# Long arithmetic progressions in the primes

Australian Mathematical Society Meeting

26 September 2006

Terence Tao (UCLA)

Additive patterns in the primes

- Many classical questions concerning additive patterns in the primes remain unsolved, e.g.:
- **Twin prime conjecture** (?Euclid, circa. 300 BC?): There exist infinitely many pairs $p, p+2$ of primes that are distance two apart: $(3,5)$, $(5,7)$, $(11,13)$, $(17,19), \ldots$.
- **Odd Goldbach conjecture** (1742): Every odd number $n \geq 7$ is the sum of three primes. $7 = 2+2+3$, $9 = 3+3+3$, $11 = 3+3+5$, etc.
- **Even Goldbach conjecture** (Euler, 1742): Every even number $n \geq 4$ is the sum of two primes. $4 = 2+2$, $6 = 3+3$, $8 = 3+5$, etc.
- But there have been some deep results, such as:
- **Chen's theorem** (1966): There exist infinitely many pairs $p, p+2$ where $p$ is a prime and $p+2$ is an almost prime (product of at most two primes).
- **Vinogradov's theorem** (1937): Every sufficiently large odd number $n$ is the sum of three primes.

- (Liu-Wang, 2002) Every odd number $n > 10^{1346}$ is the sum of three primes. [Also known for $n < 10^{20}$.]

- NB: multiplicative problems in the primes are significantly easier. For instance, it is obvious that there are no geometric progressions in the primes of length three or higher.

Arithmetic progressions in the primes

$$2$$

$$2, 3$$

$$3, 5, 7$$

$$5, 11, 17, 23$$

$$5, 11, 17, 23, 29$$

$$7, 37, 67, 97, 127, 157$$

$$7, 157, 307, 457, 607, 757$$

$$\ldots$$

$$5749146449311 + 26004868890n; \quad n = 0, \ldots, 20$$

$$11410337850553 + 4609098694200n; \quad n = 0, \ldots, 21$$

(Moran, Pritchard, Thyssen, 1995)

$$56211383760397 + 44546738095860n; \quad n = 0, \ldots, 22$$

(Frind, Underwood, Jobling, 2004)

It was conjectured for at least a century that there are arbitrarily long arithmetic progressions of primes; a more precise conjecture was that for any $k$, there is a progression of length $k$ of primes less than $k! + 1$.

In fact, modern heuristics predict one can lower $k! + 1$ to $(ke^{1-\gamma}/2)^{k(\frac{1}{2}+o(1))}$ (Granville 2006). We discuss upper bounds more at the end of the talk.

History of results and conjectures

- (Lagrange, Waring, 1770) An arithmetic progression of primes of length $k$ must have spacing divisible by all the primes less than $k$. [In particular, there are no infinitely long arithmetic progressions of primes.]
- **Hardy-Littlewood prime tuples conjecture** (1923) Gives an asymptotic prediction of how often a given additive prime pattern occur in the primes from 1 to $N$; would imply twin prime, Goldbach (at least for sufficiently large $n$), and give arbitrarily long progressions of primes. Totally open.
- **van der Waerden's theorem** (1927) If the integers are coloured using finitely many colours, then one of the colour classes must contain arbitrarily long arithmetic progressions. (For instance, either the primes or the non-primes contain arbitrarily long progressions.)
- **Erdős-Turán conjecture** (1936) Any set of positive integers whose sum of reciprocals diverges should contain arbitrarily long arithmetic progressions. [The sum of reciprocals of primes diverges (Euler, 1737).] Totally open; not even known if such a set must contain a progression of length three.

- (Van der Corput, 1939) There exist infinitely many arithmetic progressions of primes of length three. The Hardy-Littlewood asymptotic is also correct in this case.

- **Roth's theorem** (1956) Any subset of the integers of positive density contains infinitely many arithmetic progressions of length three. [The primes have density zero (Euler, 1737).]

- (Szemerédi, 1969) Any subset of the integers of positive density contains infinitely many arithmetic progressions of length four.

- **Szemerédi's theorem** (1975) Any subset of the integers of positive density contains arbitrarily long arithmetic progressions. [Implies van der Waerden's theorem.]

