Séminaire BOURBAKI March 2014
66ème année, 2013–2014, n$^{\text{o}}$ 1084

GAPS BETWEEN PRIME NUMBERS AND PRIMES IN ARITHMETIC PROGRESSIONS [after Y. Zhang and J. Maynard]

by **Emmanuel KOWALSKI**

... utinam intelligere possim rationacinationes pulcherrimas quae e propositione concisa $|\sum x_i y_i| \leqslant \|x\| \, \|y\|$ *fluunt...*
(after H. Cartan)

1. INTRODUCTION

Y. Zhang proved in [27], announced in May 2013, the following theorem:

THEOREM 1.1 (Zhang). — *There exists an even integer* $h \geqslant 2$ *with the property that there exist infinitely many pairs of prime numbers of the form* $(p, p+h)$*. In fact, there exists such an* h *with* $h \leqslant 70,000,000$*.*

Equivalently, if p_n*,* $n \geqslant 1$*, denotes the* n*-th prime number, we have*

$$\liminf_{n \to +\infty} (p_{n+1} - p_n) < +\infty,$$

and more precisely

$$\liminf_{n \to +\infty} (p_{n+1} - p_n) \leqslant 70,000,000.$$

The equivalence of the two formulations is clear by the pigeon-hole principle. The first one is psychologically more spectacular: it emphasizes the fact that *for the first time* in history, one has proved an *unconditional* existence result for infinitely many primes p and q constrained by a binary condition $q - p = h$.

Remarkably, this already extraordinary result was improved in spectacular fashion in October 2013 by J. Maynard [21]:[(1)]

THEOREM 1.2 (Maynard). — *There exists an even integer* $h \leqslant 600$ *with the property that there exist infinitely many pairs of prime numbers of the form* $(p, p+h)$*. In fact, for any fixed integer* $k \geqslant 2$*, there exist* k *distinct integers* $(h_1, \dots, h_k)$ *such that the set of integers* n *for which* $n + h_1$*, ...,* $n + h_k$ *are all primes is infinite.*

Equivalently, we have

$$\liminf_{n \to +\infty} (p_{n+m} - p_n) < +\infty \tag{1}$$

for any fixed $m \geqslant 1$*.*

(1) T. Tao derived similar results independently.

We emphasize that Zhang's method did not apply to (1) except for $m = 1$, and thus Maynard's work is a far-reaching extension, and not merely a strengthening, of Zhang's Theorem. It is all the more amazing that Maynard's proof is, in its technical aspects, much simpler than Zhang's. Indeed, it can be realistically presented from scratch in a one-year graduate course in analytic number theory. On the other hand, the main ingredient in Zhang's proof is a statement concerning the distribution of primes in arithmetic progressions to large moduli which has considerable independent interest (see Theorem 4.1), since it reflects information which lies beyond the immediate reach of the Generalized Riemann Hypothesis.

The goal of this text is to present the context of these spectacular results, and to sketch some of the main steps of the proofs, with an emphasis on the key ideas (and some bias related to the author's taste). There already exist at least three independent full (or almost full) expositions of the results of Zhang (by the Polymath 8 group [24], Granville [14] and Friedlander–Iwaniec [10]) and two of Maynard's which, as we indicated, are in any case technically simpler (in a blog post of Tao [26] and in [14]). There is therefore no doubt concerning the correctness of the results, and we will emphasize conceptual aspects instead of trying to write another complete proof.

We will also not attempt to describe the techniques that lead to the best possible bounds of h currently known, and to the best upper-bound for the liminf in (1) (as a function of m). These are for the essential part found in the versions of Theorems 1.1 and 1.2 proved in the course of the Polymath8 project (see [24] and the ongoing Polymath8b work).

Notation.

– p will always refer to a prime number and n to positive integers; it will be convenient to denote by $\mathbf{M}$ (like "moduli") the set of squarefree numbers (i.e., those positive integers q such that no square of a prime divides q); given $y \geqslant 2$, an integer $n \geqslant 1$ is called y-friable if and only if it has no prime factor $p \mid n$ such that $p > y$.[(2)]

– $\Lambda(n)$ denotes the von Mangoldt function, equal to $\log p$ for all prime powers p^k, $k \geqslant 1$, and to zero for other integers, and we also denote by $\theta(n)$ the function equal to $\log p$ for a prime p, and equal to 0 otherwise.

– For $X \geqslant 2$, $q \geqslant 1$ and $a \geqslant 1$ (or $a \in \mathbf{Z}/q\mathbf{Z}$), we let

$$\psi(X; q, a) = \sum_{\substack{X < n \leqslant 2X \\ n \equiv a \,(\mathrm{mod}\ q)}} \Lambda(n).$$

We extend this slightly as follows: an *idelette* a is an element of

$$\mathbb{I} = \prod_{1} (\mathbf{Z}/p\mathbf{Z})^{\times},$$

which therefore defines, for every modulus $q \in \mathbf{M}$, a unique invertible residue class a_q modulo q by the Chinese Remainder Theorem; we then put

$$\psi(X; q, a) = \psi(X; q, a_q).$$

– $\mu(n)$ denotes the Möbius function, $\varphi(n)$ the Euler function, $\tau(n)$ the "number of divisors" function and $\tau_k(n)$ its generalization to the number of representations $n = d_1 \cdots d_k$, i.e.,

$$\tau_k(n) = \sum_{d_1 \cdots d_k = n} 1, \qquad \sum_{n \geqslant 1} \tau_k(n) n^{-s} = \zeta(s)^k$$

(for instance $\tau_k(p) = k + 1$ for p prime), so that $\tau = \tau_2$.

(2) A standard terminology is "smooth" instead of friable.

– $f \star g$ denotes the Dirichlet convolution of arithmetic functions f and g, defined by

$$f \star g(n) = \sum_{d|n} f(d)g(n/d) = \sum_{ab=n} f(a)g(b)$$

for all $n \geqslant 1$.

– The notation $f = O(g)$ and $f \ll g$ are synonymous: $f(x) = O(g(x))$ for all $x \in D$ means that there exists an "implied" constant $C \geqslant 0$ (which often depends on other parameters, which are clearly mentioned) such that $|f(x)| \leqslant Cg(x)$ for all $x \in D$. This definition (for $O(\cdots)$) differs from that of Bourbaki [2, Chap. V], which is of topological nature.

– We denote by $\mathbf{1}_X$ the characteristic function of a set X. If $\boldsymbol{x} = (x_1, \dots, x_k)$ is a tuple of real or complex numbers, we write $|\boldsymbol{x}| = x_1 \cdots x_k$.

Acknowledgments. Thanks to É. Fouvry and H. Iwaniec as well as to R. de la Bretèche, Ph. Michel and P. Nelson, and to all the Polymath8 participants, especially T. Tao. Thanks also to V. Le Dret for reading the first version and correcting a number of typographical mistakes, and to R. Heath-Brown, Y. de Cornulier, B. Green and Y. Motohashi for comments and corrections.

2. THE GARDEN OF FORKING PATHS

All the results discussed in the introduction take as starting point the breakthrough of Goldston-Pintz-Yıldırım [12], who had proved that

$$\liminf_n \frac{p_{n+1} - p_n}{\log n} = 0 \tag{2}$$

and showed how the bounded gap theorem followed from well-established conjectures on primes in arithmetic progressions. The work of Goldston, Pintz and Yıldırım is described in detail (among other places) in this Seminar in [19]. We first recall briefly the skeleton of this method, using mostly the same notation as [19] to ease comparison (this previous *exposé* discusses the method in greater detail; another masterful exposition is found in [9, §7.13]). This will allow us to describe informally and quickly the main difference between the approaches of Zhang and Maynard.

Let $k \geqslant 2$ be an integer. An *admissible k-tuple* $\boldsymbol{h} = (h_1, \dots, h_k)$ is a k-tuple of integers such that

$$h_1 < h_2 < \cdots < h_k$$

and, for every prime number p, the reductions modulo p of the coordinates of $\boldsymbol{h}$ do not cover all of $\mathbf{Z}/p\mathbf{Z}$, or in other words, such that

$$\nu_{\boldsymbol{h}}(p) = |\{h_1 \pmod p, \dots, h_k \pmod p\}| < p$$

for all primes p. It is elementary, but important, to note that such k-tuples exist for all $k \geqslant 2$: an easy example is to take

$$k < h_1 < \cdots < h_k$$

such that all h_i are primes. (Indeed, for primes $p \leqslant k$, the reductions modulo p of the h_i are then never 0, and for $p > k$, there are not enough of them to cover $\mathbf{Z}/p\mathbf{Z}$.)

We fix $k \geqslant 2$ and an admissible k-tuple $\boldsymbol{h}$. For $n \geqslant 1$, we denote

$$\theta_{\boldsymbol{h}}(n) = \sum_{1 \leqslant i \leqslant k} \theta(n + h_i).$$

Next, given $X > 1$ and non-negative coefficients $(E(n))_{X<n\leqslant 2X}$ (which serve as parameters), we consider the sums

$$Q_1 = \sum_{X<n\leqslant 2X} \theta_{\boldsymbol{h}}(n)E(n), \qquad Q_2 = (\log 3X) \sum_{X<n\leqslant 2X} E(n).$$

The following observation is completely elementary:

LEMMA 2.1. — *Fix an admissible k-tuple $\boldsymbol{h}$. Let $\rho \geqslant 1$ be a real number. Assume that, for $X > 1$ large enough (depending on k and $\boldsymbol{h}$), one can find coefficients $E(n)$ as above such that $Q_1 \geqslant \rho Q_2$.*

Then for all integers m with $0 \leqslant m < \rho$, we have

$$\liminf_{n\to+\infty}(p_{m+n} - p_n) \leqslant h_k - h_1 < +\infty.$$

The main (qualitative) result of Zhang can then be expressed in the following form: for k large enough, and for any fixed admissible k-tuple $\boldsymbol{h}$, one can indeed find the coefficients $E(n)$ as above so that

$$\liminf_X \sup_{E(n)} \frac{Q_1}{Q_2} > 1,$$

(where the liminf will be abbreviated $\liminf_X Q_1/Q_2$ below).

Indeed, this allows one to apply the lemma with $m = 1$; since Zhang proved that one can take $k = 3,500,000$, a simple estimate leads to the quantitative form of Theorem 1.1.

Similarly, Maynard's result shows that, for any fixed $m \geqslant 1$, for k large enough depending on m and for any admissible k-tuple $\boldsymbol{h}$, one can find $E(n)$ such that

$$\liminf_X \frac{Q_1}{Q_2} > m,$$

which leads to Theorem 1.2. For $m = 1$, Maynard [21, Prop. 4.3 (1)] proves that $k = 105$ is suitable, and deduces that

$$\liminf_{n\to+\infty}(p_{n+1} - p_n) \leqslant 600$$

(using a computation by Engelsma of an admissible 105-tuple with "diameter" $h_{105} - h_1 = 600$).

Remark 2.2. — One should view the values of the parameter k which are allowed by the method as the most important outcome of this idea. Although there is a very interesting story in the search for "narrow" admissible k-tuples, this issue is to some extent orthogonal to gaps between primes. (See [24, §4] for a detailed discussion; interestingly, ideas related to the problem of finding *large* gaps between consecutive primes play an important role.) For small enough k, the smallest diameter $h_k - h_1$ of an admissible k-tuple $\boldsymbol{h}$ is known.

Goldston, Pintz and Yıldırım considered coefficients $E(n)$ depending on auxiliary parameters $x \geqslant 1$ (which will be a fixed power of X, that one seeks to have as big as feasible) and $\ell \geqslant 0$, given by

$$E(n) = \Big(\sum_{d|F_{\boldsymbol{h}}(n)} \lambda_d\Big)^2 \tag{3}$$

where

$$F_{\boldsymbol{h}} = (T + h_1)\cdots(T + h_k) \in \mathbf{Z}[T],$$

and the coefficients λ_d are given by

$$\lambda_d = \mu(d)\Big(\log\frac{x}{d}\Big)^{k+\ell} \tag{4}$$

for $1 \leqslant d \leqslant x$, and $\lambda_d = 0$ otherwise.[3][4] Now recall that a real number $\theta \geqslant 0$ is called an *exponent of distribution for the sequence of prime numbers* if for any $\varepsilon > 0$ and for any $A \geqslant 1$, we have

$$\sum_{q\leqslant X^{\theta-\varepsilon}} \max_{(a,q)=1} \Big|\psi(X;q,a) - \frac{\psi(X)}{\varphi(q)}\Big| \ll \frac{X}{(\log X)^A} \tag{5}$$

for all $X \geqslant 2$, where the implied constant depends only on ε and A. Taking such a real number θ and $x = X^{\beta/2}$ where $\beta < \theta$, Goldston, Pintz and Yıldırım proved that

$$\liminf_X \frac{Q_1}{Q_2} \geqslant \beta \times \frac{k(2\ell+1)}{(\ell+1)(k+2\ell+1)} \tag{6}$$

for X large enough (see [19, §5] for instance). The Bombieri-Vinogradov Theorem (or, conditionally, the Generalized Riemann Hypothesis for Dirichlet L-functions) shows that one can apply this for any $\beta < 1/2$, but for these values, we see that the right-hand side of the bound is never > 1.

What can one do in that case? We refer to [19, §6] for a discussion of the idea that leads to the main result (2) of [12], and observe instead that two paths are (in principle) possible to improve this argument and prove the bounded gap property of primes. These paths correspond to the very different nature of the two factors we wrote in (6):

(1) The limiting factor $\beta < \theta$ arises in controlling an *error term* when estimating the sum Q_1; one might try to improve on that estimation, making a larger value of θ (hence β) possible. Morally speaking, this is what Zhang did, although the details are much more intricate.[5]

(2) The factor

$$\frac{k(2\ell+1)}{(\ell+1)(k+2\ell+1)}$$

arises from the ratio of the *main terms* in the study of Q_1 and Q_2, and in particular from the choice of $E(n)$; one could hope to find a better choice of coefficients $E(n)$ that leads to a larger ratio of the main terms, so that even the choice of θ close to $1/2$ allowed by the Bombieri-Vinogradov Theorem leads to bounded gaps. This is essentially what Maynard succeeded in doing. It is really extraordinary to see that this leads to (1) for *any* $m \geqslant 1$, and in fact that the method succeeds also for *any* positive exponent of distribution $\theta > 0$.

