STRUCTURE AND RANDOMNESS IN THE PRIME NUMBERS

TERENCE TAO

Abstract. A quick tour through some topics in analytic prime number theory.

1. Introduction

The prime numbers $2, 3, 5, 7, \ldots$ are one of the oldest topics studied in mathematics.

We now have a lot of intuition as to how the primes *should* behave, and a great deal of confidence in our conjectures about the primes... but we still have a great deal of difficulty in *proving* many of these conjectures!

Ultimately, this is because the primes are believed to not obey behave *pseudorandomly* in many ways, and not to follow any simple pattern.

We have many ways of establishing that a pattern exists... but how does one demonstrate the *absence* of a pattern?

In this article I will try to convince you why the primes are believed to behave pseudorandomly, and how one could try to make this intuition rigorous. This is only a small sample of what is going on in the subject; I am omitting many major topics, such as sieve theory or exponential sums, and am glossing over many important technical details.

2. Finding primes

It is a paradoxical fact that the primes are simultaneously very numerous, and hard to find. On the one hand, we have the following ancient theorem:

Theorem 2.1 (Euclid's theorem)**.** [2] *There are infinitely many primes.*

In particular, given any k, there exists a prime with at least k digits.

But there is no known *quick* and *deterministic* way to locate such a prime! (Here, "quick" means "computable in a time which is polynomial in k".)

In particular, there is no known (deterministic) formula that can quickly generate large numbers that are guaranteed to be prime.

The largest known prime is $2^{43,112,609} - 1$ [3] - about 13 million digits long.

On the other hand, one can find primes quickly by *probabilistic* methods.

Indeed, any k-digit number can be tested for primality quickly, either by probabilistic methods[9, 11] or by deterministic methods (Agarwal-Kayal-Saxena 2002). These methods are based on variants of Fermat's little theorem, which asserts that $a^n = a \bmod n$ whenever n is prime. (Note that if n is a k-digit number, $a^n \bmod n$ can be computed quickly, by first repeatedly squaring a to compute $a^{2^j} \bmod n$ for various values of j, and then expanding n in binary and multiplying the indicated residues $a^{2^j} \bmod n$ together.)

Also, we have the following fundamental theorem:

Theorem 2.2 (Prime number theorem)**.** [8, 13] *The number of primes less than a given integer n is $(1 + o(1))\frac{n}{\log n}$, where $o(1)$ tends to zero as $n \to \infty$.*

In particular, the probability of a randomly selected k-digit number being prime is about $\frac{1}{k \log 10}$.

So one can quickly find a k-digit prime with high probability by randomly selecting k-digit numbers and testing each of them for primality.

2.3. **Is randomness really necessary?** To summarize: We do not know a quick way to find primes *determinstically.* However, we have quick ways to find primes *randomly.*

On the other hand, there are major conjectures in complexity theory, such as $P = BPP$, which assert (roughly speaking) that any problem that can be solved quickly by probabilistic methods, can also be solved quickly by deterministic methods. (Strictly speaking, the $P = BPP$ conjecture only applies to *decision problems* - problems with a yes/no answer - rather than *search problems* such as the task of finding a prime, but there are variants of $P = BPP$, such as $P = promise - BPP$, which would be applicable here.

These conjectures are closely related to the more famous conjecture $P \neq NP$, which is a USD \$ 1 million Clay Millennium prize problem.

Many other important probabilistic algorithms have been *derandomised* into deterministic ones, but this has not been done for the problem of finding primes. (A massively collaborative research project is currently underway to attempt this[10].)

3. Counting primes

We've seen that it's hard to get a hold of any single large prime. But it is easier to study the set of primes *collectively* rather than one at a time.

An analogy: it is difficult to locate and count all the grains of sand in a box, but one can get an estimate on this count by *weighing* the box, subtracting the weight of the empty box, and dividing by the average weight of a grain of sand. The point is that there is an easily measured statistic (the weight of the box) which is reflects the *collective* behaviour of the sand.

For instance, from the *fundamental theorem of arithmetic* one can establish *Euler's product formula*

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{n^s} &= \prod_{p \text{ prime}} (1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \dots) \\ &= \prod_{p \text{ prime}} (1 - \frac{1}{p^s})^{-1} \end{aligned} \tag{1}$$

for any $s > 1$ (and also for other values of s, if one defines one's terms carefully enough).

The formula (1) links the collective behaviour of the primes to the behaviour of the *Riemann zeta function*

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s},$$

thus

$$\prod_{p \text{ prime}} (1 - \frac{1}{p^s}) = \frac{1}{\zeta(s)} \tag{2}$$

One can then deduce information about the primes from information about the zeta function (and in particular, its zeroes).

