

Sieving and upper bounds for the number of twin primes

January 24, 2007

1 Introduction and heuristics

A pair of positive integers $p, p + 2$ is called “twin prime pair” (or just “twin primes”) if they are both prime numbers; for example, 3, 5 and 11, 13 are twin prime pairs. A famous unsolved problem in number theory is that of proving that there are infinitely many twin primes.

From probabilistic grounds alone we would expect that there are infinitely many: Suppose that we have a sequence of independent random variables

$$Z_1, \dots, Z_x,$$

where

$$Z_i = \begin{cases} 1, & \text{with probability } 1/\log x; \\ 0, & \text{with probability } 1 - 1/\log x. \end{cases}$$

Think of these random variable as picking out prime numbers – when Z_n is 1 we have n is prime, and if it is 0, then n is composite. Of course this is a silly thing to say, because a number is either prime or it isn’t, and so there is no probability at all. Nonetheless, this model is often indicative of roughly what is true, and in our case we would have that $n, n + 2$ are both “prime” with probability $1/\log^2 x$; so, we would expect there are about $x/\log^2 x$ twin prime pairs.

One failing of our model is that it would also predict that there are about $x/\log^2 x$ pairs $n, n + 1$ that are prime, which is clearly false (only 2, 3 is such a pair of primes). What goes awry in this example is “divisibility by 2”. We

could add into our model all such divisibility constraints, but we will instead look at the problem in a different way: First, define the polynomial

$$f(z) = z(z + 2),$$

and consider the sequence

$$f(1), f(2), \dots, f(x).$$

It is not hard to see that the number of twin prime pairs in $(\sqrt{x}, x]$ is the number of integers $n \leq x$ such that $f(n)$ is not divisible by any prime $p \leq \sqrt{x}$.

We note that $p \nmid f(n)$ if and only if

$$f(n) = n(n + 2) \not\equiv 0 \pmod{p}$$

For $p = 2$ this condition means that $n \not\equiv 0 \pmod{2}$, but for all the other primes $p \geq 3$ we have that the condition is $n \not\equiv 0, -2 \pmod{p}$.

We can think of the problem of locating twin prime pairs in $(\sqrt{x}, x]$ as a sieve process: We eliminate those n that are $0 \pmod{2}$, and then we eliminate those that $0, -2 \pmod{3}$, and so on up to $0, -2 \pmod{P}$ where P is the largest prime $\leq \sqrt{x}$. Given a prime p let

$$w(p) = \begin{cases} 1, & \text{if } p = 2; \\ 2, & \text{if } p \geq 3. \end{cases}$$

If the primes acted “independently” with respect to the sieving process, then we would expect that there are about

$$x \prod_{\substack{p \leq \sqrt{x} \\ p \text{ prime}}} \left(1 - \frac{w(p)}{p}\right) \tag{1}$$

twin prime pairs $n, n + 2 \leq x$, since each of these factors $1 - w(p)/p$ is the proportion of integers that remain after the sieving process with the prime p is applied.

It turns out that our guess as to the number of twin primes $\leq x$ given in (1) is actually wrong – it is off by a certain constant factor. The reason is that the primes do not really act independently with respect to sieving. We have already seen this in class in regards to trying to count the number of

primes in $(\sqrt{x}, x]$ using a sieve: If the primes did act independently, then we should expect that there are

$$\sim x \prod_{\substack{p \leq \sqrt{x} \\ p \text{ prime}}} \left(1 - \frac{1}{p}\right) \sim \frac{2e^{-\gamma}x}{\log x}$$

primes in $(\sqrt{x}, x]$, where γ is Euler's constant, and is given by

$$\gamma = \lim_{x \rightarrow \infty} \left(\sum_{1 \leq n \leq x} \frac{1}{n} - \log x \right).$$

Here we are using Mertens's theorem, which says that

$$\prod_{p \leq y} \left(1 - \frac{1}{p}\right) \sim \frac{e^{-\gamma}}{\log y}.$$

So, in order to have a good heuristic for the number of twin prime pairs $\leq x$ we need to multiply our guess (1) by an appropriate "correction factor". What should such a correction factor be? Well, when we were sieving to find primes the factor was

$$\lim_{x \rightarrow \infty} \frac{1}{\log x} \prod_{\substack{p \leq \sqrt{x} \\ p \text{ prime}}} \left(1 - \frac{1}{p}\right)^{-1} \sim e^{\gamma}/2.$$