[3] In [19], different coefficients (arising by a type of optimization in the Selberg-sieve style) are used, following [13]. Such choices simplify the analysis of Zhang and especially that of Maynard, see Section 8.

[4] The use of an exponent $k+\ell$, which is surprising from the point of view of sieve, arose also independently in the work [16] of Ho and Tsang, as pointed out by R. Heath-Brown.

[5] Note that only a very small improvement of θ is currently known, and even the best possible θ arbitrarily close to 1 is in some sense limited in its effect, and can only lead to (1) for $m = 1$.

3. THE GOLDSTON-PINTZ-YILDIRIM METHOD AND ITS DESCENDANTS

The crucial link between the Goldston-Pintz-Yıldırım method and the distribution of primes in arithmetic progressions comes from the choice of coefficients $E(n)$ in the outline of the previous section.

Zhang's modification of (3) restricts the support of the coefficients λ_d to integers with only small prime factors, more precisely to x^δ-friable integers (i.e., integers with no prime factor $\geqslant x^\delta$) for some parameter $\delta > 0$, which will be rather small in practice.

Maynard's change is much more radical: he replaces the single divisor d in the definition of $E(n)$ by a tuple $\boldsymbol{d} = (d_1, \dots, d_k)$ of positive integers where $d_i \mid n + h_i$, or in other words he considers

$$E(n) = \Big(\sum_{\boldsymbol{d} \mid n+\boldsymbol{h}} \lambda_{\boldsymbol{d}} \Big)^2 \tag{7}$$

where $\boldsymbol{d} = (d_1, \dots, d_k)$ is a k-tuple of positive integers, we write $n+\boldsymbol{h} = (n+h_1, \dots, n+h_k)$ and the divisibility relation in k-tuples means coordinate-wise divisibility. The choice of $\lambda_{\boldsymbol{d}}$ is then a multi-variable adaptation of (4), which we will introduce later. For the moment, we impose the conditions that $\lambda_{\boldsymbol{d}} = 0$ unless the d_i are coprime and satisfy $|\boldsymbol{d}| = d_1 \cdots d_k < x$, and unless $|\boldsymbol{d}|$ is coprime to the discriminant

$$\prod_{i \neq j} (h_i - h_j) \tag{8}$$

of $\boldsymbol{h}$. In particular, if $\lambda_{\boldsymbol{d}}$ is non-zero, the elements h_i are distinct modulo each d_i.

Remark 3.1. — Maynard observes that Selberg [25, p. 245] already suggested the use of distinct divisors d_i of $n + h_i$ in another approach to the twin-prime problem.

We begin by describing, in the general situation of Maynard's weights, how the "main term" and the "error term" are obtained. As in [19], it is convenient to deal with both quadratic forms simultaneously by considering a general sum

$$\sum_{X < n \leqslant 2X} a_n E(n),$$

where we assume that the sequence (a_n) satisfies a sieve-type condition concerning its distribution in arithmetic progressions, which we write in the form

$$\sum_{\substack{X < n \leqslant 2X \\ n \equiv t \ (d)}} a_n = g_t(d) \tilde{X} + r_d(X, t), \tag{9}$$

where $\tilde{X}$ may depend on X, but not on t or d, and $d \mapsto g_t(d)$ is a multiplicative function, which only depends on t modulo d (and may sometimes be zero).

Proposition 3.2. — *With notation as above, in particular $\lambda_{\boldsymbol{d}} = 0$ unless $\boldsymbol{d}$ has coprime entries and $|\boldsymbol{d}| < x$, we have*

$$\sum_{X < n \leqslant 2X} a_n \Big(\sum_{\boldsymbol{d} \mid n+\boldsymbol{h}} \lambda_{\boldsymbol{d}} \Big)^2 = \tilde{X} M + R,$$

where

$$M = \sum_{\boldsymbol{d},\boldsymbol{e}\ coprime} g_{\boldsymbol{h}}([\boldsymbol{d},\boldsymbol{e}])\lambda_{\boldsymbol{d}}\lambda_{\boldsymbol{e}}, \qquad R = \sum_{\boldsymbol{d},\boldsymbol{e}\ coprime} \lambda_{\boldsymbol{d}}\lambda_{\boldsymbol{e}} r_d(X, a([\boldsymbol{d},\boldsymbol{e}])),$$

with the following notation:

- *The lcm is extended to tuples by* $[\boldsymbol{d},\boldsymbol{e}] = ([d_i, e_i])_i$*;*
- *Two tuples* $\boldsymbol{d} = (d_i)$ *and* $\boldsymbol{e} = (e_j)$ *are said to be coprime if* $\gcd(d_i, e_j) = 1$ *for all* $i \neq j$*;*[(6)]
- *The function* $g_{\boldsymbol{h}}$ *is defined for* k*-tuples by*

$$g_{\boldsymbol{h}}(\boldsymbol{d}) = \prod_i g_{-h_i}(d_i); \tag{10}$$

- *For any* k*-tuple* $\boldsymbol{d}$ *with pairwise coprime coordinates, the residue class* $a(\boldsymbol{d})$ *is the unique residue class modulo* $|\boldsymbol{d}|$ *such that*

$$a(\boldsymbol{d}) \equiv -h_i \pmod{d_i}$$

for $1 \leqslant i \leqslant k$*.*

Proof. — This is elementary (compare with [19, Lemme 4.3]). Expanding the square and exchanging the sums, the expression to evaluate becomes

$$\sum_{\boldsymbol{d}}\sum_{\boldsymbol{e}} \lambda_{\boldsymbol{d}}\lambda_{\boldsymbol{e}} \sum_{\substack{\boldsymbol{d}|n+\boldsymbol{h}\\ \boldsymbol{e}|n+\boldsymbol{h}}} a_n.$$

The summation condition means that

$$\begin{cases} n \equiv -h_i \pmod{d_i} \text{ for } 1 \leqslant i \leqslant k \\ n \equiv -h_j \pmod{e_j} \text{ for } 1 \leqslant j \leqslant k. \end{cases}$$

For $\boldsymbol{d}$ and $\boldsymbol{e}$ such that $\lambda_{\boldsymbol{d}} \neq 0$ and $\lambda_{\boldsymbol{e}} \neq 0$, the d_i (resp. e_j) are pairwise coprime so this becomes

$$\begin{cases} n \equiv a(\boldsymbol{d}) \pmod{|\boldsymbol{d}|} \\ n \equiv a(\boldsymbol{e}) \pmod{|\boldsymbol{e}|}, \end{cases}$$

by definition. For $i \neq j$, these conditions further impose

$$h_i \equiv h_j \pmod{(d_i, e_j)}$$

which is impossible unless $(d_i, e_j) = 1$ for $i \neq j$, by the restriction on the support of $\lambda_{\boldsymbol{d}}$ to involve tuples with entries coprime with the discriminant (8).

Thus the sum over $\boldsymbol{d}$ and $\boldsymbol{e}$ is restricted to coprime k-tuples, and the inner sum becomes a congruence modulo $|[\boldsymbol{d},\boldsymbol{e}]| = \prod[d_i, e_i]$, precisely

$$\sum_{n \equiv a([\boldsymbol{d},\boldsymbol{e}]) \pmod{|[\boldsymbol{d},\boldsymbol{e}]|}} a_n$$

by definition of $a([\boldsymbol{d},\boldsymbol{e}])$.

Inserting (9), the term involving the error terms is exactly R, while the first term becomes

$$\tilde{X} \sum_{\boldsymbol{d},\boldsymbol{e}\ \text{coprime}} \lambda_{\boldsymbol{d}}\lambda_{\boldsymbol{e}} g_{a([\boldsymbol{d},\boldsymbol{e}])}(|[\boldsymbol{d},\boldsymbol{e}]|).$$

But since $g_t(d)$ depends only on t modulo d, and since

$$|[\boldsymbol{d},\boldsymbol{e}]| = \prod_i [d_i, e_i]$$

(6) Not for all i and j...

where the factors are coprime, we get by multiplicativity

$$g_{a([\boldsymbol{d},\boldsymbol{e}])}(|[\boldsymbol{d},\boldsymbol{e}]|) = \prod_i g_{-h_i}([d_i,e_i]) = g_{\boldsymbol{h}}([\boldsymbol{d},\boldsymbol{e}]),$$

since $a([\boldsymbol{d},\boldsymbol{e}]) \equiv -h_i \pmod{[d_i,e_i]}$ by construction. □

It is important for Zhang's method (but not for Maynard's) to observe that for a given modulus q arising in R as $q = |[\boldsymbol{d},\boldsymbol{e}]|$, in possibly many different ways, the residue classes $a \pmod q$ that may arise obey the constraint that, for all prime $p \mid q$, there is some j such that

$$a \equiv -h_j \pmod p.$$

In other words, one can say that a is a root modulo q of the polynomial $F_{\boldsymbol{h}}$. A weaker property is that $a \pmod q$ always belongs to the reductions of the set of idelettes

$$X_{\boldsymbol{h}} = \{x = (x_p) \in \mathbb{I} \mid \text{ for all } p,\ x_p \equiv -h_j \text{ for some } j\}.$$

We first apply this to compute Q_2. As in many sieve problems, a *preliminary sieve* turns out to be technically very useful (see [9] for a general discussion; this is also sometimes called the *W-trick*). This amounts, instead of just putting $a_n = 1$, to taking a_n to be the characteristic function of an arithmetic progression $n \equiv n_0 \pmod W$, where W is the product of primes $p \leqslant D$, where $D = D(X)$ grows very slowly, typically $D = \log\log\log X$. To ensure that there are integers n where $n + h_i$ are all coprime to W, we select n_0 so that $p \nmid n_0 + h_i$ for $1 \leqslant i \leqslant k$ and $p \leqslant D(X)$. There exist such n_0, simply because $\boldsymbol{h}$ is admissible.

This trick simplifies many computations, intuitively because the singular series

$$\mathfrak{S}(\boldsymbol{h}) = \prod_p \Big(1 - \frac{\nu_{\boldsymbol{h}}(p)}{p}\Big)\Big(1 - \frac{1}{p}\Big)^{-k}$$

(which appear frequently in [19]) are then "replaced" by

$$\mathfrak{S}_W(\boldsymbol{h}) = \prod_{p>D} \Big(1 - \frac{\nu_{\boldsymbol{h}}(p)}{p}\Big)\Big(1 - \frac{1}{p}\Big)^{-k} = 1 + o(1)$$

(in fact, this replacement happens almost invisibly, which explains why the change is so convenient).

To fit this preliminary sieve, we assume that $\lambda_{\boldsymbol{d}}$ is supported on integers with no prime factor less than $D = D(X)$. We then have (9) with

$$\tilde{X} = \frac{X}{W}, \quad g_t(d) = \begin{cases} \frac{1}{d} & \text{if } (d,W) = 1 \\ 0 & \text{otherwise} \end{cases}$$
$$|r_d(X,t)| \leqslant 1,$$

so that by an immediate estimate of the error term, we obtain

$$Q_2 = \tilde{X}(\log 3X) \sum_{\boldsymbol{d},\boldsymbol{e} \text{ coprime}} \frac{\lambda_{\boldsymbol{d}}\lambda_{\boldsymbol{e}}}{|[\boldsymbol{d},\boldsymbol{e}]|} + O(x^2(\log x)^{k+1}\|\lambda\|_\infty^2), \tag{11}$$

where the implied constant depends only on k.

For Q_1, we consider the sequences $b_n = a_n\theta(n+h_i)$ for $1 \leqslant i \leqslant k$, i.e., shifted prime (powers) for n restricted to $n \equiv n_0 \pmod W$. We then write (9) with

$$\tilde{X} = \frac{X}{\varphi(W)}, \quad g_t(d) = \begin{cases} \frac{1}{\varphi(dW)} & \text{if } (dW, t+h_i) = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Applying the proposition and summing over i, the main term for Q_1 becomes

$$M_1 = \frac{X}{\varphi(W)} \sum_{i=1}^{k} \sum_{\substack{\boldsymbol{d},\boldsymbol{e} \text{ coprime} \\ d_i = e_i = 1}} \frac{\lambda_{\boldsymbol{d}}\lambda_{\boldsymbol{e}}}{\varphi([\boldsymbol{d},\boldsymbol{e}])}, \tag{12}$$

where $\varphi([\boldsymbol{d},\boldsymbol{e}])$ has the obvious meaning (note that $g_{\boldsymbol{h}}([\boldsymbol{d},\boldsymbol{e}]) = 0$ unless $[d_i, e_i] = 1$). The error term, on the other hand, involves the distribution of primes in arithmetic progressions. Then, inserting the definition of $r_d(X,t)$ for the shifted primes, and applying the prime number theorem and trivial estimates of the contributions of the prime powers, we obtain for any $A \geqslant 1$ the first estimate

$$R_1 \ll \Big\{ \sum_{i=1}^{k} \sum_{\substack{\boldsymbol{d},\boldsymbol{e} \text{ coprime} \\ d_i = e_i = 1}} |\lambda_{\boldsymbol{d}}\lambda_{\boldsymbol{e}}| \Big| \boldsymbol{\psi}(X; W|[\boldsymbol{d},\boldsymbol{e}]|, a') - \frac{\boldsymbol{\psi}(X)}{\varphi(W[\boldsymbol{d},\boldsymbol{e}])} \Big| + \frac{X}{(\log X)^A} \Big\} \tag{13}$$

where a' is an idelette congruent to $a(\boldsymbol{d},\boldsymbol{e})$ modulo $|[\boldsymbol{d},\boldsymbol{e}]|$ and to n_0 modulo W, and where the implied constant depends on k and A. The real work now begins. In the next sections, we present the approaches of Zhang and Maynard independently of each other. We have chosen to proceed in chronological order, although Maynard's approach is much quicker and simpler. Hence, those readers who are (naturally) interested in understanding a short proof of bounded gaps between primes may begin with Section 8 (although we do not give full details).