- (Heath-Brown, 1981) There are infinitely many arithmetic progressions of length four, where three elements are prime and one is an almost prime (the product of two primes).

- (Balog, 1992) For any $k$, there exist $k$ distinct primes $p_1, \ldots, p_k$, all of whose averages $\frac{p_i+p_j}{2}$ are also prime.

- **Green-Tao theorem** (2004) The prime numbers contain arbitrarily long arithmetic progressions.

- (Green, T. 2004) There exist infinitely many progressions of length three of Chen primes (primes $p$ where $p + 2$ is almost prime).

- (T., 2005) The Gaussian primes contain arbitrarily shaped constellations.

- (Green, T., 2006) The Hardy-Littlewood asymptotic is correct for progressions of length four in the primes, as well as other additive patterns of similar complexity. (The analogous result for longer progressions is a work in progress.)

- (T., Ziegler, 2006) Let $P_1, \ldots, P_k$ be any integer polynomials with zero constant coefficient. Then the prime numbers contain infinitely many polynomial progressions of the form $n + P_1(r), \ldots, n + P_k(r)$.

- Unfortunately, the twin prime and even Goldbach conjectures remain wide open (the above methods all seem to require the patterns to have at least two independent parameters).

Prime counting heuristics

- Experience has shown that it is not feasible to try to find prime patterns (or even individual primes) directly, for instance by some explicit formula. Instead, one should count the number of primes or prime patterns in some range (e.g. counting the number of twin primes from 1 to $N$). The main task is to get a non-trivial lower bound on this count.

- While our ability to count patterns in the primes is still limited in many ways, our ability to conjecture what this count should be is very good (and uncannily accurate).

- A basic starting point is the **prime number theorem** (Hadamard, de la Vallée Poussin, 1896), which says that for large numbers $N$, the number of primes between 1 and $N$ is roughly $N/\log N$ (or more accurately $\int_2^N \frac{dx}{\log x}$). Another way of thinking about it is that a number randomly selected from 1 to $N$ will have a probability approximately $1/\log N$ of being prime. [Exactly what "approximately" means is a good question - closely connected to the famous **Riemann hypothesis** - but we won't discuss it here.]

This already gives us a crude heuristic for counting patterns in primes. Suppose for instance one wants to prove the twin prime conjecture. One could argue as follows:

(1) Pick a number $n$ randomly from 1 to $N$.

(2) The prime number theorem shows that the probability that $n$ is prime is roughly $1/\log N$.

(3) The prime number theorem also shows that the probability that $n+2$ is prime is also roughly $1/\log N$.

(4) Assuming that the events in (2) and (3) are approximately independent, the probability that $n, n+2$ are both prime should be $1/\log^2 N$.

(5) In other words, the number of twin primes from 1 to $N$ should be roughly $N/\log^2 N$.

(6) Since $N/\log^2 N$ goes to infinity as $N \to \infty$, there are infinitely many twin primes.

- Unfortunately, the above argument is incorrect. One easy way to see this is that the exact same argument would show that there are also infinitely many pairs of adjacent primes $n, n+1$, which is clearly false!

- The problem is that the assumption of independence is too naive - one is basically hoping that the primes from 1 to $N$ are distributed in an utterly random (or more precisely, a pseudorandom) fashion, with there being no correlation between the primality of $n$ and the primality of (say) $n+2$. But this is not the case, because of a very simple observation:

  Odd numbers are much more likely to be prime than even numbers.

- Intuitively, this means that if $n$ is prime, then $n$ is most likely odd, and so $n+2$ is odd also. This should significantly increase the probability that $n+2$ is prime - so the two events are not independent. (Conversely, it dramatically decreases the probability that $n+1$ is prime.)

- While this invalidates our earlier line of reasoning, it is not hard to modify that argument to accomodate this new observation about the primes. The idea is to use conditional probability and independence rather than absolute probability and independence. From the prime number theorem, and the fact that almost all primes are odd, we have

(a) If $n$ is a random even number from 1 to $N$, then the probability that $n$ is prime is negligible.

(b) If $n$ is a random odd number from 1 to $N$, then the probability that $n$ is prime is roughly $2/\log N$.

