

# On the Square Dependence Problem

Ernie Croot  
Andrew Granville  
Prasad Tetali

November 2, 2006

Problem: Given  $x \geq 1$ , select integers

$$1 \leq a_1, a_2, \dots \leq x$$

at random (independently, uniform distributions), until some subsequence has product equal to a square. What is the expected stopping time  $T$ ? Does  $T$  have a sharp threshold? – i.e. does there exist a function  $f(x)$  such that

$$\text{Prob}(T \in [(c-\epsilon)f(x), (c+\epsilon)f(x)]) = 1 - o_\epsilon(1) ?$$

Example: Say our sequence begins  $a_1 = 2$ ,  $a_2 = 5$ ,  $a_3 = 6$ ,  $a_4 = 3$ . If so, we stop at  $a_4$ , as

$$a_1 a_3 a_4 = 6^2.$$

Motivation: Factoring algorithm analysis.

Dixon's Algorithm to factor  $N = pq$ :

- Select  $b_1, b_2, \dots \leq N - 1$  at random, and compute

$$a_i \equiv b_i^2 \pmod{N}.$$

- Continue until

$$a_{i_1} \cdots a_{i_k} = y^2.$$

Then,  $N \mid y^2 - (b_{i_1} \cdots b_{i_k})^2$ , and there is a fair chance that  $\gcd(y - b_{i_1} \cdots b_{i_k}, N) = p$  or  $q$ .

The numbers  $a_i$  we generate in Dixon's algorithm are not uniformly distributed, as they are squares mod  $N$ . Nonetheless, our problem appears to give good results on the behavior of the stopping time of Dixon's algorithm; and, one can easily modify our method of proof to give good results in this case where the  $a_i$  are all squares mod  $N$ .

What is known ?

First, we need some definitions: Let

- $P(n)$  = the largest prime factor of  $n$  (or 1 if  $n = 1$ ).
- $\Psi(x, y) = |\{n \leq x : P(n) \leq y\}|$ .

A number  $n$  is  $y$ -smooth if  $P(n) \leq y$ ; i.e. if all prime divisors are  $\leq y$ . So,

$$\Psi(x, y) = \# \text{ } y\text{-smooths } \leq x.$$

Let  $y_0 = y_0(x)$  be the value of  $y$  that maximizes

$$\frac{\Psi(x, y)}{y},$$

and set

$$J_0(x) = \frac{x}{\Psi(x, y_0)} \frac{y_0}{\log y_0}.$$

Schroeppel proved:

$$\text{Prob}(T < (1 + \epsilon)J_0(x)) = 1 - o_\epsilon(1).$$

Pomerance proved:

$$\text{Prob}(T > J_0(x)^{1-\epsilon}) = 1 - o_\epsilon(1).$$

So,

$$\text{Prob}(T \in [J_0^{1-\epsilon}, (1+\epsilon)J_0(x)]) = 1 - o_\epsilon(1).$$

This interval has the shape

$$[z, z^{1+\epsilon}].$$

Croot, Granville, and Tetali proved:

$$\text{Prob}(T \in [0.4407J_0(x), (e^{-\gamma} + \epsilon)J_0(x)]) = 1 - o_\epsilon(1),$$

where

$$\gamma = \text{Euler's constant} = 0.577\dots$$

and so

$$e^{-\gamma} = 0.5614594\dots$$

We believe, in fact, that

$$\text{Prob}(T \in [(e^{-\gamma} - \epsilon)J_0(x), (e^{-\gamma} + \epsilon)J_0(x)]) = 1 - o_\epsilon(1);$$

that is, we believe that the process has a sharp threshold, and that  $e^{-\gamma}J_0(x)$  is the about the expected stopping time of the process.

The upper bound requires lots of very technical ideas to prove; the lower bound is not as bad. Here I will describe only how to prove

$$\text{Prob}(T \in [c_0J_0(x), c_1J_0(x)]) = 1 - o(1).$$

First, let us see how Schroppel's bound is proved:

Given

$$a_1, \dots, a_J$$

consider only those that are  $y$ -smooth. We expect that there are

$$\sim J \frac{\Psi(x, y)}{x}$$

of them, and some basic arguments (e.g. Chebychev's inequality) bear this out. If this number exceeds the number of primes  $\leq y$ , there must be a square product; i.e.

$$J \frac{\Psi(x, y)}{x} > \pi(y)(1 + o(1)) \sim \frac{y}{\log y},$$

implies a square product.

To see this, given a  $y$ -smooth number  $a_i$ , write

$$a_i = 2^{A_{i,1}} 3^{A_{i,2}} \cdots p_{\pi(y)}^{A_{i,\pi(y)}}$$

and consider the vector

$$v_i = (A_{i,1}, A_{i,2}, \dots, A_{i,\pi(y)}) \pmod{2}.$$

So,

$$a_{i_1} \cdots a_{i_k} = \square$$

if  $v_{i_1} + \cdots + v_{i_k} \equiv 0 \pmod{2}.$

So, there is a square product among  $a_1, \dots, a_J$  if the vectors  $v_i$  are dependent; and, this is certainly true if there are more than  $\pi(y)$  of them corresponding to  $y$ -smooth numbers  $a_i$ .

Schroeppel then selected  $y = y_0$  which minimizes the value of  $J$  needed to guarantee that

$$J \frac{\Psi(x, y)}{x} \gtrsim \frac{y}{\log y}.$$

So, we need to minimize

$$\frac{xy}{\Psi(x, y) \log y},$$

which is essentially the same as maximizing

$$\frac{\Psi(x, y)}{y},$$

which is how we defined  $y_0$  on a previous slide.