4. ZHANG'S THEOREM ON PRIMES IN ARITHMETIC PROGRESSIONS

At the time of the work of Goldston, Pintz and Yıldırım, part of the excitement was due to the fact that one *knew* some results concerning primes in arithmetic progressions going beyond the range of the Generalized Riemann Hypothesis, as required to make (6) successful. These results, due to Fouvry–Iwaniec [7], Fouvry [4] and Bombieri–Friedlander–Iwaniec [1] are of the following type: for some (explicit) $\theta > 1/2$ (the crucial feature), for any *well-factorable* function λ supported on $q \leqslant X^\theta$ (a concept which we will introduce a bit more precisely in Remark 5.4 in Section 5) and for any integer $a \geqslant 1$, we have

$$\sum_{q \leqslant X^\theta} \lambda(q) \Big(\boldsymbol{\psi}(X; q, a) - \frac{\boldsymbol{\psi}(X)}{\varphi(q)} \Big) \ll \frac{X}{(\log X)^A} \tag{14}$$

where the implied constant depends now on A, *but also on* a (compare with (5)).

This dependency on the residue class is the main issue, since in the application to bounded gaps, one wishes a to range over all zeros of $F_{\boldsymbol{h}}$ modulo q. (It had already been shown by Motohashi and Pintz [23] that the sieve argument of Goldston, Pintz and Yıldırım could incorporate a well-factorable function $\lambda(q)$.) The source of the dependency on a lies deep within the arguments of Bombieri, Fouvry, Friedlander and Iwaniec, from the use of the

spectral theory of automorphic forms to estimate sums of (incomplete) Kloosterman sums (see [7, Lemma 2] for instance, which is a basic result of Deshouillers and Iwaniec).

One of Zhang's insight is that many of the techniques introduced in these papers are nevertheless still useful without this restriction, and that it is possible to exploit some features of the residue classes that appear in Section 3 in order to pass beyond the exponent of distribution 1/2. In this sense, his methods are closer in part to older papers of Fouvry and Fouvry–Iwaniec, for instance [6], which did not exploit the spectral theory of automorphic forms; thus the dependency on a is milder in [6] (which also does not cover the distribution of primes, but of integers with no small prime factors), although only $1 \leqslant |a| \leqslant X$ is allowed, due to the use (see [6, p. 138, line 5]) of a summation by parts in which the archimedean size of a matters (see also [5, p. 632, last paragraph of §IV]).

Although the best known exponent in Zhang's equidistribution theorem is currently quite a bit smaller than the best known in (14) (any $\theta < 4/7$ in [1], improving $\theta < 9/17$ in [7]), this brilliantly leads to Theorem 1.1.

We state a version emerging from Polymath8, which is just slightly more general:

THEOREM 4.1 (Zhang). — *There exist explicit real numbers $\theta > 1/2$ and $\delta > 0$ such that for any $A \geqslant 1$, for any idelette $a \in \mathbb{I}$, we have*

$$\sum_{\substack{q \leqslant X^\theta \\ q \in \mathbf{M} \text{ is } X^\delta\text{-friable}}} \left| \psi(X; q, a) - \frac{\psi(X)}{\varphi(q)} \right| \ll \frac{X}{(\log X)^A} \tag{15}$$

where the implied constant depends only on A, and in particular is independent of a.

Remark 4.2. — (1) Since, for given X, only finitely many q are involved, one can equivalently say, using the Chinese Remainder Theorem, that a is an arbitrary integer coprime to all primes $\leqslant X^\delta$.

(2) We wish to emphasize that Zhang's methods and results retain considerable interest despite Maynard's discoveries that led to Theorem 1.2 (just like the methods of Wiles and Taylor-Wiles would retain their interest even if someone found an elementary proof of Fermat's Great Theorem.)

(3) Zhang's paper established this (up to minor differences) for $\delta = 1/1168$ and $\theta = 1/2+2\delta$. The Polymath8 values allow any $\theta > 1/2 + 7/300$, provided δ is small enough (compared to $\theta - 1/2$).

In other words, Zhang obtains an exponent of distribution larger than 1/2 for primes in arithmetic progressions by (i) imposing a restriction on the moduli $q \in \mathbf{M}$ (some amount of friability, or the existence of good factorizations, which are in the spirit of Iwaniec's well-factorable weights); (ii) imposing a condition on the residue classes modulo q which fits both the problem, and the conditions on q: he can (in the Polymath8 version, which is a bit finer than the original) choose arbitrary residue classes modulo all primes p, but then extends these to all moduli by the Chinese Remainder Theorem.

We can quickly explain how this implies the bounded gap property using the Goldston–Pintz–Yıldırım method. The multiplicative structure of idelettes and the uniformity in a in Theorem 4.1, play an important role.

We take $x = x^{\theta/2}$ and consider the weight $\lambda_{\boldsymbol{d}}$ defined following the original Goldston-Pintz-Yıldırım method (4) with the restriction to moduli which are x^{δ}-friable for some fixed $\delta > 0$ small enough that Theorem 4.1 holds. Then:

– The main terms (11) and (12) are evaluated; it emerges (without too much difficulty) that M_1 and M_2 satisfy

$$\liminf_X \frac{M_1}{M_2} \geqslant \beta \times \frac{k(2\ell+1)}{(\ell+1)(k+2\ell+1)} - \xi,$$

where $\xi \geqslant 0$ depends on δ, θ and the length k of the k-tuple; it is seen that ξ is extremely small provided k is large enough, even when δ and $\theta - 1/2$ are very small (in Zhang's paper, δ and $\theta - 1/2$ are about $1/1000$, $k = 3,500,000$, and $\xi < e^{-1200}$, for instance). This means that the conclusion (6) is basically unchanged, provided the "hard" error term R_1 in (13) satisfies (for these given k, δ, θ) the estimate

$$R_1 \ll \frac{X}{(\log X)}.$$

– Then, using (15), this estimate of R_1 follows from Theorem 4.1 using the remark after the proof of Proposition 3.2, using an averaging trick: roughly speaking, arguing as if $W = 1$ and $|\lambda_{\boldsymbol{d}}| \leqslant 1$ for simplicity, denoting

$$E(X; q, a) = \Big|\psi(X; q, a) - \frac{\psi(X)}{\varphi(q)}\Big|$$

one gets

$$R_1 \ll \sum_{q \leqslant x^2} \sum_{a \in X_{\boldsymbol{h}} \pmod q} E(X; q, a) \leqslant \Big(\sum_{q \leqslant x^2} |X_{\boldsymbol{h}} \pmod q| E(X; q, a)\Big)^{1/2} \times \Big(\sum_{q \leqslant x^2} \frac{1}{|X_{\boldsymbol{h}} \pmod q|} \sum_{a \in X_{\boldsymbol{h}} \pmod q} E(X; q, a)\Big)^{1/2}$$

by the Cauchy-Schwarz inequality, and the bounds

$$|X_{\boldsymbol{h}} \pmod q| \leqslant k^{\omega(q)}, \qquad E(X; q, a) \ll \frac{X}{\varphi(q)}$$

together with (15) allow one to conclude, since this bound is uniform over a.

On the other hand, the main steps (to be greatly expanded later) of the proof of Theorem 4.1 are roughly the following :

– A combinatorial and analytical decomposition of the von Mangoldt function reduces the problem to estimates of the same quality as (15) for certain arithmetic functions of the type

$$\alpha_1 \star \alpha_2, \qquad \alpha_1 \star \alpha_2 \star \alpha_3,$$

(bilinear and trilinear sums) where α_i is supported on integers $n \sim M_i$ with $M_1 M_2$ (resp. $M_1 M_2 M_3$) very close to X, and with M_1, M_2 (resp. M_1, M_2 and M_3) in relatively precise ranges, which are crucial for the success of the method. The bilinear sums and the trilinear sums are treated separately with different techniques.

– The bilinear forms are handled by means of adaptations of the *Linnik dispersion method* already used successfully by Fouvry–Iwaniec [6, 7] and Bombieri–Friedlander–Iwaniec [1]; as in those works, this leads to incomplete exponential sums over finite rings $\mathbf{Z}/q\mathbf{Z}$, and crucial estimates are derived by an application of Weil's theory of exponential sums in one variable;
– The trilinear sums are closely related to those that appear naturally in studying the exponent of distribution of the ternary divisor function

$$\tau_3(n) = \sum_{abc=n} 1,$$

which was first done (beyond the range of 1/2) by Friedlander and Iwaniec [11]. Here again, and in fact rather more directly than for bilinear sums, there arise incomplete exponential sums modulo q. However, if those are expanded entirely, they are exponential sums in three variables, and square-root cancellation is required, which is a very difficult problem in general. For Friedlander and Iwaniec, the required estimate was established by Birch and Bombieri, using the first proof by Deligne of the Riemann Hypothesis (the Weil conjectures for smooth projective algebraic varieties over finite fields). Zhang simply quoted this estimate. However, as will be explained in Section 7, it is here much more straightforward to keep the exponential sums in the form they arise originally, which is as *one-variable sums of summands which are themselves exponential sums.* When the modulus is prime, these summands are *trace functions* of étale sheaves on the affine line over $\mathbf{F}_p$, and the general formalism of Deligne's most general version of the Riemann Hypothesis (weight theory for compactly supported étale cohomology with coefficients in constructible ℓ-adic sheaves) leads to a quick (and conceptually clear) proof of the Birch–Bombieri estimate.

Remark 4.3. — There is a fair amount of flexibility available in this strategy. For instance, the Polymath8 project discovered that a sufficiently incisive treatment of the bilinear sums (depending only on Weil's theory, but more complex than Zhang's argument, and depending on some numerical bounds) allows one to avoid entirely the trilinear sums, and therefore the use of Deligne's forms of the Riemann Hypothesis over finite fields. It also found that one of the two flavors of bilinear estimates could be dispensed with (working therefore only with one type of bilinear forms and the trilinear ones).

5. COMBINATORIAL DECOMPOSITIONS OF PRIMES AND VINOGRADOV'S BILINEAR GAMBIT

An absolutely crucial ingredient, which is an extraordinary achievement of Vinogradov (one of the most *conceptual* insight of analytic number theory) is the fact that (1) the von Mangoldt function (or other functions closely related to the characteristic function of the set of primes, e.g., the Möbius function) can be decomposed as a linear combination of more elementary functions, which are (genuine) Dirichlet convolutions of at least two arithmetic functions; (2) the remarkable understanding that averages involving such convolutions, like any sum that can be expressed in terms of bilinear expressions, can often be estimated in situations where

very little direct knowledge of the factors is available, simply by exploiting the bilinear shape and the coefficients it involves.

It is worth emphasizing that these two principles are at the heart of the modern versions of the Bombieri–Vinogradov theorem, although the first proofs were different. Hence, the underlying ideas are equally vital in understanding Maynard's proof (which uses the Bombieri–Vinogradov theorem) as in understanding Zhang's equidistribution theorem, if one begins from first principles.

A basic result is the following:

PROPOSITION 5.1 (Heath-Brown identity). — *Let $J \geqslant 1$ be a fixed integer. Let $X \geqslant 1$ be an integer and let K be an arithmetic function supported on $X < n \leqslant 2X$. Let $B \geqslant 1$ be an integer. There exist $C \geqslant 1$, depending only on J and B, such that the sum*

$$\sum_n K(n)\Lambda(n), \tag{16}$$

can be written as a linear combination, with coefficients $\ll (\log X)^C$, of $\ll (\log X)^C$ sums of the type

$$\Sigma(\boldsymbol{M},\boldsymbol{N}) = \sum_{m_1,\cdots,m_J} \cdots \sum \alpha_1(m_1)\alpha_2(m_2)\cdots\alpha_J(m_J) \times \sum_{n_1,\cdots,n_J} \cdots \sum V_1(n_1)\cdots V_J(n_J)K(m_1\cdots m_J n_1\cdots n_J) \tag{17}$$

where

- *The parameters $\boldsymbol{M} = (M_1,\cdots,M_J)$ and $\boldsymbol{N} = (N_1,\cdots,N_J)$ are J-tuples of real numbers in $[1/2, 2N]^{2J}$ which satisfy*

$$N_1 \geqslant N_2 \geqslant \cdots \geqslant N_J, \qquad M_i \leqslant X^{1/J}, \qquad M_1\cdots M_J N_1\cdots N_J \asymp_J X; \tag{18}$$

- *Defining*

$$\Delta = 1 + \frac{1}{(\log X)^B},$$

  the arithmetic functions $m \mapsto \alpha_i(m)$ are bounded by 1 and supported in $[M_i, \Delta M_i]$;
- *The smooth functions $V_i(x)$ are compactly supported in $[N_i, \Delta N_i]$, and their derivatives satisfy*

$$y^k V_i^{(k)}(y) \ll (\log X)^C,$$

  for all $y \geqslant 1$, where the implicit constants depend only on k and B.

Hint. — First of all, one uses the Heath–Brown identity: for $1 \leqslant n < 2X$, we have

$$\Lambda(n) = -\sum_{j=1}^{J} (-1)^j \binom{J}{j} \sum_{m_1,\cdots,m_j \leqslant X^{1/J}} \mu(m_1)\cdots\mu(m_j) \sum_{m_1\cdots m_j n_1\cdots n_j = n} \log n_1.$$

Multiplying by $K(n)$ and summing over n, one obtains a linear combination of sums of the right shape, except that the ranges of the variables are not exactly as stated, and that the n_i variables are not weighted with compactly supported functions. Using suitable partitions of unity, these restrictions can be relaxed to derive the statement above (see [10, Prop. 3.1] for a particularly elegant implementation of these steps). □

Remark 5.2. — In a first reading (and in many applications), one can assume that $B = 0$, so that the arithmetic functions α_i and the smooth functions V_i are supported in a dyadic segment.

