Small Gaps between Primes

Nicholas Pasetto

June 2024

1 Introduction

One of the most famous and difficult problems in analytic number theory is to find the statistical properties of prime gaps. Mathematicians are interested in these gaps because they want to know whether the gaps are 'random' or have an underlying pattern. In particular, there has been much recent progress on understanding the size and distribution of small prime gaps. The prime number theorem, proved in 1896 by Jacques Hadamard, implies that the average size of a gap near n is $\log(n)$. However, the bound given by the prime number theorem is very weak compared to what actually seems to be true. Visual inspection yields countless gaps of length 2, and mathematicians have conjectured that there are infinitely many such gaps. The first step toward this result was taken by Paul Erdos, and the known bounds have been slowly strengthened in the following decades. The work of Goldston-Pintz-Yildirim is far beyond any previous bounds, and is close to proving bounded gaps between primes. Since the work of Goldston-Pintz-Yildirim, mathematicians have made more progress on the sizes of prime gaps. In 2013, Yitang Zhang proved that there are infinitely many prime gaps with length at most $7*10^7$. After a huge collaborative project, this bound was improved to 246.

2 Preliminaries

A few definitions are required before the proof of the main theorem can begin. First, let $\pi(x;q,a)$ be the number of primes $\leq x$ that are $a(\text{mod } q)$. By the Prime Number Theorem and the fact that primes are roughly evenly distributed between reduced residue classes, we would expect

$$\pi(x;q,a) \sim \frac{\mathrm{li}(x)}{\phi(q)}$$

Therefore we define the error function

$$E(x;q,a) = \pi(x;q,a) - \frac{\mathrm{li}(x)}{\phi(q)}$$

The Bombieri-Vinogradov theorem is used to deal with the error terms in the proof. It states that for any positive constant A, there exists constants B and C such that, for all x,

$$\sum_{q\leq Q} \max_{y\leq x} \max_{(a,q)=1} |E(y;q,a)| \leq C\frac{x}{(\log x)^A}$$

where $Q = \frac{\sqrt{x}}{(\log x)^B}$. For a prime p and a sequence of nonnegative integers $\mathcal{H} = \{h_1, h_2, \ldots, h_k\}$, $\nu_\langle(p)$ is defined to be the number of residue classes mod p occupied by elements of H. Also, we define the singular series

$$\mathcal{G}(\mathcal{H}) = \prod_p \left(1 - \frac{\nu_{\mathcal{H}}(p)}{p}\right)\left(1 - \frac{1}{p}\right)^{-k}$$

We will also need an analogue of the prime-counting function for prime constellations:

$$\pi_{\mathcal{H}}(x) = \#\{n \leq x \colon \forall j, x + h_j \text{ prime}\}$$

For positive integers a, b, (a,b) and $[a,b]$ denote the greatest common divisor and least common multiple of a and b respectively.

3 Proof

Now we are ready to start the proof of the main theorem,

$$\liminf_{n\to\infty} \frac{p_{n+1} - p_n}{\log p_n} = 0$$

First, we choose $k \geq 2$ and $\mathcal{H} = \{h_1, h_2, \ldots, h_k\}$ such that $G(H) \neq 0$. The idea behind the proof is to show that if $\mathcal{H}$ is sufficiently large, then some translation of $\mathcal{H}$ will contain at least two primes. To do this, we choose x and search for two closely spaced primes between x and $2x$. We wish to find a nonnegative weighting function $a(n)$ such that the following inequality holds

$$\sum_{j=1}^{k} \sum_{\substack{x\leq n\leq 2x \\ n+h_j \text{ prime}}} a(n) > \sum_{x\leq n\leq 2x} a(n)$$

Clearly, if such an a exists, then there exists $x \leq n \leq 2x$ and i, j such that $n + h_i$ and $n + h_j$ are both prime, so there must be two primes between x and $2x$ with gap at most $h_k - h_1$. This constraint is hard to work with, so we use the stronger constraint

$$\sum_{\substack{x\leq n\leq 2x \\ n+h_j \text{ prime}}} a(n) > \frac{1}{k} \sum_{x\leq n\leq 2x} a(n)$$

for all $1 \leq j \leq k$. Finding an a that satisfies the constraint is of course the main part of the proof. In Selberg's sieve, which uses a similar technique, $a(n)$ is defined to be

$$a(n) = \left(\sum_{d|(n+h_1)(n+h_2)\dots(n+h_k)} \lambda_d \right)^2$$

where $\lambda_1 = 1$ and $\lambda_d = 0$ for $d > R$. This essentially restricts the sum to divisors which are at most R. If R is small compared to x, then we can obtain good asymptotics for $a(n)$. Of course, we still need to define what λ is. Selberg defined it to be

