On the probabilistic complexity of numerically checking the binary Goldbach conjecture in certain intervals

JEAN-MARC DESHOUILLERS j-m.deshouillers@u-bordeaux2.fr
Mathématiques Stochastiques, Université Victor Segalen Bordeaux 2, F-33076 Bordeaux Cedex, France

HERMAN TE RIELE herman.te.riele@cwi.nl
CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract. A heuristic analysis is presented of the complexity of an algorithm which was applied recently [2] to verify the binary Goldbach conjecture for every integer in the interval $[4, 10^{14}]$, as well as for every integer in $[10^k, 10^k + 10^9]$, for different values of k up to 300. The analysis agrees reasonably well with the experimental observations.

Keywords: Goldbach conjecture, sum of primes, probabilistic complexity

1991 Mathematics Subject Classification: Primary 11P32. Secondary 11Y99.
1991 Computing Reviews Classification System: F.2.1.

1. Introduction

Although almost all (if not all) mathematicians think Goldbach was right when in the middle of the XVIII-th century he asserted that every even positive integer is a sum of two primes[1], the question is still open. All the numerical experiments that have been performed confirm our expectation that an even integer should have many representations as a sum of two primes and show that it is indeed easy to find such a representation. To our knowledge, the latest ones are those we performed jointly with Y. Saouter [2], where it was shown, among others, that every even integer in the interval $[4, 10^{14}]$ is a Goldbach number (i.e., a sum of two primes), as well as every integer in $[10^k, 10^k + 10^9]$, for different values of k up to 300.

Our aim is to give here a heuristic analysis of the complexity of the algorithm we used in our computation. This analysis is based on a probabilistic model for the distribution of prime numbers and leads to an expected complexity which we can describe as follows.

Heuristics *Let H be a large integer, D an integer less than H which is the product of the first odd primes, and k a positive integer. Let further N be an integer such that $[N - 2H, N]$ contains at least k prime numbers and let us denote by Q_k the set of those k largest elements. The probability that every even integer in $[N, N + 2H]$ is a sum of two primes, one of them belonging to Q_k, is heuristically*

S. Kanemitsu and K. Győry (eds.), Number Theory and Its Applications, 89–99.

$$\exp\left(-\frac{\phi(D)H}{D}\sum_{d|D}\frac{1}{\phi(d)}\left(1-\frac{g(d)\pi(2H)}{H}\right)^k\right),$$

where ϕ denotes the Euler function, $\pi(x)$ the number of primes up to x, and

$$g(d)=\prod_{p|D}\frac{p(p-2)}{(p-1)^2}\prod_{p|d}\frac{p-1}{p-2}.$$

We shall finally show that these heuristics are in good accordance with our numerical data.

2. The algorithm

Let us consider two positive even integers N and $2H$. In order to check that every even integer $N+2h$ in $[N+2, N+2H]$ is a Goldbach number, we choose a family of prime numbers smaller than, but close to N and, for each of these primes, say q, we check for which value of h the integer $N+2h-q$ is prime; if, after a certain number of tries, we have covered all the values of h, then our check has been successful.

Since q is close to N, we need not use a primality test for $N+2h-q$, but simply check whether it belongs to the list of the first primes, which is produced once and for all. It turns out to be even more economical to represent this list of primes by its characteristic or indicator function, i.e., by a set of bits, the value of each being 1 or 0, according to whether or not its address represents a prime number. More precisely, we introduce a further parameter K and define a sequence of bits $\eta=(\eta(1),\ldots,\eta(H+K))$, where $\eta(r)=1$ if and only if $2r+1$ is prime; thus

$$\eta=(1,1,1,0,1,1,0,1,1,0,1,0,0,1,\ldots),$$

where the 0's correspond to $9, 15, 21, 25, 27, \ldots$; we do not lose much when considering that K is less than H, although in practice it will be much smaller.

In a similar way, we allocate a string of H bits to represent even integers in $[N+2, N+2H]$ in a natural way, i.e., the h-th bit represents $N+2h$, and its value is initially 0, and becomes 1 when a representation of $N+2h$ as a sum of two primes has been found. At the s-th step of our algorithm, the value of the h-th bit is denoted by $\epsilon_h^{(s)}$.