If one splits the set of $2J$ variables m_i and n_j in two non-empty subsets, one sees that each sum $\Sigma(\boldsymbol{M}, \boldsymbol{N})$ can be expressed as a bilinear form

$$\Sigma(\boldsymbol{M}, \boldsymbol{N}) = \sum_m \sum_n \alpha(m)\beta(n)K(mn).$$

For instance, if the m variable arises by taking together m_i for $i \in A$ and n_j for $j \in B$, we have

$$\beta(n) = \sum_{(m_i)_{i\in A}} \sum_{(n_j)_{j\in B}} \prod_{i\in A} \alpha_i(m_i) \prod_{j\in B} V_j(n_j).$$

It is relatively clear that in writing such a bilinear expression, most control of the values of α and β is lost. More precisely, one still keeps estimates for the L^2-norms of these coefficients (or other norms), and some other information does remain beyond this, as we will see (the Siegel-Walfisz property, for instance). Moreover, in key special cases or variants of this gluing, one may retain specific information that can be exploited. Thus writing sums over primes as combinations of such bilinear expressions seems to be a huge gambit, with seemingly little compensation for the loss of control of the coefficients. This impression is however misleading: the *bilinear structure* contains in itself a source of rich information.

By definition, assuming we only have control of the L^2-norm of α and β, the best estimate that can be obtained for $\Sigma(\boldsymbol{M}, \boldsymbol{N})$ is in terms of the *norm* $\|B_{\boldsymbol{M},\boldsymbol{N}}\|$ of the bilinear form $B_{\boldsymbol{M},\boldsymbol{N}}$ with coefficients $\gamma(m, n) = K(mn)$: one has

$$|\Sigma(\boldsymbol{M}, \boldsymbol{N})| \leqslant \|B_{\boldsymbol{M},\boldsymbol{N}}\| \, \|\alpha\| \, \|\beta\|$$

where $\|\alpha\|$ and $\|\beta\|$ refer to the euclidean norms of the corresponding vectors, and $\|B_{\boldsymbol{M},\boldsymbol{N}}\|$ is the smallest real number for which this inequality holds for all α and β.

It is quite remarkable that this can be a useful bound, provided only that neither

$$M = \prod_i M_I, \quad \text{nor } N = \prod_j N_j$$

are too small (i.e., provided the expression $B_{\boldsymbol{M},\boldsymbol{N}}$ is "genuinely" bilinear).

We illustrate the most basic form of this *bilinear principle*. Variants of this argument occur many times in analytic number theory, although (as in Zhang's work) one often uses more precise structural features of the bilinear form when going through concrete estimates.

Using the Cauchy-Schwarz inequality, we write

$$|\Sigma(\boldsymbol{M}, \boldsymbol{N})|^2 \leqslant \Big(\sum_m |\beta(m)|^2\Big)\Big(\sum_m \Big|\sum_n \alpha(n)\gamma(m, n)\Big|^2\Big) = \|\beta\|^2\Sigma'(\boldsymbol{M}, \boldsymbol{N})$$

(say) and then transform $\Sigma'(\boldsymbol{M}, \boldsymbol{N})$ by expanding the square, and exchanging the sums to move the sum over n inside:

$$\Sigma'(\boldsymbol{M}, \boldsymbol{N}) = \sum_{m_1,m_2} \sum \alpha(m_1)\overline{\alpha(m_2)}\Delta_\gamma(m_1, m_2),$$

where

$$\Delta_\gamma(m_1, m_2) = \sum_n \gamma(m_1, n)\overline{\gamma(m_2, n)}.$$

The point is that $\Delta_\gamma(m_1, m_2)$ only involves the coefficients $\gamma(m, n)$ defining the bilinear form, and not the arithmetic functions α and β. In particular, this leads easily to the bound

$$\|B_{\boldsymbol{M},\boldsymbol{N}}\|^2 \leqslant \max_{m_1} \sum_{m_2} |\Delta_\gamma(m_1, m_2)|.$$

Thus the issue becomes that of estimating the *correlation sums* $\Delta_\gamma(m_1, m_2)$. We think of the coefficients $\gamma(m, n)$ as having bounded size, but oscillating rather randomly. The sums then often obey a kind of dichotomy: for the "diagonal pairs" $m_1 = m_2$, the sum is large because it is a sum of positive quantities, namely

$$\Delta_\gamma(m, m) = \sum_n |\gamma(m, n)|^2$$

will be typically of size roughly N. On the other hand, we can hope that the "non-diagonal" sums $\Delta_\gamma(m_1, m_2)$ with $m_1 \neq m_2$ should be quite a bit smaller than this trivial bound (of size N) because the terms $\gamma(m_1, n)\overline{\gamma(m_2, n)}$ have oscillating phases or signs.

In practice, in view of the great variety of situations and patterns that arise, one tends to perform these steps independently for each case, instead of simply quoting the norm of a bilinear form. For instance, the $\gamma(m, n)$ might themselves be sums

$$\gamma(m, n) = \sum_r \sum_s g(r, s, m, n)$$

and it might be worth pulling out one, or both, of the r and s sums before estimating the bilinear form. Another frequent phenomenon is that the diagonal case might contain more than the pairs (m_1, m_1), in which case the estimates usually depend on these not being too numerous. Finally, adjustments might be made (when further parameters are available) to help the contributions of the diagonal and non-diagonal terms match, since the final bound will be the worse of the two.

Example 5.3. — If we take the case $\gamma(m, n) = K(mn)$, the non-diagonal case occurs when $K(m_1 n)$ and $K(m_2 n)$ have relatively independent phases. This does not always happen, and it fails entirely when $K(n)$ is a multiplicative function (such as the Möbius function, or a Dirichlet character), in which case

$$K(m_1 n)\overline{K(m_2 n)} = K(m_1)\overline{K(m_2)}|K(n)|^2$$

if $(m_1, n) = (m_2, n) = 1$ (or if K is totally multiplicative).

In this case, however, the L-function techniques based on zero-free regions of Dirichlet L-functions do lead to some estimates (even though they are much weaker than expected, based on the Generalized Riemann Hypothesis). It is essential usually to combine the Vinogradov method with the basic equidistribution statements (see below the discussion of the Siegel-Walfisz property).

On the other hand, if we take $K(n) = f(n + 1)$, even when f is multiplicative, the principle can sometimes be implemented. Indeed, one of Vinogradov's first application was to $K(n) = \chi(n + 1)$, where χ is a Dirichlet character modulo q. The results obtained for

$$\sum_{p \leqslant X} \chi(p + 1)$$

are, in terms of the uniformity of q and X, *beyond* the reach of a treatment based on writing

$$\sum_{p\leqslant X}\chi(p+1)=\sum_{a\pmod q}\chi(a+1)\sum_{\substack{p\leqslant X\\ p\equiv a\pmod q}}1$$

and applying the Generalized Riemann Hypothesis.

Remark 5.4. — Because of this fundamental bilinear structure, many arithmetic problems involving sums

$$\sum_n\lambda(n)K(n)$$

can be attacked if they can be re-arranged as linear combinations of bilinear forms. This is one motivation for the introduction of *well-factorable* functions by Iwaniec in his work on the linear sieve (see, e.g., [9, §12.7]): roughly, λ is of this type if it is supported on integers $n\leqslant X$ and, for *any* factorization $X=MN$ (with M and N real numbers $\geqslant 1$), we can write $\lambda=\alpha\star\beta$ for some arithmetic functions α (resp. β) supported on integers $\leqslant M$ (resp. $\leqslant N$), and satisfying some natural size condition. Then the sum can be re-arranged bilinearly with complete flexibility in the respective sizes of the variables. We will see that Zhang uses very similar structures.

We now come back to the situation of primes in arithmetic progressions and to the basic sum

$$\sum_{q\leqslant Q}\Big|\psi(X;q,a)-\frac{\psi(X)}{\varphi(q)}\Big|, \tag{19}$$

of Theorem 4.1, for a fixed $a\in\mathbb{I}$, with possible restrictions on q and with Q meant to depend on X and to be as large as possible. This is usually treated as an average over q of sums over primes

$$\sum_{X<n\leqslant 2X}\Lambda(n)\boldsymbol{\delta}_{q,a}(n)$$

where

$$\boldsymbol{\delta}_{q,a}(n)=\mathbf{1}_{n\equiv a\pmod q}-\frac{1}{\varphi(q)}\mathbf{1}_{(n,q)=1}$$

is the normalized (to have average zero) characteristic function of a primitive residue class $a\in(\mathbf{Z}/q\mathbf{Z})^\times$ for $q\geqslant 1$ (or $q\in\mathbf{M}$), or it may be thought of as fitting directly the context of (16) that we described, by taking

$$K(n)=\sum_{q\leqslant Q}c(q)\boldsymbol{\delta}_{q,a}(n),\qquad\text{for } X<n\leqslant 2X$$

where the $c(q)$ are the signs of

$$\Big|\psi(X;q,a)-\frac{\psi(X)}{\varphi(q)}\Big|.$$

In this case, one might as well consider this type of expressions with arbitrary complex numbers $c(q)$ with modulus $|c(q)|\leqslant 1$. This is also convenient to introduce restrictions on the moduli q in (19), by taking $c(q)=0$ for other q, and we will take this approach (also partly for the sake of variety).

Using this viewpoint, the goal is to obtain, for fixed $a \in \mathbb{I}$ and suitable $(\boldsymbol{M}, \boldsymbol{N})$, the estimates

$$\Sigma(\boldsymbol{M}, \boldsymbol{N}) \ll X(\log X)^{-A} \tag{20}$$

for an arbitrary $A \geqslant 1$, where the implied constant depends only on A. The ranges of pairs $(\boldsymbol{M}, \boldsymbol{N})$ should be such that, considering the size of q with respect to X, they cover all the terms of the Heath-Brown identity, and therefore lead to

$$\sum_{X<n\leqslant 2X} \Lambda(n)K(n) \ll \frac{X}{(\log X)^A}.$$

We will say that $(\boldsymbol{M}, \boldsymbol{N})$ is *feasible* when (20) holds. The point is that, depending on the ranges involved, different techniques and methods are applicable, so that the estimate may arise from a variety of directions.

It will be convenient to denote

$$\mathsf{m} = \frac{\log M}{\log X}, \qquad \mathsf{n} = \frac{\log N}{\log X},$$

so that $\mathsf{m} + \mathsf{n} = 1 + o(1)$, and similarly for the range of other variables that may occur, all measured in logarithmic scale where the main asymptotic variable X has size 1.

In this language, here is roughly the result that Zhang obtains:

THEOREM 5.5 (Zhang's bilinear estimates). — *For every $\kappa > 0$ small enough, there exist explicit constants $\varpi < 1/20$ and $\delta > 0$ with the following property: with notation as above, $(\boldsymbol{M}, \boldsymbol{N})$ is feasible provided*

(1) *The coefficients $c(q)$ are supported on X^δ-friable moduli $q \in \mathbf{M}$ such that*

$$q \leqslant Q \leqslant X^{1/2+2\varpi};$$

(2) *The N variable has length satisfying*

$$\frac{1}{2} - \frac{1}{10} - \kappa + o(1) < \mathsf{n} < \frac{1}{2} + \frac{1}{10} + \kappa + o(1).$$

Note that it is intuitively clear that this result is harder as κ becomes larger.

Is this enough to cover the whole range of bilinear forms that arise for some application of Proposition 5.1? Not quite, but it turns out that if there are enough variables involved (i.e., if the proposition is applied with J large enough), then the gap is relatively small. Precisely, we have the following combinatorial lemma (in the version from [24, Lemma 5.1]):

LEMMA 5.6. — *Let $1/10 < \sigma < 1/2$ be a fixed real number, let $n \geqslant 1$, and let $t_1, \dots, t_n$ be non-negative real numbers such that $t_1 + \cdots + t_n = 1$. Then one of the following three statements holds:*

(1) *There is a t_i with $t_i \geqslant 1/2 + \sigma$.*

(2) *There is a partition $\{1, \dots, n\} = S \cup T$ such that*

$$\frac{1}{2} - \sigma < \sum_{i\in S} t_i \leqslant \sum_{i\in T} t_i < \frac{1}{2} + \sigma.$$

(3) *There exist distinct i, j, k with $2\sigma \leqslant t_i \leqslant t_j \leqslant t_k \leqslant 1/2 - \sigma$ and*

$$t_i + t_j,\ t_i + t_k,\ t_j + t_k \geqslant \frac{1}{2} + \sigma.$$

Furthermore, if $\sigma > 1/6$, then the last alternative cannot occur.

Hints. — Assume (1) and (2) are false; then say that a subset $S \subset \{1, \dots, n\}$ is *large* if the sum of the t_i, $i \in S$, is $\geqslant 1/2 + \sigma$. If S is not large ("small"), then the sum is $\leqslant 1/2 - \sigma$ (because (2) fails). Now say that i is *powerful* if there is a small S such that $S \cup \{i\}$ is large. Then prove that there exist exactly three powerful elements i, j and k, which can be arranged to satisfy (3).

Finally, if $\sigma > 1/6$ then $1/2 - \sigma < 2\sigma$, so that the conditions in (3) are incompatible. □

Remark 5.7. — The last statement is best possible in the sense that, for $1/10 < \sigma \leqslant 1/6$, taking $n = 3$ and $(t_1, t_2, t_3) = (2\sigma, 1/2 - \sigma, 1/2 - \sigma)$, only the third part of the lemma holds. Also, if $\sigma = 1/10$, then

$$(1/5 - \delta, 1/5, 1/5, 2/5 + \delta),$$

for small enough δ, is a quadruple where none of the three conclusions applies. (See [24, Remark 5.2] for more examples).

