Multiplicative independence for random integers

Carl Pomerance[1]

For Heini Halberstam on the occasion of his retirement

Abstract. A sequence $s_1, s_2, \dots, s_n$ of positive integers is said to be multiplicatively dependent if there are integers $r_1, r_2, \dots, r_n$, not all zero, such that $s_1^{r_1} s_2^{r_2} \cdots s_n^{r_n} = 1$; and otherwise we say the sequence is multiplicatively independent. If one is presented with a sequence of random integers uniformly distributed in the interval $[1, x]$, how far into the sequence would you expect to go before it becomes multiplicatively dependent? We show that the answer is about $\exp(\sqrt{2 \log x \log\log x})$ terms. We also show that the same estimate holds for a similar problem, namely how far into the sequence should you expect to go before a non-empty subsequence has product a square? This latter problem has applications to the analysis of many integer factorization problems.

1. Introduction

In many integer factorization algorithms one is presented with a stream of pseudo-random integers in an interval. The intermediate goal of the algorithm is to pick out a non-empty subsequence with product a square. With such a subsequence, and depending on exactly which factorization algorithm one employs, there is a method to construct two squares, A^2, B^2, whose difference is a multiple of N, the number being factored. And from two such squares, one then has a reasonable chance at factorization via the gcd computation $(A - B, N)$. For more details and references consult the survey articles [6], [7].

Though most practical factorization algorithms have only heuristic analyses, it seems reasonable to at least try to give a rigorous analysis of the expected stopping time of the above procedure when the stream of integers is assumed to be random with uniform distribution in an initial interval. To be precise, let x be a large integer, and let $\mathcal{S} = s_1, s_2, \dots, s_x$ be a sequence of length x with terms in $\{1, 2, \dots, x\}$. Since $\mathcal{S}$ either contains the term 1 or it contains two identical terms, it is clear that some non-empty subsequence of $\mathcal{S}$ has product a square. Let $D_2(\mathcal{S})$ be the least integer D such that some non-empty subsequence of $s_1, s_2, \dots, s_D$ has product a square. If the terms s_n of $\mathcal{S}$ are independent, uniformly distributed random integers in $[1, x]$, what can be said of the statistic $D_2(\mathcal{S})$?

[1]The author is supported in part by an NSF grant.

A sequence of positive integers is said to be *multiplicatively independent* if the sequence of their logarithms is linearly independent over the rationals. So for a sequence $\mathcal{S} = s_1, s_2, \dots, s_x$ of integers in $[1, x]$, we may also consider the statistic $D(\mathcal{S})$, which is the least D such that $s_1, s_2, \dots, s_D$ is multiplicatively dependent (that is, *not* multiplicatively independent). As above, any sequence containing 1 or containing two identical terms is multiplicatively dependent, so $D(\mathcal{S})$ always exists. The sequence $s_1, s_2, \dots, s_D$ is multiplicatively dependent if and only if there are rationals $r_1, r_2, \dots, r_D$, not all of them 0, such that $s_1^{r_1} s_2^{r_2} \cdots s_D^{r_D} = 1$. By taking this equation to an appropriate power, we may assume that these exponents are integers with gcd 1. We conclude that not all of the exponents are even, and that the product of those s_i with odd exponents is a square. We have proved that for any integer sequence $\mathcal{S}$ of length x with terms in $[1, x]$,

$$D_2(\mathcal{S}) \le D(\mathcal{S}). \tag{1}$$

We shall say a sequence $s_1, s_2, \dots, s_D$ is *square dependent* if a non-empty subsequence has product a square. We have shown that multiplicative dependence implies square dependence.

The following theorem was announced in [8].

Theorem 1. *Let ε be an arbitrary but fixed positive number. If the terms of a sequence $\mathcal{S}$ of length x are independent random integers chosen uniformly in $[1, x]$, then the probability that both $D(\mathcal{S})$ and $D_2(\mathcal{S})$ are in the interval*

$$\left(\exp\left((\sqrt{2}-\varepsilon)\sqrt{\log x \log\log x}\right), \exp\left((\sqrt{2}+\varepsilon)\sqrt{\log x \log\log x}\right)\right)$$

tends to 1 as $x \to \infty$.