We want a rough idea of the size of

$$J_0(x) := \frac{x}{\Psi(x, y_0)} \frac{y_0}{\log y_0}.$$

To do this, we apply the following theorem:

**Theorem.** Suppose

$$y = L(x, c) := \exp(c\sqrt{\log x \log \log x}).$$

Then,

$$\Psi(x, y) = xL(x, -1/2c + o(1)).$$

So,

$$\frac{\Psi(x, y)}{y} = xL(x, -c - 1/2c + o(1))$$

is maximized if  $c + 1/2c$  is minimized, and this happens when  $c = 1/\sqrt{2}$ , giving

$$y_0 = L(x, 1/\sqrt{2} + o(1)), \text{ and } J_0(x) = y_0^{2+o(1)}.$$

To produce the lower bound on  $T$ , one of the crucial ideas is to not express things in terms of  $x$  and  $y$ , but instead express things in terms of  $y_0$  and  $J_0$ . Specifically, what we do is find  $\eta$  such that if

$$J = \eta J_0(x),$$

then the expected number of square products among

$$a_1, \dots, a_J$$

is  $o(1)$ .

Another idea we use for the lower bound is to throw away the contribution from primes  $p \leq y$ : We have that the probability that randomly chosen

$$b_1, \dots, b_k \leq x$$

satisfies

$$b_1 \cdots b_k = \square$$

is at most the probability that

$$q_1 \cdots q_k = \square,$$

where

$$b_i = s_i q_i, P(s_i) \leq y, p(q_i) > y \text{ or } q_i = 1,$$

where  $P(m)$  is the largest prime divisor of  $m$  ( $P(1) = 1$ ), and  $p(m)$  is the smallest prime divisor of  $m$  ( $p(1) = 1$ ).

Thus, we have that the probability that  $b_1 \cdots b_k$  is a square is at most

$$\sum_{\substack{q_1, \dots, q_k \geq 1 \\ q_1 \cdots q_k = \square \\ q_i = 1 \text{ or } p(q_i) > y}} \prod_{i=1}^k \frac{\Psi(x/q_i, y)}{x} \\ \leq \left( \frac{(1 + o(1))\Psi(x, y)}{x} \right)^k \sum_{n=1 \text{ or } p(n) > y} \frac{\tau_k(n^2)}{n^{2\alpha}},$$

for a certain  $\alpha = \alpha(x, y)$ , where  $\tau_k(m)$  is the number of ways of writing  $m = d_1 \cdots d_k$ , where  $d_1, \dots, d_k$  are positive integers.

Where does  $\alpha$  come from ? It comes from the “saddle point method”, and it turns out that

$$\Psi(x/d, y) \leq \frac{1 + o(1)}{d^\alpha} \Psi(x, y),$$

for any  $1 < d \leq x$ .

So, since there are  $\binom{J}{k}$   $k$ -tuples of numbers among  $a_1, \dots, a_J$ , the expected number of subsequences of size  $k$  that product to a square is, for  $J = \eta J_0$ ,

$$\begin{aligned} &\leq \binom{J}{k} \left( \frac{(1 + o(1)) \Psi(x, y)}{x} \right)^k \sum_{\substack{n=1 \text{ or} \\ p(n) > y}} \frac{\tau_k(n^2)}{n^{2\alpha}} \\ &\ll \left( (e + o(1)) \frac{\eta y}{k \log y_0} \frac{\Psi(x, y)/y}{\Psi(x, y_0)/y_0} \right)^k \\ &\quad \times \prod_{p > y} \left( 1 + \frac{\tau_k(p^2)}{p^{2\alpha}} + \frac{\tau_k(p^4)}{p^{4\alpha}} + \dots \right). \end{aligned}$$

What we now do is break this into a number of different ranges of  $k$ , as upper bounds vary according to these ranges; however, the most significant range is

$$y_0^{1/4} \leq k \leq J_0 = y_0^{2+o(1)}.$$

For this range, the Euler product satisfies

$$\prod_{p>y} \left( 1 + \frac{\tau_k(p^2)}{p^{2\alpha}} + \dots \right) = e^{O(k)}.$$

So, the number of  $k$ -tuples is

$$\left( (e\eta + o(1)) \frac{y}{k \log y_0} \frac{\Psi(x, y)/y}{\Psi(x, y_0)/y_0} \right)^k e^{O(k)}.$$

Now, if we choose  $y$  such that  $\pi(y) = k$ , then it is not difficult to show (using PNT) that

$$\frac{y}{k \log y_0} \in [1/4 + o(1), 2 + o(1)];$$

and so this ratio is at most  $2 + o(1)$ . Furthermore the ratio involving  $\Psi(x, y)$  and  $\Psi(x, y_0)$  is at most 1, by the definition of  $y_0$ .

Putting everything together, we find that the number of  $k$ -tuples is at most  $(\eta C)^k$ , for a certain  $C$ . So, letting  $\eta < 1/C$ , we expect that there are  $o(1)$  such  $k$ -tuples; and, in fact, we will get that there are only  $o(1)$  of them for all the different values of  $k$  under consideration. It follows that if  $\eta < 1/C$ , and we choose  $\eta J_0(x)$  numbers, then almost surely there is no subsequence that has product equal to a square.