To apply this fact, assuming Theorem 5.5 is known, we pick $\sigma = 1/10 + \kappa$ for $\kappa > 0$ small enough that it applies. We then invoke the Heath-Brown identity with $J > 1/(2\sigma)$ (e.g., $J = 10$), and then consider the lemma with $n = 2J$. Defining $t_i = \log M_i / \log X$, $t_{J+i} = \log N_i / \log X$ for $1 \leqslant i \leqslant J$ (with minor adjustment to ensure that the sum is exactly 1), we see that Case (2) corresponds precisely to the range of bilinear forms covered by Theorem 5.5 when arranging the variables according to whether $i \in S$ or $i \in T$.

Moreover, with this choice of J, the first case, where some t_i is $\geqslant 1/2 + \sigma$, is in fact the easiest to handle. Indeed, note this can happen only if $i > J$: we have $t_i \geqslant 1/2 + \sigma > 2\sigma > 1/J$, and by (18), this is incompatible with the lengths of the m-variables.

Now, referring back again to Proposition 5.1, the fact that $i > J$ means that i corresponds to a "long" variable n_j with a smooth weight $V(n_j)$, perturbed by the convolution of the remaining variables. But this variable n_j is then very well distributed in arithmetic progressions to moduli q as large as M_j (essentially) and the sum over the remaining variables is handled trivially (see [24, p. 79], for instance).

From this, we conclude that Theorem 5.5 fails to cover all necessary ranges to derive Theorem 4.1 from Proposition 5.1, simply because of the appearance of the constant $1/10$ in this theorem, instead of a number $< 1/6$ (which would eliminate the third case of the lemma). As it turns out, Theorem 5.5 *does* hold with $1/6$ replacing $1/10$ (this is an outcome of [24]), but Zhang did not prove this. He therefore had to handle the last situation.

In terms of arithmetic functions, denoting by α_1, α_2, α_3 the functions in Proposition 5.1 associated to the variables with indices i, j, k, and by β the convolution of all the remaining variables, this means that one must handle sums of the type

$$\sum_n (\beta \star \alpha_1 \star \alpha_2 \star \alpha_3)(n) K(n) \tag{21}$$

where α_i is supported around N_i, and β around M with

$$\mathsf{n}_1 + \mathsf{n}_2 + \mathsf{n}_3 + \mathsf{m} = 1 + o(1), \quad \mathsf{n}_i + \mathsf{n}_j \geqslant \frac{1}{2} + \frac{1}{10} + \kappa + o(1), \tag{22}$$

$$\frac{1}{5} + 2\kappa + o(1) < \mathsf{n}_i < \frac{1}{2} - \frac{1}{10} - \kappa + o(1) \tag{23}$$

for $1 \leqslant i \neq j \leqslant 3$ (resp. $1 \leqslant i \leqslant 3$), where $\kappa > 0$ is the same number arising from the range of bilinear forms treated in Theorem 5.5.

We will discuss the proof of Theorem 5.5 in the next section, and then discuss how Zhang handled (21) in Section 7.

6. BILINEAR DISCREPANCY ESTIMATES

Our objective in this section is to present some basic ideas that lead to Theorem 5.5. We especially want to clarify where the assumptions on the ranges of the variables play a role, how the restriction to a fixed idelette appears, and how (some degree of) friability of the moduli q involved plays a crucial role.

There is a first important point, already hinted that: in the bilinear forms arising from the Heath-Brown identity, the function β do retain an important property: they satisfy the *Siegel-Walfisz property*, in the sense that (provided the length of the variable n is not too short) the values $\beta(n)$ are *individually* equidistributed modulo q, uniformly for q of size $\ll (\log X)^B$ for every $B \geqslant 1$. Precisely, if $N \gg X^\varepsilon$ for some fixed $\varepsilon > 0$, they satisfy the bounds

$$\sum_{n\equiv a \pmod q} \beta(n) - \frac{1}{\varphi(q)} \sum_n \beta(n) \ll \frac{X}{(\log X)^A} \tag{24}$$

for any $q \geqslant 1$ and $(a, q) = 1$, and for any $A \geqslant 1$, where the implied constant depends on A. (Note that this bound is trivial if q is larger than a power of $\log X$.) This (standard) Siegel-Walfisz property arises in the present setting from the following general facts:

- By (a variant of) the Siegel-Walfisz form of the Prime Number Theorem (see, e.g., [17, Cor. 5.29]), the Möbius function satisfies the Siegel-Walfisz property;
- Smooth functions like $V(n)$ in Proposition 5.1 always satisfy even stronger forms of equidistribution;
- An elementary argument shows that, provided β_1 is supported on $[N, 2N]$ with $X^\varepsilon \leqslant N \leqslant X$ for some fixed $\varepsilon > 0$, and satisfies the Siegel-Walfisz property, then $\beta_1 \star \beta_2$ has the Siegel-Walfisz property (for arbitrary β_2 supported on $[M, 2M]$ with $NM \ll X$; see, e.g., [24, Lemma 5.4]).

As a preliminary step (and as a first illustration of the use of the Heath-Brown identity and bilinear techniques) one can prove now quite easily that if $q \leqslant Q \leqslant X^{1/2}/(\log X)^B$ for some integer $B \geqslant 1$ (depending on A for which we seek (20)), then $(\boldsymbol{M}, \boldsymbol{N})$ is feasible (with terminology as in Section 5) for the sum provided $N \gg X^\varepsilon$ for some fixed $\varepsilon > 0$ and β satisfies the Siegel-Walfisz property (24); essentially, this follows by writing

$$\begin{aligned}\Sigma(\boldsymbol{M}, \boldsymbol{N}) &= \sum_q c(q) \sum_{m,n} \alpha(m)\beta(n)\Big(\mathbf{1}_{mn\equiv a \pmod q} - \frac{1}{\varphi(q)}\mathbf{1}_{(mn,q)=1}\Big)\\ &= \sum_q \frac{c(q)}{\varphi(q)} \sideset{}{^*}\sum_{\chi \pmod q} \overline{\chi(a)}\Big(\sum_m \alpha(m)\chi(m)\Big)\Big(\sum_n \beta(n)\chi(n)\Big)\end{aligned}$$

where χ runs over non-trivial Dirichlet characters modulo q, and then reducing to primitive characters, and applying either the Siegel–Walfisz Property for β (for small moduli) or the *large sieve inequality* (to α and β), namely for any complex numbers $(\gamma(n))$, we have

$$\sum_{q\leqslant Q} \sideset{}{^*}\sum_{\chi \pmod q} \Big|\sum_{N_1 < n \leqslant N_1+N_2} \gamma(n)\chi(n)\Big|^2 \leqslant (N_2 + Q^2)\sum_n |\gamma(n)|^2,$$

where χ runs over primitive Dirichlet characters modulo q (for details, see [17, Th. 17.4]). Using this argument, it is an easy matter to deduce the Bombieri-Vinogradov Theorem: for every $A \geqslant 1$, there exists $B \geqslant 1$ such that

$$\sum_{q \leqslant X^{1/2}/(\log X)^B} \max_{(a,q)=1} \left| \boldsymbol{\psi}(X; q, a) - \frac{\boldsymbol{\psi}(X)}{\varphi(q)} \right| \ll \frac{X}{(\log X)^{1/2}} \tag{25}$$

where the implied constant depends only on A. (This approach to the Bombieri-Vinogradov Theorem goes back to ideas of Gallagher and later Motohashi).

Remark 6.1. — The large sieve inequality itself is an example of the bilinear principle, see [17, §7] for a detailed discussion and many variants.

This first step is already useful as it reduces the proof of Theorem 5.5 to the cases where $Q \geqslant X^{1/2}/(\log X)^B$ for some B depending on the A for which we want to prove (20). Since we wish to use Theorem 5.5 to prove Theorem 4.1 for some (arbitrarily small) $\varpi > 0$ and $\delta > 0$, this means that we are in a borderline situation, where only a small amount needs to be gained in order to succeed. This is a point to keep in mind in the transformations that follow.

The next step exploits the friability of the moduli, and is similar to using a well-factorable weight $c(q)$ (see Remark 5.4, and see also [5, Cor. 5]). Precisely, we may restrict the support of $c(q)$ to moduli q of the type

$$q = rs$$

where r and s are in dyadic segments $R < r \leqslant 2R$ and $S < s \leqslant 2S$, with $RS = Q$. (Indeed, since any q appearing in the bilinear form is X^δ-friable, for any choices of R and S we can factor

$$q = rs$$

where r satisfies

$$X^{-\delta} R \leqslant r \leqslant R,$$

and we then rearrange these factors (r, s) dyadically). Note that we have some choice since a single q has such factorization with r in any interval of this type (again, in the spirit of well-factorability). We exploit this by taking $R = N^{1-\nu}$ for some small $\nu > 0$, so that R is rather close to N.

Now another application of the Siegel-Walfisz property, used to "detect" the main term, allows one to replace the function $\boldsymbol{\delta}_{q,a}(n)$ in $K(n)$ by

$$\boldsymbol{\delta}(n) = \mathbf{1}_{\substack{n \equiv a \pmod r \\ n \equiv b_1 \pmod q}} - \mathbf{1}_{\substack{n \equiv a \pmod r \\ n \equiv b_2 \pmod q}}$$

for some idelettes b_1 and b_2. (This is somewhat cosmetic, but by separating the main term, it simplifies the later application of the dispersion method by making the sum more symmetric).

We then begin the actual work. This is similar to the basic bilinear principle, but we have many "hidden" parameters, and we can (and must) exploit them before or when applying the Cauchy-Schwarz inequality. Here we pull outside the sum over r and m, and write therefore

$$|\Sigma(\boldsymbol{M}, \boldsymbol{N})| \leqslant \sum_r \sum_m \alpha(m) \Big(\sum_s c(rs) \sum_{mn \equiv a \pmod{rs}} \beta(n) \boldsymbol{\delta}(mn) \Big)$$

before applying the Cauchy-Schwarz inequality:

$$|\Sigma(\boldsymbol{M},\boldsymbol{N})|^2 \leqslant R\|\alpha\|^2\Big(\sum_r\sum_m\Big|\sum_s c(rs)\sum_{mn\equiv a\ (\mathrm{mod}\ rs)}\beta(n)\boldsymbol{\delta}(mn)\Big|^2\Big).$$

We then square the last sum: it is a combination of four sums of the type

$$\Sigma'(\boldsymbol{M},\boldsymbol{N}) = \sum_r\sum_m\sum_{s_1,s_2} c(rs_1)c(rs_2)\sum_{n_1,n_2}\beta(n_1)\overline{\beta(n_2)}\mathbf{1}_{\substack{mn_1\equiv a\ (\mathrm{mod}\ r)\\ mn_1\equiv b_1\ (\mathrm{mod}\ s_1)}}\mathbf{1}_{\substack{mn_2\equiv a\ (\mathrm{mod}\ r)\\ mn_2\equiv b_1\ (\mathrm{mod}\ s_2)}}\cdot$$

Now observe a crucial consequence of the arrangement of the variables and of the multiplicative structure of the idelette (already present and exploited by Fouvry and Iwaniec [6, 7]): the inner sum over n_1 and n_2 is restricted by the condition

$$n_1 \equiv n_2 \ (\mathrm{mod}\ r).$$

Note the important fact that *if* the residue class a in Theorem 4.1 was allowed to depend on q, then this extra restriction would disappear, and this line of argument would collapse.

Since we think of $R = N^{1-\nu}$ as smaller but quite close to N, this congruence is a strong constraint. In fact, we write

$$n_2 = n_1 + \ell r, \qquad \ell \leqslant \frac{N}{R}$$

and sum over a fixed value of ℓ, i.e., we consider now

$$\Sigma_\ell(\boldsymbol{M},\boldsymbol{N}) = \sum_r\sum_m\sum_{s_1,s_2} c(rs_1)c(rs_2)\sum_n\gamma(r,n)\mathbf{1}_{\substack{mn\equiv a\ (\mathrm{mod}\ r)\\ mn\equiv b_1\ (\mathrm{mod}\ s_1)}}\mathbf{1}_{m(n+\ell r)\equiv b_1\ (\mathrm{mod}\ s_2)},$$

where $\gamma(r,n) = \beta(n)\overline{\beta(n+\ell r)}$, which we view again as (mostly) arbitrary coefficients with basic control of their average size.

We can glimpse here the most important gain, that also relates to the factorability of the moduli q and to the arrangement of the bilinear argument. The sums over m and n are of length M and N unchanged from the start, and of size close to $\sqrt{X}$, hence close to $Q = RS$. The congruence conditions on m and n now involve the modulus $[rs_1, rs_2] = rs_1s_2$ which is of size $RS^2 = Q^2/R$, at least if s_1 and s_2 are coprime. If R is quite large – it will be relatively close to N – this modulus is much smaller than the modulus $q_1q_2 = r^2s_1s_2$ that would arise if one argued in the same way but without the factorization (i.e., with $R = 1$). The coprimality of s_1 and s_2 in this intuitive explanation is of course not always true; the gcd of s_1 and s_2 must be brought into the picture, and a different treatment is needed when it is large. For simplicity, we simply deal with the coprime case (the reader may check, for instance, that at least the diagonal case $s_1 = s_2$ can be dealt with easily).

Since we began with sums over m and n of length N close to (but possibly smaller than) the square root of X, this means that the resulting exponential sums will be, for R not too small, of length larger than the square root of the modulus. But such exponential sums of "algebraic" nature can be estimated efficiently, and in great generality, exactly when this length condition is met, by exploiting the Riemann Hypothesis over finite fields. In this case, only the Riemann Hypothesis for curves is needed, since the sum involves a single variable, and the results of A. Weil are therefore enough. In the next section, we will see that the trilinear sums are much more delicate in this respect.