Now we have a revised count for twin primes:

(1) Pick a number $n$ randomly from 1 to $N$. Approximately 1/2 of the time $n$ will be even; 1/2 of the time $n$ is odd.

(234a) If $n$ is even, then $n$ and $n+2$ have only a negligible chance of being prime, so the probability that $n, n+2$ are both prime should also be negligible (in fact it is zero).

(234b) If $n$ is odd, then $n$ and $n+2$ each have a probability of about $2/\log N$ of being prime, so (assuming "conditional independence") the probability that $n, n+2$ are both prime in this case should be about $4/\log^2 N$.

(5) Putting this all together (using Bayes' formula), the number of twin primes from 1 to $N$ should be roughly

$$N \times [\frac{1}{2} \times 0 + \frac{1}{2} \times \frac{4}{\log^2 N}] = 2\frac{N}{\log^2 N}.$$

(6) This still goes to infinity as $N \to \infty$, so there should still be infinitely many twin primes.

- Of course, this argument is also incorrect (though it is "less incorrect" than the previous one). For instance, it would predict infinitely many prime triples of the form $n, n+2, n+4$. The reason is that we are not incorporating some additional structural facts about the primes, in this case

  Numbers equal to 1 or 2 (mod 3) are much more likely to be prime than numbers equal to $0 (\mathrm{mod}\ 3)$.

- In fact, we have a more precise statement (Dirichlet 1837, Siegel-Walfisz, 1963):

(a) If $n$ is a random number from 1 to $N$ with $n = 0 (\mathrm{mod}\ 2)$ or $n = 0 (\mathrm{mod}\ 3)$, then $n$ has a negligible probability of being prime.

(b) If $n$ is a random number from 1 to $N$ with $n = 1 (\mathrm{mod}\ 2)$ and $n = 1 (\mathrm{mod}\ 3)$, then $n$ has probability roughly $3/\log N$ of being prime.

(c) If $n$ is a random number from 1 to $N$ with $n = 1 (\mathrm{mod}\ 2)$ and $n = 2 (\mathrm{mod}\ 3)$, then $n$ has probability roughly $3/\log N$ of being prime.

- Using this new information, we can revise our count of twin primes from $2\frac{N}{\log^2 N}$ to $\frac{3}{2}\frac{N}{\log^2 N}$.

- Of course, we can continue adjusting this count using mod 5 information, mod 7 information, etc. and obtain the following sequence of heuristics:

| Information used | Predicted # twins |
|---|---|
| Prime number theorem | $\approx \frac{N}{\log^2 N}$ |
| # primes mod 2 | $\approx 2\frac{N}{\log^2 N}$ |
| # primes mod $2, 3$ | $\approx 1.5\frac{N}{\log^2 N}$ |
| # primes mod $2, 3, 5$ | $\approx 1.41\frac{N}{\log^2 N}$ |
| # primes mod $2, 3, 5, 7$ | $\approx 1.37\frac{N}{\log^2 N}$ |
| # primes mod $2, 3, \ldots, 97$ | $\approx 1.32\frac{N}{\log^2 N}$ |

- One quickly observes that each new modulus is causing less and less of an adjustment, and the prediction for the number of twin primes less than $N$ in fact converges to $2\Pi_2\frac{N}{\log^2 N}$, where $\Pi_2$ is the twin prime constant

$$\Pi_2 = \prod_{p \text{ odd prime}} (1 - \frac{1}{(p-1)^2}) = 0.660161858\ldots.$$

- (Technical point) Actually $2\Pi_2 \int_2^N \frac{dx}{\log^2 x}$ is a slightly better prediction, as it uses the additional fact that small numbers are a bit more likely to be prime than large numbers. But this is a relatively minor correction.

This prediction is surprisingly good:

| $N$ | $2\Pi_2 \frac{N}{\log^2 N}$ | $2\Pi_2 \int_2^N \frac{dx}{\log^2 x}$ | Actual # twins $\leq N$ |
|---|---|---|---|
| $10^6$ | 6917 | 8248 | 8168 |
| $10^8$ | 389107 | 440368 | 440312 |
| $10^{10}$ | 2490284 | 27411417 | 27412679 |
| $10^{12}$ | $1.72936 \times 10^9$ | $1.87061 \times 10^9$ | $1.87059 \times 10^9$ |

Our heuristic analysis hinges on the presumption that, apart from the obvious structure in the primes (that the primes are mostly odd, mostly coprime to three, etc.), that the primes behave as if they were randomly distributed; in other words, there is no additional "secret" or "exotic" structure in the primes that would significantly affect such counts as the number of twin primes less than $N$.

