Correlations and symbol entropy of the distances between consecutive prime numbers and their increments

Pradeep Kumar, Plamen Ch. Ivanov, H. Eugene Stanley

Center for Polymer Studies and Department of Physics,
Boston University, Boston, MA 02215

Abstract

The difference between two consecutive prime numbers is called the distance between primes. We study the statistical properties of the distances, and also of their increments (the difference between two consecutive distances). Using a sequence comprising the first $N_p = 5 \times 10^7$ prime numbers, we obtain an empirical form for the symbol entropy for the set of distances as a function of sequence length $N_p$. We find a similar form for the symbol entropy for the sequence of distances as a function of sequence length $N_p$. We find a similar form for the symbol entropy for the sequence of corresponding increments. For a given $N_p$, we show that the entropy of the sequence of increments is always greater than the entropy of the sequence of distances, consistent with a greater variability in the values of the increments. We find that the histogram of the increments follows an exponential distribution, with superposed periodic behavior with period-three oscillations. We observe that the magnitude of these increments follows a logarithmic trend as a function of $N_p$, similar to the trend previously known for distances. We also investigate the correlations between both distances and increments; we find that at small and intermediate scales the distances exhibit a weak power-law anticorrelation followed by a crossover to strongly correlated behavior at large scales. This crossover is due to the logarithmic trend in the sequence of distances. For increments, we find that (i) the sequence of increments is strongly anticorrelated at all scales, (ii) due to the logarithmic trend, their magnitudes exhibit a crossover from weakly anticorrelated behavior at small scales to strongly correlated behavior at large scales, and (iii) their signs are anticorrelated at small scales and uncorrelated at large scales.

PACS numbers:

Keywords:
I. INTRODUCTION

Many physical and biological systems exhibit patterns where prime numbers play an important role. Examples range from the periodic orbits of a system in quantum chaos and the life cycles of species [1–7] to a potential for which the quantum energy levels of a particle can be mapped onto the sequence of primes [8]. A gas of independent bosons with energies equal to the logarithm of consecutive primes possesses a canonical partition function coinciding with the Riemann function [9]. The partition function of a system with energies equal to the distances between two consecutive prime numbers behaves like a sequence of non-interacting harmonic oscillators [10]. A power-law behavior in the distribution of primes and correlations in prime numbers has been found [11], along with multifractal features in the distances between consecutive primes [12]. Here, we focus on the statistical properties of the distances $\delta_i$ between consecutive prime numbers and the increments $\Delta_i$ in these distances.

The outline of this paper is as follows. In Sec. II we present an empirical formula for the information entropy of the sequence of distances (defined to be the difference between consecutive prime numbers). We also compare the entropy of the sequence of distances with the entropy of the sequence of increments (defined to be the difference between two consecutive distances). We find empirically that both the entropy of the sequence of distances and the entropy of the sequence of increments vary as $\log(\log N_p)$.

In Sec. III we study the distribution properties of the increments. We find a consistent regular pattern of period-three oscillation in the occurrence frequencies of the increments. In Sec. IV we investigate the correlation properties of distances as well as their increments, and we find a weak anticorrelated behavior for the distances and strong anticorrelated behavior for the increments. We find pronounced crossovers in the scaling of the distances and the magnitude of increments, and attribute these crossovers to logarithmic trends in the data. In Sec. V we define a one-dimensional walk on the sign of increments, where the walker takes a positive step when the increment is positive and a negative step when the increment is negative. We find that such a walk is uncorrelated for any value of $N_p$. 
II. SYMBOL ENTROPIES FOR DISTANCES BETWEEN CONSECUTIVE PRIMES AND THEIR INCREMENTS

In Fig. 1 we show the first $5 \times 10^4$ distances $\delta_i$ and their corresponding increments $\Delta_i$. We first ask whether there are any specific patterns in the values of the distances and their increments. To this end, we study the symbol entropies of the sequence of distances and increments, and ask how they depend on $N_p$.