We denote by $q_1>q_2>\cdots>q_s>\cdots$ the set of consecutive primes which are at most equal to $N-3$. The algorithm, which seems to have been introduced for the first time by Mok-Kong Shen [3] runs as follows:

Step 0 Write $\epsilon_1^{(0)}=\cdots=\epsilon_H^{(0)}=0$.

...

Step s (When *Step* $s-1$ has been performed and not all the $\epsilon_h^{(s-1)}$ are equal to 1);
$s.1$. Compute q_s; if $N-q_s>2K+1$, send a message that the algorithm was not successful in checking that all even integers under consideration are Goldbach numbers and stop the execution; otherwise, continue;
$s.2$. for $1\le h\le H$, let

$$\epsilon_h^{(s)} := \max\left(\epsilon_h^{(s-1)}, \eta\left(\frac{N-q_s-1}{2}+h\right)\right);$$

$s.3$. if all the $\epsilon_h^{(s)}$ are equal to 1, send a message that all even integers under consideration are Goldbach numbers and stop the execution; otherwise, execute *Step* $(s+1)$.

Since there are only finitely many primes less than N, the algorithm bears its name: it ultimately terminates. Moreover, since it is a "diplomatic" algorithm in the terminology of [1] (in that it may say "yes" or "maybe" but never "no"), we simply have to check that it is reliable: indeed, if at the end the value of the bits ϵ_h is 1, it means that for some s, its value changed from 0 to 1, i.e., $\eta((N-q_s-1)/2+h)=1$, which means that $N-q_s+2h$ is a prime, whence $N+2h$ is a Goldbach number. We also notice that checking whether $N-q_s > 2K+1$ insures that $(N-q_s-1)/2+h$ belongs to the range of valid indices for η.

In order to compute q_s, we consider successively the odd integers less than q_{s-1} and check whether they are strong pseudo-primes (this is rather quick); when a strong pseudo-prime has been found, then we certificate that it is a prime (this takes substantially longer). Sub-step $s.2$ is a mere "shift and or" operation on H bits and sub-step $s.3$ is a check on H bit values.

In practice, with values like $N \sim 10^{100}$, $2H \sim 10^8$ and $K \sim 10^6$, the algorithm says "yes", thus its actual complexity is directly related (and grosso modo proportional) to the number of steps actually performed. For $H = 5 \times 10^7$ and various values of N between 5×10^{11} and 6×10^{11}, we observed that the number of steps is 207 in the mean, the median being 205. Let us explain why.

3. A simplified probabilistic model

In a first instance, we assume the small primes, i.e. those counted by η, to be randomly distributed, and we assume that the operations performed at different steps are independent. Let us notice that the number of small primes which is considered at the s-th step is $M_s = \pi(N-q_s+2H) - \pi(N-q_s-2)$, where as usual $\pi(x)$ denotes the number of primes up to x; when K is small compared to H (which is the case in our actual computations), M_s is close to $\pi(2H)$, which is denoted by M in the sequel.

Our question now becomes the following: we have H baskets; we repeat successively and independently the following operation: we select randomly M baskets and put a ball in each selected basket. How many operations are needed to fill all the baskets? We shall prove the following

PROPOSITION 1 *The probability that, after k operations, all the baskets are filled is*

$$P(H,M,k) := \sum_{t=0}^{\infty}(-1)^t \left(\binom{H-t}{M} \Big/ \binom{H}{M}\right)^k \binom{H}{t}. \tag{3.1}$$

Proof: Let $t \geq 0$ and $1 \leq i_1 < i_2 < \cdots < i_t \leq H$ be t pairwise distinct given integers in $[1,H]$; the probability that, during one operation, we do not fill the

baskets numbered by $i_1, \dots, i_t$, is $\binom{H-t}{M} \Big/ \binom{H}{M}$. By the independence of the successive operations, the probability that, during k steps, we avoid the baskets numbered by $i_1, \dots, i_t$ is $\left(\binom{H-t}{M} \Big/ \binom{H}{M}\right)^k$. By the sieving inclusion-exclusion principle, the probability that, after k operations, all the baskets are filled, is:

$$\sum_{t=0}^{\infty} (-1)^t \sum_{1 \le i_1 < i_2 < \cdots < i_t \le H} \left(\binom{H-t}{M} \Big/ \binom{H}{M}\right)^k, \tag{3.2}$$

whence (3.1). ∎

We now look at a median value of k, say k_m, defined in such a way that in 50% of the cases, if we execute k_m successive steps, then all the baskets are filled.