There is a way to make this presumption rigorous; this is known as the **Hardy-Littlewood prime tuples conjecture**. This very strong conjecture would allow us to count virtually any type of arithmetic pattern in the primes, settling many open questions; but we have no way of attacking this conjecture with current technology. (How could one disprove a "conspiracy" among the primes?)

Almost primes

- While we cannot settle many questions about the primes, we have a much better understanding of the almost primes - numbers which are the product of only a small number of primes. Roughly speaking, the analogue of the Hardy-Littlewood prime tuples conjecture is known for almost primes (this fact is known as the **fundamental lemma of sieve theory**).

- To explain why almost primes are easier to control than genuine primes, let us recall the classical sieve of Eratosthenes. This sieve lets us locate all the primes between, say, $N/2$ and $N$ for some large number $N$, by the following procedure.

(0) Start with all the numbers from $N/2$ to $N$.

(2) Eliminate (or "sieve out") all multiples of 2.

(3) Eliminate all multiples of 3.

(5) Eliminate all multiples of 5. ...

($\sqrt{N}$) After all multiples of primes less than $\sqrt{N}$ are sieved out, one is left with the primes from $N/2$ to $N$.

- One can think of this sieve as "sculpting" the primes out of a big block (the integers from $N/2$ to $N$). At the beginning of the process, one removes very large and "smooth" pieces from this block (the multiples of 2, multiples of 3, etc.), and it is easy to see what is going on. For instance, we initially have roughly $\frac{N}{2}$ elements; after removing the even numbers we should have roughly $\frac{1}{2} \cdot \frac{N}{2}$ elements; after then removing the multiples of 3 we should have roughly $\frac{2}{3} \cdot \frac{1}{2} \cdot \frac{N}{2}$ elements (by the **Chinese remainder theorem**) and so forth. It is also easy to count the number of twins, arithmetic progressions, etc. at the very early stages of this process.

- However, at later stages of the sieve (e.g. at the steps between $\sqrt{N}/2$ and $\sqrt{N}$) one is performing a very large number of tiny modifications to the sculpture, removing small amounts of non-primes at a time in what appears to be a rather random process. At this stage we tend to lose all control of what is happening to this sculpture - for instance, nobody has figured out how to use sieve ideas to give a proof of the prime number theorem, let alone anything more sophisticated such as count twin primes.

- If however one stops the sieve before going all the way up to $\sqrt{N}$ - say if one only sieves up to level $N^{1/100}$ instead - then we have been able to keep control of everything. (Actually we have to make the sieve more fancy to do so, but let us ignore this technicality.) The catch, of course, is that the sieve now contains almost primes in addition to genuine primes - in this case, we still have numbers with up to 100 prime factors. However, it turns out that the set of almost primes are only mildly larger than the set of actual primes; whereas the number of actual primes from 1 to $N$ is roughly $N/\log N$, the number of almost primes obtained by sieving up to level $N^{1/100}$ is something like $100N/\log N$.

- Recent work of Goldston and Yıldırım has made these heuristics quite precise; as one striking application, it is now known that the gap $p_{n+1} - p_n$ between adjacent primes can be as small as $o(\log p_n)$ infinitely often (Goldston, Yıldırım, Pintz, 2005)

To summarise so far:

- Prime numbers have some obvious structure (they are mostly odd, coprime to 3, etc.) We don't know if they also have some additional exotic structure. Because of this, we have been unable to settle many questions about primes.
- Almost primes (such as those generated by a partial sieve of Eratosthenes) have the same obvious structure as the primes, but are known to be otherwise randomly distributed (in the sense that things like the number of almost twin primes matches the natural heuristics). The number of almost primes exceeds the number of actual primes by a constant factor (such as 100, depending on one's precise definition of "almost prime").
- In general, nobody has figured out how to use this information on almost primes to deduce information on actual primes. However, it turns out that for a few special types of patterns - most notably arithmetic progressions - it is possible to pull this off. This is because of a deep and powerful theorem known as **Szemerédi's theorem**.

