Experimental Results on Additive 2-Bases

By M. L. Stein and P. R. Stein

1. Introduction. This paper describes the construction of binary additive bases for all even numbers in some finite interval $2 \leqq 2n \leqq N$. The construction makes use of a simple algorithm, first introduced by the authors in an earlier paper [1]. In the present paper the algorithm is applied to the sequence of primes, and several distinct "sparse" prime bases, constructed with its help, are described. As a by-product of this work, the verification of the Goldbach conjecture has been extended up through all even numbers $2n \leqq 10^7$.† The algorithm has also been applied to several random sequences of odd integers chosen so that their distribution is approximately that of the primes. Although the algorithm cannot, at present, be treated theoretically, even with regard to its asymptotic behavior, one may make plausible conjectures about it on the basis of various distinctive gross features; we hope to discuss this in a separate paper.

2. Definition of the Algorithm. Given a sequence of odd integers $\{a_i\}$, we wish to select a subsequence $\{b_i\}$ of these—generally with as few elements as will serve—so that every even number $2n$ within certain limits, $N_0 \leqq 2n \leqq N_1$, can be written in the form:

$$2n = b_i + b_j . \tag{2.1}$$

For example, the sequence $\{a_i\}$ may consist of the prime numbers greater than or equal to some initial prime $p_0 = a_0$. In this case we can take the sequence $\{a_i\}$ to be as large as we like, introducing new members as they are needed; the number will depend on the upper limit N_1. We may equally well choose different sequences of odd numbers for our $\{a_i\}$—for example the sieve numbers known as "lucky numbers" [2], or, in fact, any set of odd numbers which can be generated according to a well-defined prescription.

Since our object is to produce a binary basis that will be in some sense "sparse," the following procedure immediately suggests itself. Let $b_0 = a_0$. The first even number that can be expressed in the form (2.1) is $2n_0 = 2b_0$. For our second element we take $b_1 = a_1$. To continue, form the even numbers $2b_0$, $b_0 + b_1$, $2b_1$. Let $2n^*$ be the smallest even number $> 2b_0$ which does not belong to this set. We then look for the *largest* $a \in \{a_i\}$ such that either $b_0 + a = 2n^*$ or $b_1 + a = 2n^*$. If no such element exists, we move on to $2n^* + 2$, etc. More generally, given the partial basis $\{b_0, b_1, b_2, \cdots, b_k\}$, we form all the sums:

$$S_{ij} = b_i + b_j , \qquad i \leqq j \leqq k. \tag{2.2}$$

Next we find the smallest even number $2n^* > N_0$ (where N_0 is the lower limit of the range) which does not belong to the set $\{S_{ij}\}$. We then set

$$b_{k+1} = \operatorname*{Max}_{(a)} [a + b_m = 2n^*], \qquad a \in \{a_i\}, 0 \leqq m \leqq k. \tag{2.3}$$

Received December 21, 1964. This work was done under the auspices of the Atomic Energy Commission.

† See Note added, page 433 of this issue.

If this process is to yield a basis for the even numbers, then it must always be possible to find the "next" value b_{k+1}. If, for some even number $2n^*$, it proves impossible to satisfy the equation $a + b_i = 2n^*$ $(i \leqq k)$, we say that the algorithm "fails." In this case we replace $2n^*$ by $2n^* + 2$ in equation (2.3) and proceed. If the equation cannot be satisfied for the even number $2n^* + 2$, we replace the latter by $2n^* + 4$, etc. Eventually a new basis element b_{k+1} will be generated and the algorithm can be iterated.

It is clear that while the upper limit N_1 can be chosen in advance, the lower limit is a function of the set $\{a_i\}$; in fact, it is just the smallest even number such that the sequence $\{b_i\}$ forms a true basis for all the even numbers in the range $N_0 \leqq 2n \leqq N_1$. For "reasonable" sequences $\{a_i\}$ we expect that N_0 will be "close" to $2a_0$; what this means in practice will become clear from a study of the numerical examples (see Tables I and III). If $\{a_i\}$ is the set of all primes (beginning with $a_0 = 3$) we would conjecture that the sequence $\{b_i\}$ generated by our algorithm is a binary additive basis for all even numbers $2n \geqq 6$. At present, nothing further can be said about this "sharpened" form of Goldbach's conjecture; all our calculations show is that, with $\{a_i\}$ taken to be the set of primes less than 10^7, the $\{b_i\}$ generated by our algorithm is, in fact, a basis for all the even numbers $6 \leqq 2n \leqq 10^7$.

