Acta Math., 201 (2008), 213–305
DOI: 10.1007/s11511-008-0032-5

The primes contain arbitrarily long polynomial progressions

by

Terence Tao

University of California, Los Angeles
Los Angeles, CA, U.S.A.

Tamar Ziegler

Technion – Israel Institute of Technology
Haifa, Israel

Contents

1. Introduction

In 1975, Szemerédi [30] proved that any subset A of integers of positive upper density, that is $\limsup_{N\to\infty} |A\cap[N]|/|[N]|>0$, contains arbitrarily long arithmetic progressions. Throughout this paper $[N]$ denotes the discrete interval $[N]:=\{1,...,N\}$, and $|X|$ denotes the cardinality of a finite set X. Shortly afterwards, Furstenberg [10] gave an

The second author was partially supported by NSF grant DMS-0111298. This work was initiated at a workshop held at the CRM in Montreal. The authors would like to thank the CRM for their hospitality.

ergodic-theory proof of Szemerédi's theorem. Furstenberg observed that questions about configurations in subsets of positive density in the integers correspond to recurrence questions for sets of positive measure in a probability measure preserving system. This observation is now known as the *Furstenberg correspondence principle*.

In 1978, Sárközy [28](1) (using the Hardy–Littlewood circle method) and Furstenberg [11] (using the correspondence principle, and ergodic theoretic methods) proved independently that for any polynomial(2) $P \in \mathbf{Z}[\mathbf{m}]$ with $P(0)=0$, any set $A \subset \mathbf{Z}$ of positive density contains a pair of points x, y with difference $y-x=P(m)$ for some positive integer $m \geqslant 1$. In 1996 Bergelson and Leibman [6] proved, by purely ergodic theoretic means,(3) a vast generalization of the Fustenberg–Sárközy theorem, establishing the existence of arbitrarily long polynomial progressions in sets of positive density in the integers.

THEOREM 1.1. (Polynomial Szemerédi theorem [6]) *Let $A \subset \mathbf{Z}$ be a set of positive upper density, i.e. $\limsup_{N\to\infty} |A \cap [N]|/|[N]| > 0$. Then, given any integer-valued polynomials $P_1, ..., P_k \in \mathbf{Z}[\mathbf{m}]$ in one unknown $\mathbf{m}$ with $P_1(0)=...=P_k(0)=0$, the set A contains infinitely many progressions of the form $x+P_1(m), ..., x+P_k(m)$ with $m>0$.*

Remark 1.2. By shifting x appropriately, one may assume without loss of generality that one of the polynomials P_j vanishes, say $P_1=0$. We shall rely on this ability to normalize one polynomial of our choosing to be zero, at several points in the proof, most notably in the "PET induction" step in §5.10. The arguments in [6] also establish a generalization of this theorem to higher dimensions, which will be important to us to obtain a certain uniformly quantitative version of this theorem later (see Theorem 3.2 and Appendix B).

The ergodic theoretic methods, to this day, have the limitation of only being able to handle sets of positive density in the integers, although this density is allowed to be arbitrarily small. However in 2004, Green and Tao [18] discovered a *transference principle* which allowed one (at least in principle) to reduce questions about configurations in special sets of zero density (such as the primes $\mathcal{P} := \{2, 3, 5, 7, ...\}$) to questions about sets of positive density in the integers. This opened the door to transferring the Szemerédi-type results, which are known for sets of positive upper density in the integers, to the

(1) Sárközy actually proved a stronger theorem for the polynomial $P=\mathbf{m}^2$ providing an upper bound for the density of a set A for which $A-A$ does not contain a perfect square. His estimate was later improved by Pintz, Steiger and Szemerédi in [24], and then generalized in [2] for $P=\mathbf{m}^k$ and then in [29] for arbitrary P with $P(0)=0$.

(2) We use $\mathbf{Z}[\mathbf{m}]$ to denote the space of polynomials of one variable $\mathbf{m}$ with integer-valued coefficients; see §2 for further notation along these lines.

(3) Unlike Szemerédi's theorem or Sárközy's theorem, no non-ergodic proof of the Bergelson–Leibman theorem in its full generality is currently known. However, in this direction Green [16] proved, by Fourier-analytic methods, that any set of integers of positive density contains a triple $\{x, x+n, x+2n\}$, where n is a non-zero sum of two squares.

prime numbers. Applying this transference principle to Szemerédi's theorem, Green and Tao showed that there are arbitrarily long arithmetic progressions in the prime numbers.([4])

In this paper we prove a transference principle for polynomial configurations, which then allows us to use (a uniformly quantitative version of) the Bergelson–Leibman theorem to prove the existence of arbitrarily long *polynomial* progressions in the primes, or more generally in large subsets of the primes. More precisely, the main result of this paper is the following.

THEOREM 1.3. (Polynomial Szemerédi theorem for the primes) *Let $A \subset \mathcal{P}$ be a set of primes of positive relative upper density in the primes, i.e.*

$$\limsup_{N\to\infty} \frac{|A\cap[N]|}{|\mathcal{P}\cap[N]|} > 0.$$

Then, given any $P_1, ..., P_k \in \mathbf{Z}[\mathbf{m}]$ with $P_1(0)=...=P_k(0)=0$, the set A contains infinitely many progressions of the form $x+P_1(m), ..., x+P_k(m)$ with $m>0$.

Remarks 1.4. The main result of [18] corresponds to Theorem 1.3 in the linear case $P_j := (j-1)\mathbf{m}$. The case $k=2$ of this theorem follows from the results of [24], [2] and [29], which in fact address arbitrary sets of integers with logarithmic-type sparsity, and whose proof is more direct, proceeding via the Hardy–Littlewood circle method and not via the transference principle. As a by-product of our proof, we shall also be able to impose the bound $m \leqslant x^{\varepsilon}$ for any fixed $\varepsilon>0$, and thus (by diagonalization) that $m=x^{o(1)}$; see Remark 2.4. Our results for the case $A=\mathcal{P}$ are consistent with what is predicted by the Bateman–Horn conjecture [3], which remains totally open in general (though see [19] for some partial progress in the linear case).

Remark 1.5. In view of the generalization of Theorem 1.1 to higher dimensions in [6], it is reasonable to conjecture that an analogous result to Theorem 1.3 also holds in higher dimensions, and thus any subset of $\mathcal{P}^d$ of positive relative upper density should contain infinitely many polynomial constellations, for any choice of polynomials which vanish at the origin. This is however still open even in the linear case, the key difficulty being that the tensor product of pseudorandom measures is not pseudorandom. In view of [31] however, it should be possible (though time-consuming) to obtain a counterpart to Theorem 1.3 for the Gaussian primes.

([4]) Shortly afterwards, the transference principle was also combined in [31] with the multidimensional Szemerédi theorem [12] (or more precisely a hypergraph lemma related to this theorem, see [34]) to establish arbitrarily shaped constellations in the *Gaussian* primes. A much simpler transference principle is also available for dense subsets of *genuinely random* sparse sets; see [35].

Remark 1.6. The arguments in this paper are mostly quantitative and finitary, and in particular avoid the use of the axiom of choice. However, our proof relies crucially on the Bergelson–Leibman theorem (Theorem 1.1) and more precisely on a certain multidimensional generalization of that theorem [6, Theorem A_0]. At present, the only known proof of that theorem (in [6]) requires Zorn's lemma and thus our results here are also currently dependent on the axiom of choice. However, it is expected that the Bergelson–Leibman theorem will eventually be proven by other means which do not require the axiom of choice; for instance, the 1-dimensional version of this theorem (i.e. Theorem 1.1) can be established via the machinery of characteristic factors and Gowers–Host–Kra seminorms, by modifying the arguments in [9] and [21], and this does not require the axiom of choice; this already allows us to establish Theorem 1.3 without the axiom of choice in the homogeneous case when $P_j(\mathbf{m})=c_j\mathbf{m}^d$ for all $j=1,...,k$ and some constants $c_1,...,c_k,d$ (since in this case the W factor in Theorem 3.2 can be easily eliminated without introducing additional dimensions). In a similar spirit, our arguments do not currently provide any effective bound for the first appearance of a pattern $x+P_1(m),...,x+P_k(m)$ in the set A, but one expects that the Bergelson–Leibman theorem will eventually be proven with an effective bound (e.g. by extending the arguments in [15]), in which case Theorem 1.3 will automatically come with an effective bound also.

The philosophy of the proof is similar to the one in [18]. The first key idea is to think of the primes (or any large subset thereof) as a set of positive relative density in the set of *almost primes*, which (after some application of sieve theory, as in the work of Goldston and Yıldırım [14]) can be shown to exhibit a somewhat pseudorandom behavior. Actually, for technical reasons, it is more convenient to work not with the sets of primes and almost primes, but rather with certain normalized weight functions $0\leqslant f\leqslant\nu$ which are([5]) supported (or concentrated) on the primes and almost primes, respectively, with ν obeying certain *pseudorandom measure*([6]) properties. The functions f and ν are unbounded, but have bounded expectation (mean). A major step in the argument is a *Koopman–von Neumann-type structure theorem* which decomposes f (outside of a small exceptional set) as a sum $f=f_{U^\perp}+f_U$, where $f_{U^\perp}$ is a non-negative *bounded* function with large expectation, and f_U is an error which is unbounded but is so *uniform* (in a Gowers-type sense) that it has a negligible impact on the (weighted) count of polynomial progressions. The remaining component $f_{U^\perp}$ of f, being bounded, non-negative, and of large mean, can then be handled by (a quantitative version of) the Bergelson–Leibman theorem.

([5]) This is an oversimplification, ignoring the "W-trick" necessary to eliminate local obstructions to uniformity; see §2 for full details.

([6]) The term measure is a bit misleading. It is better to think of ν as the Radon–Nikodym derivative of a measure. Still, we stick to this terminology so as not to confuse the reader who is familiar with [18].

Remark 1.7. As remarked in [18], the above transference arguments can be categorized as a kind of "finitary" ergodic theory. In the language of traditional (infinitary) ergodic theory, $f_{U^\perp}$ is analogous to a conditional expectation of f relative to a suitable *characteristic factor* for the polynomial average being considered. Based on this analogy, and on the description of this characteristic factor in terms of nilsystems (see [21] and [23]), one would hope that $f_{U^\perp}$ could be constructed out of nilsequences. In the case of linear averages, this correspondence has already some roots in reality; see [17]. In the special case $A=\mathcal{P}$, one can then hope to use analytic number theory methods to show that $f_{U^\perp}$ is essentially constant, which would lead to a more precise version of Theorem 1.3 in which one obtains a precise *asymptotic* for the number of polynomial progressions in the primes, with x and n confined to various ranges. In the case of progressions of length 4 (or for more general linear patterns, assuming certain unproven conjectures), such an asymptotic was already established in [19]. While we expect similar asymptotics to hold for polynomial progressions, we do not pursue this interesting question here.([7])

As we have already mentioned, the proof of Theorem 1.3 closely follows the arguments in [18]. However, some significant new difficulties arise when adapting those arguments([8]) to the polynomial setting. The most fundamental such difficulty arises in one of the very first steps of the argument in [18], in which one localizes the pattern $x+P_1(m),...,x+P_k(m)$ to a finite interval $[N]=\{1,...,N\}$. In the linear case $P_j=(j-1)\mathbf{m}$ this localization restricts both x and m to be of size $O(N)$. However, in the polynomial case, while the base point x is still restricted to size $O(N)$, the shift parameter m is now restricted to a much smaller range $O(M)$, where $M:=N^{\eta_0}$ and $0<\eta_0<1$ is a small constant depending on $P_1,...,P_k$ (one can take for instance $\eta_0:=1/2d$, where d is the largest degree of the polynomials $P_1,...,P_k$; by taking a little more care, one can increase this to $\eta_0=1/d$). This eventually forces us to deal with localized averages of the form([9])

$$\mathbf{E}_{m\in[M]}\int_X T^{P_1(m)}f\,...\,T^{P_k(m)}f, \tag{1}$$

([7]) One fundamental new difficulty that arises in the polynomial case is that it seems that one needs to control *short* correlation sums between primes and nilsequences, such as on intervals of the form $[x,x+x^\varepsilon]$, instead of the long correlation sums (such as on $[x,2x]$) which appear in the linear theory. Even assuming strong conjectures such as the generalized Riemann hypotesis, it is not clear how to obtain such control.

([8]) If the measure ν for the almost primes enjoyed infinitely many pseudorandomness conditions, then one could adapt the arguments in [35] to obtain Theorem 1.3 rather quickly. Unfortunately, in order for f to have non-zero mean, one needs to select a moderately large sieve level $R=N^{\eta_2}$ for the measure ν, which means that one can only impose finitely many (though arbitarily large) such pseudorandomness conditions on ν. This necessitates the use of the (lengthier) arguments in [18] rather than [35].

([9]) This is an oversimplification, as we are ignoring the need to first invoke the "W-trick" to eliminate local obstructions from small moduli, and thus ensure that the almost primes behave pseudorandomly. See Theorem 2.3 for the precise claim we need.

where $X:=\mathbf{Z}_N:=\mathbf{Z}/N\mathbf{Z}$ is the cyclic group with N elements, $f:X\to\mathbf{R}^+$ is a weight function associated with the set A, and $Tg(x):=g(x-1)$ is the shift operator on X. Here we use the ergodic theory-like([10]) notation

$$\mathbf{E}_{n\in Y}F(n):=\frac{1}{|Y|}\sum_{n\in Y}F(n)$$

for any finite non-empty set Y, and

$$\int_X f:=\mathbf{E}_{x\in X}f(x)=\frac{1}{N}\sum_{x\in X}f(x). \tag{2}$$

We shall normalize f to have mean $\int_X f=\eta_3$, and will also have the pointwise bound $0\leqslant f\leqslant\nu$ for some "pseudorandom measure" ν associated with the almost primes at a sieve level $R:=N^{\eta_2}$ for some([11]) $0<\eta_3\ll\eta_2\ll\eta_0$ (so M is asymptotically larger than any fixed power of R). The functions f and ν will be defined formally in (11) and (72), respectively, but for now let us simply remark that we will have the bound $\int_X\nu=1+o(1)$, together with many higher-order correlation estimates on ν.

Let us defer the (sieve-theoretic) discussion of the pseudorandomness of ν for the moment, and focus instead on the (finitary) ergodic theory components of the argument. If we were in the linear regime $M=N$ used in [18] (with N assumed prime for simplicity), then repeated applications of the Cauchy–Schwarz inequality (using the PET induction method) would eventually let us control the average (1) in terms of *Gowers uniformity norms* such as

$$\|f\|_{U^d(\mathbf{Z}_N)}:=\left(\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[N]^d}\int_X\prod_{\omega\in\{0,1\}^d}T^{m_1^{(\omega_1)}+\ldots+m_d^{(\omega_d)}}f\right)^{1/2^d}$$

for some sufficiently large d (depending on $P_1,\ldots,P_k$; eventually they will be of size $O(1/\eta_1)$ for some $\eta_2\ll\eta_1\ll\eta_0$), where $\omega=(\omega_1,\ldots,\omega_d)$ and $\vec{m}^{(j)}=(m_1^{(j)},\ldots,m_d^{(j)})$, $j=0,1$. If instead we were in the pseudo-infinitary regime $M=M(N)$ for some slowly growing function $M:\mathbf{Z}^+\to\mathbf{Z}^+$, repeated applications of the van der Corput lemma and PET induction would allow one to control these averages by the *Gowers–Host–Kra seminorms*

([10]) Traditional ergodic theory would deal with the case where the underlying measure space $\mathbf{Z}_N$ is infinite and the shift range M is going to infinity, and thus informally $N=\infty$ and $M\to\infty$. Unraveling the Furstenberg correspondence principle, this is equivalent to the setting where N is finite (but going to infinity) and $M=\omega(N)$ is a very slowly growing function of N. In [18] one is instead working in the regime where $M=N$ are going to infinity at the *same* rate. The situation here is thus an intermediate regime where $M=N^{\eta_0}$ goes to infinity at a polynomially slower rate than N. In the linear setting, all of these regimes can be equated using the random dilation trick of Varnavides [38], but this trick is only available in the polynomial setting if one moves to higher dimensions, see Appendix B.

([11]) The "missing" values of η, such as η_1, will be described more fully in §2.

$\|f\|_d$ from [21], which in our finitary setting would be something like

$$\|f\|_d := \left(\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[M_1]\times\ldots\times[M_d]}\int_X \prod_{\omega\in\{0,1\}^d} T^{m_1^{(\omega_1)}+\ldots+m_d^{(\omega_d)}} f\right)^{1/2^d},$$

where $M_1,...,M_d$ are slowly growing functions of N which we shall deliberately keep unspecified.([12]) In our intermediate setting $M=N^{\eta_0}$, however, neither of these two quantities seem to be exactly appropriate. Instead, after applying the van der Corput lemma and PET induction one ends up considering an averaged localized Gowers norm of the form([13])

$$\|f\|_{U^{\vec{Q}([H]^t)}_{\sqrt{M}}} := \left(\mathbf{E}_{\vec{h}\in[H]^t}\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^d}\int_X \prod_{\omega\in\{0,1\}^d} T^{Q_1(\vec{h})m_1^{(\omega_1)}+\ldots+Q_d(\vec{h})m_d^{(\omega_d)}} f\right)^{1/2^d},$$

where $H=N^{\eta_7}$ is a small power of N (much smaller than M or R), t is a natural number depending only on $P_1,...,P_d$ (and of size $O(1/\eta_1)$), and $Q_1,...,Q_d\in\mathbf{Z}[\mathbf{h}_1,...,\mathbf{h}_t]$ are certain polynomials (of t variables $\mathbf{h}_1,...,\mathbf{h}_t$) which depend on $P_1,...,P_d$. Indeed, we will eventually be able (see Theorem 4.5) to establish a polynomial analogue of the *generalized von Neumann theorem* in [18], which roughly speaking will assert that (if ν is sufficiently pseudorandom) any component of f which is "locally Gowers-uniform" in the sense that the above norm is small, and which is bounded pointwise by $O(\nu+1)$, will have a negligible impact on the average (1). To exploit this fact, we shall essentially repeat the arguments in [18] (with some notational changes to deal with the presence of the polynomials Q_j and the short shift ranges) to establish (assuming ν is sufficiently pseudorandom) an analogue of the *Koopman–von Neumann-type structure theorem* in that paper, namely a decomposition $f=f_{U^\perp}+f_U$ (outside of a small exceptional set), where $f_{U^\perp}$ is bounded by $O(1)$, is non-negative and has mean roughly δ, and f_U is locally Gowers-uniform and thus has a negligible impact on (1). Combining this with a suitable quantitative version (Theorem 3.2) of the Bergelson–Leibman theorem, one can then conclude Theorem 1.3.

We have not yet discussed how one constructs the measure ν and establishes the required pseudorandomness properties. We shall construct ν as a truncated divisor sum at level $R=N^{\eta_2}$, although instead of using the Goldston–Yıldırım divisor sum as in [14] and [18], we shall use a smoother truncation (as in [36], [20] and [19]), as it is slightly easier to estimate.([14]) The pseudorandomness conditions then reduce, after standard

([12]) In the traditional ergodic setting $N=\infty$, $M\to\infty$, one would take multiple limit superiors as $M_1,...,M_d\to\infty$, choosing the order in which these parameters go to infinity carefully; see [21].

([13]) Again, this is a slight oversimplification as we are ignoring the effects of the "W-trick".

([14]) However, in contrast to the arguments in [20] and [19], we will not be able to completely localize the estimations on the Riemann-zeta function $\zeta(s)$ to a neighborhood of the pole $s=1$, for rather minor technical reasons, and so will continue to need the classical estimates (108) on $\zeta(s)$ near the line $\mathrm{Re}s=1$.

sieve theory manipulations, to the entirely local problem of understanding the pseudorandomness of the functions $\Lambda_p: F_p \to \mathbf{R}^+$ on finite fields F_p, defined for all primes p by $\Lambda_p(x) := p/(p-1)$ when $x \neq 0 \bmod p$ and $\Lambda_p(x) = 0$ otherwise. Our pseudorandomness conditions shall involve polynomials, and so one is soon faced with the standard arithmetic problem of counting the number of points over F_p of an algebraic variety. Fortunately, the polynomials that we shall encounter will be *linear* in one or more of the variables of interest, which allows us to obtain a satisfactory count of these points without requiring deeper tools from arithmetic such as class field theory or the Weil conjectures.

1.8. Acknowledgements

The authors thank Brian Conrad for valuable discussions concerning algebraic varieties, Peter Sarnak for encouragement, Vitaly Bergelson and Ben Green for help with the references, and Elon Lindenstrauss, Akshay Venkatesh and Lior Silberman for helpful conversations. We also thank the anonymous referee for useful suggestions and corrections.

2. Notation and initial preparation

In this section we shall fix some important notation, conventions, and assumptions which will then be used throughout the proof of Theorem 1.3. Indeed, all of the sub-theorems and lemmas used to prove Theorem 1.3 will be understood to use the conventions and assumptions in this section. We thus recommend that the reader go through this section carefully before moving on to the other sections of the paper.