For instance, from the divergence of the harmonic series $\sum_{n=1}^\infty \frac{1}{n} = +\infty$ we see that $\frac{1}{\zeta(s)}$ goes to zero as s approaches 1 (from the right, at least). From this and (2) we already recover Euclid's theorem (Theorem 2.1), and in fact obtain the stronger result of Euler that the sum $\sum_p \frac{1}{p}$ of reciprocals of primes diverges also.

In a similar spirit, one can use the techniques of complex analysis, combined with the (non-trivial) fact that $\zeta(s)$ has no zeroes when $\mathrm{Re}(s) \geq 1$, to establish the prime number theorem (Theorem 2.2); indeed, this is how the theorem was originally proved (and one can conversely use the prime number theorem to deduce the fact about the zeroes of ζ).

The famous *Riemann hypothesis* asserts that $\zeta(s)$ has no zeroes when[1] $\mathrm{Re}(s) > 1/2$. It implies a much stronger version of the prime number theorem, namely that the number of primes less than an integer $n > 1$ is given by the more precise formula $\int_0^n \frac{dx}{\log x} + O(n^{1/2}\log n)$, where $O(n^{1/2}\log n)$ is a quantity which is bounded in magnitude by $Cn^{1/2}\log n$ for some absolute constant C (for instance, one can take $C = \frac{1}{8\pi}$ once n is at least 2657 [12]). The hypothesis has many other consequences in number theory; it is another of the USD $ 1 million Clay Millennium prize problems. More generally, much of what we know about the primes has come from an extensive study of the properties of the Riemann zeta function and its relatives, although there are also some questions about primes that remain out of reach even assuming strong conjectures such as the Riemann hypothesis.

4. Modeling primes

A fruitful way to think about the set of primes is as a *pseudorandom set* - a set of numbers which is not actually random, but behaves like one.

For instance, the prime number theorem asserts, roughly speaking, that a randomly chosen large integer n has a probability of about $1/\log n$ of being prime. One can then *model* the set of primes by replacing them with a random set of integers, in which each integer $n > 1$ is selected with an independent probability of $1/\log n$; this is *Cramér's random model.*

[1] A technical point: the sum $\sum_{n=1}^\infty \frac{1}{n^s}$ does not converge in the classical sense when $\mathrm{Re}(s) < 1$, so one has to interpret this sum in a fancier way, or else use a different definition of $\zeta(s)$ in this case; but I will not discuss these subtleties here.

This model is too crude, because it misses some obvious structure in the primes, such as the fact that most primes are odd. But one can improve the model to address this, by picking a model where odd integers n are selected with an independent probability of $2/\log n$ and even integers are selected with probability 0.

One can also take into account other obvious structure in the primes, such as the fact that most primes are not divisible by 3, not divisible by 5, etc. This leads to fancier random models which we believe to accurately predict the asymptotic behaviour of primes.

For example, suppose we want to predict the number of twin primes $n, n+2$ less than a given threshold N. Using the Cramér random model, we expect, for any given n, that $n, n+2$ will simultaneously be prime with probability $\frac{1}{\log n \log(n+2)}$, so we expect the number of twin primes to be about

$$\sum_{n=1}^{N} \frac{1}{\log n \log(n+2)} \approx \frac{N}{\log^2 N}.$$

This prediction is inaccurate; for instance, the same argument would also predict plenty of pairs of *consecutive* primes $n, n+1$, which is absurd. But if one uses the refined model where odd integers are prime with an independent probability of $2/\log N$ and even integers are prime with probability 0, one gets the slightly different prediction

$$\sum_{1 \leq n \leq N : n \text{ odd}} \frac{2}{\log n} \times \frac{2}{\log(n+2)} \approx 2\frac{N}{\log^2 N}.$$

More generally, if one assumes that all numbers n divisible by some prime less than a small threshold w are prime with probability zero, and are prime with a probability of $\prod_{p<w}(1-\frac{1}{p})^{-1} \times \frac{1}{\log N}$ otherwise, one is eventually led to the prediction

$$2 \prod_{p<w, p \text{ an odd prime}} (1-\frac{1}{p^2}) \times \frac{N}{\log^2 N}$$

Sending $w \to \infty$, one is led to the asymptotic prediction

$$\Pi_2 \frac{N}{\log^2 N}$$

for the number of twin primes less than N, where Π_2 is the *twin prime constant*

$$\Pi_2 := 2 \prod_{p \geq 3 \text{ prime}} 1 - \frac{1}{(p-1)^2} \approx 1.32032\ldots.$$

For $N = 10^{10}$, this prediction is accurate to four decimal places, and is believed to be asymptotically correct. (This is part of a more general conjecture, known as the *Hardy-Littlewood prime tuples conjecture.*)

Similar arguments based on random models give convincing heuristic support for many other conjectures in number theory, and are backed up by extensive numerical calculations.