Suppose that, for some N_1 and a given set $\{a_i\}$, we have generated a basis $\{b_i\}$ for all the evens $N_0 \leqq 2n \leqq N_1$. Let us now fix N_0 and extend the upper limit N_1. (Note that we are in effect redefining N_0.) We cannot say that the algorithm will not fail somewhere between N_1 and the new upper limit $N_2 > N_1$. If it does, however, our prescription still allows us to extend the sequence $\{b_i\}$, which would then no longer constitute a true basis. If the set $\{a_i\}$ is infinite we may let the upper limit approach infinity. The sequence $\{b_i\}$ is still well defined, and it would make sense to ask for an asymptotic formula for its density. A satisfactory treatment of this problem seems very desirable.

For the cases studied in this paper, every sequence $\{b_i\}$ has a density much less than that of the original sequence $\{a_i\}$. This would seem to justify using the term "sparse" to characterize the $\{b_i\}$. Accordingly, in the sequel we shall refer to the $\{b_i\}$ as "S-sequences" (or "S-bases") and to our algorithm as the "S-algorithm."

3. Results for the Prime Case. It is apparent from the above discussion that the S-sequence is uniquely determined by the set $\{a_i\}$. In particular, the elements of $\{b_i\}$ depend critically on the value of a_0. Let $\{a_i\}$ be the sequence of (odd) primes, starting with a particular prime $a_0 = p_0$. Then different sequences $\{b_i\}$ will be produced by different choices of p_0. In the sequel we shall distinguish these different sequences by adding a subscript; thus the sequence corresponding to a particular choice of p_0 will be denoted by $\{b_i\}_{p_0}$. As an example, take the two sequences corresponding to $p_0 = 3$, $p_0 = 11$, respectively. The first 22 terms of $\{b_i\}_3$ are: 3, 5, 7, 13, 19, 23, 31, 37, 43, 47, 53, 61, 79, 83, 109, 113, 101, 131, 139, 157, 167, 199. For $\{b_i\}_{11}$ the first 22 terms are: 11, 13, 17, 19, 29, 31, 41, 43, 53, 37, 59, 79, 73, 113, 109, 103, 107, 151, 163, 167, 179, 191. Note that the elements b_i are not produced in strictly ascending order. We have called this phenomenon "backtracking." Thus, in the case $p_0 = 3$, the first 16 terms constitute a 2-basis for the even numbers $6 \leqq 2n \leqq 122$. To express $2n = 124$ in the required form we must introduce the prime $b_{16} = 101$: $124 = 101 + 23$. While backtracking appears to persist even when one goes to higher values, it is not very prevalent. For example in the S-sequences for $p_0 = 1, 3$, and 5, the number of elements less than 5 million which are generated "out of order" is, respectively, 5.55%, 6.23% and 5.22% of the total.

Let $B(p_0; x)$ be the number of primes $\leqq x$ in the S-sequence $\{b_i\}_{p_0}$. It is of interest to compare the values of $B(p_0; x)$ for a set of equally spaced x-values and different p_0. In Table I such a comparison is exhibited at intervals $\Delta x = 200{,}000$, $x \leqq 5 \times 10^6$. In each case, the $\{b_i\}_{p_0}$ was found to constitute a true binary additive basis for all $N_0 \leqq 2n \leqq x$; in other words, with the exception of a few early values $(2n < N_0)$, no failure was observed in any case.

Leaving aside for the moment the anomalous case $p_0 = 7$ (see Section 5), it is quite remarkable how small the variation of $B(p_0; x)$ with p_0 is for fixed x; in the second half of the table ($x \geqq 24 \times 10^5$) the absolute spread is less than 1.5% of the lowest value for each x listed. This is perhaps all the more remarkable in view of the fact that the various S-bases $\{b_i\}_{p_0}$ are very nearly pairwise disjoint, the number of primes common to two different sequences being typically between 3% and 4% of the total number in the shorter sequence. Let $C(u, v; x)$ be the number of primes common to the two S-sequences $\{b_i\}_u$ and $\{b_i\}_v$ for the range $2n \leqq x$. Tables II-a and II-b give a partial tabulation of $C(u, v; x)$ for the range $2n \leqq 5 \times 10^6$. In two cases—$p_0 = 1$ and $p_0 = 3$—the S-sequence has been calculated up to $2n = 10^7$. In these cases, we find $B(1; 10^7) = 10474$, $B(3; 10^7) = 10576$. For this range, the number of primes common to these two sequences is $C(1, 3; 10^7) = 288$. In passing we remark that this calculation verifies the Goldbach conjecture for all even numbers $2n \leqq 10^7$; to achieve this basis, less than 1.6% of the available primes are required.