Throughout this paper we fix the set $A \subset \mathcal{P}$ and the polynomials $P_1, ..., P_k \in \mathbf{Z}[\mathbf{m}]$ appearing in Theorem 1.3. Henceforth we shall assume that the polynomials are all distinct, since duplicate polynomials clearly have no impact on the conclusion of Theorem 1.3. Since we are also assuming that $P_j(0) = 0$ for all j, we conclude that

$$P_j - P_{j'} \text{ is non-constant for all } 1 \leqslant j < j' \leqslant k. \tag{3}$$

By hypothesis, the upper density

$$\delta_0 := \limsup_{N' \to \infty} \frac{|A \cap [N']|}{|\mathcal{P} \cap [N']|}$$

is strictly positive. We shall allow all implicit constants to depend on the quantities $\delta_0, P_1, ..., P_k$.

By the prime number theorem

$$|\mathcal{P}\cap[N']| = (1+o(1))\frac{N'}{\log N'}, \tag{4}$$

we can find an infinite sequence of integers N' going to infinity such that

$$|A\cap[N']| > \frac{1}{2}\delta_0\frac{N'}{\log N'}. \tag{5}$$

Henceforth the parameter N' is always understood to obey (5).

We shall need eight (!) small quantities

$$0 < \eta_7 \ll \eta_6 \ll \eta_5 \ll \eta_4 \ll \eta_3 \ll \eta_2 \ll \eta_1 \ll \eta_0 \ll 1$$

which depend on δ_0 and on $P_1, ..., P_k$. All of the assertions in this paper shall be made under the implicit assumption that η_0 is sufficiently small depending on $\delta_0, P_1, ..., P_k$; that η_1 is sufficiently small depending on $\delta_0, P_1, ..., P_k, \eta_0$; and so forth down to η_7, which is assumed sufficiently small depending on $\delta_0, P_1, ..., P_k, \eta_0, ..., \eta_6$ and should thus be viewed as being extremely microscopic in size. For the convenience of the reader, we briefly and informally describe the purpose of each of the η_j's, their approximate size, and the importance of being that size, as follows.

• The parameter η_0 controls the coarse-scale $M := N^{\eta_0}$. It can be set equal to $1/2d$, where d is the largest degree of the polynomials $P_1, ..., P_k$. If one desires the quantity m in Theorem 1.3 to be smaller than x^ε, then one can achieve this by choosing η_0 to be less than ε. The smallness of η_0 is necessary in order to deduce Theorem 1.3 from Theorem 2.3 below.

• The parameter η_1 (or more precisely its reciprocal $1/\eta_1$) controls the degree of pseudorandomness needed on a certain measure ν to appear later. Due to the highly recursive nature of the "PET induction" step (§5.10), it will need to be rather small; it is essentially the reciprocal of an Ackermann function of the maximum degree d and the number of polynomials k. The smallness of η_1 is needed in order to estimate all the correlations of ν which arise in the proofs of Theorems 4.5 and 4.7.

• The parameter η_2 controls the sieve level $R := N^{\eta_2}$. It can be taken to be $c\eta_1/d$ for some small absolute constant $c>0$. It needs to be small relative to η_1 in order that the inradius bound of Proposition 10.1 is satisfied.

• The parameter η_3 measures the density of the function f. It is basically of the form $c\delta_0\eta_2$ for some small absolute constant $c>0$. It needs to be small relative to η_2 in order to establish the mean bound (12).

• The parameter η_4 measures the degree of accuracy required in the Koopman–von Neumann-type structure theorem (Theorem 4.7). It needs to be substantially smaller

than η_3 to make the proof of Theorem 3.16 in §4.9 work. The exact dependence on η_3 involves the quantitative bounds arising from the Bergelson–Leibman theorem (see Theorem 3.2). In particular, as the only known proof of this theorem is infinitary, no explicit bounds for η_4 in terms of η_3 are currently available.

• The parameter η_5 controls the permissible error allowed when approximating indicator functions by a smoother object, such as a polynomial; it needs to be small relative to η_4 in order to make the proof of the abstract structure theorem (Theorem 7.1) work correctly. It can probably be taken to be roughly of the form $\exp(-C/\eta_4^C)$ for some absolute constant $C>0$, though we do not attempt to make η_5 explicit here.

• The parameter η_6 (or more precisely $1/\eta_6$) controls the maximum degree, dimension, and number of the polynomials that are encountered in the argument. It needs to be small relative to η_5 in order for the polynomials arising in the proof of Proposition 7.4 to obey the orthogonality hypothesis (49) of Theorem 7.1. It can in principle be computed in terms of η_5 by using a sufficiently quantitative version of the Weierstrass approximation theorem, though we do not do so here.

• The parameter η_7 controls the fine scale $H:=N^{\eta_7}$, which arises during the "van der Corput" stage of the proof in §5.10. It needs to be small relative to η_6 in order that the "clearing denominators" step in the proof of Proposition 6.5 works correctly. It is probably safe to take η_7 to be η_6^{100} although we shall not explicitly do this. On the other hand, η_7 cannot vanish entirely, due to the need to average out the influence of "bad primes" in Corollary 11.2 and Theorem 12.1.

It is crucial to the argument that the parameters are ordered in exactly the above way. In particular, the fine scale $H=N^{\eta_7}$ needs to be much smaller than the coarse-scale $M=N^{\eta_0}$.

We use the following asymptotic notation:

• We use $X=O(Y)$, $X\ll Y$ or $Y\gg X$ to denote the estimate $|X|\leqslant CY$ for some quantity $0<C<\infty$ which can depend on $\delta_0, P_1, ..., P_k$. If we need C to also depend on additional parameters, we denote this by subscripts, e.g. $X=O_K(Y)$ means that $|X|\leqslant C_K Y$ for some C_K depending on $\delta_0, P_1, ..., P_k, K$.

• We use $X=o(Y)$ to denote the estimate $|X|\leqslant c(N')Y$, where c is a quantity depending on $\delta_0, P_1, ..., P_k, \eta_0, ..., \eta_7, N'$ which goes to zero as $N'\to\infty$ for each fixed choice of $\delta_0, P_1, ..., P_k, \eta_0, ..., \eta_7$. If we need $c(N')$ to depend on additional parameters, we denote this by subscripts, e.g. $X=o_K(Y)$ denotes the estimate $|X|\leqslant c_K(N')Y$, where $c_K(N')$ is a quantity which goes to zero as $N'\to\infty$ for each fixed choice of $\delta_0, P_1, ..., P_k, \eta_0, ..., \eta_7, K$.

We shall implicitly assume throughout that N' is sufficiently large depending on $\delta_0, P_1, ..., P_k, \eta_0, ..., \eta_7$; in particular, all quantities of the form $o(1)$ will be small.

Next, we perform the "W-trick" from [18] to eliminate obstructions to uniformity

arising from small moduli. We shall need a slowly growing function $w=w(N')$ of N'. For sake of concreteness([15]) we shall set

$$w := \tfrac{1}{10} \log\log\log N'.$$

We then define the quantity W by

$$W := \prod_{p<w} p$$

and the integer([16]) N by

$$N := \left\lfloor \frac{N'}{2W} \right\rfloor. \tag{6}$$

Here and in the sequel, all products over p are understood to range over primes, and $\lfloor x \rfloor$ is the greatest integer less than or equal to x. The quantity W will be used to eliminate the local obstructions to pseudorandomness arising from small prime moduli; one can think of W (or more precisely the cyclic group $\mathbf{Z}_W$) as the finitary counterpart of the "profinite factor" generated by the periodic functions in infinitary ergodic theory. From the prime number theorem (4), one sees that

$$W \ll \log\log N \tag{7}$$

and

$$N' = N^{1+o(1)}. \tag{8}$$

In particular, the asymptotic limit $N' \to \infty$ is equivalent to the asymptotic limit $N \to \infty$ for the purposes of the $o(\cdot)$ notation, and so we shall now treat N as the underlying asymptotic parameter instead of N'.

From (5), (6) and (7) we have

$$\big|A \cap \big[\tfrac{1}{2} WN\big] \setminus [w]\big| \gg W \frac{N}{\log N}$$

([15]) Actually, the arguments here work for *any* choice of function $w: \mathbf{Z}^+ \to \mathbf{Z}^+$ which is bounded by $\frac{1}{10} \log\log\log N'$ and which goes to infinity as $N' \to \infty$. This is important if one wants an explicit lower bound on the number of polynomial progressions in a certain range.

([16]) Unlike previous work such as [18], we will not need to assume that N is prime (which is the finitary equivalent of the underlying space X being totally ergodic), although it would not be hard to ensure that this were the case if desired. This is ultimately because we shall clear denominators as soon as they threaten to occur, and so there will be no need to perform division in $X=\mathbf{Z}_N$. On the other hand, this clearing of denominators will mean that many (fine) multiplicative factors such as $Q(\vec{h})$ shall attach themselves to the (coarse-scale) shifts one is averaging over. In any case, the "W-trick" of passing from the integers $\mathbf{Z}$ to a residue class $W \cdot \mathbf{Z} + b$ can already be viewed as a kind of reduction to the totally ergodic setting, as it eliminates the effects of small periods.

(recall that implicit constants can depend on δ_0). On the other hand, since A consists entirely of primes, all the elements in $A\backslash[w]$ are coprime to W. By the pigeonhole principle,([17]) we may thus find $b=b(N)\in[W]$ coprime to W such that

$$\big|\big\{x\in\big[\tfrac{1}{2}N\big]:Wx+b\in A\big\}\big|\gg\frac{W}{\phi(W)}\frac{N}{\log N},\tag{9}$$

where $\phi(W)=\prod_{p<w}(p-1)$ is the Euler totient function of W, i.e. the number of elements of $[W]$ which are coprime to W.

Let us fix this b. We introduce the underlying measure space $X:=\mathbf{Z}_N=\mathbf{Z}/N\mathbf{Z}$, with the uniform probability measure given by (2). We also introduce the *coarse-scale* $M:=N^{\eta_0}$, the *sieve level* $R:=N^{\eta_2}$, and the *fine scale* $H:=N^{\eta_7}$. It will be important to observe the following size hierarchy:

$$1\ll W\ll W^{1/\eta_6}\ll H\ll H^{1/\eta_6}\ll R\ll R^{1/\eta_1}\ll M\ll N=|X|.\tag{10}$$

Indeed, each quantity on this hierarchy is larger than any bounded power of the preceding quantity, for suitable choices of the η parameters, for instance $R^{O(1/\eta_1)}\leqslant M^{1/4}$.

Remark 2.1. In the linear case [18] we have $M=N$, while the parameter H is not present (or can be thought of as $O(1)$). We shall informally refer to parameters of size([18]) $O(M)$ as *coarse-scale parameters*, and parameters of size H as *fine-scale parameters*; we shall use the symbol m to denote coarse-scale parameters and h for fine-scale parameters (reserving x for elements of X). Note that because the sieve level R is intermediate between these parameters, we will be able to easily average the pseudorandom measure ν over coarse-scale parameters, but not over fine parameters. Fortunately, our averages will always involve at least one coarse-scale parameter, and after performing the coarse-scale averages first, we will have enough control on main terms and error terms to then perform the fine averages. The need to keep the fine parameters short arises because at one key "Weierstrass approximation" stage to the argument, we shall need to control the product of an extremely large number (about $O(1/\eta_6)$, in fact) of averages (or more precisely "dual functions"), and this will cause many fine parameters to be multiplied together in order to clear denominators. This is still tolerable because H remains smaller than R, M and N even after being raised to a power $O(1/\eta_6)$. Note that it is key here that the number of powers $O(1/\eta_6)$ does not depend on η_7. It will therefore be important

([17]) In the case $A=\mathcal{P}$, we may use the prime number theorem in arithmetic progressions (or the Siegel–Walfisz theorem) to choose b, for instance to set $b=1$. However, we will not need to exploit this ability to fix b here.

([18]) Later on we shall also encounter some parameters of size $O(\sqrt{M})$ or $O(M^{1/4})$, which we shall also consider to be coarse-scale.

to keep large parts of our argument uniform in the choice η_7, although we can and will allow η_7 to influence $o(1)$ error terms. The quantity H (and thus η_7) will not actually make an impact on the argument until §4, when the local Gowers norms are introduced.

We define the standard shift operator $T: X \to X$ on X by $Tx := x+1$, with the associated action on functions $g: X \to \mathbf{R}$ by $Tg := g \circ T^{-1}$, and thus $T^n g(x) = g(x-n)$ for any $n \in \mathbf{Z}$. We introduce the normalized counting function $f: X \to \mathbf{R}^+$ by setting

$$f(x) := \begin{cases} \dfrac{\phi(W)}{W} \log R, & \text{whenever } x \in \left[\tfrac{1}{2}N\right] \text{ and } Wx+b \in A, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

where we identify $\left[\frac{1}{2}N\right]$ with a subset of $\mathbf{Z}_N$ in the usual manner. The use of $\log R$ instead of $\log N$ as a normalizing factor is necessary in order to bound f pointwise by the pseudorandom measure ν which we shall encounter in later sections; the ratio η_2 between $\log R$ and $\log N$ represents the relative density between the primes and the almost primes. Observe from (9) that f has relatively large mean:

$$\int_X f \gg \eta_2.$$

In particular we have

$$\int_X f \geqslant \eta_3. \tag{12}$$

Remark 2.2. We will eventually need to take η_2 (and hence η_3) to be quite small, in order to ensure that the measure ν obeys all the required pseudorandomness properties (this is controlled by the parameter η_1, which has not yet made a formal appearance). Fortunately, the Bergelson–Leibman theorem (Theorem 1.1, or more precisely Theorem 3.2 below) works for sets of arbitrarily small positive density, or equivalently for (bounded) functions of arbitrarily small positive mean.([19]) This allows us to rely on fairly crude constructions for ν which will be easier to estimate. This is in contrast to the recent work of Goldston, Yıldırım and Pintz [13] on prime gaps, in which it was vitally important that the density of the prime counting function relative to the almost prime counting function be as high as possible, which in turn required a near-optimal (and thus highly delicate) construction of the almost prime counting function.

([19]) As in [18], the exact quantitative bound provided by this theorem (or more precisely Theorem 3.2) will not be relevant for qualitative results such as Theorem 1.3. Of course, such bounds would be important if one wanted to know how soon the first polynomial progression in the primes (or a dense subset thereof) occurs; for instance such bounds influence how small η_4 and thus all subsequent η's need to be, which in turn influences the exact size of the final $o(1)$ error in Theorem 2.3. Unfortunately, as the only known proof of Theorem 1.1 proceeds via infinitary ergodic theory, no explicit bounds are currently known, however it is reasonable to expect (in view of results such as [15] and [33]) that effective bounds will eventually become available.

To prove Theorem 1.3, it will suffice to prove the following quantitative estimate.

THEOREM 2.3. (Polynomial Szemerédi theorem in the primes, quantitative version) *Let the notation and assumptions be as above. Then we have*

$$\mathbf{E}_{m\in[M]}\int_X T^{P_1(Wm)/W}f\,...\,T^{P_k(Wm)/W}f \geqslant \tfrac{1}{2}c\big(\tfrac{1}{2}\eta_3\big)-o(1), \tag{13}$$

where the function $c(\cdot)$ is the one appearing in Theorem 3.2. (*Observe that since $P_j(0)=0$ for all $j\in[k]$, the polynomial $P_k(W\mathbf{m})/W$ has integer coefficients.*)

Indeed, suppose that the estimate (13) held. Then by expanding all the averages and using (11), we conclude that

$$\big|\big\{(x,m)\in X\times[M]: x+P_j(Wm)/W\in\big[\tfrac{1}{2}N\big] \text{ and } W(x+P_j(Wm)/W)+b\in A\big\}\big| \geqslant \big(\tfrac{1}{2}c\big(\tfrac{1}{2}\eta_3\big)-o(1)\big)MN\bigg(\frac{W}{\phi(W)\log N}\bigg)^k.$$

Here we are using the fact (from (10)) that $P_j(Wm)$ is much less than $\frac{1}{2}N$ for $m\in[M]$, and so one cannot "wrap around" the cyclic group $\mathbf{Z}_N$. Observe that each element in the set on the left-hand side yields a different pair $(x',m'):=(Wx+b,Wm)$ with the property that $x'+P_1(m'),...,x'+P_k(m')\in A$. On the other hand, as $N\to\infty$, the right-hand side goes to infinity. The claim follows.

Remark 2.4. The above argument in fact proves slightly more than is stated by Theorem 1.3. Indeed, it establishes a large number of pairs (x',m') with $x'\in[N]$, $m'\in[M]$ and $x'+P_1(m'),...,x'+P_k(m')\in A$; more precisely, there are at least[20] $cNM/\log^k N$ such pairs for some c depending on $\delta_0,P_1,...,P_k,\eta_0,...,\eta_7$.[21] By throwing away the contribution of those x' of size $\ll N$ (which can be done either by modifying f in the obvious manner, or by using a standard upper bound sieve to estimate this component), one can in fact assume that x' is comparable to N. Similarly, one may assume m' to be comparable to N^{η_0}. The upshot of this is that for any given $\eta_0>0$ one in fact obtains infinitely many "short" polynomial progressions $x'+P_1(m'),...,x'+P_k(m')$ with

[20] To obtain such a bound it is important to remember that we can take w, and hence W, to be as slowly growing as one pleases; see [18] for further discussion. Note that if $A=\mathcal{P}$ is the full set of primes then the Bateman–Horn conjecture [3] predicts an asymptotic of the form $(\gamma+o(1))NM/\log^k N$ for an explicitly computable γ; we do not come close to verifying this conjecture here.

[21] The arguments in this paper can be easily generalized to give a lower bound of $cNM_1\,...\,M_r/\log^k N$ on the number of tuples $(x',m'_1,...,m'_r)$ with

$$x'+P_1(m'_1,...,m'_r),...,x'+P_k(m'_1,...,m'_r)\in A,$$

$x'\in[N]$, $m'_l\in[M_l]$ for $l\in[r]$, and $P_j\in\mathbf{Z}[\mathbf{m}_1,...,\mathbf{m}_r]$ for $j\in[k]$. To obtain this, one would only need to slightly modify the arguments in §5 (see Remark 5.19), whereas the rest of the proof remains the same.

m' comparable to $(x')^{\eta_0}$. One can take smaller and smaller values of η_0 and diagonalize to obtain the same statement with the bound $m'=(x')^{o(1)}$. This stronger version of Theorem 1.3 is already new in the linear case $P_j=(j-1)\mathbf{m}$, although it is not too hard to modify the arguments in [18] to establish it. Note that an inspection of the Furstenberg correspondence principle reveals that the Bergelson–Liebman theorem (Theorem 1.1) has an even stronger statement in this direction, namely that if A has positive upper density in the *integers* $\mathbf{Z}$ rather than the primes $\mathcal{P}$, then there exists a *fixed* $m\neq 0$ for which the set $\{x: x+P_j(m)\in A \text{ for all } j\in[k]\}$ is infinite (in fact, it can be chosen to have positive upper density). Such a statement might possibly be true for primes (or dense subsets of the primes) but is well beyond the technology of this paper. For instance, to establish such a statement even in the simple case $(P_1,P_2)=(0,\mathbf{m})$ is tantamount to asserting that the primes have bounded gaps arbitrarily often, which is still not known unconditionally even after the recent breakthroughs in [13]. On the other hand, it may be possible to establish such a result with a logarithmic dependence between x' and m', e.g. $m'\ll \log^{O(1)} x'$. We do not pursue this issue here.

It remains to prove Theorem 2.3. This shall occupy the remainder of the paper. The proof is lengthy, but splits into many non-interacting parts; see Figure 1 for a diagram of the logical dependencies of this paper.

2.5. Miscellaneous notation

To conclude this section, we record some additional notation which will be used heavily throughout this paper.

We have already used the notation $\mathbf{Z}[\mathbf{m}]$ to denote the ring of integer-coefficient polynomials in one indeterminate([22]) $\mathbf{m}$. More generally one can consider $\mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_d]$, the ring of integer-coefficient polynomials in d indeterminates $\mathbf{x}_1,...,\mathbf{x}_d$. More generally still we have $\mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_d]^D$, the space of D-tuples of polynomials in $\mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_d]$; note that each element of this space defines a polynomial map from $\mathbf{Z}^d$ to $\mathbf{Z}^D$. Thus we shall think of elements of $\mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_d]^D$ as D-dimensionally-valued integer-coefficient polynomials over d variables. The *degree* of a monomial $\mathbf{x}_1^{n_1}...\mathbf{x}_d^{n_d}$ is $n_1+...+n_d$; the degree of a polynomial in $\mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_d]^D$ is the highest degree of any monomial which appears in any component of the polynomial; we adopt the convention that the zero polynomial has degree $-\infty$. We say that two D-dimensionally-valued polynomials $\vec{P},\vec{Q}\in\mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_d]^D$ are *parallel* if $n\vec{P}=m\vec{Q}$ for some non-zero integers n and m.