5. Finding patterns in primes

Of course, the primes are a deterministic set of integers, not a random one, so the predictions given by random models are not rigorous. But can they be made so?

There has been some progress in doing this. One approach is to try to classify all the possible ways in which a set could *fail* to be pseudorandom (i.e. it does something noticeably different from what a random set would do), and then show that the primes do not behave in any of these ways.

For instance, consider the **odd Goldbach conjecture**: every odd integer larger than five is the sum of three primes. If, for instance, all large primes happened to have their last digit equal to one, then Goldbach's conjecture could well fail for some large odd integers whose last digit was different from three. Thus we see that the conjecture could fail if there was a sufficiently strange "conspiracy" among the primes.

However, one can rule out this particular conspiracy by using the *prime number theorem in arithmetic progressions*, which tells us that (among other things) there are many primes whose last digit is different from 1. (The proof of this theorem is based on the proof of the classical prime number theorem.)

Moreover, by using the techniques of *Fourier analysis* (or more precisely, the *Hardy-Littlewood circle method*), we can show that *all* the conspiracies which could conceivably sink Goldbach's conjecture (for large integers, at least) are broadly of this type: an unexpected "bias" for the primes to prefer one remainder modulo 10 (or modulo another base, which need not be an integer), over another.

Vinogradov[14] eliminated each of these potential conspiracies, and established *Vinogradov's theorem*: every sufficiently large odd integer is the sum of three primes. This method has since been extended by many authors, to cover many other types of patterns; for instance, related techniques were used by Ben Green and myself[4] to establish that the primes contained arbitrarily long arithmetic progressions, and in subsequent work of Ben Green, myself, and Tamar Ziegler [5], [6], [7] to count a wide range of other additive patterns also. (Very roughly speaking, known techniques can count additive patterns that involve two independent parameters, such as arithmetic progressions $a, a+r, \ldots, a+(k-1)r$ of a fixed length k.)

Unfortunately, "one-parameter" patterns, such as twins $n, n+2$, remain stubbornly beyond current technology. There is still much to be done in the subject!

References

[1] M. Agrawal, N. Kayal, N. Saxena, *PRIMES is in P*, Annals of Mathematics **160** (2004), no. 2, pp. 781-793.

[2] Euclid, *The Elements*, *circa* 300 BCE.

[3] Great Internet Mersenne Prime Search, 2008. http://www.mersenne.org

[4] B. Green, T. Tao, *The primes contain arbitrarily long arithmetic progressions*, Annals of Math. **167** (2008), 481-547

[5] B. Green, T. Tao, *Linear equations in primes*, preprint.

[6] B. Green, T. Tao, *The Möbius function is asymptotically orthogonal to nilsequences*, preprint.

[7] B. Green, T. Tao, T. Ziegler, *The inverse conjecture for the Gowers norm*, preprint.

[8] J. Hadamard, *Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques*, Bull. Soc. Math. France, 24 (1896), 199-220.

[9] G. Miller, *Riemann's Hypothesis and Tests for Primality*, Journal of Computer and System Sciences **13** (1976), no. 3, 300-317.

[10] Polymath4 project: Deterministic way to find primes. http://michaelnielsen.org/polymath1/index.php?title=Finding_primes

[11] M. Rabin, *Probabilistic algorithm for testing primality*, Journal of Number Theory **12** (1980), 128-138.

[12] L. Schoenfeld, *Sharper bounds for the Chebyshev functions $\theta(x)$ and $\psi(x)$. II*, Mathematics of Computation 30 (1976), 337-360.

[13] C.-J. de la Vallée Poussin, *Recherches analytiques la thorie des nombres premiers.* Ann. Soc. scient. Bruxelles 20 (1896), 183–256.

[14] I. M. Vinogradov, *The Method of Trigonometrical Sums in the Theory of Numbers* (Russian). Trav. Inst. Math. Stekloff 10 (1937).

Department of Mathematics, UCLA, Los Angeles CA 90095-1555

E-mail address: tao@math.ucla.edu