However, when we sieve for when both n and $n + 2$ are prime, the factor should be the square of that (basically, one correction factor for each of the numbers n and $n + 2$), which is

$$\lim_{x \rightarrow \infty} \frac{1}{\log^2 x} \prod_{\substack{p \leq \sqrt{x} \\ p \text{ prime}}} \left(1 - \frac{1}{p}\right)^{-2}.$$

If we now multiply this by our guess (1), we get

$$\begin{aligned} \sim \frac{2x}{\log^2 x} \prod_{\substack{3 \leq p \leq \sqrt{x} \\ p \text{ prime}}} \frac{\left(1 - \frac{2}{p}\right)}{\left(1 - \frac{1}{p}\right)^2} &= \frac{2x}{\log^2 x} \prod_{\substack{3 \leq p \leq \sqrt{x} \\ p \text{ prime}}} \left(1 - \frac{1}{p^2(1-1/p)^2}\right) \\ &= \frac{2x}{\log^2 x} \prod_{\substack{3 \leq p \leq \sqrt{x} \\ p \text{ prime}}} \left(1 - \frac{1}{(p-1)^2}\right). \end{aligned}$$

Since this product over primes converges if we let $x \rightarrow \infty$, we have the following heuristic for the number of twin prime pairs $\leq x$:

$$\sim \frac{2x}{\log^2 x} \prod_{\substack{p \geq 3 \\ p \text{ prime}}} \left(1 - \frac{1}{(p-1)^2}\right).$$

This heuristic was first worked out by Hardy and Littlewood, and is part of a more general conjecture called the ‘‘Hardy-Littlewood Conjecture’’.

2 Sums of reciprocals of twin primes

If the twin primes had the above counting function, then we would have that

$$\sum_{\substack{p, p+2 \\ \text{prime}}} \frac{1}{p} \text{ converges,} \tag{2}$$

as can be seen via an integral test upon noting that

$$\int_2^\infty \frac{dt}{t \log^2 t} = \frac{1}{\log 2}.$$

In fact, even if we had a much worse upper bound on the number of twin primes, we would get that this sum of reciprocals converges; in particular, an upper bound of something like

$$\pi_2(x) < \frac{x}{(\log x)(\log \log x)^2},$$

where $\pi_2(x)$ is the number of twin prime pairs $\leq x$, would be enough to show (2).

2.1 A simpleminded approach

A simple approach is to do sieving in such a way that you get the exact count for the number of integers left once the sieve terminates: An obvious way to do this is to let N be the product of primes $2, 3, \dots, p_k$, where p_k is the largest prime such that $2 \cdot 3 \cdots p_k \leq x$.¹ Then, we do the following

¹Note here that, by the Prime Number Theorem,

$$p_k \sim \log x.$$

- Take all the numbers $1, 2, \dots, N$, and eliminate those n in the list for which $n(n+2)$ is divisible by 2 – that is, we eliminate all those n that are even.
- Then, taken the remaining numbers in the list, and remove all those n such that $n(n+2)$ is divisible by 3 – that is, eliminate all those n that are $\equiv 0, -2 \pmod{3}$.

⋮

- Finally, remove all those n where $p_k | n(n+2)$.

Whatever numbers that remain, they must include the set of twin primes that lie in $(\sqrt{N}, N]$; so, the number of integers that remain gives us an upper bound on the number of twin primes in $(\sqrt{N}, N]$. The question is: Just how many numbers will be left after we perform this sieve ?

Well, the numbers that remain after the sieve has finished will be all those n such that

$$\begin{aligned}
 n &\not\equiv 0 \pmod{2} \\
 &\not\equiv 0, -2 \pmod{3} \\
 &\not\equiv 0, -2 \pmod{5} \\
 &\vdots \\
 &\not\equiv 0, -2 \pmod{p_k}.
 \end{aligned}$$

Since N is the product of all these primes we have by the Chinese Remainder Theorem that are

$$N \prod_{\substack{2 \leq p \leq p_k \\ p \text{ prime}}} \left(1 - \frac{w(p)}{p}\right), \tag{3}$$

numbers that remain.