([22]) We shall use boldface letters to denote abstract indeterminates, reserving the non-boldface letters for concrete realizations of these indeterminates, which in this paper will always be in the ring of integers $\mathbf{Z}$.

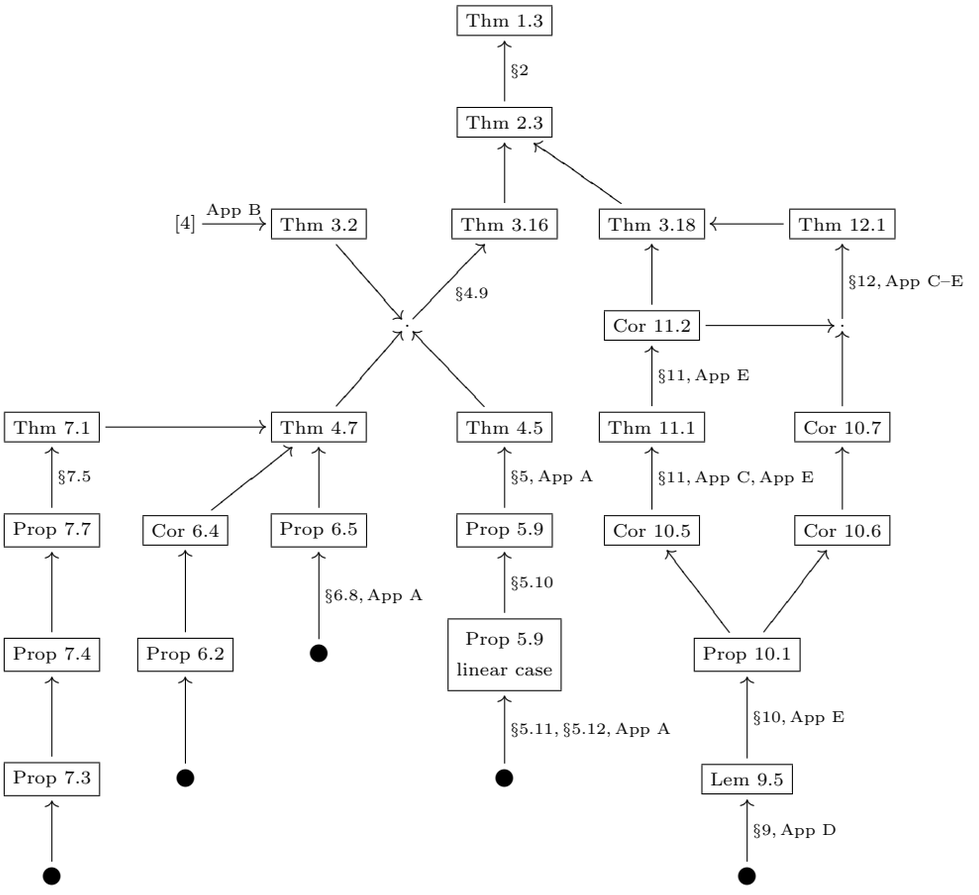

Figure 1. The main theorems in this paper and their logical dependencies. The labels on the arrows indicate the section(s) where the implication is proven, and which appendices are used; if no section is indicated, the result is proven immediately after it is stated. Self-contained arguments are indicated using a filled-in circle.

If $\vec{n}=(n_1,...,n_D)$ and $\vec{m}=(m_1,...,m_D)$ are two vectors in $\mathbf{Z}^d$, we use

$$\vec{n}\cdot\vec{m} := n_1m_1+...+n_Dm_D \in \mathbf{Z}$$

to denote their dot product.

If $f: X\to\mathbf{R}$ and $g: X\to\mathbf{R}$ are two functions, we say that f is *pointwise bounded* by g, and write $f\leqslant g$, if we have $f(x)\leqslant g(x)$ for all $x\in X$. Similarly, if $g: X\to\mathbf{R}^+$ is non-negative, we write $f=O(g)$ if we have $f(x)=O(g(x))$ uniformly for all $x\in X$. If $A\subset X$, we use $1_A: X\to\{0,1\}$ to denote the indicator (characteristic) function of A; thus $1_A(x)=1$ when $x\in A$ and $1_A(x)=0$ when $x\notin A$. Given any statement P, we use 1_P to denote 1 when P is true and 0 when P is false. Thus, for instance, $1_A(x)=1_{x\in A}$.

We define a *convex body* to be an open bounded convex subset of a Euclidean space $\mathbf{R}^d$. We define the *inradius* of a convex body to be the radius of the largest ball that is contained inside the body; this will be a convenient measure of how "large" a body is.([23])

3. Three pillars of the proof

As in [18], our proof of Theorem 2.3 rests on three independent pillars—a quantitative Szemerédi-type theorem (proven by traditional ergodic theory), a transference principle (proven by finitary ergodic theory), and the construction of a pseudorandom majorant ν for f (with the pseudorandomness proven by sieve theory). In this section we describe each of these pillars separately, and state where they are proven.

3.1. The quantitative Szemerédi-type theorem

Theorem 2.3 concerns a multiple polynomial average of an unbounded function f. To control such an object, we first need to establish an estimate for *bounded* functions g. This is achieved as follows (cf. [18, Proposition 2.3]).

THEOREM 3.2. (Polynomial Szemerédi theorem, quantitative version) *Let the notation and assumptions be as in the previous section. Let $\delta>0$, and let $g: X\to\mathbf{R}$ be any function obeying the pointwise bound $0\leqslant g\leqslant 1+o(1)$ and the mean bound $\int_X g\geqslant\delta-o(1)$. Then we have*

$$\mathbf{E}_{m\in[M]}\int_X T^{P_1(Wm)/W}g\,...\,T^{P_k(Wm)/W}g \geqslant c(\delta)-o(1) \tag{14}$$

for some $c(\delta)>0$ depending on $\delta, P_1, ..., P_k$, but independent of N and W.

([23]) In our paper there will only be essentially two types of convex bodies: "coarse-scale" convex bodies with inradius at least $M^{1/4}$, and "fine-scale" convex bodies, with inradius at least $\gg H$. In almost all cases, the convex bodies will in fact simply be rectangular boxes.

It is not hard to see that this theorem implies Theorem 1.1. The converse is not immediately obvious (the key point being, of course, that the bound $c(\delta)$ in (14) is uniform in both N and W); however, it is not hard to deduce Theorem 3.2 from (a multidimensional version of) Theorem 1.1 and the Furstenberg correspondence principle; one can also use the uniform version of the Bergelson–Leibman theorem proved in [5]. As the arguments here are fairly standard, and are unrelated to those in the remainder of the paper, we defer the proof of Theorem 3.2 to Appendix B.

3.3. Pseudorandom measures

To describe the other two pillars of the argument it is necessary for the measure ν to make its appearance. (The precise properties of ν, however, will not actually be used until §5 and §6.)

Definition 3.4. (Measure, [18]) A *measure* is a non-negative function $\nu: X \to \mathbf{R}^+$ with the total mass estimate

$$\int_X \nu = 1 + o(1) \tag{15}$$

and the crude pointwise bound

$$\nu \ll_\varepsilon N^\varepsilon \tag{16}$$

for any $\varepsilon > 0$.

Remark 3.5. As remarked in [18], it is really $\nu\mu_X$ which is a measure rather than ν, where μ_X is the uniform probability measure on X; ν should be more accurately referred to as a "probability density" or "weight function". However, we retain the terminology "measure" for compatibility with [18]. The condition (16) is needed here to discard certain error terms arising from the boundary effects of shift ranges (such as those arising from the van der Corput lemma). This condition does not prominently feature in [18], as the shifts range over all of $\mathbf{Z}_N$, which has no boundary. Fortunately, (16) is very easy to establish for the majorant that we shall end up using. We note though that while the right-hand side of (16) does not look too large, we cannot possibly afford to allow factors such as N^ε to multiply into error terms such as $o(1)$, as these terms will almost certainly cease to be small at that point. Hence we can only really use (16) in situations where we already have a polynomial gain in N, which can for instance arise by exploiting the gaps in (10).

The simplest example of a measure is the constant measure $\nu \equiv 1$. Another model example worth keeping in mind is the random measure where $\nu(x) = \log R$ with independent probability $1/\log R$ for each $x \in X$, and $\nu(x) = 0$ otherwise. The following definitions

attempt to capture certain aspects of this random measure, which will eventually be satisfied by a certain truncated divisor sum concentrated on almost primes. These definitions are rather technical, and their precise form is only needed in later sections of the paper. They are somewhat artificial in nature, being a compromise between the type of control needed to establish the relative polynomial Szemerédi theorem (Theorem 3.16) and the type of control that can be easily verified for truncated divisor sums (Theorem 3.18). It may well be that a simpler notion of pseudorandomness can be given.

Definition 3.6. (Polynomial forms condition) Let $\nu: X \to \mathbf{R}^+$ be a measure. We say that ν obeys the *polynomial forms condition* if, given any $0 \leqslant J, d \leqslant 1/\eta_1$, any polynomials $Q_1, ..., Q_J \in \mathbf{Z}[\mathbf{m}_1, ..., \mathbf{m}_d]$ of d unknowns of total degree at most $1/\eta_1$, with coefficients at most W^{1/η_1} and such that $Q_j - Q_{j'}$ is non-constant for every distinct $j, j' \in [J]$, for every $\varepsilon > 0$, and for every convex body $\Omega \subset \mathbf{R}^d$ with inradius at least N^ε, and contained in the ball $B(0, M^2)$, we have the bound

$$\mathbf{E}_{\vec{h} \in \Omega \cap \mathbf{Z}^d} \int_X \prod_{j \in [J]} T^{Q_j(\vec{h})} \nu = 1 + o_\varepsilon(1). \tag{17}$$

Note the first appearance of the parameter η_1, which is controlling the degree of the pseudorandomness here. Note also that the bound is uniform in the coefficients of the polynomials $Q_1, ..., Q_J$.

Examples 3.7. The mean bound (15) is a special case of (17); another simple example is

$$\mathbf{E}_{h \in [H]} \int_X \nu T^h \nu T^{Wh^2} \nu = 1 + o(1).$$

Observe that the smaller one makes η_1, the stronger the polynomial forms condition becomes.

Remark 3.8. Definition 3.6 is a partial analogue of the "linear forms condition" in [18]. The parameter η_1 is playing multiple roles, controlling the degree, dimension, number and size of the polynomials in question. It would be more natural to split this parameter into four parameters to control each of these attributes separately, but we have chosen to artificially unify these four parameters in order to simplify the notation slightly. The parameter ε will eventually be set to be essentially η_7, but we leave it arbitrary here, to emphasize that the definition of pseudorandomness does not depend on the choice of η_7 (or H). This will be important later, basically because we need to select ν (or more precisely η_2 (or R), which is involved in the construction of ν) before we are allowed to choose η_7.

The next condition is in a similar spirit, but considerably more complicated; it allows for arbitrarily many factors in the average, as long as they have a partly linear structure, and they are organized into relatively small groups, with a separate coarse-scale averaging applied to each of the groups.

Definition 3.9. (Polynomial correlation condition) Let $\nu: X \to \mathbf{R}^+$ be a measure. We say that ν obeys the *polynomial correlation condition* if, given any $0 \leqslant D, J, L \leqslant 1/\eta_1$, any integers $D', D'', K > 0$, and any $\varepsilon > 0$, and given any vector-valued polynomials

$$\vec{P}_j \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_{D''}]^D \quad \text{and} \quad \vec{Q}_{j,k}, \vec{S}_l \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_{D''}]^{D'}$$

of degree at most $1/\eta_1$ for $j \in [J]$, $k \in [K]$ and $l \in [L]$ obeying the following non-degeneracy conditions:

- for any distinct $j, j' \in [J]$ and any $k \in [K]$, the $(D+D')$-dimensionally-valued polynomials $(\vec{P}_j, \vec{Q}_{j,k})$ and $(\vec{P}_{j'}, \vec{Q}_{j',k})$ are not parallel,
- the coefficients of $\vec{P}_j$ and $\vec{S}_l$ are bounded in magnitude by W^{1/η_1},
- the D'-dimensionally-valued polynomials $\vec{S}_l$ are distinct as l varies in $[L]$,

and given any convex body $\Omega \subset \mathbf{R}^D$ with inradius at least $M^{1/4}$ and convex bodies $\Omega' \subset \mathbf{R}^{D'}$ and $\Omega'' \subset \mathbf{R}^{D''}$ with inradii at least N^ε, with all the convex bodies contained in $B(0, M^2)$, then

$$\mathbf{E}_{\vec{n} \in \Omega' \cap \mathbf{Z}^{D'}} \mathbf{E}_{\vec{h} \in \Omega'' \cap \mathbf{Z}^{D''}} \int_X \Big(\prod_{k \in [K]} \mathbf{E}_{\vec{m} \in \Omega \cap \mathbf{Z}^D} \prod_{j \in [J]} T^{\vec{P}_j(\vec{h}) \cdot \vec{m} + \vec{Q}_{j,k}(\vec{h}) \cdot \vec{n}} \nu \Big) \prod_{l \in [L]} T^{\vec{S}_l(\vec{h}) \cdot \vec{n}} \nu = 1 + o_{D', D'', K, \varepsilon}(1). \tag{18}$$

Remark 3.10. It will be essential here that D', D'' and K can be arbitrarily large;([24]) otherwise, this condition becomes essentially a special case of the polynomial forms condition. Indeed, in our argument, these quantities will get as large as $O(1/\eta_6)$, which is far larger than $1/\eta_1$. As in the preceding definition, ε will eventually be set to equal essentially η_7, but we refrain from doing so here to keep the definition of pseudorandomness independent of η_7, to avoid the appearance of circularity in the argument.

Remark 3.11. The correlation condition (18) would follow from the polynomial forms condition (17), if we had the pointwise bounds

$$\mathbf{E}_{\vec{m} \in \Omega \cap \mathbf{Z}^D} \prod_{j \in [J]} T^{\vec{P}_j(\vec{h}) \cdot \vec{m} + \vec{Q}_{j,k}(\vec{h}) \cdot \vec{n}} \nu = 1 + o(1) \tag{19}$$

([24]) An analogous phenomenon occurs in the correlation condition in [18], where it was essential that the exponent q appearing in that condition (which is roughly analogous to K here) could be arbitrarily large.

for each $k \in [K]$ and all $\vec{h}$ and $\vec{n}$. Unfortunately, such a bound is too optimistic to be true: for instance, if $\vec{P}_j(\vec{h}) = Q_{j,k}(\vec{h}) = 0$, then the left-hand side is an average of ν^J, which is almost certainly much larger than 1. In the number-theoretic applications in which ν is supposed to concentrate on almost primes, one also has similar problems when $\vec{P}_j(\vec{h})$ and $Q_{j,k}(\vec{h})$ are non-zero but very smooth (i.e. they have many small prime factors slightly larger than w). In [18] these smooth cases were modeled by a weight function τ, which obeyed arbitrarily large moment conditions which led to integral estimates analogous to (18). In this paper we have found it more convenient to not explicitly create the weight function, instead placing the integral estimate (18) in the correlation condition hypothesis directly. In fact, one can view (18) as an assertion that (19) holds "asymptotically almost everywhere" (cf. Proposition 6.2 below).

Remark 3.12. One could generalize (18) slightly by allowing the number of terms J in the j product to depend on k, but we will not need this strengthening and in any event it follows automatically from (18), by a Hölder inequality argument similar to that used in Lemma 3.14 below.

Definition 3.13. (Pseudorandom measure) A *pseudorandom measure* is any measure ν which obeys both the polynomial forms condition and the correlation condition.

The following lemma (cf. [18, Lemma 3.4]) is useful.

LEMMA 3.14. *If ν is a pseudorandom measure, then so is $\nu_{1/2} := \frac{1}{2}(1+\nu)$ (possibly with slightly different decay rates for the $o(1)$ error terms).*

Proof. It is clear that $\nu_{1/2}$ satisfies (15) and (16). Because ν obeys the polynomial forms condition (17), one can easily verify using the binomial formula that $\nu_{1/2}$ does also. Now we turn to the polynomial correlation condition, which requires a little more care. Setting $\vec{Q}_{j,k}$ to be independent of k, we obtain that

$$\mathbf{E}_{\vec{n}\in\Omega'\cap\mathbf{Z}^{D'}}\mathbf{E}_{\vec{h}\in\Omega''\cap\mathbf{Z}^{D''}}\int_X\Big(\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}\prod_{j\in[J]}T^{\vec{P}_j(\vec{h})\cdot\vec{m}+\vec{Q}_j(\vec{h})\cdot\vec{n}}\nu\Big)^K\prod_{l\in[L]}T^{\vec{S}_l(\vec{h})\cdot\vec{n}}\nu = 1+o_{D',D'',K,\varepsilon}(1)$$

for all $K \geqslant 0$ and $\vec{P}_j$, $\vec{Q}_j$ and $\vec{S}_l$ obeying the hypotheses of the correlation condition. By the binomial formula, this implies that

$$\mathbf{E}_{\vec{n}\in\Omega'\cap\mathbf{Z}^{D'}}\mathbf{E}_{\vec{h}\in\Omega''\cap\mathbf{Z}^{D''}}\int_X\Big(\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}\prod_{j\in[J]}T^{\vec{P}_j(\vec{h})\cdot\vec{m}+\vec{Q}_j(\vec{h})\cdot\vec{n}}\nu-1\Big)^K\prod_{l\in[L]}T^{\vec{S}_l(\vec{h})\cdot\vec{n}}\nu = 0^K+o_{D',D'',K,\varepsilon}(1).$$

(Recall of course that $0^0=1$.) Let us take K to be a large *even* integer. Another application of the binomial formula allows one to replace the final ν by $\nu_{1/2}$. By the triangle inequality in a weighted Lebesgue norm l^K, we may then replace the other occurrences of ν by $\nu_{1/2}$ also:

$$\mathbf{E}_{\vec{n}\in\Omega'\cap\mathbf{Z}^{D'}}\mathbf{E}_{\vec{h}\in\Omega''\cap\mathbf{Z}^{D''}}\int_X\Big(\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}\prod_{j\in[J]}T^{\vec{P}_j(\vec{h})\cdot\vec{m}+\vec{Q}_j(\vec{h})\cdot\vec{n}}\nu_{1/2}-1\Big)^K \times\prod_{l\in[L]}T^{\vec{S}_l(\vec{h})\cdot\vec{n}}\nu_{1/2}=0^K+o_{D',D'',K,\varepsilon}(1).$$

This was only proven for even K, but follows also for odd K by the Cauchy–Schwarz inequality (94). By Hölder's inequality, we obtain a similar statement when the Q_j's are now allowed to vary in k:

$$\mathbf{E}_{\vec{n}\in\Omega'\cap\mathbf{Z}^{D'}}\mathbf{E}_{\vec{h}\in\Omega''\cap\mathbf{Z}^{D''}}\int_X\Big(\prod_{k\in[K]}\Big(\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}\prod_{j\in[J]}T^{\vec{P}_j(\vec{h})\cdot\vec{m}+\vec{Q}_{j,k}(\vec{h})\cdot\vec{n}}\nu_{1/2}-1\Big)\Big) \times\prod_{l\in[L]}T^{\vec{S}_l(\vec{h})\cdot\vec{n}}\nu_{1/2}=0^K+o_{D',D'',K,\varepsilon}(1).$$

Applying the binomial formula again, we see that $\nu_{1/2}$ obeys (18) as desired. □

3.15. The transference principle

We can now state the second pillar of our argument (cf. [18, Theorem 3.5]).

THEOREM 3.16. (Relative polynomial Szemerédi theorem) *Let the notation and assumptions be as in §2. Then, given any pseudorandom measure ν and any $g\colon X\to\mathbf{R}$ obeying the pointwise bound $0\leqslant g\leqslant\nu$ and the mean bound*

$$\int_X g\geqslant\eta_3, \tag{20}$$

we have

$$\mathbf{E}_{m\in[M]}\int_X T^{P_1(Wm)/W}g\dots T^{P_k(Wm)/W}g\geqslant\tfrac{1}{2}c\big(\tfrac{1}{2}\eta_3\big)-o(1), \tag{21}$$

where $c(\cdot)$ is the function appearing in Theorem 3.2.

Apart from inessential factors of 2 (and the substantially worse decay rates concealed within the $o(1)$ notation), this theorem is significantly stronger than Theorem 3.2, which is essentially the special case $\nu=1$. In fact, we shall derive Theorem 3.16 from Theorem 3.2 using the transference principle technology from [18]. The argument is lengthy and will occupy §§4–7.

3.17. Construction of the majorant

To conclude Theorem 2.3 from Theorem 3.16 and (12) it clearly suffices to show the following (cf. [18, Proposition 9.1]).

THEOREM 3.18. (Existence of pseudorandom majorant) *Let the notation and assumptions be as in* §2. *Then, there exists a pseudorandom measure* ν *such that the function* f *defined in* (11) *enjoys the pointwise bound* $0 \leqslant f \leqslant \nu$.