$$\lambda_d = \mu(d) \left(\frac{\log R/d}{\log R} \right)^k$$

This is the optimal choice for the Selberg sieve. For the GPY sieve, we choose

$$\lambda_d = \mu(d) P \left(\frac{\log R/d}{\log R} \right)$$

for some polynomial P which will be chosen later. Two restrictions on P are required to get reasonable asymptotics: $P(1) = 1$ and, for all $0 \leq j < k$,

$$P^{(j)}(0) = 0$$

If we could find a polynomial P that satisfies all the necessary constraints, we would prove bounded gaps between primes. This is far stronger than the main theorem, and unfortunately, there is no such P. To prove the main theorem, we choose $\varepsilon > 0$, and let $h = \varepsilon \log x$. Now we need to find a gap between primes with length at most h. To prove this, we will sum over all $\mathcal{H}$, using a different weight function for each one. Define

$$a(n; \{h_1, \dots, h_k\}) = \left(\sum_{d|(n+h_1)(n+h_2)\dots(n+h_k)} \lambda_d \right)^2$$

with λ_d the same as before. Now, if there is some $\mathcal{H}$ such that

$$\sum_{x \leq n \leq 2x} \sum_{\substack{1 \leq l \leq h \\ n+l \text{ prime}}} a(n; \mathcal{H}) > \sum_{x \leq n \leq 2x} a(n; \mathcal{H})$$

then there are two primes between x and $2x$ with gap at most h. Summing over all $\mathcal{H}$, we need to prove that

$$\sum_{1 \leq h_1 < h_2 \dots < h_k \leq h} \sum_{1 \leq l \leq h} \sum_{\substack{x \leq n \leq 2x \\ n+l \text{ prime}}} a(n; \{h_1, \dots, h_k\}) > \sum_{1 \leq h_1 < h_2 \dots < h_k \leq h} \sum_{x \leq n \leq 2x} a(n; \{h_1, \dots, h_k\})$$

Now we want to find out how large $\sum_{x\leq n\leq 2x} a(n;\mathcal{H})$ is. We have

$$\begin{aligned}\sum_{x\leq n\leq 2x} a(n;H) &= \sum_{x\leq n\leq 2x}\left(\sum_{d|(n+h_1)\ldots(n+h_k)}\lambda_d\right)^2 = \sum_{x\leq n\leq 2x}\sum_{\substack{d_1|(n+h_1)\ldots(n+h_k)\\ d_2|(n+h_1)\ldots(n+h_k)}}\lambda_{d_1}\lambda_{d_2}\\ &= \sum_{d_1,d_2\leq R}\lambda_{d_1}\lambda_{d_2}\sum_{\substack{x\leq n\leq 2x\\ [d_1,d_2]|(n+h_1)\ldots(n+h_k)}} 1\end{aligned}$$

The constraint on n is periodic with period $[d_1,d_2]$, so we define

$$f(m) = \sum_{\substack{x\leq n<x+m\\ m|(n+h_1)\ldots(n+h_k)}} 1$$

Therefore we have

$$\sum_{x\leq n\leq 2x} a(n;\mathcal{H}) = \sum_{d_1,d_2\leq R}\lambda_{d_1}\lambda_{d_2} f([d_1,d_2])\left(\frac{x}{[d_1,d_2]}+O(1)\right)$$

If R is smaller than $\sqrt{x}$, then $[d_1,d_2]$ will be smaller than x and $\frac{x}{[d_1,d_2]}$ will dominate the $O(1)$ term. This gives

$$\sum_{x\leq n\leq 2x} a(n;H) \sim x\sum_{d_1,d_2\leq R}\lambda_{d_1}\lambda_{d_2}\frac{f([d_1,d_2])}{[d_1,d_2]}$$

Using the constraints on the polynomial P, a tedious computation shows that the above is asymptotic to

$$\frac{x}{(\log R)^k}G(H)\int_0^1\frac{y^{k-1}}{(k-1)!}P^{(k)}(1-y)^2dy$$

Summing over all $\mathcal{H}$ gives

$$\begin{aligned}&\sum_{1\leq h_1<h_2\cdots<h_k\leq h}\ \sum_{x\leq n\leq 2x} a(n;\{h_1,\ldots,h_k\})\\ &\sim\frac{x}{(\log R)^k}\int_0^1\frac{y^{k-1}}{(k-1)!}P^{(k)}(1-y)^2dy\sum_{1\leq h_1<h_2\cdots<h_k\leq h}G(\{h_1,\ldots,h_k\})\end{aligned}$$