More precisely, the simplest *completion technique* can be expressed as follows:

LEMMA 6.2. — *Let $q \geqslant 1$ be an integer, let $\varphi : \mathbf{Z}/q\mathbf{Z} \longrightarrow \mathbf{C}$ be a function extended to $\mathbf{Z}$ by periodicity.*[7] *For $Y \geqslant 1$, we have*

$$\Big|\sum_{1\leqslant n\leqslant Y} \varphi(n)\Big| \leqslant \sqrt{q}(\log 3q)\Big(1+\frac{Y}{q}\Big)\max_{h\in\mathbf{Z}/q\mathbf{Z}}|\hat{\varphi}(h)|$$

where

$$\hat{\varphi}(h) = \frac{1}{\sqrt{q}}\sum_{x\in\mathbf{Z}/q\mathbf{Z}}\varphi(x)e\Big(-\frac{hx}{q}\Big)$$

is the discrete Fourier transform of φ.

Proof. — By periodicity the sum is bounded by $\lfloor Y/q\rfloor$ times the "complete" sum over $\mathbf{Z}/q\mathbf{Z}$, which is $\hat{\varphi}(0)$, plus the maximum of the sums of length Y with $1 \leqslant Y \leqslant q$. For the latter, denoting by $\mathbf{1}$ the characteristic function of this interval, the discrete Plancherel formula gives

$$\Big|\sum_{1\leqslant n\leqslant Y}\varphi(n)\Big| = \Big|\sum_{h\in\mathbf{Z}/q\mathbf{Z}}\hat{\varphi}(h)\hat{\mathbf{1}}(h)\Big| \leqslant \sum_{h\in\mathbf{Z}/q\mathbf{Z}}|\hat{\mathbf{1}}(h)|\times\max_{h\in\mathbf{Z}/q\mathbf{Z}}|\hat{\varphi}(h)|$$

and a computation with geometric series shows that the L^1-norm of the Fourier transform of $\mathbf{1}$ is $\leqslant \sqrt{q}(1+\log q)$ (independently of the length of the interval). □

Remark 6.3. — The intuition behind this method is that, for many functions φ, the Fourier transform should be uniformly bounded by a constant because each value $\hat{\varphi}(h)$ is (up to dividing by $\sqrt{q}$) an exponential sum of length q involving a (typically) oscillatory function that should (at least theoretically) exhibit square-root cancellation. One of Weil's greatest achievements was to show that this intuition holds, e.g., when $\varphi(n) = e(f(n)/p)$ for some fixed rational function $f \in \mathbf{Q}(X)$, unless it is obviously degenerate in some sense (essentially, unless it is constant modulo p).

We apply this first to the m variable, since the sum over m is free of interfering coefficients. Precisely, we separate the zero frequency (which gives main terms which cancel out when rebuilding $\Sigma'(\boldsymbol{M},\boldsymbol{N})$) and thus must understand sums of the type

$$\Sigma'' = \sum_r\sum_{s_1,s_2}c(rs_1)c(rs_2)\sum_n\gamma(r,n)e\Big(\frac{\xi(n)h}{r[s_1,s_2]}\Big)$$

for non-zero $h \in \mathbf{Z}/r[s_1,s_2]\mathbf{Z}$ and for the residue class $\xi(n)$ modulo $r[s_1,s_2]$ characterized by

$$\xi(n) \equiv \frac{b_1}{n} \pmod{s_1}, \quad \xi(n) \equiv \frac{b_2}{n+\ell r} \pmod{s_2}, \quad \xi(n) \equiv \frac{r}{n} \pmod{r} \tag{26}$$

(of course, this depends also on (b_1,b_2,r), but we will sum over n for each of these parameters separately).

Now we must deal with the sum over n. This has the unknown coefficients $\gamma(r,n)$ attached, so we apply once more the bilinear principle: moving the variables n and r outside and applying the Cauchy-Schwarz inequality, we get

$$|\Sigma''|^2 \leqslant \Sigma_1\Sigma_2$$

where

$$\Sigma_1 = \sum_r\sum_n|\gamma(r,n)|^2$$

[7] No confusion with the Euler function should arise.

is easily handled to show that $\Sigma_1 \ll RN(\log N)^C$ for some $C \geqslant 1$ (depending only on $A \geqslant 1$), while (after opening the square and exchanging the summation over n) we have

$$|\Sigma_2| \leqslant \sum_{s_1,s_2,s_3,s_4} \sum \Big| \sum_n e\Big(\frac{\xi(n)h}{r[s_1,s_2]} - \frac{\xi(n)h}{r[s_3,s_4]}\Big)\Big|.$$

The sum over n is an algebraic exponential sum of length N with modulus $r[s_1,s_2,s_3,s_4]$ which is about RS^4 when the s_i are coprime, which is again a simplifying assumption. Again because we selected R to be close to N, and S quite small, these parameters match closely and Lemma 6.2 can be applied. One has to deal with technical details involving cases where the phase of the sum degenerates, but in the end, after applying the Chinese Remainder Theorem, the most important estimate is the following:

PROPOSITION 6.4. — *Let p be a prime number, and*[8] *$\psi(x) = e(x/p)$ an additive character modulo p. Let (a,b,c,d) be elements of $\mathbf{F}_p$ with $d \neq 0$ and not all of (a,b,c) zero. We then have*

$$\Big| \sum_{x \neq -d,0} \psi\Big(\frac{a}{x+d} + \frac{b}{x} + cx\Big)\Big| \leqslant 3\sqrt{p}.$$

Remark 6.5. — The condition (26) explains the form of the exponential sum, the linear phase $\psi(cx)$ arising from the Fourier transform in Lemma 6.2.

Sketch of proof. — This is a basic consequence of the Weil bound for exponential sums in one variable. We will however explain how to derive this (and many similar bounds) from the formalism of étale sheaves and Deligne's version of the Riemann Hypothesis over finite fields [3] (compare with the proof of Theorem 7.3 below). Fixing a prime $\ell \neq p$, and an isomorphism $\iota : \bar{\mathbf{Q}}_\ell \longrightarrow \mathbf{C}$ (which we will use as an identification when considering ℓ-adic numbers as complex numbers), there exists a lisse ℓ-adic sheaf $\mathcal{L}$ of rank 1 on $U = \mathbf{A}^1 - \{0,-d\}$ such that (the image under ι of) the trace of the geometric Frobenius of $\mathbf{F}_p$ acting on the stalk over x is equal to

$$\psi\Big(\frac{a}{x+d} + \frac{b}{x} + cx\Big).$$

This sheaf is geometrically non-trivial under the assumption that (a,b,c) are not all 0 (because it is then ramified at $-d$, 0 or ∞).

Because the sheaf is lisse on U and geometrically non-trivial, we have

$$H_c^2(U \times \bar{\mathbf{F}}_p, \mathcal{L}) = H_c^0(U \times \bar{\mathbf{F}}_p, \mathcal{L}) = 0$$

(see,e.g., [3, (1.4.1)b]). The Grothendieck-Lefschetz trace formula (see, e.g., [18, 2.3.2]) gives

$$\sum_{x \in \mathbf{F}_p} \psi\Big(\frac{a}{x+d} + \frac{b}{x} + cx\Big) = -\operatorname{Tr}(\mathrm{Fr} \mid H_c^1(U \times \bar{\mathbf{F}}_p, \mathcal{L})),$$

and by Deligne's general form of the Riemann Hypothesis [3, Th. 3.3.1], which implies that all eigenvalues of the Frobenius acting on $H_c^1(U \times \bar{\mathbf{F}}_p, \mathcal{L})$ have modulus at most $p^{1/2}$, we get

$$\Big| \sum_{x \in \mathbf{F}_p} \psi\Big(\frac{a}{x+d} + \frac{b}{x} + cx\Big)\Big| \leqslant (\dim H_c^1(U \times \bar{\mathbf{F}}_p, \mathcal{L}))p^{1/2}.$$

Now we use the Euler-Poincaré characteristic formula (see, e.g, [18, 2.3.1]) which, since the H_c^0 and H_c^2 cohomology groups vanish, gives

$$\dim H_c^1(U \times \bar{\mathbf{F}}_p, \mathcal{L}) = -\chi_c(U \times \bar{\mathbf{F}}_p, \mathcal{L}) = \chi_c(U \times \bar{\mathbf{F}}_p) - \mathrm{Swan}_0(\mathcal{L}) - \mathrm{Swan}_{-d}(\mathcal{L}) - \mathrm{Swan}_\infty(\mathcal{L}).$$

[8] No confusion with the prime counting function will arise.

Since $U = \mathbf{P}^1 - \{0, -d, \infty\}$, its Euler-Poincaré characteristic is -1, and since we have

$$\mathrm{Swan}_x(\mathcal{L}) = 1$$

for $x \in \{0, -d, \infty\}$ if (respectively) b, a or c is non-zero, and otherwise the Swan conductor at x is 0 (this corresponds to the Swan conductor for an additive character of a function with a simple pole), we get

$$\dim H_c^1(U \times \bar{\mathbf{F}}_p, \mathcal{L}) = 3,$$

provided all of (a, b, c) are non-zero, and otherwise the dimension is $\leqslant 2$. This finishes the proof. □

We have left out quite a few technical details. The final outcome is an estimate that shows that $(\boldsymbol{M}, \boldsymbol{N})$ is feasible for $q \leqslant Q \leqslant X^{1/2+2\varpi}$ for some $\varpi > 0$, provided the length of the shortest of the two variables (say N, if $N \leqslant M$) is not too small, namely if $\mathsf{n} \geqslant 1/2 - 2\varpi - \delta'$ for some arbitrarily small fixed $\delta' > 0$. This does not cover the whole range in Theorem 5.5. Nevertheless (as found out by the Polymath8 project), it is possible to improve the argument by exploiting a finer version of the completion technique, based on more friability properties of the moduli, in such a way that this gap is filled (this is the so-called q-van der Corput method, see [24, Cor. 6.15]). This leaves only the trilinear estimates to be dealt with in order to finish the proof of Zhang's Theorem.

Without this improvement (and this is the situation in Zhang's own version), one has to find a different arrangement of the final steps where the sum over n was treated. There are a number of variants of this step; Zhang [27, §11] essentially arranges the sum (in skeleton form indicating only the order of the variables, and assuming $N \leqslant M$ as before) as

$$\sum_{s_1} \sum_{n} \Big| \sum_{s_2} \Big|$$

(for each fixed r) and applies Cauchy-Schwarz to the s_1, n sums, while [24] has three different approaches (one involves the same finer completion technique mentioned above, the second exploits further friability properties of s_1 to again create a new splitting $s_1 = u_1 v_1$ which can be exploited and the third and most efficient depends on fine properties of the *family* of exponential sums in Proposition 6.4, and therefore goes well beyond Weil's method, towards the Grothendieck-Deligne formalism of étale sheaves and the general form of the Riemann Hypothesis over finite fields). At least in principle, the ideas are relatively similar to those already discussed, and we won't say more.

7. THE TERNARY DIVISOR FUNCTION

The final step in Zhang's work is the treatment of the sums (21), namely

$$\sum_{n \equiv a \pmod q} (\beta \star \alpha_1 \star \alpha_2 \star \alpha_3)(n) K(n) = \sum_{q \leqslant Q} c(q) \sum_{m n_1 n_2 n_3 \equiv a \pmod q} \beta(m) \alpha_1(n_1) \alpha_2(n_2) \alpha_3(n_3)$$

with the restrictions (22) concerning the length of the variables. Using a general version of the Bombieri-Vinogradov Theorem, we can assume that $Q \geqslant X^{1/2}/(\log X)^B$ as in the bilinear case, and we want to extend the range of Q to $Q \leqslant X^{1/2+2\varpi}$ for some $\varpi > 0$, at least if $c(q)$ is supported on suitably friable moduli.

Reviewing the situation at the end of Section 5, one might feel uneasy, noting that one could well imagine being able to give non-trivial results for these sums, but with a slightly

narrower range than required for the gap to be covered (i.e., only for variables satisfying (22) with $\kappa > 0$ replaced by a smaller number $\kappa' > 0$).

However, the sums above, as they arise from Proposition 5.1, have the following special features:

– The variable m is relatively short, since $\mathsf{m} + \mathsf{n}_1 + \mathsf{n}_2 + \mathsf{n}_3 = 1 + o(1)$ and
$$\mathsf{n}_1 + \mathsf{n}_2 + \mathsf{n}_3 = \frac{1}{2}\Big(\mathsf{n}_1 + \mathsf{n}_2 + \mathsf{n}_1 + \mathsf{n}_3 + \mathsf{n}_2 + \mathsf{n}_3\Big) \geqslant \frac{3}{4} + \frac{3}{20} - \frac{3\kappa}{2} + o(1)$$
by (22), and we can assume that $\kappa > 0$ is fixed, but arbitrarily small (in view of Theorem 5.5);

– The functions $\alpha_i(n) = V_i(n)$ are smooth (each arises from a single variable n_j in Proposition 5.1, since n is too large to be compatible with the lengths $\leqslant 1/J$ of the m_i variables).

Intuitively, one can then feel that one has a chance, because this suggests that (21) might well become *easier* (under (22)) when κ gets larger, since the variable which is the combination of the n_i variables is then *longer*. Since this is the opposite of the situation in Theorem 5.5 (as noted after its statement), a healthy optimism is *de rigueur*. In fact, Zhang's work establishes the following type of result (see [24, Th. 9.1] for a general version):

THEOREM 7.1 (Zhang's trilinear estimate). — *For all $\kappa > 0$, there exist explicit $\delta > 0$ and $\varpi > 0$ such that, for all $A \geqslant 1$, we have*
$$\sum_{q\leqslant Q} c(q)\Big(\sum_{mn_1n_2n_3\equiv a \pmod q} \beta(m)V_1(n_1)V_2(n_2)V_3(n_3) - \frac{1}{\varphi(q)} \sum_{(mn_1n_2n_3,q)=1} \beta(m)V_1(n_1)V_2(n_2)V_3(n_3)\Big) \ll \frac{N}{(\log N)^A}$$
whenever
$$Q \leqslant N^{1/2+2\varpi},$$
for any idelette a, and for any arithmetic function β and smooth functions V_i as above, in particular satisfying the support conditions (22) and the growth conditions of divisor-type functions, and for any function $c(q)$ with $|c(q)| \leqslant 1$ supported on N^δ-friable moduli.