This is the third pillar of the argument. The majorant ν acts as an "enveloping sieve" for the primes (or more precisely, for the primes equal to b modulo W), in the sense of [25] and [26]. It is defined explicitly in §8. However, for the purposes of the proof of the other pillars of the argument (Theorems 3.2 and 3.16) it will not be necessary to know the precise definition of ν, only that ν majorizes f and is pseudorandom. In order to establish this pseudorandomness, it is necessary that η_2 is small compared to η_1. On the other hand, observe that ν does not depend on H, and thus it is insensitive to the choice of η_7.

The proof of Theorem 3.18 follows similar lines to those in [18] and [19], except that the "local" or "singular series" calculation is more complicated, as one is now forced to count solutions to one or more *polynomial* equations over F_p, rather than linear equations. Fortunately, it turns out that the polynomials involved happen to be *linear* in at least one "coarse-scale" variable, and so the number of solutions can be counted relatively easily, without recourse to any deep arithmetic facts (such as the Weil conjectures). We establish Theorem 3.18 in §§8–12, using some basic facts about convex bodies, solutions to polynomial equations in F_p, and distribution of prime numbers which are recalled in Appendices C–E, respectively.

4. Overview of the proof of the transference principle

We now begin the proof of the relative polynomial Szemerédi theorem (Theorem 3.16). As in [18], this theorem will follow quickly from three simpler components. The first is the uniformly quantitative version of the ordinary polynomial Szemerédi theorem, Theorem 3.2, which will be proven in Appendix B. The second is a "polynomial generalized von Neumann theorem" (Theorem 4.5) which allows us to neglect the contribution of sufficiently "locally Gowers-uniform" contributions to (21). The third is a "local Koopman–von Neumann structure theorem" (Theorem 4.7) which decomposes a function $0 \leqslant f \leqslant \nu$ (outside of a negligible set) into a bounded positive component $f_{U^\perp}$ and a locally Gowers-uniform error f_U. The purpose of this section is to formally state the latter two

components and show how they imply Theorem 3.16; the proofs of these components will then occupy subsequent sections of the paper.

The pseudorandom measure ν plays no role in the ordinary polynomial Szemerédi theorem (Theorem 3.2). In the von Neumann theorem (Theorem 4.5) the pseudorandomness of ν is exploited via the polynomial forms condition (Definition 3.6). In the structure theorem (Theorem 4.7) it is instead the polynomial correlation condition (Definition 3.9) which delivers the benefits of pseudorandomness.

4.1. Local Gowers norms

As mentioned in the introduction, a key ingredient in the proof of Theorem 3.16 will be the introduction of a norm $\|\cdot\|_{U_M^{\vec{Q}([H]^t,W)}}$ which controls averages such as those in (21). It is here that the parameter η_7 makes its first appearance, via the shift range H. The purpose of this subsection is to define these norms formally.

Let $f\colon X\to\mathbf{R}$ be a function. For any $d\geqslant 1$, recall that the (global) *Gowers uniformity norm* $\|f\|_{U^d}$ of f is defined by the formula

$$\|f\|_{U^d}^{2^d} := \mathbf{E}_{m_1,\dots,m_d\in\mathbf{Z}_N}\int_X \prod_{(\omega_1,\dots,\omega_d)\in\{0,1\}^d} T^{\omega_1 m_1+\dots+\omega_d m_d} f.$$

An equivalent definition is

$$\|f\|_{U^d}^{2^d} := \mathbf{E}_{m_1^{(0)},\dots,m_d^{(0)},m_1^{(1)},\dots,m_d^{(1)}\in\mathbf{Z}_N}\int_X \prod_{(\omega_1,\dots,\omega_d)\in\{0,1\}^d} T^{m_1^{(\omega_1)}+\dots+m_d^{(\omega_d)}} f,$$

as can be seen by making the substitutions $m_j^{(1)}:=m_j^{(0)}+m_j$, $j\in[d]$, and shifting the integral by $m_1^{(0)}+\dots+m_d^{(0)}$.

We will not directly use the global Gowers norms in this paper, because the range of the shifts m in those norms is too large for our applications. Instead, we shall need local versions of this norm. For any steps $a_1,\dots,a_d\in\mathbf{Z}$, we define the *local Gowers uniformity norm* $U_{\sqrt{M}}^{a_1,\dots,a_d}$ by([25])

$$\|f\|_{U_{\sqrt{M}}^{a_1,\dots,a_d}}^{2^d} := \mathbf{E}_{m_1^{(0)},\dots,m_d^{(0)},m_1^{(1)},\dots,m_d^{(1)}\in[\sqrt{M}]}\int_X \prod_{(\omega_1,\dots,\omega_d)\in\{0,1\}^d} T^{m_1^{(\omega_1)}a_1+\dots+m_d^{(\omega_d)}a_d} f. \tag{22}$$

Thus, for instance, when $\sqrt{M}=N$ and $a_1,\dots,a_d$ are invertible in $\mathbf{Z}_N^\times$, then the $U_{\sqrt{M}}^{a_1,\dots,a_d}$ norm is the same as the U^d norm. When $\sqrt{M}$ is much smaller than N, however, there

([25]) We will need to pass from shifts of size $O(M)$ to shifts of size $O(\sqrt{M})$ to avoid dealing with certain boundary terms (similar to those arising in the van der Corput lemma).

appears to be no obvious comparison between these two norms. It is not immediately obvious that the local Gowers norm is indeed a norm, but we shall show this in Appendix A, where basic properties of these norms are established. In practice, we shall take $a_1, ..., a_d$ to be rather small compared to R or M, indeed these steps will have size $O(H^{O(1)})$.

Remark 4.2. One can generalize this norm to complex-valued functions f by conjugating those factors of f for which $\omega_1+...+\omega_d$ is odd. If we then set $f=e(\phi)=e^{2\pi i\phi}$ for some phase function $\phi: X\to\mathbf{R}/\mathbf{Z}$, then the local Gowers $\|f\|_{U^{a_1,...,a_d}_{\sqrt{M}}}$ norm is informally measuring the extent to which the d-fold difference

$$\sum_{\omega_1,...,\omega_d\in\{0,1\}} (-1)^{\omega_1+...+\omega_d}\phi(x+m_1^{(\omega_1)}a_1+...+m_d^{(\omega_d)}a_d)$$

is close to zero, where x ranges over X, and $m_j^{(0)}$ and $m_j^{(1)}$ range over $[M]$ for $j\in[d]$. Even more informally, these norms are measuring the extent to which ϕ "behaves like" a polynomial of degree less than d on arithmetic progressions of the form

$$\big\{x+m_1a_1+...+m_da_d : m_1, ..., m_d \in \big[\sqrt{M}\,\big]\big\},$$

where $x\in X$ is arbitrary. The global Gowers norm U^d, in contrast, measures similar behavior over the entire space X.

We shall estimate the Gowers-uniform contributions to (21), via repeated application of the van der Corput lemma, using the standard *polynomial exhaustion theorem* (PET) induction scheme [4]. This will eventually allow us to control these contributions, not by a single local Gowers-uniform norm, but rather by an *average* of such norms, in which the shifts $h_1, ..., h_d$ are fine and parameterized by a certain polynomial. More precisely, given any $t\geqslant 0$ and any d-tuple $\vec{Q}=(Q_1, ..., Q_d)\in\mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{W}]^d$ of polynomials, we define the *averaged local Gowers uniformity norm* $U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}$ by the formula

$$\|f\|^{2^d}_{U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}} := \mathbf{E}_{\vec{h}\in[H]^t}\|f\|^{2^d}_{U^{Q_1(\vec{h},W),...,Q_d(\vec{h},W)}_{\sqrt{M}}}. \tag{23}$$

Inserting (22), we thus have

$$\|f\|^{2^d}_{U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}} := \mathbf{E}_{\vec{h}\in[H]^t}\mathbf{E}_{m_1^{(0)},...,m_d^{(0)},m_1^{(1)},...,m_d^{(1)}\in[\sqrt{M}\,]} \int_X \prod_{(\omega_1,...,\omega_d)\in\{0,1\}^d} T^{m_1^{(\omega_1)}Q_1(\vec{h},W)+...+m_d^{(\omega_d)}Q_d(\vec{h},W)}f. \tag{24}$$

In Appendix A we show that the local Gowers uniformity norms are indeed norms; by the triangle inequality in l^{2^d}, this implies that the averaged local Gowers uniformity norms are also norms. To avoid degeneracies, we will assume that none of the polynomials $Q_1, ..., Q_d$ vanish.

Remark 4.3. The distinction between local Gowers uniform norms and their averaged counterparts is a necessary feature of our "quantitative" setting. In the "qualitative" setting of traditional (infinitary) ergodic theory (where X is infinite), there is no need for this sort of distinction; if the local Gowers uniformity norms go to zero as $M \to \infty$ for the shifts $h_1 = ... = h_d = 1$, then it is not hard (using various forms of the Cauchy–Schwarz–Gowers inequality, such as those in Appendix A) to show that the same is true for any other fixed choice of shifts $h_1, ..., h_d$, and hence the averaged norms will also go to zero as $M \to \infty$ for any fixed choice of $\vec{Q}$ and H. The converse implications are also easy to establish. Thus one can use a single local Gowers uniformity norm, $U^{1,...,1}_{\sqrt{M}}$, to control everything in the limit $M \to \infty$ with H bounded; this then corresponds to the Gowers–Host–Kra seminorms used in [21] and [23] to control polynomial averages. However, in our more quantitative setting, where H is allowed to grow like a (very small) power of N, we cannot afford to use the above equivalences (as they will amplify the $o(1)$ errors in our arguments to be unacceptably large), and so must turn instead to the more complicated-seeming averaged local Gowers uniformity norms.

4.4. The polynomial generalized von Neumann theorem

We are now ready to state the second main component of the proof of Theorem 3.16 (the first component being Theorem 3.2).

THEOREM 4.5. (Polynomial generalized von Neumann theorem) *Let the notation and assumptions be as in* §2. *Then, there exist* $d \geqslant 2$, $t \geqslant 0$ *of size* $O(1)$ *and a* d*-tuple* $\vec{Q} \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{W}]^d$ *of degree* $O(1)$ *with coefficients* $O(1)$, *and with none of the components of* $\vec{Q}$ *vanishing, as well as a constant* $c > 0$ *depending only on* $P_1, ..., P_k$, *such that the inequality*

$$\left| \mathbf{E}_{m \in [M]} \int_X T^{P_1(Wm)/W} g_1 \, ... \, T^{P_k(Wm)/W} g_k \right| \ll \min_{j \in [k]} \|g_j\|^c_{U^{\vec{Q}([H]^t, W)}_{\sqrt{M}}} + o(1) \qquad (25)$$

holds for any functions $g_1, ..., g_k : X \to \mathbf{R}$ *obeying the pointwise bound* $|g_j| \leqslant 1 + \nu$ *for all* $j \in [k]$ *and* $x \in X$, *and some pseudorandom measure* ν.

This theorem is a local polynomial analogue of [18, Proposition 5.3]. It will be proven by a vast number of applications of the van der Corput lemma and the Cauchy–Schwarz inequality following the standard PET induction scheme; the idea is to first apply the van der Corput lemma repeatedly to linearize the polynomials $P_1, ..., P_k$, and then apply the Cauchy–Schwarz inequality repeatedly to estimate the linearized averages by local Gowers norms. The presence of the measure ν will cause a large number of shifts of ν to appear as weights, but these will ultimately be controllable via the polynomial forms

condition (Definition 3.6). The final values of d and t obtained will be very large (indeed, they exhibit Ackermann-type behavior in the maximal degree of $P_1, ..., P_k$) but can be chosen to be small compared to $1/\eta_1$, which controls the pseudorandomness of ν.

The proof of Theorem 4.5 is elementary but rather lengthy (and notation intensive), and shall occupy all of §5. The ν=1 case of this theorem is a finitary version of a similar result in [21], while the linear case of this theorem (when the $P_j - P_{j'}$ are all linear) is essentially in [18]. Indeed, the proof of this theorem will use a combination of the techniques from both of these papers.

4.6. The local Koopman–von Neumann theorem

The third major component of the proof of Theorem 3.16 is the following structure theorem.

THEOREM 4.7. (Structure theorem) *Let the notation and assumptions be as in* §2. *Let* $t \geqslant 0$ *and* $d \geqslant 2$ *be of size* $O(1)$, *and let* $\vec{Q} \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{W}]^d$ *be polynomials of degree* $O(1)$ *with coefficients* $O(1)$ (*and with none of the components of* $\vec{Q}$ *vanishing*). *Then, given any pseudorandom measure* ν *and any* $g: X \to \mathbf{R}^+$ *with the pointwise bound* $0 \leqslant g \leqslant \nu$, *there exist functions* $g_{U^\perp}, g_U: X \to \mathbf{R}$ *with the pointwise bound*

$$0 \leqslant g_{U^\perp}(x) + g_U(x) \leqslant g(x) \tag{26}$$

of g *obeying the following estimates*:

- (*boundedness of the structured component*)

$$0 \leqslant g_{U^\perp}(x) \leqslant 1 \quad \text{for all } x \in X; \tag{27}$$

- ($g_{U^\perp}$ *captures most of the mass*)

$$\int_X g_{U^\perp} \geqslant \int_X g - O(\eta_4) - o(1); \tag{28}$$

- (*uniformity of the unstructured component*)

$$\|g_U\|_{U^{\vec{Q}([H]^t, W)}_{\sqrt{M}}} \leqslant \eta_4^{1/2^d} + o(1). \tag{29}$$

Remark 4.8. Note the first apperance of the parameter η_4, which is controlling the accuracy of this structure theorem. One can make this accuracy as strong as desired, but at the cost of pushing η_7 (and thus H) down, which will ultimately worsen many of the $o(1)$ errors appearing here and elsewhere.

Theorem 4.7 is the most technical and difficult component of the entire paper, and is proven in §6 and §7. It is a "finitary ergodic theory" argument which relies on iterating a certain "dichotomy between structure and randomness". Here, the randomness is measured using the local Gowers uniformity norm $U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}$. To measure the structured component, we need the machinery of *dual functions*, as in [18], together with an energy incrementation argument which we formalize abstractly in Theorem 7.1. A key point will be that $\nu-1$ is "orthogonal" to these dual functions in a rather strong sense (see Proposition 6.5), which will be the key to approximating functions bounded by ν with functions bounded by 1. This will be accomplished by a rather tricky series of applications of the Cauchy–Schwarz inequality and will rely heavily on the polynomial correlation condition (Definition 3.9).

4.9. Proof of Theorem 3.16

Using Theorems 4.5 and 4.7, we can now quickly prove Theorem 3.16 (and hence Theorem 1.3, assuming Theorem 3.18), following the same argument as in [18].

Let the notation and assumptions be as in §2. Let ν be a pseudorandom measure, and let $g\colon X\to\mathbf{R}$ obey the pointwise bound $0\leqslant g\leqslant\nu$ and (20).

Let $d>0$, $t>0$ and $\vec{Q}$ be as in Theorem 4.5; these expressions depend only on $P_1,...,P_k$, and so we do not need to explicitly track their influence on the $O(\cdot)$ and $o(\cdot)$ notation. Applying Theorem 4.7, we thus obtain functions g_U and $g_{U^\perp}$ obeying the properties claimed in that theorem. From (26) we have

$$\mathbf{E}_{m\in[M]}\int_X T^{P_1(Wm)/W}g\,...\,T^{P_k(Wm)/W}g \geqslant \mathbf{E}_{m\in[M]}\int_X T^{P_1(Wm)/W}(g_{U^\perp}+g_U)\,...\,T^{P_k(Wm)/W}(g_{U^\perp}+g_U).$$

We expand the right-hand side into $2^k=O(1)$ terms. Consider any of the 2^k-1 of these terms which involve at least one factor of g_U. From (26) and (27) we know that g_U and $g_{U^\perp}$ are both bounded pointwise in magnitude by $\nu+1+o(1)$, which is $O(\nu+1)$ when N is large enough. Thus, by Theorem 4.5 and (29), the contribution of all of these terms can be bounded in magnitude by a constant multiple of

$$\|g_U\|^c_{U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}}+o(1)\ll\eta_4^{c/2^d}+o(1)$$

for some $c>0$ depending only on $P_1,...,P_k$. On the other hand, from (28), (20) and the choice of parameters, we have that

$$\int_X g_{U^\perp}\geqslant\tfrac{1}{2}\eta_3.$$

Applying this, (27) and Theorem 3.2 yields

$$\mathbf{E}_{m\in[M]}\int_X T^{P_1(Wm)/W}g_{U^\perp}\dots T^{P_k(Wm)/W}g_{U^\perp}\geqslant c\big(\tfrac{1}{2}\eta_3\big)>0.$$

Putting all this together, we conclude that

$$\mathbf{E}_{m\in[M]}\int_X T^{P_1(Wm)/W}g\dots T^{P_k(Wm)/W}g\geqslant c\big(\tfrac{1}{2}\eta_3\big)-O(\eta_4^{c/2^d})-o(1).$$

As η_4 is chosen small compared to η_3, Theorem 3.16 follows.

5. Proof of the generalized von Neumann theorem

In this section we prove Theorem 4.5. In a nutshell, our argument here will be a rigorous implementation of the following scheme:

polynomial average

$\ll$ weighted linear average$+o(1)$	(van der Corput)
$\ll$ weighted parallelopiped average$+o(1)$	(weighted gen. von Neumann)
$\ll$ unweighted parallelopiped average$+o(1)$	(Cauchy–Schwarz).

The argument is based upon that used to prove [18, Proposition 5.3], namely repeated application of the Cauchy–Schwarz inequality to replace various functions g_j by ν (or $\nu+1$), followed by application of the polynomial forms condition (Definition 3.6) to replace the resulting polynomial averages of ν by $1+o(1)$. The major new ingredient in the argument compared to [18] will be the *polynomial exhaustion theorem* (PET) induction scheme (used for instance in [6]) in order to estimate the polynomial average in (25) by a linear average similar to that treated in [18, Proposition 5.3]. After using PET induction to achieve this linearization, the rest of the proof is broadly similar to that in [18, Proposition 5.3], except for the fact that the shift parameters are restricted to be of size M or $\sqrt{M}$ rather than N, and that there is also some additional averaging over short shift parameters of size $O(H)$.

The arguments are elementary and straightforward, but will require a rather large amount of new notation in order to keep track of all the weights and factors created by applications of the Cauchy–Schwarz inequality. Fortunately, none of this notation will be needed in any other section; indeed, this section can be read independently of the rest of the paper (although it of course relies on the material in earlier sections, and also on Appendix A).

We begin with some simple reductions. First observe (as in [18]) that Lemma 3.14 allows us to replace the hypotheses $|g_j| \leqslant 1+\nu$ by the slightly stronger $|g_j| \leqslant \nu$, at the (acceptable) cost of worsening the implicit constant in (25) by a factor of 2^k. Next, we claim that it suffices to find d, t, c and $\vec{Q} \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{W}]^d$ for which we have the weaker estimate

$$\left| \mathbf{E}_{m \in [M]} \int_X T^{P_1(Wm)/W} g_1 \, ... \, T^{P_k(Wm)/W} g_k \right| \ll \|g_1\|^c_{U^{\vec{Q}([H]^t, W)}_{\sqrt{M}}} + o(1) \tag{30}$$

(i.e. we only control the average using the norm of g_1, rather than the best norm of all of the g_i's). Indeed, if we could show this, then by symmetry we could find d_j, t_j, and $\vec{Q}_j \in \mathbf{Z}[h_1, ..., \mathbf{h}_{t_j}, \mathbf{W}]^{d_j}$, $j \in [k]$, such that

$$\left| \mathbf{E}_{m \in [M]} \int_X T^{P_1(Wm)/W} g_1 \, ... \, T^{P_k(Wm)/W} g_k \right| \ll \|g_j\|^{c_j}_{U^{\vec{Q}_j([H]^{t_j}, W)}_{\sqrt{M}}} + o(1),$$

whenever $j \in [k]$ and ν is pseudorandom. The claim then follows by using Lemma A.3 to obtain a local Gowers norm $U^{\vec{Q}([H]^t, W)}_{\sqrt{M}}$ which dominates each of the individual norms $U^{\vec{Q}_j([H]^{t_j}, W)}_{\sqrt{M}}$, and taking $c := \min_{j \in [k]} c_j$. (Note that the pointwise bound $|g_j| \leqslant \nu$ and the polynomial forms condition easily imply that the $U^{\vec{Q}_j([H]^{t_j}, W)}_{\sqrt{M}}$ norm of g_j is $O(1)$.)

It remains to prove (30). It should come as no surprise to the experts that this type of "generalized von Neumann" theorem will be proven via a large number of applications of van der Corput's lemma and the Cauchy–Schwarz inequality. In order to keep track of the intermediate multilinear expressions which arise during this process, it is convenient to prove a substantial generalization of this estimate. We first need the notion of a polynomial system, and averages associated with such systems.

