

Long arithmetic progressions of primes

Ben Green

ABSTRACT. This is an article for a general mathematical audience on the author's work, joint with Terence Tao, establishing that there are arbitrarily long arithmetic progressions of primes.

1. Introduction and history

This is a description of recent work of the author and Terence Tao [GTc] on primes in arithmetic progression. It is based on seminars given for a general mathematical audience in a variety of institutions in the UK, France, the Czech Republic, Canada and the US.

Perhaps curiously, the order of presentation is much closer to the order in which we discovered the various ingredients of the argument than it is to the layout in [GTc]. We hope that both expert and lay readers might benefit from contrasting this account with [GTc] as well as the expository accounts by Kra [Kra06] and Tao [Tao06a, Tao06b].

As we remarked, this article is based on lectures given to a general audience. It was often necessary, when giving these lectures, to say things which were not strictly speaking true for the sake of clarity of exposition. We have retained this style here. However, it being undesirable to commit false statements to print, we have added numerous footnotes alerting readers to points where we have oversimplified, and directing them to places in the literature where fully rigorous arguments can be found.

Our result is:

THEOREM 1.1 (G.-Tao). *The primes contain arbitrarily long arithmetic progressions.*

Let us start by explaining that the truth of this statement is not in the least surprising. For a start, it is rather easy to write down a progression of five primes (for example 5, 11, 17, 23, 29), and in 2004 Frind, Jobling and Underwood produced

2000 *Mathematics Subject Classification*. Primary 11N13, Secondary 11B25.

the example

$$56211383760397 + 44546738095860k; \quad k = 0, 1, \dots, 22.$$

of 23 primes in arithmetic progression. A very crude heuristic model for the primes may be developed based on the prime number theorem, which states that $\pi(N)$, the number of primes less than or equal to N , is asymptotic to $N/\log N$. We may alternatively express this as

$$\mathbb{P}(x \text{ is prime} \mid 1 \leq x \leq N) \sim 1/\log N.$$

Consider now the collection of all arithmetic progressions

$$x, x + d, \dots, x + (k - 1)d$$

with $x, d \in \{1, \dots, N\}$. Select x and d at random from amongst the N^2 possible choices, and write E_j for the event that $x + jd$ is prime, for $j = 0, 1, \dots, k - 1$. The prime number theorem tells us that

$$\mathbb{P}(E_j) \approx 1/\log N.$$

If the events E_j were independent we should therefore have

$$\mathbb{P}(x, x + d, \dots, x + (k - 1)d \text{ are all prime}) = \mathbb{P}\left(\bigwedge_{j=0}^{k-1} E_j\right) \approx 1/(\log N)^k.$$

We might then conclude that

$$\#\{x, d \in \{1, \dots, N\} : x, x + d, \dots, x + (k - 1)d \text{ are all prime}\} \approx \frac{N^2}{(\log N)^k}.$$

For fixed k , and in fact for k nearly as large as $2 \log N / \log \log N$, this is an increasing function of N . This suggests that there are infinitely many k -term arithmetic progressions of primes for any fixed k , and thus arbitrarily long such progressions.

Of course, the assumption that the events E_j are independent was totally unjustified. If E_0, E_1 and E_2 all hold then one may infer that x is odd and d is even, which increases the chance that E_3 also holds by a factor of two. There are, however, more sophisticated heuristic arguments available, which take account of the fact that the primes $> q$ fall only in those residue classes $a \pmod{q}$ with a coprime to q . There are very general conjectures of Hardy-Littlewood which derive from such heuristics, and a special case of these conjectures applies to our problem. It turns out that the extremely naïve heuristic we gave above only misses the mark by a constant factor:

CONJECTURE 1.2 (Hardy-Littlewood conjecture on k -term APs). *For each k we have*

$$\#\{x, d \in \{1, \dots, N\} : x, x + d, \dots, x + (k - 1)d \text{ are all prime}\} = \frac{\gamma_k N^2}{(\log N)^k} (1 + o(1)),$$

where

$$\gamma_k = \prod_p \alpha_p^{(k)}$$

is a certain product of “local densities” which is rapidly convergent and positive.

We have

$$\alpha_p^{(k)} = \begin{cases} \frac{1}{p} \left(\frac{p}{p-1}\right)^{k-1} & \text{if } p \leq k \\ \left(1 - \frac{k-1}{p}\right) \left(\frac{p}{p-1}\right)^{k-1} & \text{if } p \geq k. \end{cases}$$

In particular we compute¹

$$\gamma_3 = 2 \prod_{p \geq 3} \left(1 - \frac{1}{(p-1)^2}\right) \approx 1.32032$$

and

$$\gamma_4 = \frac{9}{2} \prod_{p \geq 5} \left(1 - \frac{3p-1}{(p-1)^3}\right) \approx 2.85825.$$

What we actually prove is a somewhat more precise version of Theorem 1.1, which gives a lower bound falling short of the Hardy-Littlewood conjecture by just a constant factor.