4. Random Odds. As remarked in Section 2, the S-algorithm is not restricted to the sequence of primes. For example, in [1] we reported the construction of an S-basis for the evens $2n \leqq 350{,}000$ which was composed of lucky numbers. This basis was found to consist of 1672 luckies out of a total of 27420 luckies in the range. More recently, we have applied the algorithm to sets of odd numbers with approximately "prime-like" distribution. These sets were generated as follows. Let p_i be the ith prime. We chose at random 360 odd numbers equally distributed in the interval 3 to p_{360}, then 360 more in the interval p_{361} to p_{720}, and so forth up to p_{78498}, the last prime less than 10^6 (the number of odds in the final interval was suitably adjusted). Five such random sets were generated; we shall denote them by the symbols RO(1), RO(2), $\cdots$, RO(5); for the conclusions we will draw here it is not necessary to specify them more fully. To each of these sets we then applied the S-algorithm. To facilitate comparison with the "standard" S-sequence $\{b_i\}_3$, we forced the first odd in each case to be $a_0 = 3$. Let us denote by $RB_j(x)$ the number of elements $\leqq x$ in the S-sequence generated from the set RO(j). In Table III we have tabulated $RB_j(x)$ for our five sets at ten equally spaced values of $x \leqq 10^6$. The last column gives $B(3; x)$ for comparison.

The agreement for given x is remarkable, especially in view of the fact that there is no connection between the random sets RO(j) beyond their common prime-like distribution. As one might expect, the S-sequences corresponding to any two

Table I

$10^{-5}x$	$B(1;x)$ $(N_0 = 2)$	$B(3;x)$ $(N_0 = 6)$	$B(5;x)$ $(N_0 = 10)$	$B(7;x)$ $(N_0 = 18)$	$B(11;x)$ $(N_0 = 22)$	$B(13;x)$ $(N_0 = 30)$
2	1235	1245	1263	1287	1236	1233
4	1820	1837	1844	1970	1822	1813
6	2265	2288	2300	2566	2268	2264
8	2658	2681	2688	3089	2663	2654
10	3000	3029	3027	3588	3005	3008
12	3307	3337	3360	4074	3300	3315
14	3598	3623	3640	4508	3599	3613
16	3868	3909	3919	4955	3881	3882
18	4123	4178	4178	5344	4130	4131
20	4377	4421	4416	5747	4382	4390
22	4612	4661	4656	6133	4615	4610
24	4826	4883	4882	6514	4837	4832
26	5057	5108	5087	6875	5048	5050
28	5268	5318	5298	7216	5251	5250
30	5472	5517	5505	7568	5454	5458
32	5656	5710	5699	7892	5647	5643
34	5857	5880	5891	8210	5832	5828
36	6045	6078	6071	8523	6021	6023
38	6218	6271	6260	8820	6204	6205
40	6405	6452	6426	9132	6383	6387
42	6568	6609	6598	9450	6548	6550
44	6733	6787	6749	9741	6719	6707
46	6896	6948	6920	10020	6889	6877
48	7057	7110	7086	10299	7045	7029
50	7211	7274	7256	10579	7220	7186

$10^{-5}x$	$B(71;x)$ $(N_0 = 166)$	$B(73;x)$ $(N_0 = 166)$	$B(79;x)$ $(N_0 = 176)$	$B(83;x)$ $(N_0 = 190)$	$B(89;x)$ $(N_0 = 190)$	$B(97;x)$ $(N_0 = 198)$
2	1231	1239	1247	1239	1246	1239
4	1803	1814	1809	1821	1811	1815
6	2259	2266	2267	2277	2263	2270
8	2644	2652	2657	2661	2644	2660
10	2984	2996	2991	3002	2994	2989
12	3294	3308	3317	3319	3308	3302
14	3580	3593	3609	3595	3606	3594
16	3864	3861	3885	3884	3875	3867
18	4125	4120	4139	4127	4128	4117
20	4359	4368	4371	4374	4368	4366
22	4603	4609	4603	4610	4614	4589
24	4831	4824	4832	4835	4829	4813
26	5042	5045	5046	5046	5035	5024
28	5254	5252	5255	5247	5246	5236
30	5455	5460	5449	5459	5448	5437
32	5636	5642	5638	5658	5640	5627
34	5835	5831	5817	5840	5833	5814
36	6021	6017	6012	6015	6020	6008
38	6201	6190	6192	6215	6188	6180
40	6376	6378	6374	6377	6369	6358
42	6539	6537	6544	6553	6534	6528
44	6720	6707	6704	6721	6703	6696
46	6876	6866	6872	6876	6869	6860
48	7031	7021	7033	7028	7027	7028
50	7181	7177	7185	7182	7186	7178