Definition 5.1. (Polynomial system) A *polynomial system* $\mathcal{S}$ consists of the following objects:

- An integer $D \geqslant 0$, which we call the *number of fine degrees of freedom*;
- A non-empty finite index set A (the elements of which we shall refer to as *nodes* of the system);
- A polynomial $R_\alpha \in \mathbf{Z}[\mathbf{m}, \mathbf{h}_1, ..., \mathbf{h}_D, \mathbf{W}]$ in $D+2$ variables attached to each node $\alpha \in A$;
- A *distinguished node* $\alpha_0 \in A$;
- A (possibly empty) collection $A' \subset A \backslash \{\alpha_0\}$ of *inactive nodes*. The nodes in $A \backslash A'$ will be referred to as *active*. Thus for instance the distinguished node α_0 is always active.

We say that a node $\alpha \in A$ is *linear* if $R_\alpha - R_{\alpha_0}$ is at most linear in $\mathbf{m}$, thus the distinguished node is always linear. We say that the entire system $\mathcal{S}$ is linear if every active node is linear. We make the following non-degeneracy assumptions:

- If α and β are distinct nodes in A, then $R_\alpha - R_\beta$ is not constant in $\mathbf{m}, \mathbf{h}_1, ..., \mathbf{h}_D$.

• If α and β are distinct *linear* nodes in A, then $R_\alpha - R_\beta$ is not constant in $\mathbf{m}$.

Given any two nodes α and β, we define the *distance* $d(\alpha,\beta)$ between the two nodes to be the $\mathbf{m}$-degree of the polynomial $R_\alpha - R_\beta$ (which is non-zero by hypothesis); thus this distance is symmetric, non-negative, and obeys the non-Archimedean triangle inequality

$$d(\alpha,\gamma) \leqslant \max(d(\alpha,\beta), d(\beta,\gamma)).$$

Note that α is linear if and only if $d(\alpha,\alpha_1)\leqslant 1$, and furthermore we have $d(\alpha,\beta)=1$ for any two distinct linear nodes α and β.

Example 5.2. Take $D:=0$, $A:=\{1,2,3\}$, with $R_1:=0$, $R_2:=\mathbf{m}$ and $R_3:=\mathbf{m}^2$, and let 3 be the distinguished node. Then the node 3 is linear and the other two are non-linear. (If the distinguished node was 1 or 2, the situation would be reversed.)

Remark 5.3. The non-Archimedean semi-metric is naturally identifiable with a tree whose leaves are the nodes of $\mathcal{S}$, and whose intermediate nodes are balls with respect to this semi-metric; the distance between two nodes is then the height of their join. It is this tree structure (and the distinction of nodes into active, inactive and distinguished) that shall implicitly govern the dynamics of the PET induction scheme which we shall shortly perform. We will however omit the details, as we shall not explicitly use this tree structure in this paper.

Definition 5.4. (Realizations and averages) Let $\mathcal{S}$ be a polynomial system and ν be a measure. We define a *ν-realization* $\vec{f}=(f_\alpha)_{\alpha\in A}$ of $\mathcal{S}$ to be an assignment of a function $f_\alpha: X\to\mathbf{R}$ to each node α with the following properties:

• for any node α, we have the pointwise bound $|f_\alpha|\leqslant\nu$;

• for any inactive node α, we have $f_\alpha=\nu$.

We refer to the function f_{α_0} attached to the distinguished node α_0 as the *distinguished function*. We define the *average* $\Lambda_{\mathcal{S}}(\vec{f})\in\mathbf{R}$ of a system $\mathcal{S}$ and its ν-realization $\vec{f}$ to be the quantity

$$\Lambda_{\mathcal{S}}(\vec{f}) := \mathbf{E}_{h_1,\ldots,h_D\in[H]}\mathbf{E}_{m\in[M]}\int_X \prod_{\alpha\in A} T^{R_\alpha(m,h_1,\ldots,h_D,W)} f_\alpha.$$

Example 5.5. If $\mathcal{S}$ is the system in Example 5.2, then

$$\Lambda_{\mathcal{S}}(\vec{f}) = \mathbf{E}_{m\in[M]}\int_X f_1 T^m f_2 T^{m^2} f_3. \tag{31}$$

Example 5.6. The average

$$\mathbf{E}_{h,h'\in[H]}\mathbf{E}_{m\in[M]}\int_X \nu T^{m+h} f_2 T^{m+h'} f_2 T^{(m+h)^2} f_3 T^{(m+h')^2} f_3$$

can be written in the form $\Lambda_{\mathcal{S}}(\vec{f})$ with distinguished function f_3, where $\mathcal{S}$ is a system with $D{:=}2$, $A{:=}\{1,2,2',3,3'\}$ with the node 1 inactive and distinguished node 3, with $R_1{:=}0$, $R_2{:=}\mathbf{m}{+}\mathbf{h}_1$, $R_{2'}{:=}\mathbf{m}{+}\mathbf{h}_2$, $R_3{:=}(\mathbf{m}{+}\mathbf{h}_1)^2$ and $R_4{:=}(\mathbf{m}{+}\mathbf{h}_2)^2$, and $\vec{f}$ is given by $f_1{:=}\nu$, $f_{2'}{:=}f_2$ and $f_{3'}{:=}f_3$.

Example 5.7. (Base example) Let $\mathcal{S}$ be the system with $D{:=}0$, $A{:=}\{1,...,k\}$, $\alpha_0{:=}1$, $A'{=}\varnothing$ (thus all nodes are active) and $Q_j{:=}P_j(\mathbf{W}\mathbf{m})/\mathbf{W}$. We observe from (3) that this is indeed a system. Then $\vec{f}{:=}(g_1,...,g_k)$ is a ν-realization of $\mathcal{S}$ with distinguished function g_1, and

$$\Lambda_{\mathcal{S}}(\vec{f}) = \mathbf{E}_{m\in[M]} \int_X T^{P_1(Wm)/W} g_1 \dots T^{P_k(Wm)/W} g_k.$$

This system $\mathcal{S}$ is linear if and only if the polynomials $P_j - P_{j'}$ are all linear.

Remark 5.8. (Translation invariance) Given a polynomial system $\mathcal{S}$ and a polynomial $R{\in}\mathbf{Z}[\mathbf{m},\mathbf{h}_1,...,\mathbf{h}_t,\mathbf{W}]$, we can define the shifted polynomial system $\mathcal{S}{-}R$ by replacing each of the polynomials R_α by $R_\alpha{-}R$; it is easy to verify that this does not affect any of the characteristics of the system, and in particular we have $\Lambda_{\mathcal{S}}(\vec{f}){=}\Lambda_{\mathcal{S}-R}(\vec{f})$ for any ν-realization $\vec{f}$ of $\mathcal{S}$ (and hence of $\mathcal{S}{-}R$). This translation invariance gives us the freedom to set any single polynomial R_α of our choosing to equal 0; indeed, we shall exploit this freedom whenever we wish to use van der Corput's lemma or the Cauchy–Schwarz inequality to deactivate any given node.

The estimate (30) then follows immediately from the following proposition.

PROPOSITION 5.9. (Generalized von Neumann theorem for polynomial systems) *Let $\mathcal{S}$ be a polynomial system with distinguished node α_0. Then, if η_1 is sufficiently small depending on $\mathcal{S}$ and α_0, there exist $d,t{\geqslant}0$, $\vec{Q}{\in}\mathbf{Z}[\mathbf{h}_1,...,\mathbf{h}_t,\mathbf{W}]^d$ and $c{>}0$ depending only on $\mathcal{S}$ and α_0 such that one has the bound*

$$|\Lambda_{\mathcal{S}}(\vec{f})| \ll_{\mathcal{S}} \|f_{\alpha_0}\|^c_{U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}} + o_{\mathcal{S}}(1)$$

whenever ν is a pseudorandom measure and $\vec{f}$ is a ν-realization of $\mathcal{S}$ with distinguished function f_{α_0}.

Indeed, one simply applies this proposition to Example 5.7 to conclude (30).

It remains to prove Proposition 5.9. This will be done in three stages. The first is the "linearization" stage, in which a weighted form of van der Corput's lemma and the polynomial forms condition are applied repeatedly (using the PET induction scheme) to reduce the proof of Proposition 5.9 to the case where the system $\mathcal{S}$ is linear. The second stage is the "parallelopipedization" stage, in which one uses a weighted variant of the "Cauchy–Schwarz–Gowers inequality" to estimate the average $\Lambda_{\mathcal{S}}(f)$ associated

with a linear system $\mathcal{S}$ by a weighted average of the distinguished function f_{α_0} over parallelopipeds. Finally, there is a relatively simple "Cauchy–Schwarz" stage in which the polynomial forms condition is used one last time to replace the weights by the constant 1, at which point the proof of Proposition 5.9 is complete. We remark that the latter two stages (dealing with the linear system case) also appeared in [18]; the new feature here is the initial linearization step, which can be viewed as a weighted variant of the usual polynomial generalized von Neumann theorem (see e.g. [6]). This linearization step is also the step which shall use the fine shifts $h=O(H)$ (for reasons which will be clearer in §6); this should be contrasted with the parallelopipedization step, which relies on coarse-scale shifts $m=O\big(\sqrt{M}\,\big)$.

5.10. PET induction and linearization

We now reduce Proposition 5.9 to the linear case. We shall rely heavily here on van der Corput's lemma, which in practical terms allows us to deactivate any given node at the expense of duplicating all the other nodes in the system. Since this operation tends to increase the number of active nodes in the system, it is not immediately obvious that iterating this operation will eventually simplify the system. To make this more clear we need to introduce the notion of a weight vector.

Definition 5.11. (Weight vector) A *weight vector* is an infinite vector $\vec{w}=(w_1,w_2,\dots)$ of non-negative integers w_j, with only finitely many of the w_j's being non-zero. Given two weight vectors $\vec{w}=(w_1,w_2,\dots)$ and $\vec{w}'=(w'_1,w'_2,\dots)$, we say that $\vec{w}<\vec{w}'$ if there exists $k\geqslant 1$ such that $w_k<w'_k$, and such that $w_j=w'_j$ for all $j>k$. We say that a weight vector is *linear* if $w_j=0$ for all $j\geqslant 2$.

It is well known that the space of all weight vectors forms a well-ordered set; indeed, it is isomorphic to the ordinal ω^ω. In particular, we may perform strong induction on this space. The space of linear weight vectors forms an order ideal; indeed, a weight is linear if and only if it is less than $(0,1,0,\dots)$.

Definition 5.12. (Weight) Let $\mathcal{S}$ be a polynomial system, and let α be a node in $\mathcal{S}$ (in practice this will not be the distinguished node α_0). We say that two nodes β and γ in $\mathcal{S}$ are *equivalent* relative to α if $d(\beta,\gamma)<d(\alpha,\beta)$. This is an equivalence relation on the nodes of $\mathcal{S}$, and the equivalence classes have a well-defined distance from α. We define the *weight vector* $\vec{w}_\alpha(\mathcal{S})$ of $\mathcal{S}$ relative to α by setting the jth component for any $j\geqslant 1$ to equal the number of equivalence classes at distance j from α.

Example 5.13. Consider the system in Example 5.2. The weight of this system relative to the node 1 is $(1,1,0,\dots)$, whereas the weight of the system in Example 5.6

relative to the node 2 is $(0,1,0,\dots)$ (note that the inactive node 1 is not relevant here, nor is the node $2'$ which has distance 0 from 2), which is a smaller weight than that of the previous system.

The key inductive step in the reduction to the linear case is then the following.

PROPOSITION 5.14. (Inductive step of linearization) *Let $\mathcal{S}$ be a polynomial system with distinguished node α_0 and a non-linear active node α. If η_1 is sufficiently small depending on $\mathcal{S}$, α_0 and α, then there exists a polynomial system $\mathcal{S}'$ with distinguished node α_0' and an active node α' with $\vec{w}_{\alpha'}(\mathcal{S}')<\vec{w}_\alpha(\mathcal{S})$ with the following property: given any pseudorandom measure ν and any ν-realization $\vec{f}$ of $\mathcal{S}$, there exists a ν-realization $\vec{f}'$ of $\mathcal{S}'$ with the same distinguished function (thus $f_{\alpha_0}=f'_{\alpha'_0}$) such that*

$$|\Lambda_{\mathcal{S}}(\vec{f})|^2 \ll \Lambda_{\mathcal{S}'}(\vec{f}')+o_{\mathcal{S}}(1). \tag{32}$$

Indeed, given this proposition, a strong induction on the weight vector $\vec{w}_\alpha(\mathcal{S})$ immediately implies that, in order to prove Proposition 5.9, it suffices to do so for linear systems (since these, by definition, are the only systems without non-linear active nodes).

Before we prove this proposition in general, it is instructive to give an example.

Example 5.15. Consider the expression (31) with f_1, f_2 and f_3 bounded pointwise by ν. We rewrite this expression as

$$\Lambda_{\mathcal{S}}(\vec{f}) = \int_X f_1 \mathbf{E}_{m\in[M]} T^m f_2 T^{m^2} f_3$$

and thus, by the Cauchy–Schwarz inequality (94),

$$|\Lambda_{\mathcal{S}}(\vec{f})|^2 \leqslant \left(\int_X \nu\right) \int_X \nu |\mathbf{E}_{m\in[M]} T^m f_2 T^{m^2} f_3|^2.$$

By (15), the first factor is $1+o(1)$. Also, from van der Corput's lemma (Lemma A.1), we have

$$|\mathbf{E}_{m\in[M]} T^m f_2 T^{m^2} f_3|^2 \leqslant \mathbf{E}_{h,h'\in[H]} \mathbf{E}_{m\in[M]} T^{m+h} f_2 T^{m+h'} f_2 T^{(m+h)^2} f_3 T^{(m+h')^2} f_3 + o(1).$$

We may thus conclude a bound of the form (32), where $\Lambda_{\mathcal{S}'}(\vec{f}')$ is the quantity studied in Example 5.6. Note from Example 5.13 that $\mathcal{S}'$ has a smaller weight than $\mathcal{S}$ relative to suitably chosen nodes.

Proof of Proposition 5.14. By using translation invariance (Remark 5.8), we may normalize $R_\alpha=0$. We split $A=A_0\cup A_1$, where $A_0:=\{\beta\in A: d(\alpha,\beta)=0\}$ and $A_1:=A\setminus A_0$.

Since α is non-linear, the distinguished node α_0 lies in A_1. Then R_β is independent of $\mathbf{m}$ for all $\beta \in A_0$: $R_\beta(m, h_1, ..., h_d, W) = R_\beta(h_1, ..., h_d, W)$. We can then write

$$\Lambda_{\mathcal{S}}(\vec{f}) = \mathbf{E}_{h_1,...,h_D \in [H]} \int_X F_{h_1,...,h_D} \mathbf{E}_{m \in [M]} G_{m,h_1,...,h_D},$$

where

$$F_{h_1,...,h_D} := \prod_{\beta \in A_0} T^{R_\beta(h_1,...,h_d,W)} f_\beta$$

and

$$G_{m,h_1,...,h_D} := T^{R_\beta(m,h_1,...,h_d,W)} f_\beta.$$

Since $|f_\beta|$ is bounded pointwise by ν, we have that $|F_{h_1,...,h_D}| \leqslant H_{h_1,...,h_D}$, where

$$H_{h_1,...,h_D} := \prod_{\beta \in A_0} T^{R_\beta(h_1,...,h_D,W)} \nu, \tag{33}$$

and thus, by the Cauchy–Schwarz inequality,

$$|\Lambda_{\mathcal{S}}(\vec{f})|^2 \leqslant \left(\mathbf{E}_{h_1,...,h_D \in [H]} \int_X H_{h_1,...,h_D} \right) \mathbf{E}_{h_1,...,h_D \in [H]} \int_X H_{h_1,...,h_D} |\mathbf{E}_{m \in M} G_{m,h_1,...,h_D}|^2.$$

Since ν is pseudorandom and thus obeys the polynomial forms condition, we see from Definition 3.6 and (33) (taking η_1 sufficiently small) that

$$\mathbf{E}_{h_1,...,h_D \in [H]} \int_X H_{h_1,...,h_D} = 1 + o_{\mathcal{S}}(1)$$

(note by hypothesis that the $R_\beta - R_{\beta'}$ are not constant in $\mathbf{m}, \mathbf{h}_1, ..., \mathbf{h}_D$). Since we are always assuming N to be large, the $o_{\mathcal{S}}(1)$ error is bounded. Thus we reduce to showing that

$$\mathbf{E}_{h_1,...,h_D \in [H]} \int_X H_{h_1,...,h_D} |\mathbf{E}_{m \in M} G_{m,h_1,...,h_D}|^2 \ll \Lambda_{\mathcal{S}'}(\vec{f}') + o_{\mathcal{S}}(1)$$

for some suitable $\mathcal{S}'$ and $\vec{f}'$. But by the van der Corput lemma (Lemma A.1) (using (16) to get the upper bounds on $G_{m,h_1,...,h_D}$), we have

$$|\mathbf{E}_{m \in M} G_{m,h_1,...,h_D}|^2 \ll \mathbf{E}_{h,h' \in [H]} \mathbf{E}_{m \in [M]} G_{m+h,h_1,...,h_D} G_{m+h',h_1,...,h_D} + o(1),$$

and so, to finish the proof, it suffices to verify that the expression

$$\mathbf{E}_{h_1,...,h_D,h,h' \in [H]} \mathbf{E}_{m \in [M]} \int_X H_{h_1,...,h_D} G_{m+h,h_1,...,h_D} G_{m+h',h_1,...,h_D} \tag{34}$$

is of the form $\Lambda_{\mathcal{S}'}(\vec{f}')$ for some suitable $\mathcal{S}'$ and $\vec{f}'$. By inspection, we see that we can construct $\mathcal{S}'$ and $\vec{f}'$ as follows:

- We have $D+2$ fine degrees of freedom, which we label $\mathbf{h}_1, ..., \mathbf{h}_D$, $\mathbf{h}$ and $\mathbf{h}'$.

• The nodes A' of $\mathcal{S}'$ are $A':=A_0\cup A_1\cup A_1'$, where A_1' is another copy of A_1 (disjoint from $A_0\cup A_1$), with distinguished node $\alpha_0'=\alpha_0\in A_1$.

• We choose the node α' to be an active node of $\mathcal{S}$ in A_1 which has minimal distance from α_0. (Note that A_1 always contains at least one active node, namely α_0.)

• If $\beta\in A_0$, then β is inactive in $\mathcal{S}'$, with

$$R'_\beta(\mathbf{m},\mathbf{h}_1,...,\mathbf{h}_d,\mathbf{h},\mathbf{h}',\mathbf{W}):=R_\beta(\mathbf{m},\mathbf{h}_1,...,\mathbf{h}_d,\mathbf{W})$$

and $\vec{f}'_\beta:=\nu$;

• If $\beta\in A_1$, then β is inactive in $\mathcal{S}'$ if and only if it is inactive in $\mathcal{S}$, with

$$R'_\beta(\mathbf{m},\mathbf{h}_1,...,\mathbf{h}_d,\mathbf{h},\mathbf{h}',\mathbf{W}):=R_\beta(\mathbf{m}+\mathbf{h},\mathbf{h}_1,...,\mathbf{h}_d,\mathbf{W})$$

and $\vec{f}'_\beta:=\vec{f}_\beta$;

• If $\beta'\in A_1'$ is the counterpart of some node $\beta\in A_1$, then β' is inactive in $\mathcal{S}'$ if and only if β is inactive in $\mathcal{S}$, with

$$R'_{\beta'}(\mathbf{m},\mathbf{h}_1,...,\mathbf{h}_d,\mathbf{h},\mathbf{h}',\mathbf{W}):=R_\beta(\mathbf{m}+\mathbf{h}',\mathbf{h}_1,...,\mathbf{h}_d,\mathbf{W})$$

and $\vec{f}'_{\beta'}:=\vec{f}_\beta$.

It is then straightforward to verify that $\mathcal{S}'$ is a polynomial system, that $\vec{f}'$ is a realization of $\mathcal{S}'$, and that (34) is equal to $\Lambda_{\mathcal{S}'}(\vec{f}')$. It remains to show that $\vec{w}_{\alpha'}(\mathcal{S}')<\vec{w}_\alpha(\mathcal{S})$. Let d be the degree in $\mathbf{m}$ of $R_\alpha-R_{\alpha'}$, and thus $d\geqslant 1$. One easily verifies that the jth component of $\vec{w}_{\alpha'}(\mathcal{S}')$ is equal to that of $\vec{w}_\alpha(\mathcal{S})$ for $j>d$, and equal to one less than that of $\vec{w}_\alpha(\mathcal{S})$ when $j=d$ (basically due to the deactivation of all of the nodes in A_0). The claim follows. (The behavior of these weight vectors for $j<d$ is much more complicated, but is fortunately not relevant due to our choice of ordering on weight vectors.) □

5.16. Parallelopipedization

By the preceding discussion, we see that to prove Proposition 5.9 it suffices to do so in the case where $\mathcal{S}$ is linear. To motivate the argument, let us first work through an unweighted example (with $\nu=1$).