It is not difficult to prove that (upon taking logs and playing around with series)

$$\prod_{\substack{p \leq y \\ p \text{ prime}}} \left(1 - \frac{2}{p}\right) > \kappa_1 \prod_{\substack{p \leq y \\ p \text{ prime}}} \left(1 - \frac{1}{p}\right)^2 \sim \frac{\kappa_1 e^{-2\gamma}}{\log y},$$

for a certain constant $\kappa_1 > 0$; and so, we deduce that the quantity in (3) is at least

$$\frac{\kappa_2 N}{(\log \log N)^2}, \text{ for a certain } \kappa_2 > 0. \quad (4)$$

So, even though we get the exact count of the number of integers that remain after the sieve finishes, the upper bound we get is quite poor! In fact, we can easily give a much stronger upper bound just by noting that the number of twin primes in $(\sqrt{N}, N]$ is at most the number of primes in this interval, and so by the Prime Number Theorem we would get the upper bound of $\sim N/\log N$ – much better than (4).

Clearly, what goes awry in trying to give a good upper bound on the number of twin primes using this method, is that we are not using that many primes in our sieve – we are only using the primes up to about $\log x$, instead of those up to about \sqrt{x} .

2.2 How many primes do we need to sieve with?

We saw that if we could achieve the upper bound of

$$\pi_2(x) < \frac{x}{(\log x)(\log \log x)^2},$$

then we could prove that the sum of reciprocals of twin primes converges. It is natural to consider just how many primes we would need to sieve with, if we were to use the method in the previous subsection, in order to achieve this bound. Well, we clearly would need to find y so that

$$\prod_{\substack{3 \leq p \leq y \\ p \text{ prime}}} \left(1 - \frac{2}{p}\right) < \frac{1}{(\log x)(\log \log x)^2}.$$

Since the left-hand-side is at most $c/\log^2 y$, we would need to solve for y in the equation

$$\log^2 y = c^{-1}(\log x)(\log \log x)^2.$$

So, we would need that

$$y = \exp\left((\log \log x)\sqrt{c^{-1} \log x}\right).$$

In fact, it suffices to use

$$y = B = B(x) := \exp\left((\log \log x)\sqrt{\log x}\right).$$

This is far and away more primes than the $\log x$ we used in the simple approach!

2.3 The Combinatorial Sieve

When we tried to give an upper bound on the number of integers left unsieved, we got a terrible upper bound; and, if one plays around with the idea from the previous subsection, one realizes that no simple modification of that idea will give anything better. So, what do we do now? How do we incorporate more primes into our sieve? One answer is to do “intelligent inclusion-exclusion”, in the form of what are called the Bonferroni inequalities. In order to set up the discussion, first suppose that we have a sequence of finite sets S_1, \dots, S_k . Then, as is well-known, we have the standard inclusion-exclusion formula

$$|S_1 \cup \dots \cup S_k| = \sum_{j=1}^k (-1)^{j+1} \sum_{1 \leq i_1 < i_2 < \dots < i_j \leq k} |S_{i_1} \cap \dots \cap S_{i_j}|.$$

The Bonferroni inequalities tell us that we can get an approximation to the size of this union if we truncate the sums at some point. Specifically, we have

Bonferroni Inequalities.

$$\begin{aligned} |S_1 \cup \dots \cup S_k| &\geq \sum_{1 \leq j \leq 2r} (-1)^{j+1} \sum_{1 \leq i_1 < i_2 < \dots < i_j \leq k} |S_{i_1} \cap \dots \cap S_{i_j}|; \text{ and,} \\ |S_1 \cup \dots \cup S_k| &\leq \sum_{1 \leq j \leq 2r-1} (-1)^{j+1} \sum_{1 \leq i_1 < i_2 < \dots < i_j \leq k} |S_{i_1} \cap \dots \cap S_{i_j}|. \end{aligned}$$

The proofs of these inequalities amount to considering the contribution of a single element in $S_1 \cup \dots \cup S_k$ to both sides. Every element in the union has contribution 1 to the left-hand-side, and with some work (involving some basic binomial coefficient identities, which I worked out in class) one can show

that it has contribution ≤ 1 to the right-hand-side of the first inequality, and ≥ 1 to the second inequality. Hence, we get the bounds as claimed.

