Justification de la normalisation des écarts premiers
Justification of the normalization of prime gaps

Michel Monfette mycmon@gmail.com

2026

Version française

1 Justification de la normalisation

Dans cette section, nous justifions la définition

$$g_n = \frac{p_{n+1} - p_n}{\log(p_n)},$$

où p_n désigne le n-ième nombre premier. Cette normalisation joue un rôle central dans la construction d'une grammaire stationnaire des écarts premiers.

1.1 1. Origine heuristique : théorème des nombres premiers

Le point de départ est le théorème des nombres premiers, qui affirme que

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

où $\pi(x)$ est le nombre de nombres premiers inférieurs ou égaux à x. Cette relation implique que, autour d'un grand réel x, la *densité* des nombres premiers est approximativement

$$\text{densité} \approx \frac{1}{\log x}.$$

L'écart moyen entre deux nombres premiers voisins dans un voisinage de x est donc de l'ordre de

$$\text{écart moyen} \approx \log x.$$

En appliquant cette heuristique à la suite des nombres premiers (p_n), on obtient

$$p_{n+1} - p_n \approx \log(p_n),$$

ce qui suggère que la taille typique de l'écart entre p_n et p_{n+1} est de l'ordre de $\log(p_n)$.

1.2 2. Normalisation sans dimension

La quantité brute $p_{n+1} - p_n$ croît en moyenne avec p_n. Pour comparer des écarts situés à des échelles numériques très différentes (par exemple autour de 10^4, 10^6 ou 10^8), il est naturel de les rapporter à leur taille moyenne attendue.

On définit donc

$$g_n = \frac{p_{n+1} - p_n}{\log(p_n)}.$$

Si l'approximation heuristique

$$p_{n+1} - p_n \approx \log(p_n)$$

est valide en moyenne, alors on s'attend à ce que

$$g_n \approx 1$$

pour une large proportion d'indices n. La quantité g_n est ainsi une version *normalisée* de l'écart, sans dimension, qui permet de distinguer :

- les écarts plus petits que la moyenne ($g_n \ll 1$),
- les écarts typiques ($g_n \approx 1$),
- les écarts plus grands que la moyenne ($g_n \gg 1$).

Cette normalisation rend les écarts comparables à toutes les échelles et prépare le terrain pour une discrétisation en symboles.

1.3 3. Validation empirique et stationnarité

L'analyse expérimentale montre que la distribution des valeurs de g_n est remarquablement stable lorsque l'on considère des intervalles de plus en plus grands, par exemple :

$$[10\,000, 20\,000], \quad [1\,000\,000, 2\,000\,000], \quad [10\,000\,000, 20\,000\,000].$$

Les histogrammes de g_n sur ces intervalles présentent des formes similaires, et les proportions des classes définies par la discrétisation de g_n (symboles $a, b, c, d, \dots$) restent quasi constantes.

Cette *stationnarité empirique* justifie a posteriori le choix de la normalisation :

$$g_n = \frac{p_{n+1} - p_n}{\log(p_n)}.$$

Elle permet de construire une grammaire symbolique dont les règles de transition ne dépendent plus de l'échelle numérique, mais uniquement de la dynamique interne des écarts normalisés.

English Version

2 Justification of the normalization

In this section, we justify the definition

$$g_n = \frac{p_{n+1} - p_n}{\log(p_n)},$$

where p_n denotes the n-th prime number. This normalization plays a central role in the construction of a stationary grammar of prime gaps.

2.1 1. Heuristic origin: prime number theorem

The starting point is the prime number theorem, which states that

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

where $\pi(x)$ is the number of primes less than or equal to x. This implies that, around a large real number x, the *density* of primes is approximately

$$\text{density} \approx \frac{1}{\log x}.$$

The average gap between consecutive primes near x is therefore of order

$$\text{average gap} \approx \log x.$$

Applying this heuristic to the sequence of primes (p_n) yields

$$p_{n+1} - p_n \approx \log(p_n),$$

which suggests that the typical size of the gap between p_n and p_{n+1} is of order $\log(p_n)$.

2.2 2. Dimensionless normalization

The raw quantity $p_{n+1} - p_n$ grows on average with p_n. To compare gaps at very different numerical scales (for example around 10^4, 10^6, or 10^8), it is natural to rescale them by their expected average size.

We therefore define

$$g_n = \frac{p_{n+1} - p_n}{\log(p_n)}.$$

If the heuristic approximation

$$p_{n+1} - p_n \approx \log(p_n)$$

holds on average, then we expect

$$g_n \approx 1$$

for a large proportion of indices n. The quantity g_n is thus a *normalized*, dimensionless version of the gap, which allows us to distinguish:

- gaps smaller than the average ($g_n \ll 1$),
- typical gaps ($g_n \approx 1$),
- gaps larger than the average ($g_n \gg 1$).

This normalization makes gaps comparable across all scales and prepares the ground for discretization into symbols.

2.3 3. Empirical validation and stationarity

Experimental analysis shows that the distribution of the values g_n is remarkably stable when considering larger and larger intervals, for example:

$$[10\,000, 20\,000], \quad [1\,000\,000, 2\,000\,000], \quad [10\,000\,000, 20\,000\,000].$$

Histograms of g_n over these intervals exhibit similar shapes, and the proportions of the classes defined by the discretization of g_n (symbols $a, b, c, d, \dots$) remain almost constant.

This *empirical stationarity* provides an a posteriori justification for the choice of the normalization

$$g_n = \frac{p_{n+1} - p_n}{\log(p_n)}.$$

It allows the construction of a symbolic grammar whose transition rules no longer depend on the numerical scale, but only on the internal dynamics of the normalized gaps.