Example 5.17. (Unweighted linear case) Consider the linear average

$$\Lambda_{\mathcal{S}}(\vec{f}\,)=\mathbf{E}_{h,h'\in[H]}\mathbf{E}_{m\in[M]}\int_X f_0T^{hm}f_1T^{h'm}f_2,$$

with distinguished function f_0, and with $|f_0|$, $|f_1|$ and $|f_2|$ bounded pointwise by 1. We introduce some new coarse-scale shift parameters $m_1,m_2\in[\sqrt{M}\,]$. By shifting m to $m-m_1-m_2$, one can express the above average as

$$\mathbf{E}_{h,h'\in[H]}\mathbf{E}_{m\in[M]}\int_X\mathbf{E}_{m_1,m_2\in[\sqrt{M}\,]}f_0T^{h(m-m_1-m_2)}f_1T^{h'(m-m_1-m_2)}f_2+o(1),$$

and then, shifting the integral by $T^{hm_1+h'm_2}$, we obtain

$$\mathbf{E}_{h,h'\in[H]}\mathbf{E}_{m\in[M]}\int_X \mathbf{E}_{m_1,m_2\in[\sqrt{M}]} T^{hm_1+h'm_2} f_0 T^{h(m-m_2)+h'm_2} f_1 T^{h'(m-m_1)+hm_1} f_2 + o(1).$$

The point is that $T^{h(m-m_2)+h'm_2} f_1$ does not depend on m_1, while $T^{h'(m-m_1)+hm_1} f_2$ does not depend on m_2. One can then use the Cauchy–Schwarz–Gowers inequality (see e.g. [19, Corollary B.3]) to estimate this expression by

$$\Big(\mathbf{E}_{h,h'\in[H]}\mathbf{E}_{m\in[M]}\int_X \mathbf{E}_{m_1,m_1',m_2,m_2'\in[\sqrt{M}]} T^{hm_1+h'm_2} f_0 T^{hm_1'+h'm_2} f_0 \times T^{hm_1+h'm_2'} f_0 T^{hm_1'+h'm_2'} f_0\Big)^{1/4} + o(1).$$

The main term here can then be recognized as a local Gowers norm (24).

Now we return to the general linear case. Here we will need to address the presence of many additional weights which are all shifted versions of the measure ν, which requires the repeated use of weighted Cauchy–Schwarz inequalities. See [18, §5] for a worked example of this type of computation. Our arguments here shall instead follow those of [19, Appendix C], in particular relying on the weighted generalized von Neumann inequality from that paper (reproduced here as Proposition A.2).

We turn to the details. To simplify the notation we write $\vec{h}:=(h_1,...,h_d)$ and $\vec{\mathbf{h}}:=(\mathbf{h}_1,...,\mathbf{h}_d)$. We use the translation invariance (Remark 5.8) to normalize $R_{\alpha_0}=0$. We then split $A=\{\alpha_0\}\cup A_{\mathrm{l}}\cup A_{\mathrm{nl}}$, where A_{l} consists of all the linear nodes, and A_{nl} all the non-linear (and hence inactive) nodes. By the non-degeneracy assumptions in Definition 5.1, we may write

$$R_\alpha = b_\alpha \mathbf{m} + c_\alpha$$

for all $\alpha\in A_{\mathrm{l}}$ and some $b_\alpha, c_\alpha\in\mathbf{Z}[\vec{\mathbf{h}},\mathbf{W}]$ with the b_α's all distinct and non-zero. We can then write

$$\Lambda_{\mathcal{S}}(\vec{f}) = \mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{m\in[M]}\int_X f_{\alpha_0}\Big(\prod_{\alpha\in A_{\mathrm{nl}}} T^{R_\alpha(m,\vec{h},W)}\nu\Big)\prod_{\alpha\in A_{\mathrm{l}}} T^{b_\alpha m+c_\alpha} f_\alpha.$$

We introduce some new coarse-scale shift parameters $m_\alpha\in\big[\sqrt{M}\big]$ for $\alpha\in A_{\mathrm{l}}$, and thus the vector $\vec{m}:=(m_\alpha)_{\alpha\in A_{\mathrm{l}}}$ lies in $\big[\sqrt{M}\big]^{A_{\mathrm{l}}}$. We shift m to $m-\sum_{\alpha\in A_{\mathrm{l}}} m_\alpha$ and observe that

$$\mathbf{E}_{m\in[M]} x_m = \mathbf{E}_{m\in[M]} x_{m-\sum_{\alpha\in A_{\mathrm{l}}} m_\alpha} + o(1)$$

whenever $m_\alpha\in\big[\sqrt{M}\big]$ and $x_m \ll_\varepsilon N^\varepsilon$ for all $\varepsilon>0$. Averaging this in $\vec{m}$ (cf. (96)), we obtain

$$\mathbf{E}_{m\in[M]} x_m = \mathbf{E}_{\vec{m}\in[\sqrt{M}]^{A_{\mathrm{l}}}}\mathbf{E}_{m\in[M]} x_{m-\sum_{\alpha\in A_{\mathrm{l}}} m_\alpha} + o(1).$$

Applying this (and (16)), and shifting the integral by the polynomial

$$Q_0 := \sum_{\alpha \in A_1} b_\alpha(\vec{\mathbf{h}}, \mathbf{W}) \mathbf{m}_\alpha, \tag{35}$$

we obtain

$$\Lambda_{\mathcal{S}}(\vec{f}) = \mathbf{E}_{\vec{h} \in [H]^D} \mathbf{E}_{m \in [M]} \int_X \mathbf{E}_{\vec{m} \in [\sqrt{M}]^{A_1}} f_{\alpha_0, m, \vec{h}, \vec{m}, W} \prod_{\alpha \in A_1} f_{\alpha, m, \vec{h}, \vec{m}, W} + o(1),$$

where

$$f_{\alpha_0, \vec{h}, \vec{m}, W} := T^{Q_{\alpha_0}(\vec{h}, \vec{m}, W)} \Big(f_{\alpha_0} \prod_{\alpha \in A_{\mathrm{nl}}} T^{R_\alpha(m - \sum_{\alpha \in A_1} m_\alpha, \vec{h}, W)} \nu \Big),$$
$$f_{\alpha, m, \vec{h}, \vec{m}, W} := T^{b_\alpha(\vec{h}, W) m + c_\alpha(\vec{h}, W) + \sum_{\beta \in A_1} (b_\beta(\vec{h}, W) - b_\alpha(\vec{h}, W)) m_\alpha} f_\alpha.$$

The point of all these manipulations is that for each linear node $\alpha \in A_1$, $f_{\alpha, m, \vec{h}, \vec{m}, W}$ is independent of the coarse-scale parameter m_α. Also observe the pointwise bound

$$|f_{\alpha, m, \vec{h}, \vec{m}, W}| \leqslant \nu_{\alpha, m, \vec{h}, \vec{m}, W},$$

where

$$\nu_{\alpha, m, \vec{h}, \vec{m}, W} := T^{b_\alpha(\vec{h}, W) m + c_\alpha(\vec{h}, W) + \sum_{\beta \in A_1} (b_\beta(\vec{h}, W) - b_\alpha(\vec{h}, W)) m_\beta} \nu.$$

By applying the weighted generalized von Neumann theorem (Proposition A.2) in the $\vec{m}$ variables, we thus have

$$|\Lambda_{\mathcal{S}}(\vec{f})| \leqslant \mathbf{E}_{\vec{h} \in [H]^D} \mathbf{E}_{m \in [M]} \int_X \|f_{\alpha_0, m, \vec{h}, \cdot, W}\|_{\square^{A_1}(\nu)} \prod_{\alpha \in A_1} \|\nu_{\alpha, m, \vec{h}, \cdot, W}\|^{1/2}_{\square^{A_1 \setminus \{\alpha\}}} + o(1), \tag{36}$$

where the Gowers box norms $\square^{A_1 \setminus \{\alpha\}}$ and the weighted Gowers box norms $\square^{A_1}(\nu)$ are defined([26]) in Appendix A. We now claim the estimate

$$\mathbf{E}_{\vec{h} \in [H]^D} \mathbf{E}_{m \in [M]} \int_X \|\nu_{\alpha, m, \vec{h}, \cdot, W}\|^{2^{|A_1|-1}}_{\square^{A_1 \setminus \{\alpha\}}} \ll 1 \tag{37}$$

for each $\alpha \in A_1$. Indeed, the left-hand side can be expanded as

$$\mathbf{E}_{\vec{h} \in [H]^D} \mathbf{E}_{\vec{m}^{(0)}, \vec{m}^{(1)} \in [H]^{A_1}} \mathbf{E}_{m \in [M]} \int_X \prod_{\omega \in \{0,1\}^{A_1 \setminus \{\alpha\}}} T^{b_\alpha(\vec{h}, W) m + c_\alpha(\vec{h}, W) + \sum_{\beta \in A_1} (b_\beta(\vec{h}, W) - b_\alpha(\vec{h}, W)) m_\beta^{(\omega_\beta)}} \nu.$$

([26]) These norms will only make a brief appearance here; they are not used elsewhere in the main argument.

The distinctness of the b_β's ensures that the polynomial shifts of ν here are all distinct, and so, by the polynomial forms condition (Definition 3.6), we obtain the claim (taking η_1 suitably small).

In view of (36), (37) and Hölder's inequality, we see that

$$|\Lambda_{\mathcal{S}}(\vec{f})| \ll_{\mathcal{S}} \left(\mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{m\in[M]}\int_X \|f_{\alpha_0,m,\vec{h},\cdot,W}\|_{\square^{A_1}(\nu)}^{2^{|A_1|}}\right)^{1/2^{|A_1|}} + o(1).$$

Thus, to prove Proposition 5.9, it suffices to show that

$$\mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{m\in[M]}\int_X \|f_{\alpha_0,m,\vec{h},\cdot,W}\|_{\square^{A_1}(\nu)}^{2^{|A_1|}} = \|f_{\alpha_0}\|_{U^{\vec{Q}([H]^D,W)}_{\sqrt{M}}}^{2^{|A_1|}} + o_{\mathcal{S}}(1) \tag{38}$$

for some suitable $\vec{Q}\in\mathbf{Z}[\mathbf{h}_1,...,\mathbf{h}_D,\mathbf{W}]$. The left-hand side of (38) can be expanded as a weighted average of f_{α_0} over parallelopipeds, or more precisely as

$$\mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^{A_1}}\int_X\Big(\prod_{\omega\in\{0,1\}^{A_1}} T^{Q_0(\vec{h},\vec{m}^{(\omega)},W)}f_{\alpha_0}\Big)w(\vec{h},\vec{m}^{(0)},\vec{m}^{(1)}), \tag{39}$$

where $\vec{m}^{(\omega)}:=(m_\alpha^{(\omega_\alpha)})_{\alpha\in A_1}$ and $w(\vec{h},\vec{m}^{(0)},\vec{m}^{(1)})$ is the weight

$$w(\vec{h},\vec{m}^{(0)},\vec{m}^{(1)}) := \mathbf{E}_{m\in[M]}\prod_{\omega\in\{0,1\}^{A_1}}\prod_{\alpha\in A_{\mathrm{nl}}} T^{Q_0(\vec{h},\vec{m}^{(\omega)},W)+R_\alpha(m-\sum_{\alpha\in A_1}m_\alpha,\vec{h},W)}\nu.$$

5.18. The final Cauchy–Schwarz inequality

Let us temporarily drop the weight w in (39) and consider the unweighted average

$$\mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^{A_1}}\int_X\prod_{\omega\in\{0,1\}^{A_1}} T^{Q_0(\vec{h},\vec{m}^{(\omega)},W)}f_{\alpha_0}.$$

Using (35), this is

$$\mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^{A_1}}\int_X\prod_{\omega\in\{0,1\}^{A_1}} T^{\sum_{\alpha\in A_1}b_\alpha(\vec{\mathbf{h}},\mathbf{W})m_\alpha^{(\omega_\alpha)}}f_{\alpha_0},$$

which, on comparison with (24), is indeed of the form $\|f_{\alpha_0}\|_{U^{\vec{Q}([H]^D,W)}_{\sqrt{M}}}^{2^{|A_1|}}$ for some[27] $\vec{Q}\in\mathbf{Z}[\mathbf{h}_1,...,\mathbf{h}_D,\mathbf{W}]$; note that, since the b_α's are non-zero, all the components of $\vec{Q}$ are non-zero. Thus it suffices to show that

$$\mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^{A_1}}\int_X\Big(\prod_{\omega\in\{0,1\}^{A_1}} T^{Q_0(\vec{h},\vec{m}^{(\omega)},W)}f_{\alpha_0}\Big)(w(\vec{h},\vec{m}^{(0)},\vec{m}^{(1)})-1) = o_{\mathcal{S}}(1).$$

([27]) There is the (incredibly unlikely) possibility that $D=0$ or $D=1$, but by using the monotonicity of the Gowers norms (Lemma A.3), one can easily increase D to avoid this.

Applying the Cauchy–Schwarz inequality (94) with the pointwise bound $|f_{\alpha_0}| \leqslant \nu$, we reduce to showing that

$$\mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^{A_1}}\int_X\Big(\prod_{\omega\in\{0,1\}^{A_1}}T^{Q_0(\vec{h},\vec{m}^{(\omega)},W)}\nu\Big)(w(\vec{h},\vec{m}^{(0)},\vec{m}^{(1)})-1)^j = 0^j+o_{\mathcal{S}}(1)$$

for $j=0,2$ (with the usual convention $0^0=1$) which in turn follows if we show that

$$\mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^{A_1}}\int_X\Big(\prod_{\omega\in\{0,1\}^{A_1}}T^{Q_0(\vec{h},\vec{m}^{(\omega)},W)}\nu\Big)w(\vec{h},\vec{m}^{(0)},\vec{m}^{(1)})^j = 1+o_{\mathcal{S}}(1)$$

for $j=0,1,2$. Let us just demonstrate this in the hardest case $j=2$, as it will be clear from the proof that the same argument also works for $j=0,1$ (as they involve fewer factors of ν). We expand the left-hand side as

$$\begin{aligned}\mathbf{E}_{\vec{h}\in[H]^D}\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^{A_1}}\mathbf{E}_{m,m'\in[M]}\int_X\Big(\prod_{\omega\in\{0,1\}^{A_1}}T^{Q_0(\vec{h},\vec{m}^{(\omega)},W)}\nu\Big)\\ \times\prod_{\omega\in\{0,1\}^{A_1}}\prod_{\alpha\in A_{\mathrm{nl}}}T^{Q_0(\vec{h},\vec{m}^{(\omega)},W)+R_\alpha(m-\sum_{\alpha\in A_1}m_\alpha,\vec{h},W)}\nu\\ \times T^{Q_0(\vec{h},\vec{m}^{(\omega)},W)+R_\alpha(m'-\sum_{\alpha\in A_1}m_\alpha,\vec{h},W)}\nu.\end{aligned}$$

One can then invoke the polynomial forms condition (Definition 3.6) one last time (again taking η_1 small enough) to verify that this is indeed $1+o_{\mathcal{S}}(1)$. Note that as every node in A_{nl} is non-linear, the polynomials R_α have degree at least 2, which ensures that the polynomials used to shift ν here are all distinct. This concludes the proof of Proposition 5.9 in the linear case, and hence in general, and Theorem 4.5 follows.

Remark 5.19. One can define polynomial systems and weights (Definitions 5.1 and 5.12) for systems of multivariable polynomials $R_\alpha\in\mathbf{Z}[\mathbf{m}_1,...,\mathbf{m}_r,\mathbf{h}_1,...,\mathbf{h}_D,W]$ (see for example [23]). Following the steps of the PET induction (§5.10) and parallelopipedization (§5.16), one can prove a multivariable version of the polynomial generalized von Neumann theorem (Theorem 4.5).

6. Polynomial dual functions

This section and the next will be devoted to the proof of the structure theorem, Theorem 4.7. In these sections we shall assume the notation of §2, and fix the bounded quantities $t\geqslant 0$, $d\geqslant 2$ and $\vec{Q}\in\mathbf{Z}[\mathbf{h}_1,...,\mathbf{h}_t,\mathbf{W}]^d$. As they are bounded, we may permit all

implicit constants in the $o(\cdot)$ and $O(\cdot)$ notation to depend on these quantities. We also fix the pseudorandom measure ν. We shall abbreviate

$$\|f\|_U := \|f\|_{U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}}.$$

Roughly speaking, the objective here is to split any non-negative function bounded pointwise by ν into a non-negative function bounded pointwise by 1, plus an error which is small in the $\|\cdot\|_U$ norm. For technical reasons (as in [18]), we will also need to exclude a small exceptional set of measure $o(1)$, of which more will be said later.

Following [18], our primary tool for understanding the U norm shall be via the concept of a *dual function* of a function f associated with this norm.

Definition 6.1. (Dual function) If $f: X \to \mathbf{R}$ is a function, we define the *dual function* $\mathcal{D}f: X \to \mathbf{R}$ by the formula

$$\mathcal{D}f := \mathbf{E}_{\vec{h}\in[H]^t} \mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^d} \prod_{(\omega_1,\ldots,\omega_d)\in\{0,1\}^d\setminus\{0\}^d} T^{\sum_{j\in[d]}(m_j^{(\omega_j)}-m_j^{(0)})Q_j(\vec{h},W)} f,$$

where $m^{(k)} = (m_1^{(k)}, \ldots, m_d^{(k)})$ for $k = 0, 1$.

From (24) and the translation invariance of the integral $\int_X$, we obtain the fundamental relationship

$$\|f\|_U^{2^d} = \int_X f \mathcal{D}f. \tag{40}$$

Thus we have a basic dichotomy: either f has small U norm, or else it correlates with its own dual function.([28]) As in [18], it is the iteration of this dichotomy via a stopping time argument which shall power the proof of Theorem 4.7.

For future reference, we observe the trivial but useful facts that $\mathcal{D}$ is monotone and homogeneous of degree $2^d - 1$:

$$|f| \leqslant g \text{ pointwise} \quad \Longrightarrow \quad |\mathcal{D}f| \leqslant \mathcal{D}g \text{ pointwise}; \tag{41}$$

$$\mathcal{D}(\lambda f) = \lambda^{2^d-1} \mathcal{D}f \quad \text{for all } \lambda \in \mathbf{R}. \tag{42}$$

We will need two key facts about dual functions, both of which follow primarily from the polynomial correlation condition. The first, which is fairly easy, is that dual functions are essentially bounded.

([28]) In the language of infinitary ergodic theory, it will be the dual functions which generate (in the measure-theoretic sense) the characteristic factor for the U norm. The key points will be that the dual functions are essentially bounded, and that $\nu - 1$ is essentially orthogonal to the characteristic factor.

PROPOSITION 6.2. (Essential boundedness of dual functions) *Let $f: X \to \mathbf{R}$ obey the pointwise bound $|f| \leqslant \nu+1$. Then for any integer $K \geqslant 1$ we have the moment estimates*

$$\int_X |\mathcal{D}f|^K(\nu+1) \leqslant 2(2^{2^d-1})^K + o_K(1). \tag{43}$$

In particular, if we define the global bad set

$$\Omega_0 := \{x \in X : \mathcal{D}\nu(x) \geqslant 2^{2^d}\}, \tag{44}$$

then we have the measure bound

$$\int_X (\mathcal{D}\nu)^K 1_{\Omega_0}(\nu+1) = o_K(1) \tag{45}$$

for all $K \geqslant 0$, and the pointwise bound

$$|\mathcal{D}f|(1-1_{\Omega_0}) \leqslant 2^{2^d}. \tag{46}$$

Remark 6.3. In [18], the correlation conditions imposed on ν were strong enough so that one could bound the dual function $\mathcal{D}f$ uniformly by $2^{2^d-1}+o(1)$, thus removing the need for a global bad set Ω_0. One could do something similar here by strengthening the correlation condition. However, we were then unable to establish Theorem 3.18, i.e. we were unable to construct a measure concentrated on almost primes which obeyed this stronger correlation condition. The basic difficulty is that the polynomials in $\vec{Q}$ could contain a number of common factors which could significantly distort functions such as $\mathcal{D}\nu$ at some rare points (such as the origin). Fortunately, the presence of a small global bad set does not significantly impact our analysis (similarly to how sets of measure zero have no impact on ergodic theory), especially given that it does not depend on f. In practice, K will get as large as $1/\eta_6$, but no greater.

