J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD
© Indian Institute of Science 2022.

REVIEW ARTICLE

Analytic Number Theory in the Last Decade

Ritabrata Munshi[*]

Abstract | This is a short survey of some of the most important milestones achieved in Analytic Number Theory during 2011–2020.

1 Introduction

In the last decade 2011–2020, we witnessed an all-round spectacular progress in Analytic Number Theory. It was marked with resolutions of some outstanding classical problems, such as the ternary Goldbach problem and the Vinogradov mean value theorem. There was fantastic progress towards the twin prime conjecture. In addition, results were established, breaking long-standing barriers in the field of multiplicative arithmetic functions and the analytic theory of automorphic L-functions. Overall, it had been an exciting decade for analytic number theorists.

2 Gaps Between Primes

Perhaps, the most compelling conjecture about prime numbers is the Twin Prime Conjecture, which says that there are infinitely many primes p such that $p+2$ is also a prime. This problem has fascinated generations of mathematicians (professionals and amateurs alike) and has led to developments of several important branches of analytic and combinatorial number theory, e.g. the sieve methods. It is a well-known result in sieve theory that there are infinitely many primes p, such that $p+2$ has at most two prime factors. A slightly different way of looking at the twin prime problem would be to study pairs of consecutive primes with small gaps. The prime number theorem (PNT) says that the average gaps between consecutive primes with roughly $\log x$ many digits is $\log x$.

Consider

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

where p_n denotes the n-th prime number. Then, PNT implies that $\Delta \leq 1$, whereas the twin prime conjecture implies that $\Delta = 0$ (and much more). Erdös in 1940s proved unconditionally that $\Delta < 1$, and the bound was subsequently improved over the years. But $\Delta = 0$ remained a far cry, till the breakthrough work of Goldston, Pintz and Yildirim (GPY)[5] in the first decade of this millennium. In a follow-up paper, they established a much stronger estimate

$$\liminf_{n\to\infty} \frac{p_{n+1}-p_n}{\sqrt{\log p_n}(\log\log p_n)^2} < \infty.$$

However, the main contribution of the GPY series was a conditional result with far reaching consequences. Consider a set of integers

$$\mathcal{H} = \{h_1 < h_2 < \cdots < h_k\},$$

and for every prime p let $\nu_p(\mathcal{H})$ be the number of distinct residues classes in $\mathcal{H}$ modulo p. We say that $\mathcal{H}$ is admissible if $\nu_p(\mathcal{H}) < p$ for all primes p. The prime k-tuple conjecture of Hardy and Littlewood says that if $\mathcal{H}$ is admissible then there are infinitely many integers n such that all the components of

$$n+\mathcal{H} = \{n+h_1, n+h_2, \ldots, n+h_k\}$$

are primes. The twin prime conjecture corresponds to the special case with $\mathcal{H} = \{0, 2\}$.

GPY approaches the problem through the Selberg sieve. Roughly speaking, the sieve removes those n's where any component in $n+\mathcal{H}$ has a small prime factor and one is left only with tuples where all the components have large prime factors. As such, the remaining tuples have few prime factors distributed among the components, and we obtain a sequence where we are likely to find primes with small gaps. For example, if we had a k-tuple where the components together have at most $2k-2$ prime factors, this would force at least two components of the tuple to be prime. Hence, we would have produced two primes whose difference is at most the width of the tuple. Alas, sieves never work this well and this strategy is doomed to fail. We need to add some new inputs to demonstrate existence of primes in the sieved tuples.

[1] *Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India.* **ritabratamunshi@gmail.com*

The extra input that GPY employ is the uniformity of the distribution of primes in arithmetic progressions with large moduli. We say that the primes have level of distribution θ if for any $\varepsilon > 0$, one has

$$\sum_{q \leq N^{\theta-\varepsilon}} \max_{(a,q)=1} \left| \sum_{\substack{p \leq N \\ p \equiv a \bmod q}} \log p - \frac{N}{\phi(q)} \right| \ll \frac{N}{(\log N)^A},$$

for any $A > 0$. Of course one wants to take θ as large as possible. The famous Bombieri–Vinogradov theorem says that primes have level of distribution 1/2. This is also the best one can achieve assuming the generalised Riemann Hypothesis. However, one can go even beyond the Riemann Hypothesis, and a conjecture of Elliot–Halberstam predicts that the level of distribution is 1. Now, the GPY strategy is to find the proportion of times the Selberg sieve produces almost prime tuples with one particular component fixed to be prime. If this proportion turns out to be bigger than $1/k$ (when we are considering k-tuples), then the k events

$$E_j = \{\text{j-th component in the k-tuple n} + \mathcal{H} \text{ is prime}\}$$
$$j = 1, 2, \ldots, k$$

cannot all be disjoint from each other and there must be some overlap where two of the components are primes simultaneously. This produces two primes in the given tuple. Therefore, the success of the method depends on the chance of the event E_j for $1 \leq j \leq k$, and the Selberg sieve coefficients need to be chosen optimally. The level of distribution of the primes now enters the picture, as the sieve weights are constrained by the level of distribution. To get better result in Selberg sieve, one needs to improve the level of distribution of the primes.

The main contribution of the GPY series is the following. Suppose the primes have level of distribution $\theta > 1/2$, then there exists an explicitly computable constant $c(\theta)$ such that for any admissible k-tuple $\mathcal{H}$ with $k \geq c(\theta)$, the set $n + \mathcal{H}$ has at least two primes for infinitely many integers n. In particular, if one assumes the Elliot–Halberstam conjecture, then using the admissible 6-tuple

$$\mathcal{H} = \{0, 4, 6, 10, 12, 16\}$$

one can show that the difference between two consecutive primes is less than 16 infinitely often. With this GPY brought the bounded gap conjecture, that there exists a constant c such $p_{n+1} - p_n \leq c$ for infinitely many n, within a 'hair's breadth'. All one had to do was to break the Bombieri–Vinogradov barrier. But this has eluded the best minds for decades.