We shall let

$$L = L(x) = \exp\left(\sqrt{\log x \log\log x}\right).$$

Thus, Theorem 1 says that the normal values of $D(\mathcal{S})$ and $D_2(\mathcal{S})$ are $L^{\sqrt{2}+o(1)}$ as $x \to \infty$. We shall not only prove Theorem 1, but also prove that the same estimate holds for the expected values of $D(\mathcal{S})$ and $D_2(\mathcal{S})$.

In light of (1), to prove Theorem 1 it will suffice to show that the probability that $D(\mathcal{S}) < L^{\sqrt{2}+\varepsilon}$ tends to 1 as $x \to \infty$ and that the probability that $D_2(\mathcal{S}) > L^{\sqrt{2}-\varepsilon}$ also tends to 1 as $x \to \infty$. We respectively call these two assertions the "upper bound" problem and the "lower bound" problem. The proof of the upper bound is known in principle, going back at least to unpublished letters of R. Schroeppel from 1977. In addition, published proofs of the upper bound are essentially in [2] and [7]. Since none of these sources does exactly what is claimed here, we present this upper bound proof, yet again,

for completeness. The lower bound has been implicitly conjectured in many heuristic factorization analyses but a proof has never been given. This turns out to be a fairly hard problem and is the heart of this paper.

Our main theorem is not directly applicable to the analysis of existing factorization algorithms, but the method of its proof is. In particular from the proof it can be shown that various rigorously analyzed algorithms actually run in the time advertised as the running time upper bound. One example is the factorization algorithm in [6]. This algorithm, which on input of a composite number n, is expected to find a nontrivial factorization of n in at most $L(n)^{1+o(1)}$ bit operations. With the method of this paper, it can be shown that this algorithm is actually expected to require $L(n)^{1+o(1)}$ bit operations when n is an odd composite that is not a power. (For even composites and powers, the running time is less.) Similar assertions can be made about various rigorous discrete logarithm algorithms, such as by Lovorn-Bender and Pomerance. And also for the ERH conditional algorithm of Hafner and McCurley [5] for computing invariants of the class group of an imaginary quadratic number field.

2. The upper bound problem

In this section we prove that if S is a random sequence of integers from $[1, x]$ of length at most $L^{\sqrt{2}+\epsilon}$, then with probability tending to 1 as $x \to \infty$, the sequence S is multiplicatively dependent. We use the concept of a *smooth number*. Say a natural number is y-smooth if it has no prime factor exceeding y. Let $\pi(y)$ denote the number of primes up to y. It is easy to see that any sequence of natural numbers containing $\pi(y)+1$ terms that are y-smooth must be multiplicatively dependent.

Let $\psi(x, y)$ denote the number of y-smooth integers up to x. Our principal tool in both the upper and lower bound problems is the following result.

Theorem 2.1. *For real numbers x, y with $x \geq y \geq 2$, let $u = u(x, y) = \log x / \log y$. We have*

$$\psi(x, y) = x/u^{(1+o(1))u}$$

uniformly as $u \to \infty$, $u \leq \log x / \log \log x$.

This result is due to de Bruijn [1] and Canfield, Erdös and Pomerance [3]. We are especially interested in this result when $y = L^a$ for some fixed $a > 0$. In this case we have

$$\psi(x, L^a) = x/L^{1/(2a)+o_a(1)} \text{ as } x \to \infty.$$

Further, if a is chosen bounded away from 0 and ∞, then the functions $o_a(1)$

tend to 0 uniformly in a as $x \to \infty$.

