^x \frac{dt}{(logt)^2}}{\frac{x}{log\, x}}$$

For practical purposes, over large ranges of x, this ratio typically falls in the range of 5%-6%. Empirical studies and computational data on prime numbers suggest that the percentage of prime gaps that are exactly 2 among all prime gaps is approximately 5% for large ranges of primes. After having proved that states that he percentage of prime gaps of lengths 2 and 4 asymptotically converges to the same value, we conclude:

Percentage of prime gaps of length 2=Percentage of prime gaps of length 4 ≈ 5% to 6%

**Conjecture 1:** The Probability Mass Function (PMF) of prime gaps for gaps of size 6 and greater follows an infinite combination of geometric distributions determined by the prime factorization of even numbers. More precisely, the Probability Mass Function (PMF) of prime gaps is uniform for gaps of size 2 and 4, and for gaps of size 6 and greater, it follows an infinite combination of geometric distributions determined by the prime factorization of even numbers.

Based on our results, hypothesize that the distribution of prime gaps, denoted by the Probability Mass Function (PMF) $P(g_n)$, can be expressed as an infinite sum of weighted geometric distributions. Each geometric distribution corresponds to a specific category, defined by the prime factorization of the even number representing the prime gap. The more categories we include in our model, the more precise our approximation of the prime gap distribution becomes. In the limit, as the number of categories approaches infinity, the model converges to the exact distribution of prime gaps.

Let C represent the set of all possible prime factorizations of even numbers, and let Ci denote a specific factorization category. Then, the PMF $P(g_n = 2k)$ can be expressed as:

$$P(g_n = 2k) = \sum_{C_i \in C} w_i . P_i(g_n = 2k)$$

- $P_i(g_n = 2k)$ is the geometric PMF for the i-th factorization category Ci.
- $w_i$ is the weight assigned to the i-th category, representing the proportion of prime gaps that fall into this category.
- The sum extends over all possible categories $C_i$, corresponding to all possible prime factorizations of even numbers.

As the number of categories in C increases, the sum becomes more inclusive of the different possible factorizations of prime gaps. In the limit, as the number of categories approaches infinity, the model becomes fully precise, capturing the exact distribution of prime gaps:

$$\lim_{|c|\to\infty} P(g_n = 2k) = \sum_{C_i \in C} w_i . P_i(g_n = 2k)$$

This model suggests that prime gaps are deeply connected to the arithmetic structure of integers. Their distribution can be understood as a complex, yet structured, combination of simpler geometric distributions. The more we account for the different prime factorizations, the more accurately we model the true distribution of prime gaps.

Thus, we conjecture:

$$P(g_n = 2k) = \begin{cases} \frac{1}{2} & if\ k = 1 \vee 2, \\ \sum_{ci \in c} w_i . P_i(g_n = 2k) & if\ k \geq 3, \end{cases}$$

for:

$$P_i(g_n = 2k) = \left(1 - \frac{2}{E(n)}\right)^{k-1} . \frac{2}{E(n)}$$

Where $E(n)$ is the expectation of the prime gaps, and $w_i$ is the weight associated with the i-th prime factorization category $C_i$.

## CONCLUSION

This study has provided significant insights into the distribution of prime gaps, emphasizing the importance of models that incorporate arithmetic and combinatorial properties. We categorized models into three groups: Basic, modulo 6 congruence, and prime factorization models. Among these, prime factorization models consistently yielded the most accurate predictions, particularly for larger prime gaps, high lighting the intricate arithmetic structures that influence prime gaps.

A critical advancement in this research is the refinement of the expectation of prime gaps, $E(n)$. The refined function, $E(n) \approx 1.08556674 \cdot \ln(n) + 0.4495518$, offers a more precise estimate than the traditional logarithmic approximation, improving model accuracy, especially when combined with a piece wise approach.

The geometric Probability Mass Function (PMF) structure effectively captures the exponential decay of prime gaps, with larger gaps becoming increasingly rare. This study led to conjecture 1, which posits that while the PMF of prime gaps is uniform for gaps of size 2 and 4, it follows an infinite combination of geometric distributions for gaps of size 6 and greater, determined by prime factorization. This conjecture highlights the dual nature of prime gaps-simple and uniform for small gaps, but complex and structured for larger ones.

Prime factorization serves as a critical factor in understanding these larger gaps, suggesting that the distribution of prime gaps is not random but deeply connected to the arithmetic properties of integers. The more complex the factorization, the more intricate the gap's distribution becomes. In conclusion, conjecture 1 offers a unifying framework that links small, uniform gaps with larger, more complex gaps through the lens of prime factorization. Future research should focus on testing and refining this conjecture, potentially uncovering deeper connections between prime numbers and their gaps.

## REFERENCES

1. Euclid, Elements. Clark Univers. 1996.
2. Hardy GH, Littlewood JE. Some problems of 'Partitio numerorum'; III: On the expression of a number as a sum of primes. Acta Math. 1923;44(1):1-70.
3. Erdos P. On the difference of consecutive primes. Q J Math. 1935;6(1):124-128.
4. Goldston DA, Motohashi Y, Pintz J, Yıldırım CY. Small gaps between primes exist. Proc Japan Acad Ser. A Math. Sci. 82(4): 61-65
5. Zhang Y. Bounded gaps between primes. Ann Math. 2014;179:1121-1174.

6. Hadamard J. On the distribution of zeros of the function $\zeta(s)$ and its arithmetic consequences. Bull Soc Math Fr. 1896; 24:199-220.
7. Abdullah D, Rahim R, Apdilah D, Efendi S, Tulus T, Suwilo S. Prime numbers comparison using sieve of eratosthenes and sieve of sundaram algorithm. J Phys Conf Ser. 2018;978:012123.
8. Tao T. Every odd number greater than 1 is the sum of at most five primes. Math Comp. 2014;83(286):997-1038.
9. Cramér H. On the order of magnitude of the difference between consecutive prime numbers. Acta Arith. 1936;2: 23-46.