ON THE NUMBER OF PRIMES IN A SEQUENCE

YVES GALLOT

1. Introduction

When we search for prime numbers in a sequence, we would like to estimate how many are prime in a fixed range with a simple method before starting a long computation. Today, our mathematical knowledge in this domain is about zero. Only the problem of the sequences of the form $a \cdot n + b$ was solved by Hadamard and La Vallée-Poussin on the based of genius Riemann work. For the rest, we have conjectures: but many of them are strong and today very well verified by computations. The purpose of this paper is to report known results to estimate the number of primes in a sequence.

Let $S_{f(n)}(N)$ the finite sequence $f(1), f(2), ..., f(N)$.

$\pi_{f(n)}(N)$ denotes the number of prime numbers in $S_{f(n)}(N)$.

2. The form $a \cdot n + b$

Theorem 2.1 (Prime number theorem)**.**

$$\pi_n(N) \sim \sum_{n=2}^{N} \frac{1}{\log n} \sim \int_2^N \frac{dt}{\log t} \sim \frac{N}{\log N}$$

It was established, independently, by La Vallée-Poussin and Hadamard in 1896 (see, for example, [7]).

Theorem 2.2. *If* $(a, b) = 1$ *then*

$$\pi_{a \cdot n+b}(N) \sim \frac{1}{\varphi(a)} \int_2^{aN+b} \frac{dt}{\log t} \sim \frac{a}{\varphi(a)} \int_2^N \frac{dt}{\log t}.$$

It was proved by La Vallée-Poussin in 1896 by combining prime number theorem and Dirichlet theorem. Note that current notation is different:
let $\pi_{a,b}(x) := |\{p \le x : p \equiv b \pmod a\}|$, then $\pi_{a,b}(x) \sim \frac{1}{\varphi(a)} \int_2^x \frac{dt}{\log t}$.

The notation

$$\pi_{a \cdot n+b}(N) \sim \frac{a}{\varphi(a)} \sum_{\substack{n=1 \\ an+b>1}}^{N} \frac{1}{\log(an+b)}$$

is more adapted to a generalization.

Date: May 5, 2001; revised December 5, 2001.

2000 *Mathematics Subject Classification.* Primary 11Y11; Secondary 11A41.

Key words and phrases. prime numbers.

3. Polynomials

In 1962, Bateman and Horn indicated a quantitative form [3] of the famous "Hypothesis H" of Schinzel and Sierpiński [10] (see [9, Ch. 6] for interesting details). If we just consider one irreducible polynomial and change notation, then we obtain:

Conjecture 3.1. *Let $f(n)$ be an irreducible polynomial, with integral coefficients and a positive leading coefficient and let $w(p)$ be the number of solutions of the congruence $f(x) \equiv 0 \pmod p$. Then*

$$\pi_{f(n)}(N) \sim C_f \sum_{\substack{n=1 \\ f(n)>1}}^{N} \frac{1}{\log f(n)} \sim \frac{C_f}{\deg f} \int_2^N \frac{dt}{\log t}$$

where

$$C_f = \prod_{p\ prime} \frac{1 - w(p)/p}{1 - 1/p}$$

Many computations verified that the number of primes of polynomial forms agree well with the conjecture up to a fixed N. Some values of C_f were computed precisely: for example Shanks computed $C_{n^2+1} = 1.37281346...$ [11].

4. Weight of sequences

We would like to extend this conjecture to any function, then we define:

Definition 4.1. Let

$$C_f(N) = \frac{\pi_{f(n)}(N)}{\sum_{\substack{n=1 \\ f(n)>1}}^{N} \frac{1}{\log f(n)}}$$

$C_f(N)$ is called the weight of the sequence $S_{f(n)}(N)$. If $C_f = \lim_{N\to\infty} C_f(N)$ exists, it is called the weight of the infinite sequence $S_{f(n)}$.

In 1959, Riesel found a non trivial function such that $C_{f(n)} = 0$. This function is $509203 \cdot 2^n - 1$. In 1960, Sierpiński proved that there exist infinitely many integers k such that $k \cdot 2^n + 1$ is composite for every positive n (see [9, Ch. 5.VII]). In 1947, Mills discovered that we can construct a number θ, which is equal to 1.3064..., such that $f(n) = [\theta^{3^n}]$ is prime for every n (see [9, Ch. 3.II]). Then C_f can be infinite.