If $\mathcal{S}$ is a random sequence of $[L^{\sqrt{2}+\epsilon}]$ integers drawn from $[1,x]$, then it follows from Theorem 2.1 with $y = L^{\sqrt{2}/2}$, that with probability $\to 1$ as $x \to \infty$, $\mathcal{S}$ contains at least y terms that are y-smooth. Since $y \geq \pi(y) + 1$ it follows that with probability $\to 1$ as $x \to \infty$ that $\mathcal{S}$ is multiplicatively dependent. This concludes our proof of the upper bound.

We now consider an upper bound for the expected value of $D(\mathcal{S})$. We wish to show that

$$\sum D(\mathcal{S}) \leq x^x L^{\sqrt{2}+o(1)}$$

as $x \to \infty$, where the sum is over the x^x integer sequences $\mathcal{S}$ of length x with terms in $[1,x]$. To this end it will suffice to show that for each fixed $\epsilon > 0$ and all sufficiently large x depending on the choice of ϵ,

$$\sum_{D(\mathcal{S}) > L^{\sqrt{2}+\epsilon}} D(\mathcal{S}) \leq x^x.$$

Since $D(\mathcal{S}) \leq x$ for each $\mathcal{S}$, it will suffice to show that the probability that $D(\mathcal{S}) > L^{\sqrt{2}+\epsilon}$ is $\leq x^{-1}$ for all large x. Let $y = L^{\sqrt{2}/2}$ and let $z = [L^{\sqrt{2}+\epsilon}]$. Note that every sequence $\mathcal{S}$ in the above sum must have fewer than y terms that are y-smooth among its first z terms. Let $\alpha = \psi(x,y)/x$, the probability that a random integer in $[1,x]$ is y-smooth. By Theorem 2.1, we have $\alpha = y^{-1+o(1)}$ as $x \to \infty$. Thus, the probability that a sequence $\mathcal{S}$ of length x and with fewer than y terms that are y-smooth among the first z terms is

$$\begin{aligned}\sum_{0 \leq i < y} \binom{z}{i} \alpha^i (1-\alpha)^{z-i} &= (1-\alpha)^z \sum_{0 \leq i < y} \binom{z}{i} \left(\frac{\alpha}{1-\alpha}\right)^i \\ &< e^{-\alpha z} \sum_{0 \leq i < y} \frac{1}{i!} \left(\frac{\alpha z}{1-\alpha}\right)^i \\ &< e^{-\alpha z} (\alpha z)^y e^{1/(1-\alpha)} < e^{-\alpha z/2} < x^{-1}\end{aligned}$$

for large x. So the expected value of $D(\mathcal{S})$ is $\leq L^{\sqrt{2}+o(1)}$ for $x \to \infty$.

3. The lower bound problem

In this section we shall show that a random sequence of integers in $[1,x]$ of length at most $L^{\sqrt{2}-\epsilon}$ is almost surely not square dependent. We begin our proof with the following observation.

Proposition 3.1. *Let $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$ be subsets of $\{1,2,\ldots,x\}$. Consider a random integer sequence drawn from $[1,x]$ of length at most l. The probability that it contains k distinct terms $m_1, m_2, \ldots, m_k$ with $m_1 \in \mathcal{A}_1, m_2 \in$*

$\mathcal{A}_2, \dots, m_k \in \mathcal{A}_k$ *is at most*

$$l^k x^{-k} \# \mathcal{A}_1 \# \mathcal{A}_2 \cdots \# \mathcal{A}_k.$$

We note that the numbers $m_1, \dots, m_k$ in the proposition, though they are distinct terms in the sequence, need not be distinct integers.

Proposition 3.2. *Consider a random integer sequence drawn from* $[1, x]$ *of length at most* $L^{1.5}$. *The probability it has a term divisible by the square of a prime* $p > L^3$ *or two terms each divisible by a prime* $p > L^3$ *is* $O(1/\sqrt{\log x})$.