PROPOSITION 2 *Let $1 \le M < H$ be integers. Then for any integer*

$$k \ge \frac{\log(0.47) - \log H}{\log(1 - M/H)}, \tag{3.3}$$

we have $P(H, M, k) \ge 0.50$.

Proof: We readily see that each term in the series (in fact a finite sum) in the RHS of (3.1) has absolute value

$$\frac{H^t}{t!} \left((1 - \frac{M}{H}) \cdots (1 - \frac{M}{H - t + 1}) \right)^k (1 - \frac{1}{H}) \cdots (1 - \frac{t-1}{H})$$

which is at most $\frac{H^t}{t!} \left(1 - \frac{M}{H}\right)^{kt}$ for any t (and close to it for small t). Let us write

$$P(H, M, k) = 1 - H(1 - \frac{M}{H})^k + \frac{H^2}{2}(1 - \frac{M}{H})^k (1 - \frac{M}{H-1})^k (1 - \frac{1}{H}) + R.$$

We let $u = H(1 - M/H)^k$ and use the relation $|e^u - 1 - u - u^2/2| \le u^3 e^u/6$ for $u > 0$; this leads to

$$|R| \le \sum_{t \ge 3} \frac{H^t}{t!} \left(1 - \frac{M}{H}\right)^{kt} \le \frac{u^3 e^u}{6}.$$

By relation (3.3) we have $u \le 0.47$, whence $|R| \le 0.03$. We thus have $P(H, M, k) \ge 1 - u - |R| \ge 0.9501$. We may further notice that the positive term which we ignored is about $u^2/2 = 0.0012....$ ∎

As a numerical application, we consider the cases when

$$2H = 10^9, M = \pi(10^9) = 50847534$$

which leads to a median value $k_m = 194$, and $2H = 10^8$, $M = 5761455$, leading to $k_m = 151$. These values are substantially smaller than the median values we actually observed. As we shall see, the discrepancy comes from the fact that our model does not take into account the irregularities of distribution of primes in arithmetic progressions: only the easiest modulus 2 has been (trivially) taken care of.

4. An arithmetic refinement of the basic probabilistic model

Prime numbers are not evenly distributed in arithmetic progressions modulo a given integer: it is clear that any arithmetic progression $(an+b)_n$ with $\gcd(a,b) > 1$ can contain at most one prime number; however, the primes are evenly distributed in arithmetic progressions $(an+b)_n$ with $\gcd(a,b) = 1$. The prime number theorem for arithmetic progressions implies that when $\gcd(a,b) = 1$ we have

$$\text{Card}\{p \leq x \text{ s.t. } p \equiv b \bmod a\} = \frac{\pi(x)}{\phi(a)}(1+o(1)) \text{ as } x \to \infty. \tag{4.4}$$

A trivial consequence of this fact is that even integers are more easily represented as sum of two primes than odd integers; a hardly less trivial one is that integers which are divisible by 6 are more frequently represented as sum of two primes than other even integers. Let us see why: by (4.4), the primes which are congruent to 1 mod 6 occur with the same frequency as those which are congruent to -1 mod 6; but if we wish to represent n as p_1+p_2, when $n \equiv 0$ mod 6, we may choose $p_1 \equiv 1$ mod 6 and $p_2 \equiv -1$ mod 6 or $p_1 \equiv -1$ mod 6 and $p_2 \equiv 1$ mod 6, whereas when $n \equiv 2$ mod 6 we must choose $p_1 \equiv p_2 \equiv 1$ mod 6. This latter configuration is exactly as frequent as each of the former ones, exactly in the same way that when two coins are tossed, the probability of getting two heads is half that of obtaining one head and one tail. In a similar way, for an odd prime l, the number of solutions of the congruence $x+y \equiv n \bmod l$, with x and y coprime to l, is $l-1$ when $l|n$, and $l-2$ otherwise.