The symbol entropy of the sequence of the distances is defined as [13]

$$S_\delta(N_p) \equiv - \sum_i p(\delta_i) \log p(\delta_i). \quad (1)$$

Here $\delta_i$ is the sequence of distances and

$$p(\delta_i) \equiv \frac{f(\delta_i)}{N_\delta}, \quad (2)$$

where $f(\delta_i)$ is the occurrence frequency of the distance $\delta_i$, and $N_\delta$ is the number of consecutive prime pairs, i.e.,

$$N_\delta \equiv \sum_i f(\delta_i) = N_p - 1. \quad (3)$$

The symbol entropy of a sequence of discrete symbols or states provides a measure of the diversity (or the information content) of the sequence [13]. The greater the diversity of the symbols in a sequence, the greater the entropy, while a sequence with regular patterns has a small value for the entropy.

In Fig. 2(a) we show the symbol entropy $S_\delta$ for the sequence of distances $\delta_i$ as a function of $N_p$. For small $N_p$, we find that $S_\delta$ increases rapidly. For large $N_p$, we find that the rate of increase becomes very small. For finite $N_p$, the symbol entropy for the sequence of distances can be well fit by the form

$$S_\delta(N_p) = A_\delta \log(\log N_p) + B_\delta. \quad (4)$$

We repeat the above analysis for the sequence of increments $\Delta_i$, and define the analogs of Eqs. (1)–(3) for

$$S_\Delta(N_p) \equiv - \sum_i p(\Delta_i) \log p(\Delta_i). \quad (5)$$

We find a functional form similar to Eq. (4) [Fig. 2(b)] for $S_\Delta$. Further, we find $S_\Delta > S_\delta$, suggesting that the sequence of increments displays a greater variability than the sequence of distances.
III. DISTRIBUTION OF INCREMENTS

While the occurrence frequency $f(\delta_i)$ of distances has been studied [10, 14–16], the occurrence frequency $f(\Delta_i)$ of increments $\Delta_i$ between consecutive distances has not. We find that $f(\Delta_i)$ [Fig. 3(a)] exhibits large peaks for given values of the increments, and exhibits medium and small peaks for other values, and that these peaks display period-three oscillation. Specifically, we find that the increments with values of $6k + 2$ ($k = 0, 1, 2, 3, ...$) have the highest occurrence frequency, followed by increments with values of $6k + 4$. Values of $6k$ are relatively rare and correspond to the small peaks in the distribution. This regularity is present for both positive and negative increments and does not depend on the length $N_p$ of the sequence.

We also find that $f(\Delta_i)$ decreases exponentially and that this exponential behavior is pronounced for both large and small peaks, forming a “double-tent” shape [Fig. 3(b)]. The exponential distribution has superposed periodic behavior with period-three oscillation. An exponential distribution with period-six oscillation was previously found for $f(\delta_i)$ [10]. Remarkably, we find that for a given sequence length $N_p$ the occurrence frequency of a positive increment is almost the same as the occurrence frequency of its negative counterpart [Fig. 3(c)] suggesting a symmetric form of the distribution $f(\Delta_i)$.

IV. CORRELATIONS IN DISTANCES BETWEEN CONSECUTIVE PRIME NUMBERS AND THEIR INCREMENTS

To study the organization in the sequence of distances $\delta_i$ and increments $\Delta_i$ along the sequence of prime numbers, we investigate the correlations and scaling properties of these sequences. A broad range of physical, biological, and social systems, as well as various mathematical objects, have no characteristic length scale and exhibit fractal scaling properties characterized by long-range power-law correlations [17–23]. The outputs of such systems are often inhomogeneous and nonstationary. Inhomogeneous patterns and complex fluctuations are also observed in the sequence of distances between consecutive prime numbers and their increments [Fig. 1].

Traditional approaches such as the power spectrum and Hurst analysis are not well suited for nonstationary fluctuating signals with embedded polynomial trends. Detrended fluctua-
tion analysis (DFA) was developed to accurately quantify long-range power-law correlations, and the performance of DFA for signals possessing different types of nonstationarity has been studied [24–29]. DFA compares favorably to other existing methods [29, 30], and has been applied in a wide range of fields [31–42].

Briefly, the DFA method involves the following steps:

1. Consider a series \( u(i) \) where \( i = 1, \ldots, N \) and \( N \) is the length of the series. We first integrate the series \( u(i) \) and subtract the mean \( \langle u \rangle \)

\[
y(k) = \sum_{i=1}^{k} [u(i) - \langle u \rangle],
\]

where

\[
\langle u \rangle = \frac{1}{N} \sum_{i=1}^{N} u(i).
\]

2. Divide the integrated signal \( y(k) \) into boxes of equal size \( n \) (scale of analysis).

3. In each box of size \( n \), fit \( y(k) \) using a polynomial function, \( y_{\text{fit}}(k) \), called the local trend. For order-\( \ell \) DFA (DFA-1 if \( \ell = 1 \), DFA-2 if \( \ell = 2 \) etc.), an \( \ell \)-order polynomial function is used.