In late April of 2013, Yitang Zhang, a less known professor of Mathematics at the University of New Hampshire, submitted a paper[19] to the Annals of Mathematics claiming to prove that there are infinitely many pairs of primes that differ by less than 70 million. The proof was thoroughly crosschecked and verified by experts, and soon the news spread like wildfire, rocking the world of Mathematics.

Theorem 1 [19] *Let p_n denote the n-th prime. Then, for infinitely many n, we have*

$$p_{n+1} - p_n \leq 70{,}000{,}000.$$

Zhang's theorem is a giant leap forward in the direction of the twin prime conjecture. Therefore, how did Zhang overcome the Bombieri–Vinogradov barrier? Actually, he did not, he found a clever way to bypass it completely. It has long been known that it is possible to get a level of distribution larger than 1/2 provided certain restrictions are placed on the residue classes. These were worked out by Bombieri, Friedlander and Iwaniec, and also by Fouvry and Iwaniec, in a well-known series of papers in 1980s. However, all these theorems require that the residue class $a \bmod q$ be fixed. But this restriction is incompatible with the GPY strategy. However, in 2008, Motohashi and Pintz made a fundamental observation, which was the basis of Zhang's approach. Let us try to make it little more precise. The main idea of GPY is to study the sum

$$\sum_{N < n \leq 2N} \left(\sum_{i=1}^{k} \theta(n + h_i) - \log 3N \right) \Lambda(n; \mathcal{H})^2,$$

where

$$\theta(n) = \begin{cases} \log n & \text{if n is a prime} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\Lambda(n; \mathcal{H}) = \sum_{d \mid (n+h_1)(n+h_2)\ldots(n+h_k)} \lambda_d,$$

where λ_d are the sieve weights. Motohashi–Pintz observed that the GPY result does not weaken

a bit if one restricts the sum to smooth d, i.e. d without large prime divisor. The fundamental idea of Zhang is to restrict the residue classes a mod q to run over the roots of the polynomial

$$\prod_{h\neq h'\in\mathcal{H}}(x+h-h')$$

modulo q. Now if q is prime then we only have bounded number of residue classes, and hence we have a larger level of distribution. However, for q composite, we may have many more. Zhang applies the Chinese remainder theorem to control this, and this turns out to be sufficient to control the error term. In other words, Zhang shows that in the Selberg sieve one can restrict the sieving to divisors with no large prime factors (i.e. smooth divisors). This though decreases the effectiveness of the sieve in GPY but the effect is small. On the other hand, Zhang increases the sieving range corresponding to a level of distribution $1/2+1/584$. The gain turns out to be large enough to overcome the loss from smoothing. To control the error terms Zhang makes the fundamental observation, which he himself calls 'the most novel part of the proof', that since the divisors in the sieving have no large prime factors and, therefore, may themselves be factored into factors of various sizes with considerable flexibility. One should note here that Zhang does not improve the level of distribution of primes, he only offers a way to bypass the issue.

Zhang's work[19] generated a lot of research in the last decade, and still continues to do so in the present. Several researchers have come up with simplifications of Zhang's arguments resulting in squeezing the gap even further. A special mention should be made here of Maynard's contributions. He has worked out a new way of proving bounded gap[10]. Maynard also uses Selberg sieve, but his weights are more flexible and take the form

$$\Lambda(n;\mathcal{H})=\sum_{d_i|(n+h_i);1\leq i\leq k}\lambda_{d_1,d_2,\ldots,d_k}.$$

Using these weights, he reduced the gap between primes to 600, and under the Elliot–Halberstam conjecture this gap further reduces to 12. In follow-up papers, Maynard has applied his ideas in other contexts as well. In particular, he has proved existence of primes with large gaps settling a well-known problem posed by Erdös[11] (also see[4]). Zhang's work also inspired a Polymath project led by Tao[15], that came up with several refinements of Zhang's work and reduced the gap to 246.

3 Ternary Goldbach Problem

Both the ternary Goldbach conjecture and the binary (or strong) Goldbach conjecture have their origin in a famous exchange of letters between Euler and Goldbach in 1742. While the strong form, that every even integer larger than 2 can be written as a sum of two primes, has been elusive, progress has been made towards the weaker conjecture that every odd integer larger than 5 can be written as a sum of three primes since 1920s. First, in 1923, assuming the Generalised Riemann Hypothesis, Hardy and Littlewood proved that every sufficiently large odd integer can be written as a sum of three primes. Then, Vinogradov used his powerful techniques to make the result unconditional in 1937. More precisely, it followed that for any odd integer $n\geq V$, with $V=3^{3^{15}}$, the equation

$$n=p_1+p_2+p_3$$

has solutions p_i in the set of prime numbers. Since $3^{3^{15}}\approx 10^{6,846,168}$ is a huge number, verifying the conjecture for smaller integers up to V is computationally impossible, even with our present day super computers. The bound V was improved to $e^{e^{16.038}}$ by Borozdkin in 1956, which in turn was substantially improved to $e^{e^{11.503}}$ by Chen and Wang in 1989. The best-known bound till the recent breakthrough by Helfgott, stood at e^{3100} which was obtained by Liu and Wang in 2002. In his recent work, Helfgott (see[6]) reduced the bound to 10^{27}, i.e. he showed that any odd integer $n>10^{27}$ is a sum of three primes. In another pre-print, Helfgott together with Platt verified using computers that every odd integer smaller than 8.845×10^{30} can be written as a sum of three primes.[A] Together, they settle the ternary Goldbach problem completely.