Our derivation of the heuristics given in the first section will depend on the consideration of the average number of representations of an even integer as a sum of two primes which takes into account the remark we just stated. If we denote by $r_2(2n)$ the number of representations of $2n$ as a sum of two primes, and recall that l is an odd prime, the previous remark leads to

$$\sum_{\substack{2n \equiv a \bmod l \\ 2n \leq t}} r_2(2n) \sim \frac{l-2}{l-1} \sum_{\substack{2n \equiv 0 \bmod l \\ 2n \leq t}} r_2(2n) \text{ when } l \nmid a \text{ and } t \to \infty.$$

More generally, we have for any squarefree odd D:

$$\sum_{\substack{2n \equiv a \bmod D \\ 2n \leq t}} r_2(2n) \sim \prod_{\substack{l|D \\ l \nmid a}} \frac{l-2}{l-1} \sum_{\substack{2n \equiv 0 \bmod D \\ 2n \leq t}} r_2(2n), \text{ as } t \to \infty;$$

since

$$\sum_{2n\le t} r_2(2n) = \sum_{p_1+p_2\le t} 1 = \sum_{p_1\le t} \pi(t-p_1) \sim \frac{1}{2}\pi^2(t),$$

a little manipulation of arithmetic functions leads to

$$\sum_{\substack{2n\equiv a \bmod D \\ 2n\le t}} r_2(2n) \sim \frac{1}{2}\prod_{l\mid D}\left(1-\frac{1}{(l-1)^2}\right)\prod_{l\mid \gcd(a,D)}\frac{l-1}{l-2}\pi^2(t).$$

We now include this arithmetical feature into the probabilistic model used in the previous section. Let $\mathcal{L} = \{3 < \cdots < l < \cdots < L\}$ be a set of consecutive primes and let D denote their product. We assume that N and H are sufficiently large and that D divides H. We consider a collection of H baskets which are split into $\mathcal{H}_1, \ldots, \mathcal{H}_d, \ldots, \mathcal{H}_D$, where d denotes a divisor of D, with the condition that $H_d = |\mathcal{H}_d| = \phi(D/d)H/D$. ($\mathcal{H}_d$ corresponds to the even integers a in $[2N+2, 2N+2H]$ such that $\gcd(a, D) = d$). We choose an integer M which is a multiple of $\phi(D)$, and for each d we let

$$M_d := \frac{M}{\phi(D)}\prod_{\substack{l\mid D \\ l\nmid d}}(l-2) = \frac{D}{\phi(D)}\prod_{\substack{l\mid D \\ l\nmid d}}\left(\frac{l-2}{l-1}\right)\frac{M}{H}H_d;$$

from the second expression, it is clear that M_d is proportional to the size of $\mathcal{H}_d$, corrected by the average representability as a sum of two primes of an even integer which is congruent to d mod D. Moreover, one easily sees from the first expression that the sum of the M_d's is M.

We now proceed in the following way: for each d, we select randomly M_d baskets in $\mathcal{H}_d$ and we put a ball in the chosen baskets. We repeat this operation successively and independently. We denote by $P(H, M, D, k)$ the probability that, after k operations, all the baskets are filled.

Let $\boldsymbol{t} = (t_1, \ldots, t_D)$ be a set of non-negative integers and $1 \le i_1^{(d)} < \cdots < i_{t_d}^{(d)} \le H_d$ be, for each d dividing D, a set of pairwise distinct given integers in $[1, H]$; the probability that, during one operation, we do not fill the baskets numbered by the family $\boldsymbol{i} = (i_j^{(d)})_{d,j}$ is

$$\prod_d \binom{H_d - t_d}{M_d}\binom{H_d}{M_d}^{-1}.$$

By the independence of the successive operations, the probability that, during k steps, we avoid the baskets numbered by the family $\boldsymbol{i}$ is

$$\left(\prod_d \binom{H_d - t_d}{M_d}\binom{H_d}{M_d}^{-1}\right)^k.$$

By the sieving principle, the probability that, after k operations, all the baskets are filled is

$$\sum_{\tau=0}^{\infty}(-1)^{\tau}\sum_{\substack{\boldsymbol{t}=(t_1,\ldots,t_D)\\ t_1+\ldots+t_D=\tau}}\sum_{\boldsymbol{i}\text{ ass. to }\boldsymbol{t}}\left(\prod_d\binom{H_d-t_d}{M_d}\binom{H_d}{M_d}^{-1}\right)^k,$$

where " $\boldsymbol{i}$ ass. to $\boldsymbol{t}$ " means that the last summation is performed over all the families of indices $\boldsymbol{i}=(i_j^{(d)})$, where for each d, there are exactly t_d indices $i_j^{(d)}$. The number of such families of indices is

$$\prod_d\binom{H_d}{t_d}.$$

We thus have

$$P(H,M,D,k)=\sum_{\tau=0}^{\infty}(-1)^{\tau}\sum_{|\boldsymbol{t}|=\tau}\left(\prod_d\binom{H_d-t_d}{M_d}\binom{H_d}{M_d}^{-1}\right)^k\prod_d\binom{H_d}{t_d}. \quad (4.5)$$