4. Detrend the integrated time series \( y(k) \) by subtracting the “local” trend \( y_{\text{fit}}(k) \) in each box, and then calculate the detrended fluctuation function

\[
Y(k) \equiv y(k) - y_{\text{fit}}(k).
\]

5. For a given box size \( n \), calculate the root mean square (rms) fluctuation

\[
F(n) \equiv \sqrt{\frac{1}{N} \sum_{k=1}^{N} [Y(k)]^2}
\]

and repeat the above computation for different box sizes \( n \) (different scales) to provide a relationship between \( F(n) \) and \( n \).

A power-law relation

\[
F(n) \sim n^\alpha
\]

between \( F(n) \) and the box size \( n \) yields the scaling (correlation) exponent \( \alpha \), which represents the correlation properties of the signal. If \( \alpha = 0.5 \), there is no correlation and the signal
is uncorrelated (white noise); if $\alpha < 0.5$, the signal is anticorrelated; if $\alpha > 0.5$, there are positive correlations in the signal. For integrated white noise (random walk) the exponent $\alpha = 1.5$. The DFA method accurately quantifies $\alpha$ within the error bars of $\pm 0.02$ for an optimal range of scales $n$ [29].

In Fig. 4(a) we present results from the DFA analysis for the sequence $\delta_i$. We observe two distinct scaling regimes at small and intermediate scales characterized by scaling exponent $\alpha = 0.45 \pm 0.02$, and at large scales with exponent $\alpha_\ell = 1.5$. To test if the observed scaling regimes are due to intrinsic correlations in the sequence of distances, or are caused by embedded trends in the data, we generate an anticorrelated noise with $\alpha = 0.45$ with a superposed logarithmic trend. This artificial signal exhibits identical scaling behavior, with $\alpha_\ell = 1.5$ at large scales, and an identical shift of the position of the crossover scale $n_\times$ with increasing length of the sequence as observed for the sequence of $\delta_i$. Further, our tests on different realizations of anticorrelated noise with different values for the correlation exponent $\alpha$ and with the superimposed logarithmic trend show a similar crossover to $\alpha_\ell = 1.5$ at large scales, suggesting that the scaling behavior at large scales is indeed due to the logarithmic trend [Fig. 4(b)]. Our results for the local exponent for the sequence of distances over the range of small and intermediate scales show $\alpha_{loc} = 0.45$, which is significantly smaller than $\alpha_{loc} = 0.50$, obtained for realizations of white noise [Fig. 4(c)]. Thus our DFA results suggest that at small and intermediate scales there are weak but genuine anticorrelations following a power law scaling behavior, while at large scales the observed behavior is an artifact of a logarithmic trend in the values of the distances between consecutive prime numbers.

Recent studies have shown that signals with identical correlations in their fluctuations may exhibit different correlations for the magnitude and sign of the fluctuations [43, 44], and that the correlations in the sequences of the magnitude and sign provide additional information not contained in the correlation properties of the original signal. We next investigate the correlation properties of (i) the increments $\Delta_i$, (ii) their magnitude $|\Delta_i|$, and (iii) their signs $sgn(\Delta_i)$ [Fig. 4(d)].

(i) For the increments, we find that the scaling exponent $\alpha \approx 0$ and the scaling behavior does not exhibit any crossover for any value of $N_p$, suggesting that the increment sequence is strongly anticorrelated.

(ii) For the magnitude of the increments, we observe an anticorrelated behavior at small
and intermediate scales with $\alpha = 0.45 \pm 0.02$ followed by a crossover to $\alpha_l = 1.5$ at large scales [Fig. 4(d)]. This crossover behavior is similar to that observed for the $\delta_i$ sequence [Fig. 4(a)], and is due to a logarithmic trend in the $\Delta_i$, which we find to be identical to the trend in the $\delta_i$ [Fig. 5].