Let us now briefly explain the approach of Vinogradov. We are interested in the function

$$R(n)=\sum_{p_1,p_2,p_3\ \text{prime}}\sum\sum_{p_1+p_2+p_3=n}(\log p_1)(\log p_2)(\log p_3).$$

To settle the ternary Goldbach problem, one seeks to prove that $R(n)>0$ for any odd $n>5$. (For the strong Goldbach, one needs to prove that $R(n)>0$ for all integers $n>5$. For even n, one of

[A] It should be noted that Helfgott's work is yet to appear in print, though his pre-prints are available online since 2012.

the primes is forced to be 2, making the problem many fold harder.) Given n, we define the (truncated) generating function

$$f(x) = \sum_{\substack{p \text{ prime} \\ p < n}} (\log p) e(px),$$

so that

$$R(n) = \int_0^1 f(x)^3 e(-nx) \mathrm{d}x.$$

This is where the circle method kicks in. The main contribution to the integral comes from the short arcs around the points where the generating function peaks. This is the major arc contribution, which we need to evaluate asymptotically (or at least get a good lower bound). The rest of the 'circle' is the minor arc, where one seeks to prove a strong upper bound. In Vinogradov's approach, one fixes a positive integer B, and takes

$$P = (\log n)^B.$$

Then, the major arc is given by

$$\mathfrak{M} = \bigcup_{\substack{1 \le a \le q \le P \\ (a,q)=1}} \left[\frac{a}{q} - \frac{P}{n}, \frac{a}{q} + \frac{P}{n}\right],$$

and the minor arc is given by

$$\mathfrak{m} = \left[\frac{P}{n}, 1 + \frac{P}{n}\right] - \mathfrak{M}.$$

Using periodicity, we get

$$\begin{aligned} R(n) &= \int_{P/n}^{1+P/n} f(x)^3 e(-nx) \mathrm{d}x \\ &= \int_{\mathfrak{M}} f(x)^3 e(-nx) \mathrm{d}x \\ &\quad + \int_{\mathfrak{m}} f(x)^3 e(-nx) \mathrm{d}x. \end{aligned}$$

For bounding the minor arc from above, one now employs Vinogradov's estimates for trigonometric sums over primes. In particular one has

$$f(\alpha) \ll (\log n)^4 \left(\frac{n}{q^{1/2}} + n^{4/5} + (nq)^{1/2}\right),$$

where $|\alpha - a/q| \le q^{-2}$. Observe that the bound is non-trivial only when α is badly approximable by rational numbers with small denominators. This (together with the prime number theorem) leads to the upper bound

$$\int_{\mathfrak{m}} f(x)^3 e(-nx) \mathrm{d}x \ll \frac{n^2}{(\log n)^{\frac{B-10}{2}}}.$$

On the other hand for $\alpha \in [a/q - P/n, a/q + P/n]$ in the major arc, one uses the theory of distribution of primes in arithmetic progression to approximate $f(\alpha)$ as follows:

$$\begin{aligned} f(\alpha) =& \frac{\mu(q)}{\phi(q)} \sum_{m=1}^{n} e(m(\alpha - a/q)) \\ &+ O(n \exp(-C(\log n)^{1/2})). \end{aligned}$$

From this, one derives that

$$\begin{aligned} &\int_{\mathfrak{M}} f(x)^3 e(-nx) \mathrm{d}x = \tfrac{1}{2} n^2 \mathfrak{S}(n) \\ &+ O(n^2 (\log n)^{-B/2}), \end{aligned}$$

where the singular series is given by

$$\mathfrak{S}(n) = \prod_{p|n} (1 - (p-1)^{-2}) \prod_{p \nmid n} (1 + (p-1)^{-3}).$$

Observe that if n is even $\mathfrak{S}(n) = 0$ due to the factor coming from the prime 2. But for n odd $\mathfrak{S}(n) \gg 1$, and so for n sufficiently large odd integer we get that $R(n) > 0$. Now to find out what sufficiently large means, we need to work out the implied constants. As we noted above, Vinogradov's original approach works for $n \ge 3^{3^{15}}$, i.e. odd numbers with more than 6.8 million digits. Helfgott reduced it to numbers with 28 or more digits. As one can guess such major advancement does not come just by fine tuning the existing machinery. Helfgott introduces several refinements—smoothing exponential sums over primes with different weight functions (carefully chosen at the end to optimise the bounds) and interweaving the large sieve techniques with the circle method.

In his setup, the major arc is given by

$$\mathfrak{M} = \bigcup_{\substack{1 \le a \le q \le P \\ (a,q)=1}} \left[\frac{a}{q} - \frac{cr_0}{qx}, \frac{a}{q} + \frac{cr_0}{qx}\right],$$

where c and r_0 are constants (even when x grows) and $q \le r_0$. Next one introduces smooth weight functions in the generating function, at the cost of settling only for a lower bound for $R(n)$ instead of an asymptotic. Indeed the endgame in Helfgott's approach consists of picking the weight functions optimally. The estimation of the major arc reduces to estimation of

$$\sum_{n=1}^{\infty} \Lambda(n)\chi(n)\, e\left(\frac{\delta n}{x}\right)\eta\left(\frac{n}{x}\right),$$

where η is a smooth weight function, χ runs over Dirichlet characters modulo $q \le r_0$ and δ is small. To estimate this sum, Helfgott uses explicit computations involving L-functions. For the minor arc, following Vinogradov, one usually uses estimates for exponential sums over primes. But this by itself is not strong enough for the present purpose, and Helfgott splits the minor arc estimation into two parts. In one part, he uses point-wise estimates for exponential sums and in the other, he employs large sieve for primes to get uniform upper bounds for the L^2 norm of the exponential sums over segments of major arcs.