We can bound from above the generic term in the RHS of (4.5) by

$$\prod_d\left(1-\frac{M_d}{H_d}\right)^{kt_d}\frac{H_d^{t_d}}{t_d!},$$

and the two expressions are equivalent when M_d and H_d are sufficiently large. Under those circumstances, $P(H,M,D,k)$ is close (and can be proved to be equivalent) to

$$\sum_{\tau=0}^{\infty}(-1)^{\tau}\sum_{|\boldsymbol{t}|=\tau}\prod_d\frac{1}{t_d!}\left(H_d\left(1-\frac{M_d}{H_d}\right)^k\right)^{t_d}=$$

$$=\exp\left(-\sum_d H_d\left(1-\frac{M_d}{H_d}\right)^k\right)$$

$$=\exp\left(-\frac{\phi(D)H}{D}\sum_{d|D}\frac{1}{\phi(d)}\left(1-\frac{D}{\phi(D)}\frac{M}{H}\prod_{l|D,\ l\not|\, d}\frac{l-2}{l-1}\right)^k\right),$$

which may be rewritten as

$$P(H,M,D,k)\sim\exp\left(-\frac{\phi(D)H}{D}\sum_{d|D}\frac{1}{\phi(d)}\left(1-\prod_{l|D}\frac{l(l-2)}{(l-1)^2}\frac{M}{H}\prod_{l|d}\frac{l-1}{l-2}\right)^k\right). \quad (4.6)$$

In our probabilistic model, we assumed that D divides H, that $\phi(D)$ divides M, that the M_d baskets are chosen randomly, that the successive operations are independent .... In the concrete situation we are studying, almost none of these assumptions is fulfilled! The divisibility requirements can be relaxed in the probabilistic model: it is sufficient to assume that H and M are sufficiently large and that the distribution in the different classes is asymptotically even. However, the random choice of the "baskets", the independence of the successive choices mainly rely on the actual behaviour of primes in short intervals. Considering that *usually* this behaviour will be close to its average behaviour leads us to read (4.6) as the *heuristics* we propose.

5. Comparison of the heuristics and the observed data

We compare formula (4.6) with one of our actual jobs for the numerical verification of the Goldbach conjecture, viz. on the interval $I = [5 \times 10^{11}, 6 \times 10^{11}]$. This was split up in 1000 jobs for the intervals

$$I_j = [5 \times 10^{11} + (j-1) \times 10^8, 5 \times 10^{11} + j \times 10^8], \quad j = 1, \ldots, 1000.$$

We saved the number of steps our algorithm needed to verify the Goldbach conjecture on I_j. Table 1 gives counts of the numbers of steps observed (in column 2), for classes of length 5, starting with 170–174, up till 295–299, together with cumulative counts (in column 3). The minimum, maximum, and mean number of steps found was 171, 288 and 207, respectively. One immediate observation is that the distribution found is not symmetric.

For the comparison with formula (4.6), we list the values of

$$P(5 \times 10^7, \pi(10^8), D, k), \quad k = 175, 180, \ldots, 300$$

with $D = 3 \times 5 \times 7 \times 11 \times 13 \times 17 \times 19$, in the last column of Table 1.