5. Estimation of the weight

Let $u_f(N, P)$ be the number of elements of $C_f(N)$ which are not strictly divisible by a prime $p < P$. If $C_f(N)$ is a sequence of random numbers and $P << f(n)$, then the result should be closed to $v_f(N, P) = \prod_{p<P}(1 - 1/p) \cdot N \sim N/(e^\gamma \log P)$. If we make the assumption that the weight is mainly due to the divisibility by factors smaller than P_0, then we have:

$$C_f(N) \approx \frac{e^\gamma \log P_0 \times u_f(N, P_0)}{N}$$

and if C_f exists and the limit converges quickly, we can select N_0 such that:

$$C_f \approx \frac{e^\gamma \log P_0 \times u_f(N_0, P_0)}{N_0}$$

For example, let $f(n) = n^4 + 1$, $P_0 = 1000$ and $N_0 = 10^5$. We obtain the approximation $C_f \approx 2.64$ and the precise value of C_f predicted by Bateman and Horn conjecture is 2.67896... [12].

6. The form $k \cdot 2^n \pm 1$

For fixed n, the distribution can be evaluated with La Vallée-Poussin's theorem. Then we just consider the case k fixed, which is still an open question. We make the assumption that $C_{k\cdot 2^n \pm 1}$ exists for all k and use the notation $C_{k_+} = C_{k\cdot 2^n+1}$ and $C_{k_-} = C_{k\cdot 2^n-1}$.

With the approximation $k \cdot 2^n \pm 1 \sim k \cdot 2^n$, we have:

Conjecture 6.1.

$$\pi_{k\cdot 2^n+1}(N) \sim C_{k_+} \sum_{n=1}^{N} \frac{1}{n\log 2 + \log k} \sim C_{k_+} \int_1^N \frac{dt}{t\log 2 + \log k}$$

and

$$\int_1^N \frac{dt}{t\log 2 + \log k} = \frac{\log(N\log 2 + \log k)}{\log 2} - \frac{\log(\log 2 + \log k)}{\log 2}$$

then

$$\pi_{k\cdot 2^n+1}(N) \sim C_{k_+} \log_2 \frac{N + \log_2 k}{1 + \log_2 k} \tag{6.1}$$

and

$$\pi_{k\cdot 2^n-1}(N) \sim C_{k_-} \log_2 \frac{N + \log_2 k}{1 + \log_2 k}. \tag{6.2}$$

The estimated and actual values of the numbers of primes of the form $k \cdot 2^n + 1$ for $3 \le k \le 19$ and $N = 200000$ are shown in Table 1. The weights were evaluated with $P_0 = N_0 = 10^4$.

Table 1. Comparison between the estimates of the number of primes of the form $k \cdot 2^n + 1$ and the actual number of primes found

k	weight	estimate	actual
3	2.446	40	34
5	1.017	16	19
7	2.446	38	29
9	2.689	42	49
11	1.576	24	20
13	1.088	17	16
15	3.340	51	41
17	0.755	12	16
19	0.960	15	17

We could be tempted to try to find a formula for the weight similar to the one for the polynomials. Let o_p be the order of 2 modulo p and let $w(p)$ be the number of solutions of the congruence $k \cdot 2^x + 1 \equiv 0 \pmod p$ in the range $0, 1, ..., o_p - 1$. A possible formula for the weight is $C_{k_+} =? \prod_{p \text{ prime}} \frac{1-w(p)/o_p}{1-1/p}$. But the values generated by this formula are not correct estimates. The reason is that $\gcd(o_{p_1}, o_{p_2}) \neq 1$ for many couples (p_1, p_2). Then the probabilities $1 - w(p)/o_p$ are dependent and a simple product cannot be used as numerator because of conditional probabilities.

7. Application to the Sierpiński problem

An integer k such that $k \cdot 2^n + 1$ is composite for every $n \geq 1$ is called a Sierpiński number. It is conjectured that the integer $k = 78557$ is the smallest Sierpiński number. To prove the conjecture, it suffices to exhibit a prime $k \cdot 2^n + 1$ for each $k < 78557$ (see [9, Ch. 5.VII] and [6] for details). Today, this had been done for all except for 17 values.

Wilfrid Keller defined the frequency f_m to be the number of k giving their first prime $k \cdot 2^n + 1$ for an exponent n in the interval $2^m \leq n < 2^{m+1}$ [2][8]. Jack Brennen proposed a method to compute the probability that at least one prime of the form $k \cdot 2^n + 1$ exists for each of the remaining k values, with $n \leq N$ [4]. The author extended the computation to the 39278 candidates and used it to estimate the frequencies f_m.