This result has the desired feature of being valid for all κ that allows us to bridge the gap left by Zhang's bilinear estimates, and therefore gives Theorem 4.1.

We explain some ideas of the proof. If we work with a fixed q close to $X^{1/2}$, and a fixed m, we have a sum of the type
$$T_m(a;q) = \sum_{mn\equiv a \pmod q} \alpha_1 \star \alpha_2 \star \alpha_3(n) = \sum_{mn_1n_2n_3\equiv a \pmod q} \alpha_1(n_1)\alpha_2(n_2)\alpha_3(n_3)$$
with α_i smooth and compactly supported on $[N_i/2, N_i]$, where the modulus is not that much larger than the square root of the length $N_1N_2N_3$. In other words, we can (in fact, must) use results or ideas concerning the distribution of the ternary divisor function in arithmetic progressions to large moduli. (It is because of this essential component of the sum that we speak of trilinear estimates).

This is again a very good sign, because of the deep work of Friedlander and Iwaniec [11], who had indeed obtained an exponent of distribution larger than $1/2$ for this arithmetic function. Their work immediately implies (in fact, is proved by establishing) that for all $\varepsilon > 0$, all moduli $1 \leqslant q \leqslant X^{1/2+1/230-\varepsilon}$ and all primitive residue classes a modulo q, and for

all $A \geqslant 1$, we have, for $\alpha_i = V_i(n/N_i)$ (with V_i satisfying the same smoothness conditions as the functions in Proposition 5.1), the bound

$$\sum_{n\equiv a \pmod q} \alpha_1 \star \alpha_2 \star \alpha_3(n) - \frac{1}{\varphi(q)} \sum_n \alpha_1 \star \alpha_2 \star \alpha_3(n) \ll \frac{1}{q} \frac{X}{(\log X)^A}$$

where the implied constant depends only on A and ε. This result, although it is very strong (note the uniformity with respect to a, as well as the fact that averaging over the modulus is not needed) is insufficient to finish the proof of (21) (and therefore that of Theorem 4.1, in the original version), but again the friability of the moduli can be used to extend the exponent of distribution to close the gap.

Remark 7.2. — The exponent $1/2 + 1/230$ of Friedlander and Iwaniec was improved to $1/2 + 1/82$ by Heath-Brown [15], and recently (and independently of considerations of gaps between primes) to $1/2 + 1/46$ by Fouvry, Michel and the author [8], the latter restricted to the (most difficult) case of prime moduli.

We sketch here quite informally an argument different from that of Zhang (which is closer to [11]); as in the case of the bilinear sums, there are different variants available.

We write $Q = RS$ and we may consider moduli $q = rs$ with r around R and s around S, with $R < S$, and the splitting may be chosen using the friability. We begin with m fixed.

Although some variables n_i might be shorter than the square root of the modulus, it is symmetrically advantageous to perform completion of the three variables at once. Since we have a smooth weight, we use the Poisson summation formula and obtain something like

$$T_m(a;q) \approx \frac{N}{q^3} \sum_{1\leqslant|h_i|\leqslant H_i} \sum\sum\sum \Big(\sum_{\substack{n_1,n_2,n_3 \pmod q \\ mn_1n_2n_3 = a \pmod q}} \sum\sum\sum e\Big(\frac{h_1n_1 + h_2n_2 + h_3n_3}{q}\Big)\Big) + (\text{remainder})$$

where $H_i \asymp Q/N_i$ and the remainder accounts for the contributions of the "high" as well as the "degenerate" frequencies (h_1, h_2, h_3) with $h_1h_2h_3 = 0$. The case $(h_1, h_2, h_3) = (0, 0, 0)$ will lead to the main term (and hence cancels out), while the others are easier to estimate.

For h_i (and m) coprime with q, borrowing one factor $1/q$, we have

$$\frac{1}{q} \sum_{\substack{n_1,n_2,n_3 \pmod q \\ mn_1n_2n_3 = a \pmod q}} e\Big(\frac{h_1n_1 + h_2n_2 + h_3n_3}{q}\Big) = \mathrm{Kl}_3(ah_1h_2h_3m^{-1}; q)$$

where $\mathrm{Kl}_3(x;q)$ is a hyper-Kloosterman sum in two variables modulo q:

$$\mathrm{Kl}_3(x;q) = \frac{1}{q} \sum_{\substack{\alpha.\beta,\gamma\in\mathbf{Z}/q\mathbf{Z} \\ \alpha\beta\gamma = x}} \sum\sum e\Big(\frac{\alpha+\beta+\gamma}{q}\Big).$$

Thus, up to the expected main term and negligible or simpler contributions, we get

$$T_m(a;q) \approx \frac{N}{q^2} \sum_{1\leqslant|h_i|\leqslant H_i} \sum\sum\sum \mathrm{Kl}_3(ah_1h_2h_3m^{-1}; q).$$

The total number of points in the sums is $H = H_1H_2H_3 \asymp Q^3/N$ which is quite large, but the summation is not free. We express the right-hand side as

$$\frac{N}{q^2} \sum_{1\leqslant|h_i|\leqslant H_i} \sum\sum\sum \mathrm{Kl}_3(ah_1h_2h_3m^{-1}; q) = \frac{N}{q^2} \sum_{1\leqslant|h|\leqslant H} \tilde{\tau}_3(h)\, \mathrm{Kl}_3(ahm^{-1}; q)$$

where $\tilde{\tau}_3(h) \leqslant \tau_3(h)$ is viewed again as an almost unknown sequence of coefficients, with average size under control.

We now use the splitting $q = rs$, where $(r, s) = 1$ since q is squarefree. We sum over r, for a fixed value of s (we could also sum over m at this point with the coefficients $\beta(m)$), and use the bilinear principle for the h variable in order to eliminate the unknown coefficients. The inner sums (analogue of $\Delta_\gamma(m_1, m_2)$ in the description of Section 5) are then of the type

$$\sum_{1\leqslant|h|\leqslant H} \mathrm{Kl}_3(ahm^{-1}; r_1 s)\overline{\mathrm{Kl}_3(ahm^{-1}; r_2 s)}.$$

Comparing the length of h and the modulus $[r_1, r_2]s$, we can see again that we have reached a situation where the completion technique (Lemma 6.2) can lead to a non-trivial bound for these sums, outside of a well-understood diagonal situation, provided we can control the size of the discrete Fourier transform of the function modulo $[r_2, r_2]s$ given by

$$\varphi(h) = \mathrm{Kl}_3(ahm^{-1}; r_1 s)\overline{\mathrm{Kl}_3(ahm^{-1}; r_2 s)}.$$

Thus, in the end, a crucial ingredient is the following exponential sum estimate, obtained as an application of the Riemann Hypothesis over finite fields.

THEOREM 7.3. — *There exists an absolute constant $C \geqslant 1$ such that for any prime p, any $(a, b, c) \in \mathbf{F}_p^\times \times \mathbf{F}_p^\times \times \mathbf{F}_p$ with $a \neq b$, we have*

$$\Big|\sum_{x\in\mathbf{F}_p} \mathrm{Kl}_3(ax; p)\overline{\mathrm{Kl}_3(bx; p)}e\Big(\frac{cx}{p}\Big)\Big| \leqslant Cp^{1/2}.$$

Remark 7.4. — Although Friedlander and Iwaniec [11] (and then Zhang) proceed along slightly different lines, they end up with an exponential sum which is, in fact, equivalent (see the proof). The first treatment of the sum as we stated it is due to Michel [22] (who considered more general cases).

Proof. — We give a full proof to emphasize that, with today's technology in hand and a modicum of knowledge of the relevant étale formalism, this is again an application of Deligne's work [3], and that it is (almost) as elementary as any application of Weil's bounds for one-variable character sums. Indeed, the proof follows closely in outline the one we gave for the one-variable sum in Proposition 6.4. We will even obtain an explicit constant $C = 8$ in this manner, but this value should not be considered important.

We write $\mathrm{Kl}_3(x; p) = \mathrm{Kl}_3(x)$ for simplicity. We begin by expanding the Kloosterman sums and exchanging the sums, obtaining easily by orthogonality of characters the formula

$$\sum_{x\in\mathbf{F}_p} \mathrm{Kl}_3(ax)\overline{\mathrm{Kl}_3(bx)}e\Big(\frac{cx}{p}\Big) = \sum_{t\neq 0,-c} \mathrm{Kl}_2\Big(\frac{a}{t}\Big)\,\mathrm{Kl}_2\Big(\frac{b}{t+c}\Big) = \sum_{y\neq 0,-a/c} \mathrm{Kl}_2(y)\,\mathrm{Kl}_2\Big(\frac{by}{cy+a}\Big) \tag{27}$$

where, for $a \in \mathbf{F}_p$, the sum

$$\mathrm{Kl}_2(a) = \frac{1}{\sqrt{p}} \sum_{\substack{x,y\in\mathbf{F}_p^\times \\ xy=a}} e\Big(\frac{x+y}{p}\Big) \in \mathbf{R}$$

is a standard Kloosterman sum. This last sum of Kloosterman sums is the one that arises in [11]; the estimate that follows is then proved by Birch and Bombieri in the Appendix to [11]. They give two different proofs, both relying of course on Deligne's work, which are also different than the one we give.

Now, as in Proposition 6.4, we fix a prime $\ell \neq p$, and fix an isomorphism $\iota : \bar{\mathbf{Q}}_\ell \longrightarrow \mathbf{C}$. There is an additive character $\psi : \mathbf{F}_p \longrightarrow \bar{\mathbf{Q}}_\ell^\times$ such that $\iota(\psi_\ell(x)) = e(x/p)$. Furthermore, by another important work of Deligne, there exists a lisse geometrically irreducible ℓ-adic sheaf $\mathcal{K}\ell_2$ of rank 2 on

the multiplicative group over $\mathbf{F}_p$ such that for $a \in \mathbf{F}_p^\times$, the trace of the geometric Frobenius of $\mathbf{F}_p$ acting on the stalk over a is equal to $-\operatorname{Kl}_2(a)$. The pullback of $\mathcal{K}\ell_2$ to any open dense subset of the multiplicative group is still geometrically irreducible, and it is known that this sheaf is self-dual, tamely ramified at 0 and totally wildly ramified at ∞ with only break $1/2$, hence Swan conductor 1. It is pointwise ι-pure of weight 0 (see [18] for this result, and much more concerning Kloosterman sheaves), the latter result being itself an application of Weil's theory for Kloosterman sums.

As a corollary, given (a, b, c) as before and the matrix $\gamma = \begin{pmatrix} b & 0 \\ c & a \end{pmatrix} \in \mathrm{PGL}_2(\mathbf{F}_p)$, which is not the identity, the sheaf

$$\mathcal{F} = \mathcal{K}\ell_2 \otimes \gamma^*\mathcal{K}\ell_2$$

is lisse and pointwise pure of weight 0 on $U = \mathbf{P}^1 - \{0, -c/a, \infty\}$, and has trace of Frobenius equal to

$$\operatorname{Kl}_2(x)\operatorname{Kl}_2(\gamma \cdot x)$$

for all $x \in U(\mathbf{F}_p)$. Hence the rightmost sum in (27) is simply the sum of the local traces of this sheaf over $U(\mathbf{F}_p)$.

The sheaf $\mathcal{F}$ is of rank 4. It is tamely ramified at 0 (because $\mathcal{K}\ell_2$ and $\gamma^*\mathcal{K}\ell_2$ are tame at 0) and totally wildly ramified at $-a/c$ and ∞, with Swan conductor 2 at each of these singularities (it has unique break $1/2$ with multiplicity 4 at these points).

Again as we did before, we get

$$H_c^0(U \times \bar{\mathbf{F}}_p, \mathcal{F}) = 0$$

because $\mathcal{F}$ is lisse on U and also

$$H_c^2(U \times \bar{\mathbf{F}}_p, \mathcal{F}) = 0$$

(by [3, (1.4.1)b], either because the two tensor factors are geometrically irreducible on U and have different singularities, if $c \neq 0$, or by a computation of Katz [18, Prop. 10.4.1] if $c = 0$, where one uses the fact that $a \neq b$).

