A Symbolic Grammar for Normalized Prime Gaps: An Empirical Markov Model

Michel Monfette mycmon@gmail.com (Quebec, Canada)

January 2026

Abstract

This short note presents an empirical structure observed in normalized prime gaps. After normalizing gaps by $\log(p)$, the resulting values cluster into stable ranges that can be discretized into a small symbolic alphabet. The frequencies of bigrams and trigrams in this symbolic sequence remain stable across large intervals. A simple first-order Markov model is constructed from these motifs. Combined with elementary arithmetic filters (mod 30 residues and removal of small prime multiples), the model generates simulated values that match actual prime gaps significantly better than random baselines. This work does not propose a theory of primes; it simply documents an empirical structure that may be of independent interest.

1 Introduction

Prime gaps have been studied extensively from analytic and probabilistic perspectives. This work explores them from a different angle: symbolic discretization. The idea originated from a geometric intuition: primes greater than 5 occupy only eight residue classes modulo 30. This suggested that normalized prime gaps might exhibit stable symbolic patterns.

The goal of this note is to describe this empirical structure and share it with the mathematical community.

2 Normalization of Prime Gaps

For consecutive primes p_n and p_{n+1}, define the gap:

$$\Delta_n = p_{n+1} - p_n.$$

Normalize it using:

$$g_n = \frac{\Delta_n}{\log(p_n)}.$$

This normalization makes gaps comparable across large intervals. The values g_n tend to cluster into a small number of ranges.

3 Symbolic Grammar

The normalized values are discretized into four symbols:

- a: very small gap
- b: small gap
- c: typical gap
- d: large gap

Across multiple intervals (1–2000, 20k–50k, 1M–2M), the most frequent motifs are:

Bigrams

$$bb,\ ab,\ ba,\ bc,\ cb,\ aa,\ ca$$

Trigrams

$$bbb,\ bba,\ bab,\ abb,\ bcb,\ bbc$$

Tetragrams

$$bbbb,\ bbab,\ bcbb,\ bbba,\ cbba,\ abbb$$

These motifs remain stable across scales, suggesting a symbolic structure.

4 Markov Model

A first-order Markov chain is built from bigram frequencies:

$$P(s_{n+1} \mid s_n).$$

Each symbol corresponds to a typical normalized gap value. Simulated sequences of symbols can be converted back into numerical gaps.

5 Arithmetic Filters

Two simple filters improve realism:

Allowed residues modulo 30

Only the residues

$$1,\ 7,\ 11,\ 13,\ 17,\ 19,\ 23,\ 29$$

are allowed.

Small-prime filter

Numbers divisible by

$$7,\ 11,\ 13,\ 17,\ 19,\ 23,\ 29$$

are rejected.

These filters remove obvious composites.

6 Numerical Observations

In the interval 1–2 million, with 100 simulated values:

- 30–41% of simulated values are actual primes
- true density: 7.24%
- performance: approximately 4.1× random

7 Comparison with Granville

Granville's model predicts:

$$P(n) \approx \frac{1.2}{\ln(n)} \approx 7\%.$$

Model	Result
Symbolic grammar	30–41%
Granville	$\approx 7\%$
Random baseline	$\approx 7\%$

The symbolic model captures local motifs that Granville's global model does not attempt to describe.

8 Conclusion

This work does not propose a theory of prime numbers. It documents an empirical symbolic structure that appears stable across large intervals. The goal is simply to share this observation with the mathematical community so that the idea is not lost and may inspire further exploration.