Proof. We apply Proposition 3.1 first with $k = 1$ and $\mathcal{A}_1$ the set of multiples of p^2 in $[1, x]$, and next with $k = 2$ and $\mathcal{A}_1 = \mathcal{A}_2 =$ the set of multiples of p in $[1, x]$. We get that the probability in the proposition is at most

$$\sum_{p > L^3} \frac{L^{1.5}}{p^2} + \sum_{p > L^3} \left(\frac{L^{1.5}}{p} \right)^2,$$

where p runs over primes. But $\sum_{p>y} 1/p^2 \sim 1/(y \log y)$ as $y \to \infty$, so the proposition follows from the definition of L.

Say a positive integer m has the prime factorization $p_1 p_2 \dots p_l$, where $p_1 \geq \dots \geq p_l$. For $k = 1, \dots, l$, let $p_k(m) = p_k$, that is, $p_k(m)$ is the kth largest prime factor of m. For $k > l$, let $p_k(m) = 1$.

Proposition 3.3. *For any fixed positive integer* k, *the number of* L^3*-smooth integers* $m \leq x$ *such that* $p_k(m) \leq L^{1/3}$ *is* $x/L^{1.5+o_k(1)}$.

Proof. Let $\mathcal{B}$ denote the set of integers composed of at most $k-1$ primes in the interval $(L^{1/3}, L^3]$. Thus an integer counted in the proposition is of the form bn, where $b \in \mathcal{B}$ and n is $L^{1/3}$-smooth. Thus, the number of such integers up to x is exactly

$$\sum_{b \in \mathcal{B}} \psi(x/b, L^{1/3}).$$

Note that each $b \in \mathcal{B}$ is at most $L^{3(k-1)}$, so that from Theorem 2.1,

$$\psi(x/b, L^{1/3}) = x/(bL^{1.5+o_{k,b}(1)})$$

where the functions $o_{k,b}(1)$ tend to 0 as $x \to \infty$ uniformly for each $b \in \mathcal{B}$. But

$$1 \leq \sum_{b \in \mathcal{B}} 1/b \leq \left(1 + \sum_{p \in (L^{1/3}, L^3]} 1/p \right)^{k-1} \ll_k 1.$$

The proposition thus follows.

Proposition 3.4. *For any fixed positive integer k, we have uniformly for $a \in [1/3, 3]$, and n a positive integer $\le L^{3k}$, that the number of L^3-smooth integers $m \le x$, for which m is a multiple of n and $p_k(m) \le L^a$, is at most $x/(nL^{1/(2a)+o_k(1)})$.*

Proof. Similarly as in the proof of Proposition 3.3, let $\mathcal{B}$ denote the set of integers b composed of at most $k-1$ primes in $(L^a, L^3]$. Then an integer counted in the proposition is of the form bln, where l is L^a-smooth. Thus the number of such integers up to x is at most

$$\sum_{b \in \mathcal{B}} \psi(x/(bn), L^a).$$

Since $\sum 1/b \ll_k 1$ as in the above proof, the proposition thus follows from Theorem 2.1.

Proposition 3.5. *For any fixed positive integer k, the number of L^3-smooth integers $m \le x$ with $p_k(m)^2 | m$ is $\le x/L^{\sqrt{2}+o_k(1)}$ as $x \to \infty$.*

Proof. This result is essentially known when $k = 1$, but we shall nevertheless give the complete proof. For $i = 1, 2, \dots, \lceil (5/3) \log L \rceil$, let $I(i)$ denote the interval $(e^{i-1}L^{1/3}, e^i L^{1/3}]$ and let $N(i)$ denote the number of L^3-smooth integers $m \le x$ with $p_k(m)^2 | m$ and $p_k(m) \in I(i)$. Let $a_i = 1/3 + i/\log L$, so that $I(i) = (L^{a_{i-1}}, L^{a_i}]$. If m is counted in $N(i)$, then m is L^3-smooth, $p_k(m) \le L^{a_i}$ and there is a prime $p \in I(i)$ with $p^2 | m$. Thus, by Proposition 3.4,

$$N(i) \le \sum_{p \in I(i)} x/(p^2 L^{1/(2a_i)+o_k(1)}) = x/L^{a_i+1/(2a_i)+o_k(1)}.$$

The expression $a + 1/(2a)$ is minimized when a is $\sqrt{2}/2$, with minimum value $\sqrt{2}$, so we have for each i that

$$N(i) \le x/L^{\sqrt{2}+o_k(1)}.$$

Since there are only $O(\log L)$ values of i, and since by Proposition 3.3, the integers m with $p_k(m) \le L^{1/3}$ are negligible, we have the result.