Let $\lambda_{k,N}$ be the expected number of primes of the form $k \cdot 2^n + 1$ for fixed k and $1 \leq n < N$. It can be evaluated by (6.1). By assuming a Poisson distribution, the probability that the range contains no prime is $p_{k,N} = e^{-\lambda_{k,N}}$. Then the chance of solving Sierpiński problem at N is $P(N) = \prod_k (1 - p_{k,N})$. The expected number of remaining candidates is $E(N) = \sum_k p_{k,N}$. Then the estimate of the frequency f_m is $\hat{f}_m = E(2^m) - E(2^{m+1})$.

Table 2. Comparison between the estimates of the frequencies and the actual frequencies found (Eq. (6.1))

m	0	1	2	3	4	5	6	7	8	9
f_m	7238	10194	9582	6272	3045	1445	685	331	195	114
$\hat{f}_m$	6271	8467	8925	7008	4222	2158	1045	515	267	147
m	10	11	12	13	14	15	16	17	18	19
f_m	47	34	26	11	18	12	5	≥ 5	≥ 2	?
$\hat{f}_m$	85	52	34	22	15	11	7.8	5.7	4.3	3.3

The estimated and actual values of the frequencies are shown in Table 2. The weights were evaluated with $P_0 = N_0 = 10^4$. If we exclude the small values for m, for which the estimates are not accurate, we note that the estimates are translated in comparison with the actual frequencies. We can translate the estimates by using

$$\pi_{k \cdot 2^n + 1}(N) \sim C_{k_+} \log_2 \frac{1.5N + \log_2 k}{1 + \log_2 k} \tag{7.1}$$

in place of (6.1). The new estimated values are shown in Table 3.

Table 3. Comparison between the estimates of the frequencies and the actual frequencies found (Eq. (7.1))

m	0	1	2	3	4	5	6	7	8	9
f_m	7238	10194	9582	6272	3045	1445	685	331	195	114
$\hat{f}_m$	11031	9037	8031	5351	2886	1413	687	348	187	106
m	10	11	12	13	14	15	16	17	18	19
f_m	47	34	26	11	18	12	5	≥ 5	≥ 2	?
$\hat{f}_m$	64	40	26	18	12	8.9	6.5	4.8	3.7	2.8

Note that the author found no theoretical justification to Eq. (7.1), however it is indicated because it produces an accurate estimate.

In Table 4 we list the expected status of the current and future search.

TABLE 4. Number of remaining k values expected at N (Eq. (7.1))

N	2^{16}	2^{17}	2^{18}	2^{19}	2^{20}	2^{21}	2^{22}	2^{23}	2^{24}	2^{25}	2^{26}	2^{27}
$E(N)$	30.7	24.2	19.3	15.7	12.8	10.6	8.9	7.5	6.3	5.4	4.6	4.0

For large N, (6.1) and (7.1) give about the same chance of solving Sierpiński problem at N. Note also that for large N, the variance $V(N) = \sum_k p_{k,N}(1 - p_{k,N})$ is approximately $E(N)$.

We have a 50% chance of solving Sierpiński problem at $N = 2^{43} \approx 10^{13}$. We have a 5% chance of solving it at $N = 2^{30} \approx 10^9$. We have a 95% chance of solving it at $N = 2^{81} \approx 10^{24}$. Note also that the chances at 2^{20}, 2^{21} and 2^{22} are respectively about 10^{-6}, 10^{-5} and 10^{-4}.

The weights of the remaining k values are listed in Table 5. Note that the smallest weight is 0.044 for $k = 51173$ but hopefully $51173 \cdot 2^{29} + 1$ is prime (and $51173 \cdot 2^{3089} + 1$!).

TABLE 5. Weights of the remaining k values

k	4847	5359	10223	19249	21181	22699	24737	27653	28433
weight	0.20	0.25	0.23	0.08	0.19	0.07	0.20	0.12	0.11
k	33661	44131	46157	54767	55459	65567	67607	69109	
weight	0.19	0.31	0.11	0.28	0.25	0.09	0.07	0.18	

8. APPLICATION TO THE RIESEL PROBLEM

An integer k such that $k \cdot 2^n - 1$ is composite for every $n \geq 1$ is called a Riesel number. It is conjectured that the integer $k = 509203$ is the smallest Riesel number (see [9, Ch. 5.VII], [5] and [1] for details). Today, a prime $k \cdot 2^n - 1$ had been found for all $k < 509203$ except for 123 values.