TABLE II-a

u	v = 3	5	11	13
1	250	253	282	269
3		251	263	258
5			237	275
11				274

TABLE II-b

u	v = 73	79	83	89	97
71	264	253	280	259	290
73		284	244	269	286
79			272	269	239
83				301	300
89					281

TABLE III

$10^{-5}\,x$	$RB_1(x)$ $(N_0 = 38)$	$RB_2(x)$ $(N_0 = 16)$	$RB_3(x)$ $(N_0 = 136)$	$RB_4(x)$ $(N_0 = 52)$	$RB_5(x)$ $(N_0 = 158)$	$B(3; x)$
1	754	757	760	763	760	843
2	1115	1113	1113	1120	1116	1245
3	1392	1384	1382	1389	1383	1565
4	1619	1626	1624	1623	1618	1837
5	1826	1826	1835	1830	1826	2075
6	2025	2023	2018	2023	2016	2288
7	2200	2205	2193	2197	2186	2494
8	2370	2364	2361	2379	2361	2681
9	2530	2516	2523	2526	2516	2862
10	2676	2663	2661	2673	2666	3029

random sets RO(i), RO(j) have very few common elements. For the pair RO(1), RO(2), the S-sequences have 117 elements in common; for the other nine pairs the number of common elements varies between 23 and 39.

It is noteworthy that these prime-like random sets give rise to S-sequences markedly sparser than those produced by the primes themselves. The RO(i) are, however, prime-like only with respect to their overall density. For example, the distribution of gaps between successive elements is quite different from that which obtains for the prime sequence. In Table IV we compare the prime gap distribution for gaps of size $g \leqq 56$ (between successive primes) with the corresponding distribution for four of our random sets (range: $a_i \leqq 10^6$); in this table, $N(g)$ denotes the number of gaps of size g between successive elements. The complete absence of "modulo 6 peaks" and the consequent monotonic decrease of $N(g)$ with increasing g are just what one would expect. In view of the results presented in Table III, we may say that the primes, far from being a "privileged" sequence with regard to their efficiency as a binary additive basis for the evens, are somehow handicapped because of the distribution imposed on them by their defining sieve.

The greater "efficiency" of the S-sequences generated from our random odd sets is also mirrored in the corresponding "Goldbach frequency distribution." This distribution may be defined for prime S-sequences as follows. Let $\nu(p_0\,; 2n)$ be the number of solutions of the equation

$$2n = b_i + b_j, \qquad i \leqq j,\ b_i,\ b_j \in \{b_i\}_{p_0}. \tag{4.1}$$

TABLE IV

Number of Gaps N(g) of Size g Between Consecutive Elements

range: $a_i \leqq 10^6$

g	$N(g)$ (Primes)	$N(g)$ [RO(2)]	$N(g)$ [RO(3)]	$N(g)$ [RO(4)]	$N(g)$ [RO(5)]
2	8169	12412	12469	12325	12532
4	8143	10387	10429	10529	10411
6	13549	8827	8808	8777	8744
8	5569	7271	7298	7482	7299
10	7079	6245	6308	6289	6258
12	8005	5300	5156	5192	5274
14	4233	4355	4376	4230	4374
16	2881	3630	3666	3670	3593
18	4909	3090	3163	3078	3111
20	2401	2583	2605	2629	2627
22	2172	2316	2250	2246	2243
24	2682	1909	1819	1895	1901
26	1175	1599	1557	1581	1543
28	1234	1289	1333	1354	1301
30	1914	1181	1123	1142	1142
32	550	938	941	943	935
34	557	807	761	775	824
36	767	702	721	669	701
38	330	585	579	570	556
40	424	493	438	452	494
42	476	438	416	426	415
44	202	362	366	333	339
46	155	273	306	307	284
48	196	237	231	255	250
50	106	205	229	216	222
52	77	148	160	166	162
54	140	134	148	159	139
56	53	122	142	123	126