Proof. We begin with (43). By (41) and (42), it suffices to show that

$$\int_X \left(\mathcal{D}\frac{\nu+1}{2}\right)^K \frac{\nu+1}{2} = 1 + o_K(1);$$

in view of Lemma 3.14, it suffices to show that

$$\int_X (\mathcal{D}\nu)^K \nu = 1 + o_K(1).$$

The left-hand side can be expanded as

$$\mathbf{E}_{\vec{h}^{(1)},\dots,\vec{h}^{(K)} \in [H]^t} \int_X \prod_{k\in[K]} \left(\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)} \in [\sqrt{M}]^d} \prod_{(\omega_1,\dots,\omega_d)\in\{0,1\}^d\setminus\{0\}^d} T^{\sum_{j\in[d]}(m_j^{(\omega_j)} - m_j^{(0)})Q_j(\vec{h}^{(k)},W)}\nu \right)\nu.$$

But this is $1+o_K(1)$ from (18) (with $D=d$, $D'=0$, $D''=Kt$, $L=1$, and the $\vec{Q}_{j,k}$ and $\vec{S}_l$ vanishing). This proves (43). From Chebyshev's inequality, this implies that

$$\int_X (\mathcal{D}\nu)^{K'} 1_{\Omega_0}(\nu+1) \leqslant \frac{2(2^{2^d-1})^{K'}}{2^K} + o_K(1)$$

for any $0 \leqslant K' < K$. For fixed K', the right-hand side can be made arbitrarily small by taking K large, and then choosing N large depending on K; thus, the left-hand side is $o(1)$, which is (45). Finally, (46) follows from (41) and (44). □

The global bad set Ω_0 is somewhat annoying to deal with. Let us remove it by defining the *modified dual function* $\widetilde{\mathcal{D}}f$ of f as

$$\widetilde{\mathcal{D}}f := (1-1_{\Omega_0})\mathcal{D}f.$$

Then Proposition 6.2 and (40) immediately imply the following result.

COROLLARY 6.4. (Boundedness of modified dual function) *Let $f: X \to \mathbf{R}$ obey the pointwise bound $|f| \leqslant \nu+1$. Then $\widetilde{\mathcal{D}}f$ takes values in the interval*

$$I := [-2^{2^d}, 2^{2^d}]. \tag{47}$$

Furthermore, we have the correlation property

$$\int_X f\widetilde{\mathcal{D}}f = \|f\|_U^{2^d} + o(1). \tag{48}$$

The second important estimate is easy to state, although non-trivial to prove.

PROPOSITION 6.5. ($\nu-1$ orthogonal to products of modified dual functions) *Let $1 \leqslant K \leqslant 1/\eta_6$ be an integer, and let $f_1, ..., f_K: X \to \mathbf{R}$ be functions with the pointwise bounds $|f_k| \leqslant \nu+1$ for all $k \in [K]$. Then*

$$\int_X \widetilde{\mathcal{D}}f_1 \ldots \widetilde{\mathcal{D}}f_K(\nu-1) = o(1). \tag{49}$$

Remark 6.6. Note that (43) already gives an upper bound of $O(1)$ for (49); the whole point is thus to extract enough cancellation from the factor $\nu-1$ to upgrade this bound to $o(1)$.

The rest of the section is devoted to the proof of Proposition 6.5. The argument follows that of [18, §6], and is based on a large number of applications of the Cauchy–Schwarz inequality, and the polynomial correlation condition (Definition 3.9). The arguments here are not used again elsewhere in this paper, and so the rest of this section may be read independently of the remainder of the paper.

We begin with a very simple reduction: from (45) we can replace the modified dual functions $\widetilde{\mathcal{D}}f_k$ by their unmodified counterparts $\mathcal{D}f_k$. Our task is then to show that

$$\int_X \mathcal{D}f_1 \ldots \mathcal{D}f_K(\nu-1) = o(1). \tag{50}$$

6.7. A model example

Before we prove Proposition 6.5 in general, it is instructive to work with a simple example to illustrate the idea. Let us take an oversimplified toy model of the dual function $\mathcal{D}f$, namely

$$\mathcal{D}f := \mathbf{E}_{h\in[H]}\mathbf{E}_{m\in[\sqrt{M}]}T^{mh}f.$$

This does not quite correspond to a local Gowers norm,([29]) but will serve as an illustrative model nonetheless. Pick functions $f_1, ..., f_K$ with the pointwise bounds $|f_k|\leqslant\nu$ for $k\in[K]$ and consider the task of showing (50). We expand the left-hand side as

$$\mathbf{E}_{h_1,...,h_K\in[H]}\mathbf{E}_{m_1,...,m_K\in[\sqrt{M}]}\int_X T^{m_1h_1}f_1 \,...\, T^{m_Kh_K}f_K(\nu-1). \tag{51}$$

Note that we cannot simply take absolute values and apply the pseudorandomness conditions, as these will give bounds of the form $O(1)$ rather than $o(1)$. One could instead attempt to apply the Cauchy–Schwarz inequality many times (as in the previous section), however the fact that $K=O(1/\eta_6^2)$ could be very large compared to the pseudorandomness parameter $1/\eta_1$ defeats a naive implementation of this idea. Instead, we must perform a change of variables to introduce two new parameters $n^{(0)}$ and $n^{(1)}$ to average over (which only requires a single Cauchy–Schwarz inequality to estimate) rather than K parameters (which would essentially require K applications of the Cauchy–Schwarz inequality).

More precisely, we introduce two slightly less coarse-scale parameters $n^{(0)}, n^{(1)}\in[M^{1/4}]$ than $m_1, ..., m_K$. Define the multipliers $\hat{h}_k := \prod_{k'\in[K]\setminus k} h_{k'}$, thus $\hat{h}_k = O(H^{K-1})$, which are small compared to $M^{1/4}$ by the relative sizes of η_7, η_6 and η_2. Shifting each of the m_k's by $\hat{h}_k(n^{(1)}-n^{(0)})$ and using (16), we conclude that (51) is equal to

$$\mathbf{E}_{h_1,...,h_K\in[H]}\mathbf{E}_{m_1,...,m_K\in[\sqrt{M}]}\int_X\Big(\prod_{k\in[K]}T^{(m_k+\hat{h}_k(n^{(1)}-n^{(0)}))h_k}f_k\Big)(\nu-1)+o_K(1)$$

for all $n^{(0)}, n^{(1)}\in[M^{1/4}]$. Averaging over all $n^{(0)}$ and $n^{(1)}$, and shifting the integral by

$$n^{(0)}h_1 \,...\, h_K = \hat{h}_1 n^{(0)} h_1 = ... = \hat{h}_K n^{(0)} h_K,$$

we can thus write (51) as

$$\mathbf{E}_{h_1,...,h_K\in[H]}\mathbf{E}_{m_1,...,m_K\in[\sqrt{M}]}\mathbf{E}_{n^{(0)},n^{(1)}\in[M^{1/4}]}$$
$$\int_X T^{n^{(1)}h_1...h_K}(T^{m_1h_1}f_1 \,...\, T^{m_Kh_K}f_K)T^{n^{(0)}h_1...h_K}(\nu-1)+o_K(1),$$

([29]) However, the slight variant $\mathbf{E}_{h\in[H]}\mathbf{E}_{m,m'\in[\sqrt{M}]}T^{(m-m')h}f$ does correspond to a (very simple) local Gowers norm, with $t=d=1$ and $\vec{Q}=(\mathbf{h}_1)$.

which we may factorize as

$$\mathbf{E}_{h_1,\dots,h_K\in[H]}\int_X\Big(\mathbf{E}_{n^{(1)}\in[M^{1/4}]}\prod_{k\in[K]}\mathbf{E}_{m\in[\sqrt{M}]}T^{n^{(1)}h_1\dots h_k+mh_k}f_k\Big)\\ \times\mathbf{E}_{n^{(0)}\in[M^{1/4}]}T^{n^{(0)}h_1\dots h_K}(\nu-1)+o_K(1).$$

By the Cauchy–Schwarz inequality, it thus suffices to show that

$$\mathbf{E}_{h_1,\dots,h_K\in[H]}\int_X\Big(\mathbf{E}_{n^{(1)}\in[M^{1/4}]}\prod_{k\in[K]}\mathbf{E}_{m\in[\sqrt{M}]}T^{n^{(1)}h_1\dots h_k+mh_k}f_k\Big)^2\ll_K 1$$

and

$$\mathbf{E}_{h_1,\dots,h_K\in[H]}\int_X(\mathbf{E}_{n^{(0)}\in[M^{1/4}]}T^{n^{(0)}h_1\dots h_K}(\nu-1))^2=o_K(1).$$

To prove the first estimate, we estimate f_k by ν and expand the square to reduce to showing that

$$\mathbf{E}_{h_1,\dots,h_K\in[H]}\mathbf{E}_{n^{(1)},n^{(2)}\in[M^{1/4}]}\int_X\prod_{j=1}^{2}\prod_{k\in[K]}\mathbf{E}_{m\in[\sqrt{M}]}T^{n^{(j)}h_1\dots h_k+mh_k}\nu\ll_K 1,$$

but this follows from the correlation condition (18) (for η_1 small enough[30]). To prove the second estimate, we again expand the square and reduce to showing that

$$\mathbf{E}_{h_1,\dots,h_K\in[H]}\int_X(\mathbf{E}_{n^{(0)}\in[M^{1/4}]}T^{n^{(0)}h_1\dots h_K}\nu)^j=1+o_K(1)$$

for $j=0,1,2$, which will again follow from (18) for η_1 small enough.

6.8. Conclusion of the argument

Now we prove (50) in the general case. We may take $d\geqslant 1$, since the $d=0$ case follows from (15). We expand the left-hand side as

$$\mathbf{E}_{\vec{h}}\mathbf{E}_{\vec{m}}\int_X\Big(\prod_{k\in[K]}\prod_{\omega\in\{0,1\}^d\setminus\{0\}^d}T^{\sum_{j\in[d]}(m_{j,k}^{(\omega_j)}-m_{j,k}^{(0)})Q_j(\vec{h}^{(k)},W)}f_k\Big)(\nu-1),$$

where $\omega=(\omega_1,\dots,\omega_d)$ and we use the abbreviations

$$\mathbf{E}_{\vec{h}}:=\mathbf{E}_{\vec{h}^{(1)},\dots,\vec{h}^{(K)}\in[H]^t}\quad\text{and}\quad\mathbf{E}_{\vec{m}}:=\mathbf{E}_{m_{j,k}^{(\omega)}\in[\sqrt{M}],\omega\in\{0,1\},j\in[d],k\in[K]}.$$

---

[30] It is important to note however that η_1 does not have to be small relative to K or to parameters such as η_7.

We introduce moderately coarse-scale parameters $n_j^{(0)}, n_j^{(1)} \in [M^{1/4}]$ for $j \in [d]$, and the multipliers

$$\hat{h}_{k,j} = \hat{h}_{k,j}(\vec{h}, W) := \prod_{k' \in [K] \backslash \{k\}} Q_j(\vec{h}^{(k')}, W).$$

Observe that $\hat{h}_{k,j} = O(H^{O(K)})$, which will be much smaller than $M^{1/4}$ by the relative sizes of η_7, η_6 and η_2. Shifting each $m_{j,k}^{(1)}$ by $\hat{h}_{k,j}(n_j^{(1)} - n_j^{(0)})$ and using (16), we can then rewrite (50) as

$$\mathbf{E}_{\vec{h}} \mathbf{E}_{\vec{m}} \int_X \Bigg(\prod_{k \in [K]} \prod_{\omega \in \{0,1\}^d \backslash \{0\}^d} T^{\sum_{j \in [d]} (m_{j,k}^{(\omega_j)} - m_{j,k}^{(0)} + \hat{h}_{k,j}(n_j^{(\omega_j)} - n_j^{(0)})) Q_j(\vec{h}^{(k)}, W)} f_k \Bigg)(\nu - 1) + o_K(1)$$

for any $n_1^{(0)}, \ldots, n_d^{(0)}, n_1^{(1)}, \ldots, n_d^{(1)} \in [M^{1/4}]$. Now, from construction, we have

$$\hat{h}_{k,j} Q_j(\vec{h}^{(k)}, W) = b_j,$$

where

$$b_j = b_j(\vec{h}, W) := \prod_{k \in [K]} Q_j(\vec{h}^{(k)}, W)$$

(note that $b_j \neq 0$ by hypothesis on $\vec{Q}$), and after averaging over the n variables, we can write the left-hand side of (50) as

$$\mathbf{E}_{\vec{h}, \vec{m}, \vec{n}^{(0)}, \vec{n}_1^{(1)}} \int_X \prod_{\omega \in \{0,1\}^d} T^{\sum_{j \in [d]} (n_j^{(\omega_j)} - n_j^{(0)}) b_j} g_{\omega, \vec{h}, \vec{m}} + o_K(1),$$

where $\vec{n}^{(j)} = (n_0^{(j)}, \ldots, n_d^{(j)})$ will be understood to range over $[M^{1/4}]^d$ for $j = 0, 1$,

$$g_{\omega, \vec{h}, \vec{m}} := \prod_{k \in [K]} T^{\sum_{j \in [d]} (m_{j,k}^{(\omega_j)} - m_{j,k}^{(0)}) Q_j(\vec{h}^{(k)}, W)} f_k$$

for $\omega \in \{0,1\}^d \backslash \{0\}^d$, and

$$g_{\{0\}^d, \vec{h}, \vec{m}} := \nu - 1.$$

Shifting the integral by $T^{\sum_{j \in [d]} n_j^{(0)} b_j}$, we can rewrite this as

$$\mathbf{E}_{\vec{h}, \vec{m}} \int_X \mathbf{E}_{\vec{n}^{(0)}, \vec{n}^{(1)}} \prod_{\omega \in \{0,1\}^d} T^{\sum_{j \in [d]} n_j^{(\omega_j)} b_j} g_{\omega, \vec{h}, \vec{m}} + o_K(1).$$

Now use the Cauchy–Schwarz–Gowers inequality (99) to obtain the pointwise estimate

$$\Bigg| \mathbf{E}_{\vec{n}^{(0)}, \vec{n}^{(1)}} \prod_{\omega \in \{0,1\}^d} T^{\sum_{j \in [d]} n_j^{(\omega_j)} b_j} g_{\omega, \vec{h}, \vec{m}} \Bigg|$$
$$\leqslant \prod_{\omega' \in \{0,1\}^d} \Bigg(\mathbf{E}_{\vec{n}^{(0)}, \vec{n}^{(1)}} \prod_{\omega \in \{0,1\}^d} T^{\sum_{j \in [d]} n_j^{(\omega_j)} b_j} g_{\omega', \vec{h}, \vec{m}} \Bigg)^{1/2^d}.$$

By Hölder's inequality, we thus see that to prove (50) it suffices to show that the quantity

$$\mathbf{E}_{\vec{h},\vec{m}} \int_X \mathbf{E}_{\vec{n}^{(0)},\vec{n}^{(1)}} \prod_{\omega\in\{0,1\}^d} T^{\sum_{j\in[d]} n_j^{(\omega_j)} b_j} g_{\omega',\vec{h},\vec{m}} \tag{52}$$

is $O_K(1)$ when $\omega'\in\{0,1\}^d\backslash\{0\}^d$ and is $o_K(1)$ when $\omega'=0^d$.

Let us first deal with the case when $\omega'\neq 0^d$. Our task is to show that

$$\mathbf{E}_{\vec{h},\vec{m}} \int_X \mathbf{E}_{\vec{n}^{(0)},\vec{n}^{(1)}} \prod_{\omega\in\{0,1\}^d} T^{\sum_{j\in[d]} n_j^{(\omega_j)} b_j} \prod_{k\in[K]} T^{\sum_{j\in[d]} (m_{j,k}^{(\omega'_j)} - m_{j,k}^{(0)}) Q_j(\vec{h}^{(k)},W)} f_k = O_K(1).$$

We can bound f_k pointwise by ν, and factorize the left-hand side as

$$\mathbf{E}_{\vec{h},\vec{n}^{(0)},\vec{n}^{(1)}} \int_X \prod_{k\in[K]} \mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in[\sqrt{M}]^d} \prod_{\omega\in\{0,1\}^d} T^{\sum_{j\in[d]} (m_j^{(\omega'_j)} - m_j^{(0)}) Q_j(\vec{h}^{(k)},W) + n_j^{(\omega_j)} b_j} \nu.$$

But this is $1+o_K(1)=O_K(1)$ by (3.9) with $L=0$ (here we use the fact that the b_j's are non-zero polynomials of $\vec{h}$ and W). For this we need η_1 to be sufficiently small depending on t, d and $\vec{Q}$, but not on K.

Finally, we have to deal with the case $\omega=0^d$. Since $g_{\omega',\vec{h},\vec{m}}=\nu-1$ and $b_j=b_j(\vec{h},W)$ are independent of W, we can rewrite (52) as

$$\mathbf{E}_{\vec{h},\vec{n}^{(0)},\vec{n}^{(1)}} \int_X \prod_{\omega\in\{0,1\}^d} T^{\sum_{j\in[d]} n_j^{(\omega_j)} b_j(\vec{h},W)} (\nu-1),$$

and so, by the binomial formula, it suffices to show that

$$\mathbf{E}_{\vec{h},\vec{n}^{(0)},\vec{n}^{(1)}} \int_X \prod_{\omega\in A} T^{\sum_{j\in[d]} n_j^{(\omega_j)} b_j(\vec{h},W)} \nu = 1+o(1)$$

for all $A\subset\{0,1\}^d$. But this follows from the polynomial correlation condition (18) (with $K=0$), again taking η_1 sufficiently small depending on t, d and $\vec{Q}$. This concludes the proof of (50), and hence of Proposition 6.5.

7. Proof of the structure theorem

We can now complete the proof of the structure theorem by using the arguments of [18, §7 and §8] more or less verbatim. In fact these arguments can be abstracted as follows.

THEOREM 7.1. (Abstract structure theorem) *Let I be an interval bounded by $O(1)$. Let $\nu: X \to \mathbf{R}^+$ be any measure, and let $f \mapsto \widetilde{\mathcal{D}} f$ be a (non-linear) operator obeying the following properties*:

- *if the pointwise bound $|f| \leqslant \nu + 1$ holds, then $\widetilde{\mathcal{D}} f: X \to I$ takes values in I, in particular*

$$\widetilde{\mathcal{D}} f = O(1); \tag{53}$$

- *if $1 \leqslant K \leqslant 1/\eta_6$ and $f_1, ..., f_K: X \to \mathbf{R}$ are functions satisfying the pointwise bound $|f_k| \leqslant \nu + 1$, $k \in [K]$, then (49) holds.*

Then, for any $g: X \to \mathbf{R}^+$ with the pointwise bound $0 \leqslant g \leqslant \nu$, there exist functions $g_{U^\perp}, g_U: X \to \mathbf{R}$ obeying estimates (26), (27) and (28), such that

$$\left| \int_X g_U \widetilde{\mathcal{D}} g_U \right| \leqslant \eta_4. \tag{54}$$

Indeed, Theorem 4.7 immediately follows by applying Theorem 7.1, with (29) following from (54) and (48).

In the remainder of this section we prove Theorem 7.1. Henceforth we fix I, ν and $\widetilde{\mathcal{D}}$ obeying the hypotheses of the theorem.

7.2. Factors

As in [18], we shall recall the very useful notion of *factor* from ergodic theory, though for our applications we actually only need the finitary version of this concept.

Let us set $\mathbf{X}$ to be the probability space $\mathbf{X} = (X, \mathcal{B}_X, \mu_X)$, where $X = \mathbf{Z}_N$, $\mathcal{B}_X = 2^X$ is the power set of X and μ_X is the uniform probability measure on $\mathbf{X}$. We define a *factor*([31]) to be a quadruple $\mathbf{Y} = (Y, \mathcal{B}_Y, \mu_Y, \pi_Y)$, where $(Y, \mathcal{B}_Y, \mu_Y)$ is a probability space (and thus $\mathcal{B}_Y$ is a σ-algebra on Y and μ_Y is a probability measure on $\mathcal{B}_Y$) together with a measurable map $\pi: X \to Y$ such that $(\pi_Y)_* \mu_X = \mu_Y$, or in other words $\mu_X(\pi_Y^{-1}(E)) = \mu_Y(E)$ for all $E \in \mathcal{B}_Y$. The factor map π_Y induces the pullback map $\pi_Y^*: L^2(\mathbf{Y}) \to L^2(\mathbf{X})$ and its adjoint $(\pi_Y)_*: L^2(\mathbf{X}) \to L^2(\mathbf{Y})$, where $L^2(\mathbf{X})$ is the usual Lebesgue space of square-integrable functions on $\mathbf{X}$. We refer to the projection $\pi_Y^*(\pi_Y)_*: L^2(\mathbf{X}) \to L^2(\mathbf{X})$ as the *conditional expectation operator*, and denote $\pi_Y^*(\pi_Y)_*(f)$ by $\mathbf{E}(f|\mathbf{Y})$; this is a linear self-adjoint orthogonal projection from $L^2(\mathbf{X})$ to $\pi_Y^* L^2(\mathbf{Y})$.

([31]) In infinitary ergodic theory one also requires the probability spaces $\mathbf{X}$ and $\mathbf{Y}$ to be invariant under the shift T, and for the factor map π to respect the shift. In the finitary setting it is unrealistic to demand these shift-invariances, for if N were prime then this would mean that there were no non-trivial factors whatsoever. While there are concepts of "approximate shift-invariance" which can be used as a substitute, see [33], we will fortunately not need to use them here, as the remainder of the argument does not even involve the shift T at all.