The last sum appears to be very difficult to evaluate. Fortunately, Gallagher proved that

$$\sum_{1\leq h_1<h_2\cdots<h_k\leq h}\mathcal{G}(\{h_1,\ldots,h_k\})\sim\sum_{1\leq h_1<h_2\cdots<h_k\leq h}1=\frac{h^k}{k!}$$

as $h\to\infty$. This finally gives an asymptotic

$$\sum_{1\leq h_1<h_2\cdots<h_k\leq h}\ \sum_{x\leq n\leq 2x} a(n;\{h_1,\ldots,h_k\})\sim\frac{x}{(\log R)^k}\frac{h^k}{k!}\int_0^1\frac{y^{k-1}}{(k-1)!}P^{(k)}(1-y)^2dy$$

Now we need an asymptotic for $\sum_{1\le l\le h}\sum_{\substack{x\le n\le 2x\\ n+l \text{ prime}}} a(n;\mathcal{H})$. Splitting the sum into two cases where $l=h_j$ and $l\neq h_j$, We have

$$\sum_{1\le l\le h}\sum_{\substack{x\le n\le 2x\\ n+l \text{ prime}}} a(n;\mathcal{H})=\sum_{j=1}^{k}\sum_{\substack{x\le n\le 2x\\ n+h_j \text{ prime}}} a(n;\mathcal{H})+\sum_{\substack{1\le l\le h\\ l\neq h_j}}\sum_{\substack{x\le n\le 2x\\ n+l \text{ prime}}} a(n;\mathcal{H})$$

Evaluating the first sum gives

$$\begin{aligned}\sum_{\substack{x\le n\le 2x\\ n+h_j \text{ prime}}} a(n;\mathcal{H}) &= \sum_{\substack{x\le n\le 2x\\ n+h_j \text{ prime}}}\left(\sum_{d|(n+h_1)\dots(n+h_k)}\lambda_d\right)^2\\ &=\sum_{\substack{x\le n\le 2x\\ n+h_j \text{ prime}}}\sum_{\substack{d_1|(n+h_1)\dots(n+h_k)\\ d_2|(n+h_1)\dots(n+h_k)}}\lambda_{d_1}\lambda_{d_2}\\ &=\sum_{d_1,d_2\le R}\lambda_{d_1}\lambda_{d_2}\sum_{\substack{x\le n\le 2x\\ [d_1,d_2]|(n+h_1)\dots(n+h_k)\\ n+h_j \text{ prime}}} 1\end{aligned}$$

Similarly, the constraint on n is periodic with period $[d_1,d_2]$. But $n+h_j$ needs to be coprime to $[d_1,d_2]$, reducing the amount of possible residue classes for n. We define

$$L(m)=\{n\colon 0\le n<m\&m \mid (n+h_1)\dots(n+h_k)\&(n+h_j,m)=1\}$$

L is the set of possible residue classes for n mod m. But we also need that $n+h_j$ is prime. This gives

$$\begin{aligned}\sum_{d_1,d_2\le R}\lambda_{d_1}\lambda_{d_2}\sum_{\substack{x\le n\le 2x\\ [d_1,d_2]|(n+h_1)\dots(n+h_k)\\ n+h_j \text{ prime}}} 1 &= \sum_{d_1,d_2\le R}\lambda_{d_1}\lambda_{d_2}\sum_{r\in L([d_1,d_2])}\sum_{\substack{x\le n\le 2x\\ n\equiv r (\text{mod } [d_1,d_2])\\ n+h_j \text{ prime}}} 1\\ &=\sum_{d_1,d_2\le R}\lambda_{d_1}\lambda_{d_2}\sum_{r\in L([d_1,d_2])}\pi(2x;[d_1,d_2],r)-\pi(x;[d_1,d_2],r)\end{aligned}$$

Using the Bombieri-Vinogradov theorem, we can give good asymptotics for π when $R<x^{1/4}$. This gives

$$\sum_{d_1,d_2\le R}\lambda_{d_1}\lambda_{d_2}\sum_{r\in L([d_1,d_2])}\pi(2x;[d_1,d_2],r)-\pi(x;[d_1,d_2],r)\sim\frac{x}{\log x}\sum_{d_1,d_2\le R}\lambda_{d_1}\lambda_{d_2}\frac{\#L([d_1,d_2])}{\phi([d_1,d_2])}$$