The method proposed for the Sierpiński problem was applied to the Riesel problem. The estimated and actual values of the frequencies are shown in Table 6. The weights were evaluated with $P_0 = N_0 = 10^4$.

TABLE 6. Comparison between the estimates of the frequencies and the actual frequencies found (Eq. (6.2))

m	0	1	2	3	4	5	6	7	8	9
f_m	39867	59460	62311	45177	24478	11668	5360	2728	1337	785
$\hat{f}_m$	35326	50272	56873	48204	30868	16301	7954	3896	1996	1084
m	10	11	12	13	14	15	16	17	18	19
f_m	467	289	191	125	87	62	38	35	≥ 11	≥ 2
$\hat{f}_m$	624	378	240	159	108	76	54	40	30	23

We note that the estimates are translated in comparison with the actual frequencies. So we again translate the estimates by using

$$\pi_{k\cdot 2^n-1}(N) \sim C_{k_-} \log_2 \frac{1.5N + \log_2 k}{1 + \log_2 k} \tag{8.1}$$

in place of (6.2). The new estimated values are shown in Table 7. This again produces an accurate estimate.

TABLE 7. Comparison between the estimates of the frequencies and the actual frequencies found (Eq. (8.1))

m	0	1	2	3	4	5	6	7	8	9
f_m	39867	59460	62311	45177	24478	11668	5360	2728	1337	785
$\hat{f}_m$	63198	55829	53549	38265	21586	10748	5219	2617	1387	779
m	10	11	12	13	14	15	16	17	18	19
f_m	467	289	191	125	87	62	38	35	≥ 11	≥ 2
$\hat{f}_m$	463	289	188	126	88	62	45	34	25	19

In Table 8 we list the expected status of the current and future search.

TABLE 8. Number of remainning k values expected at N (Eq. (7.1))

N	2^{16}	2^{17}	2^{18}	2^{19}	2^{20}	2^{21}	2^{22}	2^{23}	2^{24}	2^{25}	2^{26}	2^{27}
$E(N)$	209	163	130	104	85	70	58	49	41	35	30	26

We have a 50% chance of solving Riesel problem at $N = 2^{70} \approx 10^{21}$. We have a 5% chance of solving it at $N = 2^{47} \approx 10^{14}$. We have a 95% chance of solving it at $N = 2^{134} \approx 10^{40}$. Note also that the chances at 2^{20}, 2^{25} and 2^{30} are respectively about 10^{-40}, 10^{-16} and 10^{-8}.

REFERENCES

1. R. Ballinger and W. Keller, *The Riesel Problem: Definition and Status*, http://www.prothsearch.net/rieselprob.html
2. R. Ballinger and W. Keller, *The Sierpiński Problem: Definition and Status*, http://www.prothsearch.net/sierp.html
3. P. T. Bateman and R. A. Horn, *A Heuristic Asymptotic Formula Concerning the Distribution of Prime Numbers*, Math. Comp. **16** (1962), 363–367.
4. J. Brennen, *communication to PrimeNumbers egroup*, 2000.
5. C. Caldwell, *The Prime Glossary: Riesel number*, http://www.utm.edu/research/primes/glossary/RieselNumber.html
6. C. Caldwell, *The Prime Glossary: Sierpiński number*, http://www.utm.edu/research/primes/glossary/SierpinskiNumber.html .
7. H. M. Edwards, *Riemann's Zeta Function*, Academic Press, New York, 1974.
8. W. Keller, *Factors of Fermat Numbers and Large Primes of the form $k\cdot 2^n+1$. II*, unpublished, 1992.
9. P. Ribenboim, *The New Book of Prime Number Records*, 3rd ed., Springer-Verlag, New York, 1995.
10. A. Schinzel and W. Sierpiński, *Sur certaines hypothèses concernant les nombres premiers*, Acta Arith. **4** (1958), 185–208, Erratum **5** (1959), 259.
11. D. Shanks, *On the Conjecture of Hardy & Littlewood concerning the Number of Primes of the Form $n^2 + a$*, Math. Comp. **14** (1960), 321–332.
12. D. Shanks, *On Numbers of the Form $n^4 + 1$*, Math. Comp. **15** (1961), 186–189.

E-mail address: galloty@wanadoo.fr