(iii) We also find that the sign series of the increments is anticorrelated with $\alpha \approx 0.2$ at small scales following a crossover to uncorrelated behavior with $\alpha_l = 0.5$ at large scales.

V. RANDOM WALKS DEFINED ON THE SEQUENCE OF INCREMENTS

To investigate the organization in the order of increments in the sequence of distances between consecutive prime numbers, we map the problem onto a random walk. A walker takes a step $\epsilon(j) = +1$ when an increment $\Delta_j$ is positive and a step $\epsilon(j) = -1$ when $\Delta_j$ is negative, regardless of the value of $\Delta_j$. If $\Delta_j = 0$, no step is taken. In Fig. 6(a) we show

$$W(M) \equiv \sum_{j=1}^{M} \epsilon(j),$$

which is the displacement of the walker from the origin after $M$ steps.

We use DFA to analyze this walk for $N_p = 10^7$ and find an uncorrelated behavior with a scaling exponent $\alpha = 1.5$ at both intermediate and large scales [Fig. 6(b)]. This shows that the differences between the numbers of positive and negative increments in sequences of prime numbers of increasing length perform an uncorrelated random walk.

Additional information is contained in a different walk, where the walker takes a positive or negative step only when the increment has a particular value $\Delta_0$. Specifically, $\epsilon_{\Delta_0}(j) = +1$ when the increment is $+\Delta_0$, and $\epsilon_{\Delta_0}(j) = -1$ when the increment is $-\Delta_0$, so that after $N_{\Delta_0}$ steps we have

$$W_{\Delta_0}(N_{\Delta_0}) \equiv \sum_{j=1}^{N_{\Delta_0}} \epsilon_{\Delta_0}(j).$$

We define such walks $W_{6k}(N_{6k}), W_{6k+2}(N_{6k+2}),$ and $W_{6k+4}(N_{6k+4})$ for each pair of increments $\pm 6, \pm (6k + 2),$ and $\pm (6k + 4)$, where $k = 0, 1, 2,$.. [Fig. 6(a)]. We find that at intermediate and large scales, these walks are also characterized by an exponent $\alpha = 1.5$, suggesting an uncorrelated walk [Fig. 6(b)] for all pairs of increments.
VI. SUMMARY

In summary, we find new statistical features in the sequence of distances between prime numbers and their increments. We find that for any sequence length $N_p$, the symbol entropy of the sequence of distances $\delta_i$ has a smaller value compared to the symbol entropy of the sequence of increments $\Delta_i$, suggesting a higher variability in the values of the increments. We find that the symbol entropy depends on $N_p$ as $\log(\log N_P)$ for both the sequence of distances and increments, and we provide analytical arguments supporting this observation. We also find a period-three oscillation in the distribution of increments and that this distribution follows a double-tent exponential form.

Further, we find that at small and intermediate scales both the distances and the magnitude of their increments exhibit weak anticorrelations described by a power law, suggesting a fractal organization in these sequences. At large scales the correlated behavior of both the distances and the magnitude of their increments is dominated by logarithmic trends. These trends leads to strong positive correlations in the sequence of distances and the magnitude of increments at large scales. In contrast, we find that the increments are strongly anticorrelated at all scales, and that the sign of the increments exhibits anticorrelated behavior at small scales with a crossover to uncorrelated behavior at intermediate and large scales.

Acknowledgments

We thank M. Wolf and M. Taqqu for useful discussions, and NIH & NSF for support.

FIG. 1: Distances $\delta_i$ between consecutive prime numbers and the differences $\Delta_i$ between consecutive distances (increments). (a) The first $5 \times 10^4 \delta_i$. (b) The first $5 \times 10^4 \Delta_i$.