THEOREM 1.3 (G.–Tao). *For each $k \geq 3$ there is a constant $\gamma'_k > 0$ such that*

$$\#\{x, d \in \{1, \dots, N\} : x, x+d, \dots, x+(k-1)d \text{ are all prime}\} \geq \frac{\gamma'_k N^2}{(\log N)^k}$$

for all $N > N_0(k)$.

The value of γ'_k we obtain is very small indeed, especially for large k .

Let us conclude this introduction with a little history of the problem. Prior to our work, the conjecture of Hardy-Littlewood was known only in the case $k = 3$, a result due to Van der Corput [vdC39] (see also [Cho44]) in 1939. For $k \geq 4$, even the existence of infinitely many k -term progressions of primes was not previously known. A result of Heath-Brown from 1981 [HB81] comes close to handling the case $k = 4$; he shows that there are infinitely many 4-tuples $q_1 < q_2 < q_3 < q_4$ in arithmetic progression, where three of the q_i are prime and the fourth is either prime or a product of two primes. This has been described as “infinitely many $3\frac{1}{2}$ -term arithmetic progressions of primes”.

2. The relative Szemerédi strategy

A number of people have noted that [GTc] manages to avoid using any deep facts about the primes. Indeed the only serious number-theoretical input is a zero-free region for ζ of “classical type”, and this was known to Hadamard and de la Vallée Poussin over 100 years ago. Even this is slightly more than absolutely necessary; one can get by with the information that ζ has an isolated pole at 1 [Taoa].

Our main advance, then, lies not in our understanding of the primes but rather in what we can say about *arithmetic progressions*. Let us begin this section by telling a little of the story of the study of arithmetic progressions from the combinatorial point of view of Erdős and Turán [ET36].

¹For a tabulation of values of γ_k , $3 \leq k \leq 20$, see [GH79]. As $k \rightarrow \infty$, $\log \gamma_k \sim k \log \log k$.

DEFINITION 2.1. Fix an integer $k \geq 3$. We define $r_k(N)$ to be the largest cardinality of a subset $A \subseteq \{1, \dots, N\}$ which does not contain k distinct elements in arithmetic progression.

Erdős and Turán asked simply: what is $r_k(N)$? To this day our knowledge on this question is very unsatisfactory, and in particular we do not know the answer to

QUESTION 2.2. Is it true that $r_k(N) < \pi(N)$ for $N > N_0(k)$?

If this is so then the primes contain k -term arithmetic progressions on density grounds alone, irrespective of any additional structure that they might have. I do not know of anyone who seriously doubts the truth of this conjecture, and indeed all known lower bounds for $r_k(N)$ are much smaller than $\pi(N)$. The most famous such bound is Behrend's assertion [Beh46] that

$$r_3(N) \gg Ne^{-c\sqrt{\log N}};$$

slightly superior lower bounds are known for $r_k(N)$, $k \geq 4$ (cf. [LL, Ran61]).

The question of Erdős and Turán became, and remains, rather notorious for its difficulty. It soon became clear that even seemingly modest bounds should be regarded as great achievements in combinatorics. The first really substantial advance was made by Klaus Roth, who proved

THEOREM 2.3 (Roth, [Rot53]). *We have $r_3(N) \ll N(\log \log N)^{-1}$.*

The key feature of this bound is that $\log \log N$ tends to infinity with N , albeit slowly². This means that if one fixes some small positive real number, such as 0.0001, and then takes a set $A \subseteq \{1, \dots, N\}$ containing at least 0.0001 N integers, then provided N is sufficiently large this set A will contain three distinct elements in arithmetic progression.

The generalisation of this statement to general k remained unproven until Szemerédi clarified the issue in 1969 for $k = 4$ and then in 1975 for general k . His result is one of the most celebrated in combinatorics.

THEOREM 2.4 (Szemerédi [Sze69, Sze75]). *We have $r_k(N) = o(N)$ for any fixed $k \geq 3$.*

Szemerédi's theorem is one of many in this branch of combinatorics for which the bounds, if they are ever worked out, are almost unimaginably weak. Although it is in principle possible to obtain an explicit function $\omega_k(N)$, tending to zero as $N \rightarrow \infty$, for which

$$r_k(N) \leq \omega_k(N)N,$$

to my knowledge no-one has done so. Such a function would certainly be worse than $1/\log_* N$ (the number of times one must apply the log function to N in order to get a number less than 2), and may even be slowly-growing compared to the inverse of the Ackermann function.

The next major advance in the subject was another proof of Szemerédi's theorem by Furstenberg [Fur77]. Furstenberg used methods of ergodic theory, and

²cf. the well-known quotation "log log log N has been proved to tend to infinity with N , but has never been observed to do so".