From the Grothendieck-Lefschetz trace formula (see, e.g., [18, 2.3.2]) we get

$$\sum_{t \neq 0, -c} \operatorname{Kl}_2\Big(\frac{a}{t}\Big)\operatorname{Kl}_2\Big(\frac{b}{t+c}\Big) = -\operatorname{Tr}(\mathrm{Fr} \mid H_c^1(U \times \bar{\mathbf{F}}_p, \mathcal{F})),$$

and then from Deligne's Riemann Hypothesis [3, Th. 3.3.1], since $\mathcal{F}$ is pointwise of weight 0, we get

$$\Big|\sum_{t \neq 0, -c} \operatorname{Kl}_2\Big(\frac{a}{t}\Big)\operatorname{Kl}_2\Big(\frac{b}{t+c}\Big)\Big| \leqslant (\dim H_c^1(U \times \bar{\mathbf{F}}_p, \mathcal{F}))p^{1/2}.$$

Finally, if $c \neq 0$, the Euler-Poincaré characteristic formula gives

$$\dim H_c^1(U \times \bar{\mathbf{F}}_p, \mathcal{F}) = -\chi_c(U \times \bar{\mathbf{F}}_p, \mathcal{F}) = \\ \operatorname{rank}(\mathcal{F})\chi_c(U \times \bar{\mathbf{F}}_p) - \operatorname{Swan}_0(\mathcal{F}) - \operatorname{Swan}_{-\alpha}(\mathcal{F}) - \operatorname{Swan}_\infty(\mathcal{F}) = 8$$

using $\chi_c(U \times \bar{\mathbf{F}}_p) = -1$ and

$$\operatorname{Swan}_0(\mathcal{F}) = 0, \qquad \operatorname{Swan}_{-\alpha}(\mathcal{F}) = \operatorname{Swan}_\infty(\mathcal{F}) = 2.$$

For $c = 0$, one gets similarly

$$\dim H_c^1(U \times \bar{\mathbf{F}}_p, \mathcal{F}) \leqslant 8.$$

□

8. MAYNARD'S THEOREM

We now come back to Maynard's theorem. Here, we estimate the error term (13) in the quadratic form Q_1, defined using the general choice (7) by means of the Bombieri-Vinogradov estimate: we take $x = X^{1/4-\varepsilon}$ for some small $\varepsilon > 0$, and we then get from (25), without further ado, the bound

$$R_1 \ll \|\lambda\|_\infty^2 \frac{X}{(\log X)^A}$$

where the implied constant depends on k and A.

On the other hand, the main terms can be evaluated asymptotically for a suitable choice of $\lambda_{\boldsymbol{d}}$. As described in [13, 19], a diagonalization procedure of Q_2 can be used to simplify and motivate the choice, but we omit this step. This leads Maynard however to define

$$\lambda_{\boldsymbol{d}} = \mu(\boldsymbol{d})|\boldsymbol{d}| \sum_{\substack{\boldsymbol{d}|\boldsymbol{r} \\ (\boldsymbol{r},W)=1}} \frac{\mu(|\boldsymbol{r}|)^2}{\varphi(\boldsymbol{r})} F\Big(\frac{\log d_1}{\log x}, \dots, \frac{\log d_k}{\log x}\Big)$$

where

$$\mu(\boldsymbol{d}) = \prod_i \mu(d_i), \qquad \varphi(\boldsymbol{d}) = \prod_i \varphi(d_i),$$

$\boldsymbol{r}$ runs over k-tuples of integers where all r_i are squarefree, pairwise coprime and coprime to W, and $d_i \mid r_i$ for all i, and where F is a piecewise smooth real-valued function defined on $[0,1]^k$ and supported on the domain

$$D_k = \{(t_1, \dots, t_k) \in [0, +\infty[^k \mid \ 0 \leqslant t_1 + \cdots + t_k \leqslant 1\}.$$

Remark 8.1. — A choice of the type

$$F(\boldsymbol{t}) = f(t_1 + \cdots + t_k)$$

for some smooth function f recovers the type of functions considered in the Goldston-Pintz-Yıldırım method.

Maynard [21, Prop. 4.1, Lemma 6.2, Lemma 6.3] proves the following asymptotic formulas, stated using the notation in (11) (in particular, W is the auxiliary product of primes $\leqslant D(X) = \log\log\log X$ used in the preliminary sieve):

Proposition 8.2 (Maynard). — *Let $\lambda_{\boldsymbol{d}}$ be defined as above for $x = X^{1/4-\varepsilon}$ for some $\varepsilon > 0$, and let F be a real-valued piecewise smooth function. We then have*

$$Q_2 = \tilde{X}\Big(\frac{\varphi(W)}{W}\Big)^k (\log x)^{k+1} M_2(F)(1 + o(1)),$$

$$Q_1 = \Big(\frac{1}{4} - \varepsilon\Big)\tilde{X}\Big(\frac{\varphi(W)}{W}\Big)^k (\log x)^{k+1} M_1(F)(1 + o(1)),$$

as $X \to +\infty$, where

$$M_2(F) = \|F\|^2 = \int_{D_k} F(\boldsymbol{t})^2 d\boldsymbol{t},$$

$$M_1(F) = \sum_{i=1}^k \int_{[0,1]^{k-1}} F_i(\boldsymbol{u})^2 d\boldsymbol{u},$$

with F_i defined on $[0,1]^{k-1}$ by

$$F_i(u_1,\dots,u_{k-1}) = \int_{[0,1]} F(u_1,\dots,u_{i-1},t,u_{i+1},\dots,u_{k-1})dt.$$

The proof of this result is not very difficult, and is to some extent just an extension of the known case of the Goldston-Pintz-Yıldırım method, which requires some care (bookkeeping and notational, for instance), but does not involve new ideas by itself (once one has thought of trying to prove it!). We refer to the very precise and clear exposition by Maynard [21, §5,§6].

The question therefore becomes that of maximizing the ratio

$$\frac{M_1(F)}{M_2(F)}$$

as F runs over admissible functions. The objective is to show that the supremum, say ρ_k, is > 4 for k large enough (to prove the bounded gap property using the Bombieri-Vinogradov theorem), and in fact that ρ_k is unbounded as k grows (to prove the sharper statement (1)). Thus, the following finishes the proof of Theorem 1.2:

Proposition 8.3. — *With notation as above, we have*

$$\lim_{k\to+\infty} \rho_k = +\infty.$$

Sketch of proof. — We use the probabilistic interpretation of Maynard's computation by Tao [26]. We consider functions defined by

$$F(\boldsymbol{t}) = \prod_{i=1}^{k} G(t_i)$$

for $\boldsymbol{t} \in D_k$, and $F(\boldsymbol{t}) = 0$ otherwise, where $G \geqslant 0$ is a function with compact support in $[0,\alpha]$ for some $\alpha \leqslant 1$, with L^2-norm equal to 1. It follows then that

$$M_2(F) \leqslant \|G\|_2^k = 1,$$

and we look for a lower-bound on $M_1(F)$. We note first that, G being symmetric, we have

$$M_1(F) = k\int_{[0,1]^{k-1}} F_k(\boldsymbol{u})^2 d\boldsymbol{u}.$$

We then note that, because of the support of G, we certainly have

$$F_k(\boldsymbol{u}) = \int_0^1 G(\boldsymbol{u},t)dt = \prod_{j=1}^{k-1} G(u_j)\int G(t)dt$$

for all $(\boldsymbol{u},t) \in D_k$ such that

$$u_1 + \cdots + u_{k-1} \leqslant 1-\alpha.$$

Denoting $\mu = \int G(t)dt$, we get therefore

$$M_1(F) \geqslant k\mu^2 \int_{u_1+\cdots+u_{k-1}\leqslant 1-\alpha} \prod_{j=1}^{k-1} G(u_j)^2 d\boldsymbol{u}.$$

Since G has L^2-norm 1, the measure

$$\prod_{j=1}^{k-1} G(u_j)^2 d\boldsymbol{u}$$

is the probability distribution of a vector $(Y_1, \dots, Y_{k-1})$ of independent random variables, identically distributed with law $G(u)^2 du$. Thus what we have proved is

$$M_1(F) \geqslant k\mu^2 \mathbf{P}(Y_1 + \cdots + Y_{k-1} \leqslant 1 - \alpha).$$

If this probability law satisfies

$$(k-1)\mathbf{E}(Y_1) \leqslant k\mathbf{E}(Y_1) < 1 - \alpha,$$

then by the law of large numbers, the condition $Y_1 + \cdots + Y_{k-1} \leqslant 1 - \alpha$ will be true with probability almost 1. Precisely, the Chebychev inequality (and a simple bound on the variance of Y_i) gives

$$\mathbf{P}(Y_1 + \cdots + Y_{k-1} \leqslant 1 - \alpha) \geqslant 1 - \frac{\alpha k \mathbf{E}(Y_1)}{(1 - \alpha - k\mathbf{E}(Y_1))^2} \geqslant 1 - \frac{\alpha}{(1 - \alpha - k\mathbf{E}(Y_1))^2}.$$

The optimization of this bound (in terms of G) is done by Maynard. We just quote his choice: take $\alpha = 1/(\log k)^3$ (for $k \geqslant 2...$) and G supported on $[0, \alpha]$ by

$$G(t) = \frac{\sqrt{k}\beta}{1 + k(\log k)t},$$

where β is the normalization factor that implies that the L^2-norm is 1. One finds that $\beta \sim (\log k)^{1/2}$, and that

$$\mathbf{E}(Y_1) = \int tG(t)^2 dt = \frac{1}{k}\Big(1 - 2\frac{\log\log k}{\log k} + O\Big(\frac{1}{\log k}\Big)\Big)$$

while

$$\mu \gg \sqrt{\frac{\log k}{k}},$$

proving finally that, for this particular choice of function, we have

$$\rho_k \geqslant M_1(F) \gg \log k.$$

□

Remark 8.4. — (1) P. Sarnak[(9)] has observed that there is a certain similarity between this question and the type of estimates related to the proof of "stability of the second kind" in quantum mechanics, which also involve variational optimization problems for functions of increasing number of variables (see [20]). Whether this observation can be used to understand the behavior of ρ_k is an interesting question.

(2) The quantitative lower bound leads to

$$\liminf_{n \to +\infty} (p_{n+m} - p_n) \ll m^3 e^{4m}$$

for $m \geqslant 1$ (see [21, Prop. 4.3 (3)]; this has been improved by the Polymath8b project).

(3) Because of the underlying simplicity of the argument and its dependency only on arbitrarily small positive exponents of distribution, it is certain that the ideas of Maynard will

(9) During a lecture by K. Soundararajan, December 2013.

have considerable influence on the study of the distribution of many arithmetic functions besides the primes.

REFERENCES

[1] E. BOMBIERI, J. FRIEDLANDER and H. IWANIEC – *Primes in arithmetic progressions to large moduli*, Acta Math. 156 (1986), 203–251.

[2] N. BOURBAKI – *Fonctions d'une variable réelle*, Paris, Hermann, 1976, réimpression Springer, 2007.

[3] P. DELIGNE – *La conjecture de Weil, II*, Publ. Math. IHÉS 52 (1980), 137–252.

[4] É. FOUVRY – *Autour du théorème de Bombieri-Vinogradov*, Acta Math. 152 (1984), 219–244.

[5] É. FOUVRY – *Autour du théorème de Bombieri-Vinogradov, II*, Ann. Sci. ENS 20 (1987), 617–640.

[6] É. FOUVRY and H. IWANIEC – *On a theorem of Bombieri-Vinogradov type*, Mathematika 27 (1980), 135–152.

[7] É. FOUVRY and H. IWANIEC – *Primes in arithmetic progressions*, Acta Arithmetica 42 (1983), 197–218.

[8] É. FOUVRY, E. KOWALSKI and Ph. MICHEL – *On the exponent of distribution of the ternary divisor function*, Mathematika, to appear.

[9] J. FRIEDLANDER and H. IWANIEC – *Opera de Cribro*, Colloquium Publ. 57, A.M.S, 2010.

[10] J. FRIEDLANDER and H. IWANIEC – *Close encounters among the primes*, prépublication, `arXiv:1312.2926v2`

[11] J. FRIEDLANDER and H. IWANIEC – *Incomplete Kloosterman sums and a divisor problem* (with an appendix by B. J. BIRCH and E. BOMBIERI), Annals of Math. 121 (1985), 319–350.

[12] D.A. GOLDSTON, J. PINTZ and C.Y. YILDIRIM – *Primes in tuples, I*, Annals of Math. 170 (2009), 819–862.

[13] D.A. GOLDSTON, S.W. GRAHAM, J. PINTZ and C.Y. YILDIRIM – *Small gaps between primes or almost primes*, Trans. Amer. Math. Soc. 361 (2009), 5285–5330.

[14] A. GRANVILLE – *Primes in intervals of bounded length*, prépublication (2013), `http://www.dms.umontreal.ca/~andrew/CEBBrochureFinal.pdf`.

[15] D.R. HEATH–BROWN – *The divisor function $d_3(n)$ in arithmetic progressions*, Acta Arithmetica 47 (1986), 29–56.

[16] K-H HO and K-M TSANG – *On almost prime k-tuples*, J. Number Theory 120 (2006), 33–46.

[17] H. IWANIEC and E. KOWALSKI – *Analytic Number Theory*, Colloquium Publ. 53, A.M.S, 2004.

[18] N.M. KATZ – *Gauss Sums, Kloosterman Sums and Monodromy Groups*, Annals of Math. Studies 116, Princeton Univ. Press (1988).

[19] E. KOWALSKI – *Écarts entre nombres premiers successifs (d'après Goldston, Pintz, Yıldırım)*, Séminaire Bourbaki (2005/06), Exp. 959, Astérisque 311 (2007), 177–210.

[20] E. LIEB and R. SEIRINGER – *The stability of matter in quantum mechanics*, Cambridge Univ. Press, 2010.

[21] J. MAYNARD – *Small gaps between primes*, prépublication, `arXiv:1311.4600`.

[22] Ph. MICHEL – *Minorations de sommes exponentielles*, Duke Math. J. 95 (1998), 227–240.

[23] Y. MOTOHASHI and J. PINTZ – *A smoothed GPY sieve*, Bull. Lond. Math. Soc. 40 (2008), 298–310.

[24] D.H.J POLYMATH – *New equidistribution theorems of Zhang type, and bounded gaps between primes*, prépublication, `arXiv:1402.0811`.

[25] A. SELBERG – *Collected works, II*, Springer 1991.

[26] T. TAO – *Bounded intervals with many primes, after Maynard*, blog post, availabel at `terrytao.wordpress.com/2013/11/19/polymath8b-bounded-intervals-with-many-primes-after-maynard/`

[27] Y. ZHANG – *Bounded gaps between primes*, Annals of Math. 179 (2014), 1121–1174.

Emmanuel KOWALSKI

ETH Zürich – DMATH
Rämistrasse 101
8092 Zürich, Switzerland
E-mail : `kowalski@math.ethz.ch`