With more calculations, the above sum can be found to be asymptotic to

$$\frac{x}{(\log x)(\log R)^{k-1}}\mathcal{G}(\mathcal{H})\int_0^1\frac{y^{k-2}}{(k-2)!}P^{(k-1)}(1-y)^2dy$$

Summing over all $\mathcal{H}$ gives

$$\sum_{1\leq h_1<h_2<\cdots<h_k\leq h}\sum_{j=1}^{k}\sum_{\substack{x\leq n\leq 2x\\ n+h_j \text{ prime}}} a(n;\{h_1,\ldots,h_k\}) \sim k\frac{x}{(\log x)(\log R)^{k-1}}\frac{h^k}{k!}\int_0^1 \frac{y^{k-2}}{(k-2)!}P^{(k-1)}(1-y)^2dy$$

for sufficiently large h. Now we must choose a specific P. In this case, $P(y) = y^{k+r}$ works well, where r is some positive integer. Now, evaluating the two expressions and dividing gives

$$\left(\frac{\log R}{\log x}\right)\left(\frac{2k(2r+1)}{(r+1)(k+2r+1)}\right) > 1$$

Since $\frac{\log R}{\log x} < \frac{1}{4}$, the second fraction needs to be greater than 4. Unfortunately, it can get arbitrarily close to 4 but never reach it. We will need to include the values of the sum when $l \neq h_j$ for any j. Specifically, we want to find asymptotics for

$$\sum_{\substack{1\leq l\leq h\\ l\neq h_j}}\sum_{\substack{x\leq n\leq 2x\\ n+l \text{ prime}}} a(n;H)$$

If $n+l$ is prime, then we have

$$a(n;\mathcal{H}) = \left(\sum_{d|(n+h_1)\ldots(n+h_k)}\lambda_d\right)^2 = \left(\sum_{d|(n+h_1)\ldots(n+h_k)(n+l)}\lambda_d\right)^2 = a(n;\mathcal{H}\cup\{l\})$$

This is because any extra divisors coming from $n+l$ must be larger than R, so λ will be 0 at those divisors. Now we can use a similar calculation to obtain

$$\sum_{\substack{1\leq l\leq h\\ l\neq h_j}}\sum_{\substack{x\leq n\leq 2x\\ n+l \text{ prime}}} a(n;\mathcal{H})$$

$$\sim \sum_{1\leq h_1<h_2<\cdots<h_k\leq h}\sum_{\substack{l\leq h\\ l\neq h_j}}\frac{x}{(\log x)(\log R)^k}\mathcal{G}(\{h_1,\ldots,h_k,l\})\int_0^1\frac{y^{k-1}}{(k-1)!}P^{(k)}(1-y)^2dy$$

Using the Gallagher's result once again, we get

$$\sim \frac{x}{(\log R)^k}\frac{h^k}{k!}\frac{h}{\log x}\int_0^1\frac{y^{k-1}}{(k-1)!}P^{(k)}(1-y)^2dy$$

Now $\frac{h}{\log x} = \varepsilon$, so this sum gives an extra ε to the ratio. Since this ratio was already seen to be arbitrarily close to 1, the main theorem is proven.

4 Conclusion

With recent improvements to the Bombieri-Vinogradov theorem, it has been proven that gaps of length at most $P = 246$ occur infinitely often. Although the world's greatest mathematicians have tried to reduce this bound, the final goal of $P = 2$, and therefore the twin prime conjecture, remains unsolved. An even stronger conjecture by Hardy-Littlewood states that for all positive integer sequences $\mathcal{H} = \{h_1, h_2, \ldots, h_k\}$, we have

$$\pi_{\mathcal{H}}(x) = (\mathcal{G}(\mathcal{H}) + o(1))\frac{x}{(\log x)^k}$$

Unfortunately, the techniques of sieve theory do not seem to be sufficient to solve the Hardy-Littlewood Conjecture. In fact, they are probably not even sufficient to prove for all r

$$\liminf_{n\to\infty} \frac{p_{n+r} - p_n}{\log p_n} = 0$$

since sieve techniques rapidly weaken as the number of primes increases. This means that a deeper theory will be required to prove results stronger than those of Goldston-Pintz-Yildirim or Yitang Zhang.

References

[1] 'Small Gaps between Prime Numbers: The Work of Goldston-Pintz-Yildirim', K. Soundararajan, *Bulletin of the American Mathematical Society*, Volume 44, Number 1, January 2007, Pages 1-18

[2] 'Primes in Tuples I', D. A. Goldston, J. Pintz, and C. Y. Yildirim, https://arxiv.org/abs/math/0508185, August 2005