Table 1. Observed and heuristic probability distribution of the number of steps needed to verify the Goldbach conjecture on 1000 intervals of length 10^8 in $[5 \times 10^{11}, 6 \times 10^{11}]$.

class	freq.	cum. freq.	k	$P(5 \times 10^7, \pi(10^8), D, k)$
0–169	0	0	170	0.000
170–174	3	3	175	0.000
175–179	13	16	180	0.000
180–184	30	46	185	0.001
185–189	72	118	190	0.010
190–194	100	218	195	0.045
195–199	154	372	200	0.125
200–204	126	498	205	0.249
205–209	114	612	210	0.395
210–214	108	720	215	0.538
215–219	67	787	220	0.660
220–224	68	855	225	0.758
225–229	50	905	230	0.831
230–234	33	938	235	0.883
235–239	21	959	240	0.920
240–244	12	971	245	0.946
245–249	10	981	250	0.964
250–254	7	988	255	0.975
255–259	2	990	260	0.984
260–264	4	994	265	0.989
265–269	2	996	270	0.993
270–274	2	998	275	0.995
275–279	0	998	280	0.997
280–284	1	999	285	0.998
285–289	1	1000	290	0.999
290–294	0	1000	295	0.999
295–299	0	1000	300	0.999

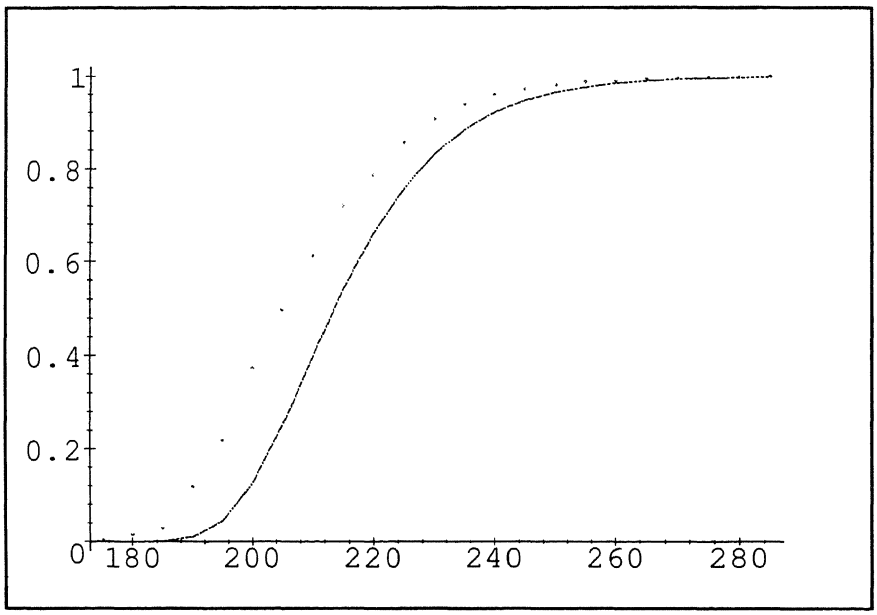

Figure 1. Plot of the data given in Table 1.

In Figure 1 we have plotted the observed frequencies (points) and the heuristic probability distribution (line). This approximates the experimental data reasonably well: at least it is not too optimistic in the estimation of the number of steps needed to verify the Goldbach conjecture in the range which we have covered in [2].

We notice that there are different presentations of our heuristics which are asymptotically equivalent, for example by extending to infinity the Eulerian convergent product in the inner part of (4.6), or by putting the Eulerian products of the inner sum as a factor of k in the exponent; moreover, in the numerical applications, we can modify the choice of D. In our case, these modifications lead to results which are all in rather good accordance with our observations; the one we present here is the one that comes out naturally without incorporating the modifications which would have led to the best agreement. For example, a value of D as small as 3×5 leads to a much better fit, a fact which is surprising to us.

Acknowledgments

We are grateful to V. Bagdonavicius and S. Malov for an interesting discussion during the preparation phase of this paper. The first named author benefited from the support of the Universités Victor Segalen Bordeaux 2, Bordeaux 1, CNRS and Rutgers University.

Notes

1. In his time, 1 was considered as a prime; with our modern definition we would replace "even positive integer" by "even integer larger than 2".

References

1. Francine Delmer and Jean-Marc Deshouillers, *On the computation of $g(k)$ in Waring's problem*, Math. Comp. **54** (1990), pp. 885–893.
2. Jean-Marc Deshouillers, Herman J.J. te Riele and Yannick Saouter, *New experimental results concerning the Goldbach conjecture*, To appear in the Proceedings of ANTS-III (Algorithmic Number Theory Symposium III, Reed College, Portland, Oregon, USA, June 21–25, 1998).
3. Mok-Kong Shen, *On Checking the Goldbach conjecture*, BIT **4** (1964), pp. 243–245.