Theorem 2 (*Helfgott* [6])*Every odd integer larger than 5 can be written as a sum of three primes.*

4 Vinogradov's Mean Value Theorem

During 1920s, Hardy and Littlewood wrote a series of papers, titled 'Some problems in "Partitio Numerorum"', reshaping the newly formed circle method to study integer solutions of polynomial equations. In particular, their powerful analytic method ushered a new era in the study of the Waring problem—finding numbers $g(k)$ (resp. $G(k)$) such that every (resp. sufficiently large) positive integer is a sum of at most that many k-th powers of integers. The circle method reduces the problem to estimation of the 'minor arc' contribution involving exponential sums. To estimate the exponential sums, Hardy–Littlewood employed the method of Weyl, and thereby obtained explicit upper bounds like

$$G(k) \le (k-2)2^{k-1} + 5.$$

Around mid-1930s, Vinogradov introduced a new method to estimate exponential sums of Weyl's type. His method is particularly effective for short sums of large amplitude. With this discovery, he not only substantially improved the known results towards the Waring's problem, e.g. he established

$$G(k) \le k(3\log k + 11),$$

but also established a new zero-free region of the Riemann zeta function which remains unchallenged even after 80 years. An important ingredient of Vinogradov's method is an estimation of the mean value of certain exponential sums $J_{\ell,k}(X)$. Vinogradov proved an upper bound for this mean value, but it was weaker than the expected bound. Improving Vinogradov's estimate came to be known as the Vinogradov's mean value problem.

Despite significant attention paid to it by a large class of mathematicians, including some of the finest analytic number theorists, the conjectural bound remained elusive till 2015, when two different proofs were announced from two different frontiers. First, Wooley using analytic number theory, made significant progress towards the resolution of the main conjecture in a series of papers between 2010 and 2015, and resolved it in full in the first non-trivial case; second, Bourgain, Demeter, and Guth using harmonic analysis, resolved the Main Conjecture in full in 2015.

The starting point of Vinogradov's work is the exponential sum

$$S_f(a,b) = \sum_{a<n<b} e(f(n)),$$

where $f(x)$ is a real valued smooth function on the interval $[N,\ 2N]$, and $N \le a < b \le 2N$. (Recall that $e(z) = \mathrm{e}^{2\pi i z}$.) Using Weyl differencing with factorisable shifts, he then reduces the problem to estimation of a bilinear form

$$S = \sum_{1\le x\le X}\ \sum_{1\le y\le X} e(F(xy)),$$

where

$$F(x) = \sum_{0\le j\le k} \alpha_j x^j,$$

is a polynomial of degree k with real coefficients α_j. An application of the Hölder's inequality now yields

$$|S|^{2\ell^2} \ll \ell^{2k} X^{4\ell(\ell-1)+k(k+1)}\, J_{\ell,k}^2(X)\, \Delta,$$

where

$$\Delta = \prod_{1\le h\le k} \frac{1}{\ell^2 X^{2h}} \sum_{|m|\le \ell X^h} \left| \sum_{|n|\le \ell X^h} e(\alpha_h mn) \right|$$

and

$$J_{\ell,k}(X) = \int_0^1 \cdots \int_0^1 \left| \sum_{1\le x\le X} e(\alpha_1 x + \cdots + \alpha_k x^k) \right|^{2\ell} \mathrm{d}\alpha_1 \ldots \mathrm{d}\alpha_k.$$

The source of cancellation in Δ is exponential sums of linear polynomials which was also the main input in Weyl's method. But Vinogradov's method reaches this step much faster than Weyl's-roughly speaking Vinogradov takes k^4 steps whereas Weyl takes 2^{k-1} steps. But the major problem now lies in the estimation of the mean value $J_{\ell,k}(X)$. This mean value can also be reinterpreted as the number of solutions to the system of homogeneous equations

$$\begin{aligned} y_1 + y_2 + \cdots + y_\ell &= y_{\ell+1} + y_{\ell+2} + \cdots + y_{2\ell} \\ y_1^2 + y_2^2 + \cdots + y_\ell^2 &= y_{\ell+1}^2 + y_{\ell+2}^2 + \cdots + y_{2\ell}^2 \\ &\cdots\cdots \\ y_1^k + y_2^k + \cdots + y_\ell^k &= y_{\ell+1}^k + y_{\ell+2}^k + \cdots + y_{2\ell}^k \end{aligned}$$

in integers $1 \le y_j \le X$. By applying simple heuristics, one can come to the following prediction:

Conjecture 1 *For all integers $\ell, k \ge 1$, we have*

$$J_{\ell,k}(X) \ll_{\ell,k,\varepsilon} X^{\varepsilon}\left(X^{\ell} + X^{2\ell - \frac{k(k+1)}{2}}\right)$$

for all $X \ge 1$ and any $\varepsilon > 0$.