In our case, we will suppose that $p_1, \dots, p_k \geq 3$ are all the primes $\leq B$, and then we define

$$S_i := \{n \leq x : p_i | n(n+2)\}.$$

Note that we are throwing away the prime 2 – the reason is to make the exposition a little easier to read (we can throw away any finite set of primes from our sieve, and it will only affect our upper bound by a constant factor).

In order to apply the Bonferroni inequalities we will need to have decent estimates for the sizes of the various set intersections: We have that

$$n \in S_{i_1} \cap \dots \cap S_{i_t} \iff n \leq x, \text{ and for every } j = 1, \dots, t, n \equiv 0, -2 \pmod{p_{i_j}}. \quad (5)$$

To count this right-most quantity, break $[1, x]$ up into consecutive intervals of width $\Delta = p_{i_1} \cdots p_{i_t}$, with possibly an incomplete interval at the end if Δ does not divide x .² The integers in each of the “complete intervals” form a complete system of residue classes modulo Δ ; and, in each of these complete intervals we have that there are exactly $2^t = \tau(\Delta)$ (recall that $\tau(n)$ is the number of divisors of n – in the case where n is a product of t distinct primes we have $\tau(n) = 2^t$) integers n satisfying the congruences in (5). The single “incomplete interval” contains at most $\tau(\Delta)$ integers n satisfying (5). Since there are $\lfloor x/\Delta \rfloor$ complete intervals, and either 0 or 1 incomplete intervals, we get that

$$|S_{i_1} \cap \dots \cap S_{i_t}| = \tau(\Delta)(\lfloor x/\Delta \rfloor + O(1)) = \frac{x\tau(\Delta)}{\Delta} + O(\tau(\Delta)).$$

Now, which of the Bonferroni inequalities do we use? Well, an upper bound for the number of twin prime pairs $n, n+2 \leq x$ (well, up to an error $O(B)$ because our sieve also eliminates twin primes pairs $\leq B$) would be the size of complement of the union $S_1 \cup \dots \cup S_k$ within $[1, x]$, because this union gives all those n that we want to sieve out. In other words, we want an upper bound for

$$x - |S_1 \cup \dots \cup S_k|, \quad (6)$$

²Actually, if $\Delta > x$, then the entire interval $[1, x]$ will be an “incomplete interval”.

which amounts to a lower bound for $|S_1 \cup \dots \cup S_k|$. So, we use the first Bonferoni inequality. Before we do that, let us note something first, which makes the identity easier to work with: We have that

$$(-1)^t \sum_{1 \leq i_1 < \dots < i_t \leq k} |S_{i_1} \cap \dots \cap S_{i_t}| = \sum_{\substack{d|p_1 \dots p_k \\ \omega(d)=t}} \mu(d) \left(\frac{x\tau(d)}{d} + O(\tau(d)) \right).$$

Recall here that “ $\omega(d) = t$ ” means that d has exactly t prime factors; and, in the case d square-free (which is true for all the d in the above sum), having t prime factors, we have that $(-1)^t = \mu(d)$.

So, we have that

$$\pi_2(x) - \pi_2(B) \leq \sum_{\substack{d|p_1 \dots p_k \\ \omega(d) \leq 2r}} \mu(d) \left(\frac{x\tau(d)}{d} + O(\tau(d)) \right). \quad (7)$$

Note that the term $d = 1$ corresponds to the term x in (6) – how convenient!

What value of r should we take to get good bounds? Well, this can be worked out precisely; however, just to make a long story short, let us just take $2r$ to be the largest even integer satisfying

$$2r \leq \frac{\sqrt{\log x}}{(\log \log x)}.$$

When we do this, the error incurred by summing the $O(\tau(d))$ term above over all $d|p_1 \dots p_k$ with $\omega(d) \leq 2r$, will be at most

$$2^{2r} \sum_{1 \leq j \leq 2r} \binom{k}{j}$$

Since k is so much larger than $2r$, the term $j = 2r$ will dominate all the other terms in the binomial coefficient sum; and so, we get that for x sufficiently large, this error is at most a constant multiple of

$$2^{2r} \frac{k^{2r}}{(2r)!} < 2^{2r} \left(\frac{ek}{2r} \right)^{2r} = \left(\frac{ek}{r} \right)^{2r} < x(e/r)^{2r}. \quad (8)$$