We now prove the lower bound estimate. Let $\varepsilon > 0$ be arbitrary, but fixed. We wish to show that it is unlikely that a random integer sequence of length $[L^{\sqrt{2}-\varepsilon}]$ drawn from $[1, x]$ is square dependent. Suppose such a sequence is square dependent. From Proposition 3.2, we may assume that each number involved in the square dependency is L^3-smooth. Let k be an arbitrary fixed

positive integer. From Proposition 3.3, we may assume that each number m involved in the square dependency has $p_k(m) > L^{1/3}$. Of the numbers involved in the square dependency, let m_0 be one with $p_k(m_0)$ maximal. Let $p_j = p_j(m_0)$ for $j = 1, 2, \dots, k$. By Proposition 3.5, we may assume that $p_1 > p_2 > \cdots > p_k$. Since there is a square dependency, there are distinct terms $m_1, m_2, \dots, m_k$ in the sequence (and distinct from m_0) with $p_1 \cdots p_k | m_1 \cdots m_k$. We thus have an ordered factorization $p_1 \cdots p_k = n_1 \cdots n_k$ (where some of the factors may be 1) such that $n_1 | m_1, \dots, n_k | m_k$. We shall also let $n_0 = p_1 \cdots p_k$, so that $n_0 | m_0$. Note that for a given choice of $p_1 \cdots p_k$, there are $O_k(1)$ factorizations $n_1 \cdots n_k$.

For a given choice of primes $p_1 > \cdots > p_k$ and an ordered factorization $n_1 \cdots n_k$ of $p_1 \cdots p_k$, we count the number of $(k+1)$-tuples $m_0, m_1, \dots, m_k$ with each m_j being L^3-smooth, each $p_k(m_j) \le p_k$, and each $n_j | m_j$. In particular, let $I(i) = (L^{a_{i-1}}, L^{a_i}]$ be as in the proof of Proposition 3.5, and let $N(i)$ be the number of ordered $(k+1)$-tuples $m_0, m_1, \dots, m_k$ corresponding to some choice of $p_1 > \cdots > p_k$ and $n_1, \cdots, n_k$ with $p_k \in I(i)$. We have from Proposition 3.4 that

$$N(i) \le \sum_{\substack{p_1 > \cdots > p_k \\ p_k \in I(i)}} \sum_{\substack{n_0 = p_1 \cdots p_k \\ n_1 \cdots n_k = n_0}} \left(\frac{x}{n_0 L^{1/(2a_i)+o_k(1)}} \cdots \frac{x}{n_k L^{1/(2a_i)+o_k(1)}} \right)$$

$$= \sum_{\substack{p_1 > \cdots > p_k \\ p_k \in I(i)}} \frac{x^{k+1}}{(p_1 \cdots p_k)^2 L^{(k+1)/(2a_i)+o_k(1)}} = \frac{x^{k+1}}{L^{ka_i+(k+1)/(2a_i)+o_k(1)}}.$$