Vinogradov proved a slightly weaker bound, namely for any positive integer m, $\ell \ge k(k+m)$ and any $X \ge k^{k(1-1/k)^{-m}}$

$$J_{\ell,k}(X) \le 2^{4\ell m} X^{2\ell - \frac{k(k+1)}{2} + \eta_{\ell,k}}$$

with

$$\eta_{\ell,k} = \tfrac{1}{2}k(k+1)(1-1/k)^m.$$

Vinogradov's work was extended by Linnik, and later by Karatsuba and Stechkin, who managed to reduce the exponent to

$$\eta_{\ell,k} = \tfrac{1}{2}k^2(1-1/k)^{[\ell/k]}.$$

Their works also give an explicit constant (in place of $2^{4\ell m}$). Since $\eta_{\ell,k} \le k^2 \mathrm{e}^{-\ell/k^2}$, decays as ℓ becomes sufficiently larger than the degree k, one get the conjectural bound for

$$\ell \ge 3k^2(\log k + O(\log\log k)).$$

In fact, one obtains an asymptotic

$$J_{\ell,k}(X) \sim C(\ell,k) X^{2\ell - \frac{k(k+1)}{2}}$$

with an explicit positive constant $C(\ell,k)$.

In 1990s, Wooley took the next giant leap, and introducing efficient differencing technique, proved the main conjecture for

$$\ell \ge k^2(\log k + 2\log\log k + O(1)).$$

This remained the record till the beginning of the last decade. In a major breakthrough in 2012, Wooley[16,17] extended the range to

$$\ell \ge k(k+1),$$

removing for the first time the extra logarithmic factor, and bringing the range just a factor of 2 away from the conjecture (the critical index being $\ell = k(k+1)/2$). His proof is based on a new technique 'efficient congruencing', which he extended and generalised in a series of papers in the first half of the last decade. This culminated in his proof of the main conjecture in full (for all ℓ) for the first non-trivial degree, i.e. $k = 3$[18]. (The main conjecture is a triviality for degrees one and two.)

Theorem 3 [18] *For $k = 3$ and any integer $\ell \ge 1$, we have*

$$J_{\ell,k}(X) \ll_{\ell,\varepsilon} X^{\varepsilon}\left(X^{\ell} + X^{2\ell - \frac{k(k+1)}{2}}\right)$$

for all $X \ge 1$ and any $\varepsilon > 0$.

In December 2015, Bourgain, Demeter and Guth[3] announced the resolution of the remaining cases of the mean value theorem.

Theorem 4 [3] *For $k \ge 4$ and any integer $\ell \ge 1$, we have*

$$J_{\ell,k}(X) \ll_{\ell,\varepsilon} X^{\varepsilon}\left(X^{\ell} + X^{2\ell - \frac{k(k+1)}{2}}\right)$$

for all $X \ge 1$ and any $\varepsilon > 0$.

The proof of this theorem is rooted in harmonic analysis. Of course, results from Fourier analysis have been used in the analytic theory of numbers since its conception. But BDG takes this synergy to a new height, employing the recently developed ℓ^2 decoupling techniques in their proof. As we will note in the last section, Bourgain[2] goes on to apply ℓ^2 decoupling to give new bounds for the Riemann zeta function.

5 Multiplicative Functions

Multiplicative functions are ubiquitous in analytic theory of numbers. One may say that arithmetic is just a study of the interactions of the additive and multiplicative structure of the integers, and as such the study of the averages of multiplicative functions is one of the main aims of number theory. For example, the randomness of the Möbius function $\mu(n)$ (or its closely related cousin—the Liouville function $\lambda(n)$) holds the key to the mystery of the distribution of the

prime numbers. It is well known that the prime number theorem

$$\lim_{x\to\infty} \frac{\#\{p < x : p \text{ prime}\}}{x/(\log x)} = 1$$

is equivalent to the assertion that

$$\sum_{n<x} \mu(n) = o(x),$$

which in turn is equivalent to

$$\sum_{n<x} \lambda(n) = o(x).$$

The Riemann hypothesis on the other hand is equivalent to assertions that the above summatory functions have square-root cancellations. One way to study this problem is to look at shorter averages

$$\frac{1}{H} \sum_{x<n<x+H} \lambda(n),$$

with $H < x$. Trivially, this sum is bounded by 1, as $\lambda(n) = (-1)^{\Omega(n)}$ (where $\Omega(n)$ is the number of prime factors of n counted with multiplicity) is bounded by 1. From the prime number theorem, we can show that the average is $o(1)$ as long as H grows linearly with x, i.e. $H > \varepsilon x$ for some $\varepsilon > 0$. But proving cancellation in the average when H grows much slowly compared with x is an incredibly difficult problem. Maier and Montgomery showed that, under the Riemann Hypothesis, cancellation kicks in as soon as H grows faster than $x^{1/2}(\log x)^c$ for some absolute constant c. Probabilistic models suggest that one can take H much smaller, e.g. $H > x^{\varepsilon}$ for any $\varepsilon > 0$, going far beyond the scope of the Riemann Hypothesis. On the other hand, Chowla has conjectured that H cannot be taken to be 'too small', as one expects long strings of consecutive integers without sign changes in the Liouville function.

The problem becomes more tractable if one introduces an extra averaging—instead of looking for cancellation in each short intervals, one now seeks cancellation in almost all short intervals. More precisely, we may consider the second moment

$$\frac{1}{X} \int_X^{2X} \left| \frac{1}{H} \sum_{x<n<x+H} \lambda(n) \right|^2 \mathrm{d}x,$$

and try to show that this goes to 0 as $X \to \infty$. Using the zero density estimates, Ramachandra proved that for any $\varepsilon > 0$, and $A > 0$ one has

$$\frac{1}{X} \int_X^{2X} \left| \frac{1}{H} \sum_{x<n<x+H} \lambda(n) \right|^2 \mathrm{d}x \ll_{\varepsilon,A} (\log X)^{-A}$$

if

$$X^{1/6+\varepsilon} < H < X.$$

In some sense, this is still the best result known in this direction, as the bound saves an arbitrary power of $\log X$. Of course, the major drawback is that one still needs to take H 'quite big'. The recent breakthrough result of Matomaki and Radziwill[8] overcomes this barrier at the cost of getting a much smaller saving.