FIG. 2: (a) Symbol entropies $S_\delta$ and $S_\Delta$ of distances $\delta_i$ between primes and their increments $\Delta_i$ as functions of the sequence length $N_p$ comprising the first $N_p$ prime numbers. For any given length $N_p$, $S_\Delta(N_p) > S_\delta(N_p)$, suggesting a greater variability in the values of increments. (b) Plots of $S_\delta(N_p)$ and $S_\Delta(N_p)$ vs. log(log $N_p$), up to $N_p = 5 \times 10^5$ (see Eq. (4) and Appendix A).
FIG. 3: (a) Histogram of increments $\Delta_i$ in the distances between consecutive prime numbers for the sequence of the first $N_p = 10^6$ primes. The bin width is unity. The occurrence frequency of increments $f(\Delta_i)$ exhibits a period-three oscillation. Increments $\Delta_i$ with values $\pm(6k + 2)$ ($k = 0, 1, 2, 3, \ldots$) occur most often, increments with values $\pm(6k)$ occur least often. This regularity is always present regardless of the sequence length $N_p$. (b) Tent-shape of the histogram of increments on a linear-log plot suggests an exponential form for large, medium, and small peaks. The top curve (corresponding to large and medium peaks) is thus $\pm 2, \pm 4, \pm 8, \pm 10, \pm 14, \pm 16, \ldots$, while the bottom curve (corresponding to small peaks) is $0, \pm 6, \pm 12, \pm 18, \ldots$. (c) Symmetry is observed in the occurrence frequency $f(\Delta_i)$ of positive increments and the corresponding negative increments.
FIG. 4: First order detrended fluctuation analysis (DFA-1) of distances between consecutive primes and their increments. (a) The fluctuation function $F(n)$ as a function of the scale $n$, for different sequence lengths $N_p = 5 \times 10^6$, $10 \times 10^6$, and $50 \times 10^6$. The exponent $\alpha = 0.45 \pm 0.02$ suggests weakly anticorrelated behavior at small and intermediate scales. The crossover scale $n_c$ shifts right with an increasing sequence length $N_p$, while $\alpha_c = 1.5$ remains constant regardless of the length of the sequence. This behavior is due to the underlying logarithmic trend in the values of the distances shown in Fig. 5. (b) Scaling obtained using DFA-1 for realizations of anticorrelated noise with different exponents $\alpha$ and with superposed logarithmic trend. At intermediate scales, the slope of the scaling curve corresponds to the exponent of the noise. At large scale there is a crossover to $\alpha_c = 1.5$ as seen in the sequence of the distances between prime numbers showing in (a), indicating that the crossover is due to the logarithmic trend. (c) For the sequence of distances the local exponent fluctuates around $\alpha_{loc} \approx 0.45 \pm 0.02$ over a broad range of scales, indicating a weak anticorrelated behavior. This is significantly different from the behavior observed for white noise where the local exponent fluctuates around $\alpha_{loc} \approx 0.5 \pm 0.01$. (d) Scaling behavior for the increments with $\alpha \approx 0$ indicates strong anticorrelations at all scales $n$. A crossover from $\alpha = 0.45 \pm 0.02$ to $\alpha_c = 1.5$ is observed for the magnitudes of the increments. This behavior is similar to that observed for the distances in (a) and suggests the presence of an identical logarithmic trend [Fig. 5]. The signs of the increments are anticorrelated at small scales ($\alpha \approx 0.2$) but uncorrelated at large scales ($\alpha = 0.5$).
FIG. 5: (a) Sequence of the first $5 \times 10^4$ distances $\delta_i$ between consecutive prime numbers and the magnitudes of their increments $\Delta_i$ after performing a moving average with a box of size 500. The sequence of the magnitudes of the increments $\Delta_i$ exhibits a logarithmic trend that is similar to the known trend found in the distances [16]. Solid lines represent logarithmic fits. (b) Log-linear plots of the data and fits in (a) suggest similar logarithmic trends with $P_\delta \approx P_{|\Delta|}$, even though $Q_\delta \neq Q_{|\Delta|}$.

![Graph](image)

FIG. 6: A random walk defined by the sequence of increments $\Delta_i$. (a) The walker displacement is plotted on the $y$-axis and the number of steps $j$ is plotted on the $x$-axis. Displacement of the walker when the walk is defined on the entire sequence of increments (top panel) [see Eq. (9)] and displacement of the walker for individual positive-negative pairs of increments $\pm 2, \pm 4, \pm 6$ and $\pm 8$ shown as $W_2, W_4, W_6$ and $W_8$ respectively (bottom panel) [see Eq. (10)]. (b) First order detrended fluctuation analysis DFA-1 of the integrated walks. The DFA exponent $\alpha = 1.5$ for the walk defined on the entire sequence of increments (top panel) as well as for the walk defined on individual pairs of positive and negative increments (bottom panel), demonstrating that the walks on individual increments, as well as the walk on entire sequence of increments, both behave like uncorrelated random walks.