For any fixed positive number k, the expression $ka + (k+1)/(2a)$ is minimized when $a = \sqrt{(k+1)/(2k)}$ and the minimum value is $\sqrt{2k(k+1)}$. Thus, summing for all values of i, we have that the number of possible $(k+1)$-tuples $m_0, m_1, \dots, m_k$ is

$$\le \frac{x^{k+1}}{L^{\sqrt{2k(k+1)}+o_k(1)}}.$$

Since our sequence has $[L^{\sqrt{2}-\epsilon}]$ terms, the probability it has such a $(k+1)$-tuple as just described is, by Proposition 3.1, at most

$$L^{(k+1)(\sqrt{2}-\epsilon)-\sqrt{2k(k+1)}+o_k(1)}.$$

Choose k so large that $\sqrt{2k/(k+1)} > \sqrt{2} - \epsilon$. For such a value of k we have

$$(k+1)(\sqrt{2}-\epsilon) - \sqrt{2k(k+1)} = (k+1)\left(\sqrt{2} - \epsilon - \sqrt{2k/(k+1)}\right) < 0,$$

so the probability that the sequence will contain such a $(k+1)$-tuple tends to 0 as $x \to \infty$. This completes our proof of the lower bound and of Theorem 1.

It is to be remarked that the lower bound trivially implies that the expected value of $D_2(\mathcal{S})$ is $\geq L^{\sqrt{2}+o(1)}$ as $x \to \infty$. Combined with our upper bound on the expected value of $D(\mathcal{S})$ from the last section, we have that both $D(\mathcal{S})$ and $D_2(\mathcal{S})$ have expected value $L^{\sqrt{2}+o(1)}$ as $x \to \infty$.

Remark. One might ask if it is possible to do better than Theorem 1. In that theorem, the stopping time is almost surely placed in an interval $(A(x), B(x))$, where $\log A(x) \sim \log B(x)$ as $x \to \infty$. Is there a genuine threshold function $T(x)$, such that almost surely $D(\mathcal{S})$ and $D_2(\mathcal{S})$ are in $((1-\varepsilon)T(x), (1+\varepsilon)T(x))$ as $x \to \infty$, for any fixed $\varepsilon > 0$? I conjecture yes, but I am unsure of what function to suggest for $T(x)$. I hope to return to this problem in a future paper.

Acknowledgements. I wish to thank Red Alford, Neil Calkin, Rod Canfield and Prasad Tetali for their interest in this paper and several helpful conversations.

References

[1] N. G. de Bruijn, *On the number of positive integers $\leq x$ and free of prime factors $> y$, II*, Nederl. Akad. Wet. Proc. Ser. A **69** = Indag. Math. **38** (1966), 239-247.

[2] J. Buhler, H. W. Lenstra, Jr. and C. Pomerance, *Factoring integers with the number field sieve*, in The development of the number field sieve (A. K. Lenstra and H. W. Lenstra, Jr., eds.), Lecture Notes in Math. **1554**, Springer-Verlag, Berlin, 1993, 50-94.

[3] E. R. Canfield, P. Erdös and C. Pomerance, *On a problem of Oppenheim concerning "factorisatio numerorum"*, J. Number Theory **17** (1983), 1-28.

[4] J. L. Hafner and K. S. McCurley, *A rigorous subexponential algorithm for computation of class groups*, J. Amer. Math. Soc. **2** (1989), 837–850.

[5] H. W. Lenstra, Jr. and C. Pomerance, *A rigorous time bound for factoring integers*, J. Amer. Math. Soc. **5** (1992), 483-516.

[6] C. Pomerance, *Factoring*, in Cryptology and computational number theory – an introduction (C. Pomerance, ed.), Proc. Symp. Appl. Math. **42**, Amer. Math. Soc., Providence, 1990, 27-47.

[7] ———, *The number field sieve*, in Mathematics of Computation 1943–1993: a half century of computational mathematics (W. Gautschi, ed.), Proc. Symp. Appl. Math. **48**, Amer. Math. Soc., Providence, 1994, 465–480.

[8] ———, *The role of smooth numbers in number theoretic algorithms*, in: Proceedings of the International Congress of Mathematicians (S. D. Chatterji, ed.), Birkhaüser Verlag, Basel, 1995, 411–422..

Carl Pomerance
Department of Mathematics
University of Georgia
Athens, Georgia 30602
carl@ada.math.uga.edu