Theorem 5 [8] *For any $2 \le H < X$, one has*

$$\frac{1}{X} \int_X^{2X} \left| \frac{1}{H} \sum_{x<n<x+H} \lambda(n) \right|^2 \mathrm{d}x \ll (\log H)^{-c}$$

for some absolute constant $c > 0$.

As a consequence, one now has

$$\frac{1}{X} \int_X^{2X} \left| \frac{1}{H} \sum_{x<n<x+H} \lambda(n) \right|^2 \mathrm{d}x = o(1)$$

as $X \to \infty$, as long as H grows to infinity with x no matter how slowly. All previous works in this direction, including that of Ramachandra, were based on complex analytic techniques using the meromorphic continuation of $1/\zeta(s)$ and $\zeta(2s)/\zeta(s)$ inside the critical strip. Matomaki–Radziwill approach is different and does not depend on meromorphic continuation. Instead, they approach the problem from the 'pretentious multiplicative number theory' viewpoint of Granville and Soundararajan.

To make this more precise, let us introduce the pretentious distance

$$\mathbb{D}(f,g;X) = \sqrt{\sum_{\substack{p<X \\ p \text{ prime}}} \left(\frac{1 - \mathrm{Re}(f(p)\overline{g(p)})}{p} \right)}$$

of two multiplicative functions f and g (with $|f(n)|, |g(n)| \le 1$ for all n) at threshold X. Let E_t denote the multiplicative function $E_t(n) = n^{it}$. Then, a classical theorem of Halasz states that for X sufficiently large, there is a constant c such that

$$\frac{1}{x} \sum_{n<x} f(n) \ll \mathrm{e}^{-c \min_{t<T} \mathbb{D}(f,E_t;x)^2} + \frac{1}{T}$$

for any $T > 0$. This inequality implies that if there is no cancellation in the left hand side then the pretentious distance between the multiplicative function from some E_t should be small. As such, we can formulate a weaker version of the Matomaki–Radziwill theorem in the following manner. Given $\varepsilon > 0$, and $1 < H < X$ sufficiently large depending on ε, if

$$\frac{1}{X}\int_X^{2X}\left|\frac{1}{H}\sum_{x<n<x+H}\lambda(n)\right|^2 \mathrm{d}x > \varepsilon^2,$$

then

$$\mathbb{D}(\lambda, E_t; X) \ll_\varepsilon 1,$$

for some $t \ll_\varepsilon X/H$. This qualitative result is not strong enough to recover the original theorem, as stated above. But there is a quantitative reformulation of the above statement which is strong enough to recover the decay rate of $(\log H)^{-c}$.

The key ingredient in the proof is Fourier analytic in nature, where one studies various norms of the corresponding Dirichlet series

$$\sum_{n=1}^{\infty}\frac{\lambda(n)}{n^s} = \frac{\zeta(2s)}{\zeta(s)}$$

at the edge of the critical strip $s = 1 + it$. Since one does not need to penetrate inside the critical strip, meromorphic continuation is not required, and as such the method applies to a broader class of multiplicative functions. Now, the theorem of Halasz, as stated above, gives a good control over the L^∞ norm of these functions. However, one needs to get good estimates for the L^1 and L^2 norms. The key idea here is to use convolution of two functions, as we have inequalities of the type

$$\|f \star g\|_{FL^1} \le \|f\|_{FL^2}\|g\|_{FL^2},$$
$$\|f \star g\|_{FL^2} \le \|f\|_{FL^2}\|g\|_{FL^\infty}.$$

The problem now is to find out a way to factorise general multiplicative functions (in particular the Liouville's function) into factors for which we will have some control. Matomaki and Radziwill use the Turán–Kubilius phenomenon, that for certain moderately wide ranges of primes $[p, q]$, almost all $n \in [p, q]$ has close to $\log\log q - \log\log p$ many prime factors. Therefore, if one introduces the arithmetic function

$$w_{p,q}(n) = \begin{cases} \frac{1}{\log\log q - \log\log p} & \text{if } n \text{ is a prime in the range } [p,q] \\ 0 & \text{otherwise,} \end{cases}$$

one roughly has

$$f \approx f \star f w_{p,q}.$$

This key factorisation is then used to derive bounds for L^1 and L^2 norms. More details can be be found in the joint work of Matomaki and Radziwill with Tao[9].

6 Subconvexity Problem

Bounding the size of automorphic L-functions inside the critical strip is a problem of utmost importance in analytic number theory. The main guiding conjecture in this field is the Generalised Lindelöf Hypothesis (GLH), a consequence of the Grand Riemann Hypothesis (GRH), that predicts that the L-function grows only mildly with respect to its conductor. More precisely, if π is an automorphic form of level q for $GL_d(\mathbb{A}_\mathbb{Q})$ with Langlands parameters $(\mu_1, \ldots, \mu_d)$, then the (analytic) conductor is defined by

$$C(\pi, t) = q\prod_{j=1}^{d}(3 + |t + \mu_j|),$$

and GLH predicts that

$$L(1/2 + it, \pi) \ll_\varepsilon C(\pi, t)^\varepsilon$$

for any $\varepsilon > 0$. From the basic theory of L-functions (i.e. analytic continuation and functional equation), one can deduce, using the Phragmen–Lindelöf convexity principle from complex analysis, that

$$L(1/2 + it, \pi) \ll_\varepsilon C(\pi, t)^{1/4+\varepsilon}.$$

Proving a bound with a smaller exponent is called the subconvexity problem. Such a bound is not only a progress towards the GLH, but often comes with deep applications especially in certain equidistribution problems.