Here we have used the fact that

$$m! > (m/e)^m, \text{ for } m \geq 1,$$

as well as the fact that

$$k^{2r} < x.$$

It is easy to see that the right-most term in (8) is $o(x/(\log x)(\log \log x)^2)$; in fact, it is much much smaller, but all we need is $o(x/(\log x)(\log \log x)^2)$. So, we have

$$\pi_2(x) \leq x \sum_{\substack{d|p_1 \cdots p_k \\ \omega(d) \leq 2r}} \frac{\mu(d)\tau(d)}{d} + o\left(\frac{x}{(\log x)(\log \log x)^2}\right).$$

We next want to extend this sum from those d with $\omega(d) \leq 2r$ to $\omega(d) \geq 0$. In order to do this, we need an upper bound for the contribution of those d with $\omega(d) \geq 2r + 1$ to this complete sum over all d : First, observe that

$$\sum_{\substack{d|p_1 \cdots p_k \\ \omega(d)=j}} \frac{\tau(d)}{d} \leq \frac{1}{j!} \left(\sum_{\substack{p \leq B \\ p \text{ prime}}} \frac{2}{p} \right)^j. \quad (9)$$

This follows since on expanding out this j th power we get a sum of terms

$$\cdots + \frac{2^j}{p_{i_1} \cdots p_{i_j}} + \cdots$$

and, for each set of primes $\{p_{i_1}, \dots, p_{i_j}\}$ we get that this product occurs in the j th power expansion $j!$ times; so, upon dividing by $j!$ we get our upper bound. The reason that we get an upper bound, and not an equality, is that this j th power will also have some terms involving the square of primes (and even higher powers) $p \leq B$.

Now, if $j \geq 2r + 1$, then our upper bound for (9) has size at most

$$\left(\frac{e \log \log x}{\sqrt{\log x} - 2} \right)^j (2 \log \log B + O(1))^j = \left(\frac{(\log \log x)^2 (e + o(1))}{\sqrt{\log x}} \right)^j.$$

(The reason for the -2 in the denominator here is that $2r \leq \sqrt{\log x} / \log \log x$, and so is potentially 2 less than this upper bound.) Summing this over all $j \geq 2r + 1$ we get a very small upper bound; in particular, it is certainly at most $o(1/(\log x)(\log \log x)^2)$ by a mile!

What this means is that

$$\sum_{d|p_1 \cdots p_k} \frac{\mu(d)\tau(d)}{d} = \sum_{\substack{d|p_1 \cdots p_k \\ \omega(d) \leq 2r}} \frac{\mu(d)\tau(d)}{d} + o\left(\frac{1}{(\log x)(\log \log x)^2}\right);$$

and so, we deduce that

$$\pi_2(x) \leq x \sum_{d|p_1 \cdots p_k} \frac{\mu(d)\tau(d)}{d} + o\left(\frac{x}{(\log x)(\log \log x)^2}\right).$$

This sum over the divisors d of $p_1 \cdots p_k$ has a very nice, simple form: It is just the product of $(1 - 2/p)$ over all $3 \leq p \leq B$, which gives

$$\pi_2(x) \leq x \prod_{\substack{3 \leq p \leq B \\ p \text{ prime}}} \left(1 - \frac{2}{p}\right) + o\left(\frac{x}{(\log x)(\log \log x)^2}\right).$$

By taking the logarithm of this product over primes, and using some basic estimates for the error in the Taylor expansion of $\log(1-t)$, and then exponentiating, it is not difficult to show that the product is $\sim c/(\log x)(\log \log x)^2$; and so,

$$\pi_2(x) \leq \frac{(c + o(1))x}{(\log x)(\log \log x)^2},$$

which is just what we needed in order to prove that the sum of reciprocals of twin primes converges.

2.4 What more can we prove?

As you should be able to tell from the argument in the previous section, we could give much better upper bounds on $\pi_2(x)$ using the Bonferroni approach, if we just sieve by more primes. What should clue you in to the fact that we have not reached the limit of this method is the fact that the error terms we got were very very small compared to the main term. For example, take the error $O(\tau(d))$, summed over all d with $\omega(d) \leq 2r$, that occurs in (7). In (8) we got that it had size at most $x(e/r)^{2r}$, and all we needed that it was $o(x/(\log x)(\log \log x)^2)$; in fact, we have that

$$x(e/r)^{2r} = x \exp(-O(\sqrt{\log x})),$$

using the fact that $2r \sim \sqrt{\log x} / \log \log x$.