By the "Goldbach frequency distribution at the point (k, x)" we mean the number of solutions $\sigma_k(p_0\,;\,x)$ of the equation

$$\nu(p_0\,;\,2n) = k, \qquad 2n \leqq x. \tag{4.2}$$

A corresponding definition holds for the random odd S-sequences, where we replace the label p_0 by an appropriate symbol characterizing the underlying random odd set. In general, the random odd S-sequences have frequency distributions which are much more peaked (as a function of k for fixed x) than those belonging to the prime S-sequences. For example, if we form the sum $\sum_{k=1}^{10} \sigma_k(p_0\,;\,10^6)$ for any of the prime S-sequences (excluding $p_0 = 7$), we find that we have included approximately 83 % of the total number of decompositions; the corresponding number for the random odd S-sequences is about 98 %.

5. The Anomalous Case. It is evident from Table I that $\{b_i\}_7$ is much denser than any of the other prime S-sequences studied. The result is so anomalous that one is

TABLE V

p_0	Number of evens $2n^*$ of the form $6m$
1	33
3	45
5	30
7	18
11	33
13	42
71	38
73	27
79	30
83	30
89	31
97	41

led to suspect a calculational error; numerous independent checks, however, have failed to turn up anything of the sort. So far as we can tell, the observed behavior is simply a numerical accident. There is, however, one property of the S-algorithm which may shed some light on the nature of this accident. It happens that, for all p_0 tried so far, the even numbers $2n^*$ which determine the successive b_i (see equation (2.3)) are very rarely divisible by 6. This is shown in Table V for the range $2n \leqq 5 \times 10^6$. This behavior itself remains to be explained, but given this observed property it is not unreasonable that a sufficiently large asymmetry in the distribution of the b_i modulo 3 will increase in magnitude rather than be damped out. Such behavior would clearly lead to a much denser sequence $\{b_i\}$ and perhaps even to eventual failure of the algorithm. As it happens, all our prime S-sequences except that for $p_0 = 7$ are evenly distributed (mod 3). The anomalous sequence, however, shows a ratio of 1.93 between primes $\equiv 2 \pmod 3$ and primes $\equiv 1 \pmod 3$. This is for the interval $2n \leqq 5 \times 10^6$. The sequence $\{b_i\}_7$ was actually computed up to $2n = 7 \times 10^6$; here $B(7; 7 \times 10^6) = 13108$ and the above-mentioned ratio has risen to 2.04. We have watched the development of this asymmetry in some detail without, however, learning anything whatsoever about the underlying reason for the anomalous behavior.

Other distinctive properties of $\{b_i\}_7$ are consistent with the observed behavior of $B(7; x)$. The Goldbach frequency distribution $\sigma_k(7; x)$ is much broader (as a function of k) than it is for any of the other cases studied. In addition, the "backtracking" phenomenon mentioned in Section 2 is much more pronounced for this case; in the range $2n \leqq 5 \times 10^6$, some 9.95% of the minimals were generated "out of order."

Note added in proof. Since this article was written, we have extended our verification of the Goldbach conjecture up to one hundred million, using a simple sieve technique quite independent of the S-sequence method reported here. As a result of this work we may state the following—not very surprising—empirical theorem.

Let $p = P(2n)$ be the *smallest* odd prime $\geqq 3$ such that $2n - p$ is a prime. Then, for $63276 \leqq 2n \leqq 10^8$, $P(2n) < \sqrt{2n}$. For this range, the maximum value of $P(2n)$ turns out to be 1093: $60119912 = 1093 + 60118819$.

At the suggestion of Dr. D. Shanks, we also carried through the verification, over the same range, of the "modified" Goldbach conjecture, namely that every even number $4n + 2$ is the sum of two primes of the form $4k + 1$ (here 1 is counted as a prime). The above theorem holds, *mutatis mutandis*, for this case also, i.e. for $1457284 \leqq 2n\ (=4m + 2) < 10^8$, $P(2n) < \sqrt{2n}$. In this case, the maximum value of $P(2n)$ over the range is 2953: $76550462 = 2953 + 76547509$.

Los Alamos Scientific Laboratory
Los Alamos, New Mexico

1. M. L. Stein & P. R. Stein, "New experimental results on the Goldbach conjecture," *Math. Mag.*, v. 38, 1965, pp. 72–80.
2. V. Gardiner, R. Lazarus, N. Metropolis & S. Ulam, "On certain sequences of integers defined by sieves," *Math. Mag.*, v. 29, 1956, pp. 117–122. MR **17**, 711.