The subconevxity problem has a long history which goes back to a little over 100 years, to the works of Weyl and Hardy–Littlewood. Though Littlewood had announced his work with Hardy in a meeting of the London Mathematical Society, their paper never appeared in print. It was only much latter, around 1926 that Landau first published a proof of what now is famously called the Weyl bound for the Riemann zeta function

$$\zeta(1/2 + it) \ll t^{1/6+\varepsilon}$$

for $t > 2$. (One can even replace t^ε by a power of $\log t$.) The convexity bound in this case is given by $t^{1/4+\varepsilon}$. Therefore, in one stroke, Weyl and Hardy–Littlewood reduced the exponent by a factor of 2/3. The same analysis also works for the Dirichlet L-functions as well and yields the bound

$$L(1/2+it,\chi) \ll_{\varepsilon,\chi} t^{1/6+\varepsilon}.$$

Here, the implied constant depends on the modulus of the Dirichlet character χ, and so the bound is only subconvex in the t-aspect. Later Burgess in 1960s introduced his ingenious technique of estimating (short) character sums of the form

$$\sum_{a<n<b} \chi(n),$$

and thus established the first subconvex bound in the level aspect

$$L(1/2,\chi) \ll_{\varepsilon} q^{3/16+\varepsilon}.$$

Here, χ is a primitive Dirichlet character modulo q. In late 1970s, Heath–Brown combined Burgess and Weyl, and established

$$L(1/2+it,\chi) \ll_{\varepsilon} (qt)^{3/16+\varepsilon},$$

for $t > 2$. In one sense, this settles the subconvexity problem for degree one L-functions, up to the strength of the exponent. Though the Weyl bound in the t-aspect has been improved gradually in the last hundred years, the Burgess bound remained untouched for about 60 years. In a breakthrough paper in 2020, Petrow–Young[14] improved Burgess and established an exponent of Weyl strength.

Theorem 6 (Petrow–Young [14]) Let χ be a primitive Dirichlet character of modulus q, then we have

$$L(1/2,\chi) \ll_{\varepsilon} q^{1/6+\varepsilon}.$$

The t-aspect subconvexity problem for the Riemann zeta function can be easily reduced to bounding exponential sums. Such sums have been the focus of research in analytic number theory for more than 100 years. The techniques developed by Weyl, van der Corput and Vinogradov still are the most fundamental in this branch. The search is always on for new exponent pairs. Hence, the bound for the zeta function has seen steady improvements over the years, albeit only gradually. Bourgain's work on ℓ^2 decoupling gave rise to a new exponent pair, and using this he improved the existing record to

$$\zeta(1/2+it) \ll t^{1/6-1/84}$$

for $t > 3$.

As we mentioned above, the level aspect problem for $L(1/2,\chi)$ is more delicate. Petrow and Young approach the problem through the Conrey–Iwaniec technique, which is a cubic moment computation. In this approach non-negativity of the L-values plays a crucial role. Conrey–Iwaniec used their method and the non-negativity of $L(1/2,f\otimes\chi)$ for χ quadratic and $SL(2,\mathbb{Z})$ form f, to get Weyl bound for quadratic characters. That the same method can be extended to non-quadratic characters came as a surprise. Petrow–Young used a clever trick to get non-negativity of the L-value even for non-quadratic characters χ.

In their first paper, they restrict to cube-free modulus q. Let χ be a primitive Dirichlet character modulo q, and let $\mathcal{H}_{it_j}(m,\bar{\chi}^2)$ be the set of Hecke–Maass cusp form of level $m|q$, central character $\bar{\chi}^2$ and spectral parameter t_j. The fundamental observation of Petrow–Young is that for $f \in \mathcal{H}_{it_j}(m,\bar{\chi}^2)$ the twisted form $f\otimes\chi$ is a self-dual new form of level q^2 and trivial central character. As such, one has the non-negativity

$$L(1/2,f\otimes\chi) \geq 0,$$

and hence one can apply the Conrey–Iwaniec method to derive subconvexity from the cubic moment. The key bound in Petrow–Young is the following:

$$\sum_{m|q}\sum_{|t_j|\leq T}\sum_{f\in\mathcal{H}_{it_j}(m,\bar{\chi}^2)} L(1/2,f\otimes\chi)^3 + \int_{-T}^{T} |L(1/2+it,\chi)|^6 \mathrm{d}t \ll T^B q^{1+\varepsilon}$$

for some $B > 0$. The most important aspect of this bound is its strength in the q-aspect, which is 'Lindelöf on average'. Now using non-negativity of $L(1/2,f\otimes\chi)$ we can drop the discrete sum on the left hand side, and conclude that

$$L(1/2,\chi) \ll q^{1/6+\varepsilon}.$$

The main ingredients in the proof are the Kuznetsov trace formula and the Riemann hypothesis for varieties over finite fields (Deligne's theorem).

The subconvexity problem for degree three L-functions remained wide open for a long time. At the beginning of the last decade, Li[7] published a breakthrough paper on subconvexity for $GL(3)$ and $GL(3)\times GL(2)$ L-functions. Li's approach was based on the Conrey–Iwaniec technique, and as such non-negativity of the L-values played a crucial role. Consequently, she had to restrict herself to only self-dual representations of $GL(3)$, or the symmetric square lifts of $GL(2)$ forms. In the first half of the last decade, I developed a new approach for subconvexity based on separation of oscillation using the circle/delta method. This led

to the first subconvex bound for generic degree three L-functions[12].