Thus, it should not be surprising that by estimating things more precisely, and by working with more primes, this “Bonferroni method” can be used to prove that

$$\pi_2(x) < \frac{cx \log \log x}{(\log x)^2}.$$

Although this upper bound is much better than what we had before, it is still not of the form $cx/(\log x)^2$ as we conjectured earlier. Well, it turns out that yet another idea is needed, in addition to the Bonferroni inequalities,³ if we are to get an upper bound of this general shape $cx/(\log x)^2$. Roughly, this extra idea is to derive Bonferroni-type inequalities, where instead of truncating the divisors d at $\omega(d) \leq 2r$ or $2r - 1$, we truncate them according to a more elaborate rule (which roughly is that our d have at most a certain number of prime factors taken from certain sequences of intervals). If we do this, after the dust has settled we obtain the following theorem:

The Combinatorial Sieve. For every $K \geq 1$, there exist constants $\delta \in (0, 1]$, $\kappa_1 > 0$, and $\kappa_2 > 0$ such that the following holds for all x sufficiently large: For each prime $p \leq x^\delta$, suppose that we distinguish $w(p)$ residue classes, where

$$0 \leq w(p) \leq \min(p - 1, K),$$

which are residue classes that we want to “sieve out by”. Then, let S be the set that results after we remove all the integers $n \in [1, x]$ that happen to lie in one of these distinguished residue classes for at least one of these primes $p \leq x^\delta$. Then,

$$|S| < \kappa_1 x \prod_{\substack{p \leq x^\delta \\ p \text{ prime}}} \left(1 - \frac{w(p)}{p}\right);$$

and

$$|S| > \kappa_2 x \prod_{\substack{p \leq x^\delta \\ p \text{ prime}}} \left(1 - \frac{w(p)}{p}\right).$$

³Well, there are other, non-combinatorial sieve methods that do not use Bonferroni inequalities, such as Selberg’s sieve or the Large Sieve, which we may talk about later on in the semester.

This parameter K is called the “sifting dimension” for the sieve; actually, I have only stated a corollary of the combinatorial sieve – technically, the sifting dimension is

$$\frac{1}{\log \log x} \sum_{\substack{p \leq x^\delta \\ p \text{ prime}}} \frac{w(p)}{p}.$$

Just to see what we can use this theorem to prove, suppose we want to give an upper bound for $\pi_2(x)$. Then, we take $K = 2$, and then for primes $p \leq x^\delta$ we have, as before,

$$w(p) = \begin{cases} 1, & \text{if } p = 2; \\ 2, & \text{if } p \geq 3. \end{cases}$$

The value of $\delta = \delta(K) = \delta(2)$ that the above theorem uses will be somewhat smaller than $1/2$, which means that the theorem only will give us an upper bound on $\pi_2(x)$ – the lower bound it gives will only be the number of integers $n \leq x$ such that $n(n+2)$ is not divisible by any prime $p \leq x^\delta$, which is a larger set than just the twin primes.

At any rate, if we compute the upper bound that the theorem gives, we find that it is

$$\leq \kappa_1 x \left(1 - \frac{1}{2}\right) \prod_{\substack{3 \leq p \leq x^\delta \\ p \text{ prime}}} \left(1 - \frac{2}{p}\right) \sim \frac{Cx}{\log^2 x},$$

for a certain $C > 0$.

Another, vastly more interesting consequence of the combinatorial sieve is the following: Suppose that $f(x) \in \mathbb{Z}[x]$ is some polynomial that does not have any fixed prime factor; that is, suppose that there is no prime p such that $p|f(n)$ for all integers n . Then, we have that there exists some integer r that depends on f (in fact, it can be bounded from above purely in terms of the degree of f), such that for infinitely many integers n , $f(n)$ has at most r prime factors. Furthermore, we can give good lower bounds (off by at most a constant factor from the conjectured true lower bound) for this count. There are other, similar applications of the combinatorial sieve. I will save some of these applications for homework problems...