- **Szemerédi's theorem** (1975) Let $A$ be a subset of the integers $\mathbf{Z}$ of positive (upper) density. Then $A$ contains arbitrarily long arithmetic progressions.

- This result was conjectured by Erdős and Turán in 1936. The remarkable thing about this theorem is that one is given almost no information on $A$ other than that it is large, and yet this is still enough to force $A$ to contain arbitrarily long progressions. In contrast, patterns such as twins $n, n+2$ do not have this property; for instance, the multiples of 3 has a density of $1/3$ but contains no twins.

- This important theorem now has many different proofs: a combinatorial proof (Szemerédi, 1975), an ergodic theory proof (Furstenberg, 1977), a Fourier-analytic proof (Gowers, 1998), and a hypergraph proof (Nagle-Rödl-Schacht-Skokan/Gowers 2005). All the proofs are non-trivial. This is ultimately because there are two extreme cases for what a set $A$ can look like (structured and pseudorandom), and in each case the arithmetic progressions must be obtained in different ways.

Progressions in the primes

- Now we can sketch how one finds progressions in the primes.

- It is convenient to work with the von Mangoldt function $\Lambda : \mathbf{N} \to \mathbf{R}$, defined by setting $\Lambda(n) := \log p$ when $n$ is a power of a prime $p$, and zero otherwise. This function can be thought of as a weight function for the primes; it is natural because of the identity

  $$\sum_{d|n} \Lambda(d) = \log n \text{ for all } n = 1, 2, 3, \ldots$$

  (this is basically the fundamental theorem of arithmetic in disguise).

- To find progressions of length $k$ in the primes, one is basically led to try to obtain lower bounds for averages such as

  $$\frac{1}{N^2} \sum_{n=1}^{N} \sum_{r=1}^{N} \Lambda(n)\Lambda(n+r)\ldots\Lambda(n+(k-1)r).$$

  In contrast, Szemerédi's theorem allows us to lower-bound quantities such as

  $$\frac{1}{N^2} \sum_{n=1}^{N} \sum_{r=1}^{N} f(n)f(n+r)\ldots f(n+(k-1)r)$$

but only when $f$ is bounded, non-negative, and has large mean. $\Lambda$ is non-negative with large mean, but is unfortunately not bounded.

- However, using a structure theorem (motivated by arguments in ergodic theory and combinatorics) one can split $\Lambda$ into a structured component $\Lambda_{U^\perp}$, and a pseudorandom component $\Lambda_U$ (the definition of these terms was motivated by arguments in Fourier analysis). The pseudorandom component $\Lambda_U$ turns out to be negligible (motivated by arguments in hypergraph theory). The task is then to understand the structured component $\Lambda_{U^\perp}$.
- Now we use the information about the almost primes. There is an analogue of the von Mangoldt function for the almost primes, which we call $\nu$; it is a little bit bigger than $\Lambda$. $\nu$ also splits into a structured part $\nu_{U^\perp}$ and a pseudorandom part $\nu_U$, but because the almost primes are so pseudorandomly distributed (except for some obvious irregularities due to mod 2, mod 3, etc. which can be easily dealt with), the structured part $\nu_{U^\perp}$ of $\nu$ turns out to be very simple - in fact, it is basically constant. There is then a comparison principle that implies that the structured part $\Lambda_{U^\perp}$ of $\Lambda$ is bounded - which lets one apply Szemerédi's

theorem.

Quantitative bounds

- All the arguments here can be made quantitative, and we in fact know that the first progression of primes of length $k$ has entries less than

$$2^{2^{2^{2^{2^{2^{2^{100k}}}}}}}.$$

- Recall that the conjectured upper bound is in fact $k! + 1$ (or $(ke^{1-\gamma}/2)^{k(\frac{1}{2}+o(1))}$).
- This is a tower of seven exponentials. Two of them come from the number-theoretic pseudorandomness bounds on the almost primes. Four come from the best known bounds on Szemerédi's theorem (from work of Gowers). The last one comes from the structure theorem.

- If the Riemann hypothesis is true, we can remove one exponential.