Theorem 7 (M. 2015) *Let π be an Hecke–Maass cusp form for $SL(3,\mathbb{Z})$. Then, we have*

$$L(1/2+it,\pi) \ll t^{3/4-1/16+\varepsilon}$$

for $t > 3$.

First, by approximate functional equation, we get that

$$L(1/2+it,\pi) \ll t^{\varepsilon} \sup_{N \ll t^{3/2}} \frac{|S(N)|}{\sqrt{N}} + t^{-2021},$$

where $S(N)$ are smooth sums of the form

$$S(N) = \sum_{n\sim N} A(1,n) n^{it},$$

where $A(m, n)$ are the Whittaker–Fourier coefficients of the form π. The subconvexity problem now boils down to showing cancellation in the sums $S(N)$ for $N \approx t^{3/2}$. Applying the delta method one gets that $S(N)$ is approximately given by

$$\frac{1}{Q^2}\int_{|x|\ll 1} \sum_{q\sim Q} g(q,x)^{\star} \sum_{a \bmod q} \sum_{n\sim N} A(1,n) e\left(\frac{an}{q} - \frac{xn}{qQ}\right) \times \sum_{m\sim N} m^{it} e\left(-\frac{am}{q} + \frac{xm}{qQ}\right) \mathrm{d}x,$$

where $g(q, x)$ are some 'well-behaved' weights. As one can see, the delta symbol has separated the Fourier coefficients $A(1, n)$ from the analytic oscillatory factor m^{it}. The separation, of course, comes at a huge cost as now we need to recover the whole length N. But now we have a lot of flexibility, as we can freely apply the Poisson and the Voronoi summation formulas. One crucial idea here is that the modulus of the delta method Q should be taken smaller than square-root of the length of the equation, thereby pushing some part of the harmonics of the delta method into the analytic side. This works like a conductor lowering mechanism, which helps to recover the initial loss and gives something more.

Various forms of the delta method have now been applied in several other subconvexity problems. The above bound in the t-aspect has been improved substantially, see, e.g. Aggarwal[1]. The method also works for twists of degree three L-functions, $L(1/2, \pi \otimes \chi)$ (see[13]). Recently, the method has been extended to cover degree six L-functions given by Rankin–Selberg convolutions of the form $GL(2) \times GL(3)$.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements

The author is grateful to the anonymous referee for her/his helpful comments. The author is supported by J.C. Bose Fellowship JCB/2021/000018, SERB DST Government of India.

Received: 6 January 2022 Accepted: 17 June 2022
Published online: 10 August 2022

References

1. Aggarwal K (2021) A new subconvex bound for $GL(3)$ L-functions in the t-aspect. Int J Number Theory 17(5):1111–1138
2. Bourgain J (2017) Decoupling, exponential sums and the Riemann zeta function. J Am Math Soc 30(1):205–224
3. Bourgain J, Demeter C, Guth L (2016) Proof of the main conjecture in Vinogradov's mean value theorem for degrees higher than three. Ann Math 184:633–682
4. Ford K, Green B, Konyagin S, Maynard J, Tao T (2018) Long gaps between primes. J Am Math Soc 31(1):65–105
5. Goldston D, Pintz J, Yildirim C (2009) Primes in tuples I. Ann Math 170:819–862
6. Helfgott H (2015) The ternary Goldbach problem. arXiv: 1501.05438 (pre-print)
7. Li X (2011) Bounds for $GL(3) \times GL(2)L$-functions and $GL(3)L$-functions. Ann Math 173:301–336
8. Matomaki K, Radziwill M (2016) Multiplicative functions in short intervals. Ann Math 183(3):1015–1056
9. Matomaki K, Radziwill M, Tao T (2020) Fourier uniformity of bounded multiplicative functions in short intervals on average. Invent Math 220(1):1–58
10. Maynard J (2015) Small gaps between primes. Ann Math 181(1):383–413
11. Maynard J (2016) Large gaps between primes. Ann Math 183(3):915–933
12. Munshi R (2015) The circle method and bounds for L-functions—III. t-aspect subconvexity for $GL(3)L$-functions. J Am Math Soc 28:913–938
13. Munshi R (2015) The circle method and bounds for L-functions—IV: subconvexity for twists of $GL(3)L$-functions. Ann Math 182(2):617–672
14. Petrow I, Young M (2020) The Weyl bound for Dirichlet L-functions of cube-free conductor. Ann Math 192(2):437–486

15. Polymath DHJ (2014) Variants of the Selberg sieve, and bounded intervals containing many primes. Res Math Sci 1(12):83
16. Wooley T (2012) Vinogradov's mean value theorem via efficient congruencing. Ann Math 175(3):1575–1627
17. Wooley T (2013) Vinogradov's mean value theorem via efficient congruencing, II. Duke Math J 162(4):673–730
18. Wooley T (2016) The cubic case of the main conjecture in Vinogradov's mean value theorem. Adv Math 294:532–561
19. Zhang Y (2014) Bounded gaps between primes. Ann Math 179(3):1121–1174

Ritabrata Munshi is known for his work in the analytic theory of L-functions. He did his bachelors and masters studies in the Indian Statistical Institute and worked on his PhD thesis under the supervision of Prof. Andrew Wiles at Princeton University. After spending four years as a postdoc at the Rutgers University and the Institute for Advanced Study, he returned to India and joined TIFR Mumbai. He is now a professor at the Indian Statistical Institute, Kolkata. He has received several awards including the SS Bhatnagar Prize, ICTP Ramanujan Prize and Infosys Science Award. He is a fellow of the Indian National Science Academy and the Indian Academy of Sciences.