The conditional expectation operator is in fact completely determined by the σ-algebra $\pi_Y^{-1}(\mathcal{B}_Y) \subset \mathcal{B}_X$. Since X is finite (with every point having positive measure), $\pi_Y^{-1}(\mathcal{B}_Y)$ is generated by a partition of X into atoms (which by abuse of notation we refer to as atoms of the factor $\mathbf{Y}$), and the conditional expectation is given explicitly by the formula

$$\mathbf{E}(f|\mathbf{Y})(x) = \mathbf{E}_{y \in \mathcal{B}(x)} f(y),$$

where $\mathcal{B}(x)$ is the unique atom of $\pi^{-1}(\mathcal{B}_Y)$ which contains x. We refer to the number of atoms of $\mathbf{Y}$ as the *complexity* of the factor $\mathbf{Y}$.([32]) By abuse of notation, we say that a function $f: X \to \mathbf{R}$ is *measurable with respect to* $\mathbf{Y}$ if it is measurable with respect to $\pi_Y^{-1}(\mathcal{B}_Y)$, or equivalently if it is constant on all atoms of $\mathbf{Y}$. Thus for instance $(\pi_Y)^* L^q(\mathbf{Y})$ consists of the functions in $L^q(\mathbf{X})$ which are measurable with respect to $\mathbf{Y}$.

If $\mathbf{Y} = (Y, \mathcal{B}_Y, \mu_Y, \pi_Y)$ and $\mathbf{Y}' = (Y', \mathcal{B}_{Y'}, \mu_{Y'}, \pi_{Y'})$ are two factors, we may form their join $\mathbf{Y} \vee \mathbf{Y}' = (Y \times Y', \mathcal{B}_Y \times \mathcal{B}_{Y'}, \mu_Y \times \mu_{Y'}, \pi_Y \oplus \pi_{Y'})$ in the obvious manner; note that the atoms of $\mathbf{Y} \vee \mathbf{Y}'$ are simply the non-empty intersections of atoms of $\mathbf{Y}$ with atoms of $\mathbf{Y}'$, and so any function which is measurable with respect to $\mathbf{Y}$ or $\mathbf{Y}'$ is automatically measurable with respect to $\mathbf{Y} \vee \mathbf{Y}'$.

Note that any function $f: \mathbf{X} \to \mathbf{R}$ automatically generates a factor $(\mathbf{R}, \mathcal{B}_{\mathbf{R}}, f_* \mu_X, f)$, where $\mathcal{B}_{\mathbf{R}}$ is the Borel σ-algebra, which is the minimal factor with respect to which f is (Borel-)measurable. In our finitary setting it turns out that we need a discretized version of this construction, which we give as follows.

PROPOSITION 7.3. (Each function generates a factor) *For any function $G: X \to I$ there exists a factor $\mathbf{Y}(G)$ with the following properties*:

- (G lies in its own factor) *for any factor $\mathbf{Y}'$,*

$$G = \mathbf{E}(G|\mathbf{Y}(G) \vee \mathbf{Y}') + O(\eta_4^2); \tag{55}$$

- (bounded complexity) $\mathbf{Y}(G)$ *has at most $O_{\eta_4}(1)$ atoms*;
- (approximation by continuous functions of G) *if A is any atom in $\mathbf{Y}(G)$, then there exists a polynomial $\Psi_A: \mathbf{R} \to \mathbf{R}$ of degree $O_{\eta_5}(1)$ with coefficients $O_{\eta_5}(1)$ such that*

$$\Psi_A(x) \in [0,1] \quad \text{for all } x \in I \tag{56}$$

and

$$\int_X |1_A - \Psi_A(G)|(\nu + 1) \ll \eta_5. \tag{57}$$

([32]) It would be more natural to work instead with the *entropy* of $\mathbf{Y}$ rather than the complexity, but the entropy is a slightly more technical concept and so we have avoided its use here for simplicity.

Proof. This is essentially [18, Proposition 7.2], but we shall give a complete proof here for the convenience of the reader.

We use the probabilistic method. Let α be a real number in the interval $[0,1]$, chosen at random. We then define the factor

$$\mathbf{Y}(G) := (\mathbf{R}, \mathcal{B}_{\eta_4^2,\alpha}, G_*\mu_X, G),$$

where $\mathcal{B}_{\eta_4^2,\alpha}$ is the σ-algebra on the real line $\mathbf{R}$ generated by the intervals

$$[(n+\alpha)\eta_4^2, (n+\alpha+1)\eta_4^2)$$

for $n \in \mathbf{Z}$. This is clearly a factor of $\mathbf{X}$, with atoms

$$A_{n,\alpha} := G^{-1}([(n+\alpha)\eta_4^2, (n+\alpha+1)\eta_4^2)).$$

Since G ranges in I, and we allow constants to depend on I, it is clear that there are at most $O_{\eta_4}(1)$ non-empty atoms and that G fluctuates by at most $O(\eta_4^2)$ on each atom, which yields the first two desired properties. It remains to verify that with positive probability, the approximation by continuous functions property holds for all atoms $A_{n,\alpha}$. By the union bound, it suffices to show that each individual atom $A_{n,\alpha}$ has the approximation property with probability $1-O(\eta_5)$.

By the Weierstrass approximation theorem, we can for each α find a polynomial $\Psi_{A_{n,\alpha}}$ obeying (56) which is equal to $1_{[(n+\alpha)\eta_4^2,(n+\alpha+1)\eta_4^2)}+O(\delta)$ outside of the set

$$E_{n,\alpha} := [(n+\alpha-\eta_5^2)\eta_4^2, (n+\alpha+\eta_5^2)\eta_4^2] \cup [(n+\alpha+1-\eta_5^2)\eta_4^2, (n+\alpha+1+\eta_5^2)\eta_4^2].$$

Simple compactness arguments allow us to take $\Psi_{A_{n,\alpha}}$ to have degree $O_{\eta_5}(1)$ and coefficients $O_{\eta_5}(1)$. Since

$$1_{A_{n,\alpha}} = 1_{[(n+\alpha)\eta_4^2,(n+\alpha+1)\eta_4^2)}(G),$$

we thus conclude (from (15)) that

$$\int_X |1_A - \Psi_{A_{n,\alpha}}(G)|(\nu+1) \ll \eta_5 + \int_X 1_{E_{n,\alpha}}(G)(\nu+1).$$

By Markov's inequality, it thus suffices to show that

$$\int_0^1 \left(\int_X 1_{E_{n,\alpha}}(G)(\nu+1) \right) d\alpha \ll \eta_5^2.$$

But this follows from Fubini's theorem, (15) and the elementary pointwise estimate

$$\int_0^1 1_{E_{n,\alpha}}(G)\, d\alpha \ll \eta_5^2.$$

□

Henceforth we set $\mathbf{Y}(G)$ to be the factor given by the above proposition. A key consequence of the hypotheses of Theorem 7.1 is that $\nu-1$ is well distributed with respect to any finite combination of these factors.

PROPOSITION 7.4. (ν uniformly distributed with respect to dual function factors) *Let $K\geqslant 1$ be an integer with $K=O_{\eta_4}(1)$, and let $f_1,...,f_K\colon X\to\mathbf{R}$ be functions with the pointwise bounds $|f_k|\leqslant\nu+1$ for all $k\in[K]$. Let $\mathbf{Y}:=\mathbf{Y}(\widetilde{\mathcal{D}}f_1)\vee...\vee\mathbf{Y}(\widetilde{\mathcal{D}}f_K)$. Then*

$$\widetilde{\mathcal{D}}f_k=\mathbf{E}(\widetilde{\mathcal{D}}f_k|\mathbf{Y})+O(\eta_4^2) \tag{58}$$

for all $k\in[K]$, there is a $\mathbf{Y}$-measurable set $\Omega\subset X$ obeying the smallness bound

$$\int_X 1_\Omega(\nu+1)\ll_{\eta_4}\eta_5^{1/2} \tag{59}$$

and we have the pointwise bound

$$|(1-1_\Omega)\mathbf{E}(\nu-1|\mathbf{Y})|\leqslant O_{\eta_4}(\eta_5^{1/2}). \tag{60}$$

Proof. We repeat the arguments from [18, Proposition 7.3]. The claim (58) follows immediately from (55), so we turn to the other two properties. Since each $\mathbf{Y}(\widetilde{\mathcal{D}}f_k)$ is generated by $O_{\eta_4}(1)$ atoms, $\mathbf{Y}$ is generated by $O_{\eta_4,K}(1)=O_{\eta_4}(1)$ atoms. Call an atom A of $\mathbf{Y}$ *small* if $\int_X 1_A(\nu+1)\leqslant\eta_5^{1/2}$, and let Ω be the union of all the small atoms, then Ω is clearly $\mathbf{Y}$-measurable and obeys (59). It remains to prove (60), or equivalently that

$$\frac{\int_X 1_A(\nu-1)}{\int_X 1_A}=\mathbf{E}_{y\in A}\nu(y)-1\ll_{\eta_4}\eta_5^{1/2}+o(1)$$

for all non-small atoms A.

Fix a non-small atom A. Since A is not small, we have

$$\int_X 1_A(\nu-1)+2\int_X 1_A=\int_X 1_A(\nu+1)>\eta_5^{1/2}.$$

Hence it will suffice to show that

$$\int_X 1_A(\nu-1)\ll_{\eta_4}\eta_5+o(1).$$

On the other hand, as A is the intersection of atoms $A_1,...,A_K$ from $\mathbf{Y}(\widetilde{\mathcal{D}}f_1),...,\mathbf{Y}(\widetilde{\mathcal{D}}f_K)$, we see from Proposition 7.3 and an easy induction argument that there exists a polynomial $\Psi\colon\mathbf{R}^K\to\mathbf{R}$ of degree $O_{\eta_5}(1)$ with coefficients $O_{\eta_5}(1)$ which maps I^K into $[0,1]$ such that

$$\int_X|1_A-\Psi(\widetilde{\mathcal{D}}f_1,...,\widetilde{\mathcal{D}}f_K)|(\nu+1)\ll_{\eta_4}\eta_5.$$

In particular,

$$\int_X (1_A - \Psi(\widetilde{\mathcal{D}} f_1, ..., \widetilde{\mathcal{D}} f_K))(\nu - 1) \ll_{\eta_4} \eta_5.$$

On the other hand, by decomposing Ψ into monomials and using (49) (assuming η_6 sufficiently small depending on η_5), we have

$$\int_X \Psi(\widetilde{\mathcal{D}} f_1, ..., \widetilde{\mathcal{D}} f_K)(\nu - 1) = o(1)$$

and the claim follows (we can absorb the $o(1)$ error by taking N large enough). □

7.5. The inductive step

The proof of the abstract structure theorem proceeds by a stopping time argument. To clarify this argument we introduce a somewhat artificial definition.

Definition 7.6. (Structured factor) A *structured factor* is a tuple

$$\mathbf{Y}_K = (\mathbf{Y}_K, K, F_1, ..., F_K, \Omega_K),$$

where $K \geqslant 0$ is an integer, $F_1, ..., F_K : X \to \mathbf{R}$ are functions with the pointwise bounds $|F_k| \leqslant \nu + 1$ for all $k \in [K]$, $\mathbf{Y}_K$ is the factor $\mathbf{Y}_K := \mathbf{Y}_K(F_1) \vee ... \vee \mathbf{Y}_K(F_K)$ and $\Omega_K \subset X$ is a $\mathbf{Y}_K$-measurable set. We refer to K as the *order* of the structured factor, and Ω_K as the *exceptional set*. We say that the structured factor has *noise level* σ for some $\sigma > 0$ if we have the smallness bound

$$\int_X 1_{\Omega_K}(\nu + 1) \leqslant \sigma$$

and the pointwise bound

$$|(1 - 1_{\Omega_K}) \mathbf{E}(\nu - 1 | \mathbf{Y}_K)| \leqslant \sigma. \tag{61}$$

If $g : X \to \mathbf{R}$ is the function in Theorem 7.1, we define the *energy* $\mathcal{E}_g(\mathbf{Y}_K)$ of the structured factor Y relative to g to be the quantity

$$\mathcal{E}_g(\mathbf{Y}_K) := \int_X (1 - 1_{\Omega_K}) \mathbf{E}(g | \mathbf{Y}_K)^2.$$

If $\mathbf{Y}_K$ has noise level $\sigma \leqslant 1$, then, since g is bounded in magnitude by ν,

$$|(1 - 1_{\Omega_K}) \mathbf{E}(g | \mathbf{Y}_K)| \leqslant (1 - 1_{\Omega_K})(\mathbf{E}(\nu - 1 | \mathbf{Y}_K) + 1) \leqslant 1 + \sigma \leqslant 2, \tag{62}$$

and so we conclude the energy bound

$$0 \leqslant \mathcal{E}_g(\mathbf{Y}_K) \leqslant 4. \tag{63}$$

This will allow us to apply an *energy increment argument* to obtain Theorem 7.1. More precisely, Theorem 7.1 is obtained from the following inductive step.

PROPOSITION 7.7. (Inductive step) *Let* $\mathbf{Y}_K=(\mathbf{Y}_K,K,F_1,...,F_K,\Omega_K)$ *be a structured factor of order* K *with noise level* $0<\sigma<\eta_4^4$. *If we set*

$$F_{K+1}:=\frac{1}{1+\sigma}(1-1_{\Omega_K})(g-\mathbf{E}(g|\mathbf{Y})) \tag{64}$$

and we suppose that

$$\left|\int_X F_{K+1}\widetilde{\mathcal{D}}F_{K+1}\right|>\eta_4, \tag{65}$$

then there exists a structured factor $\mathbf{Y}_{K+1}=(\mathbf{Y}_{K+1},K+1,F_1,...,F_K,F_{K+1},\Omega_{K+1})$ *of order* $K+1$ *with noise level* $\sigma+O_{\eta_4}(\eta_5^{1/2})$ *satisfying the energy increment property*

$$\mathcal{E}_g(\mathbf{Y}_{K+1})\geqslant\mathcal{E}_g(\mathbf{Y}_K)+c\eta_4^2 \tag{66}$$

for some constant $c>0$ (*depending only on* I).

Let us assume Proposition 7.7 for the moment and deduce Theorem 7.1. Starting with a trivial structured factor $\mathbf{Y}_0$ of order 0, and iterating Proposition 7.7 repeatedly (and using (63) to prevent the iteration for proceeding for more than $4/c\eta_4^2=O_{\eta_4}(1)$ steps), we may find a structured factor $\mathbf{Y}_K$ of order $K=O_{\eta_4}(1)$ with noise level

$$\sigma=O_{\eta_4}(\eta_5^{1/2})<\eta_4^4, \tag{67}$$

such that the function F_{K+1} defined in (64) obeys the bound

$$\left|\int_X F_{K+1}\widetilde{\mathcal{D}}F_{K+1}\right|\leqslant\eta_4.$$

If we thus set $g_U:=F_{K+1}$ and

$$g_{U^\perp}:=\frac{1}{1+\sigma}(1-1_{\Omega_K})\mathbf{E}(g|\mathbf{Y}),$$

then we easily verify (26) and (54), while (27) follows from (61), since

$$\mathbf{E}(g|\mathbf{Y})\leqslant 1+\mathbf{E}(\nu-1|\mathbf{Y}).$$

To prove (28), we see from (67) that it suffices to show that

$$\int_X(1-1_{\Omega_K})\mathbf{E}(g|\mathbf{Y})=\int_X g-O_{\eta_4}(\eta_5^{1/2}).$$

Since Ω_K is $\mathbf{Y}$-measurable, the left-hand side is $\int_X g-\int_X 1_{\Omega_K}g$. But the claim then follows from (61) and (67). This proves Theorem 7.1.

It remains to prove Proposition 7.7. Set

$$\mathbf{Y}_{K+1} := \mathbf{Y} \vee \mathbf{Y}(\widetilde{\mathcal{D}} F_{K+1}) = \mathbf{Y}(\widetilde{\mathcal{D}} F_1) \vee \ldots \vee \mathbf{Y}(\widetilde{\mathcal{D}} F_{K+1}).$$

Now, by Proposition 7.4, we can find a $\mathbf{Y}_{K+1}$-measurable set Ω obeying the smallness bound (59) and the pointwise bound

$$|(1-1_\Omega)\mathbf{E}(\nu-1|\mathbf{Y}_{K+1})| \leqslant O_{\eta_4}(\eta_5^{1/2}). \tag{68}$$

Set $\Omega_{K+1} := \Omega_K \cup \Omega$. This is still $\mathbf{Y}_{K+1}$-measurable and $\int_X \Omega_{K+1} \leqslant \sigma + O_{\eta_4}(\eta_5^{1/2})$; from (68), we thus conclude that $\mathbf{Y}_{K+1}$ has noise level $\sigma + O_{\eta_4}(\eta_5^{1/2})$. Thus the only thing left to verify is the energy increment property (66).

From (64) and (65) we have

$$\left| \int_X (1-1_{\Omega_K})(g - \mathbf{E}(g|\mathbf{Y}_K)) \widetilde{\mathcal{D}} F_{K+1} \right| \geqslant \eta_4 - O(\eta_4^2). \tag{69}$$

Now, from (53), the pointwise bound $0 \leqslant g \leqslant \nu$, (62) and (59), we have

$$\begin{aligned} \left| \int_X (1_{\Omega_{K+1}} - 1_{\Omega_K})(g - \mathbf{E}(g|\mathbf{Y}_K)) \widetilde{\mathcal{D}} F_{K+1} \right| &\leqslant O\left(\int_X (1_{\Omega_{K+1}} - 1_{\Omega_K})(\nu+1) \right) \\ &\leqslant O_{\eta_4}(\eta_5^{1/2}) = O(\eta_4^2), \end{aligned}$$

and hence, by (69),

$$\left| \int_X (1-1_{\Omega_{K+1}})(g - \mathbf{E}(g|\mathbf{Y}_K)) \widetilde{\mathcal{D}} F_{K+1} \right| \geqslant \eta_4 - O(\eta_4^2).$$

Next, from (58), the pointwise bound $0 \leqslant g \leqslant \nu$ and (15), we have

$$\left| \int_X (1-1_{\Omega_{K+1}})(g - \mathbf{E}(g|\mathbf{Y}_K))(\widetilde{\mathcal{D}} F_{K+1} - \mathbf{E}(\widetilde{\mathcal{D}} F_{K+1}|\mathbf{Y}_{K+1})) \right| \leqslant \int_X (\nu+1) O(\eta_4^2) = O(\eta_4^2),$$

and thus

$$\left| \int_X (1-1_{\Omega_{K+1}})(g - \mathbf{E}(g|\mathbf{Y}_K)) \mathbf{E}(\widetilde{\mathcal{D}} F_{K+1}|\mathbf{Y}_{K+1}) \right| \geqslant \eta_4 - O(\eta_4^2).$$

Since Ω_{K+1}, $\mathbf{E}(g|\mathbf{Y}_K)$ and $\mathbf{E}(\widetilde{\mathcal{D}} F_{K+1}|\mathbf{Y}_{K+1})$ are already $\mathbf{Y}_{K+1}$-measurable, we conclude that

$$\left| \int_X (1-1_{\Omega_{K+1}})(\mathbf{E}(g|\mathbf{Y}_{K+1}) - \mathbf{E}(g|\mathbf{Y}_K)) \mathbf{E}(\widetilde{\mathcal{D}} F_{K+1}|\mathbf{Y}_{K+1}) \right| \geqslant \eta_4 - O(\eta_4^2).$$

By (53) and the Cauchy–Schwarz inequality, we conclude that

$$\int_X (1-1_{\Omega_{K+1}}) \big| \mathbf{E}(g|\mathbf{Y}_{K+1}) - \mathbf{E}(g|\mathbf{Y}_K) \big|^2 \geqslant 2c\eta_4^2 - O(\eta_4^3) \tag{70}$$

for some $c>0$.

To pass from this to (66), first observe from (62) and (59) that

$$\int_X (1_{\Omega_{K+1}}-1_{\Omega_K})\mathbf{E}(g|\mathbf{Y}_K)^2 \ll_{\eta_4} \eta_5^{1/2},$$

and so, by the triangle inequality and (63), (66) will follow from the estimate

$$\int_X (1-1_{\Omega_{K+1}})\mathbf{E}(g|\mathbf{Y}_{K+1})^2 \geqslant \int_X (1-1_{\Omega_{K+1}})\mathbf{E}(g|\mathbf{Y}_K)^2 + 2c\eta_4^2 - O(\eta_4^3).$$

Using the identity

$$\mathbf{E}(g|\mathbf{Y}_{K+1})^2 = \mathbf{E}(g|\mathbf{Y}_K)^2 + \big|\mathbf{E}(g|\mathbf{Y}_{K+1})-\mathbf{E}(g|\mathbf{Y}_K)\big|^2 + 2\mathbf{E}(g|\mathbf{Y}_K)(\mathbf{E}(g|\mathbf{Y}_{K+1})-\mathbf{E}(g|\mathbf{Y}_K))$$

and (70), it will suffice to show that

$$\int_X (1-1_{\Omega_{K+1}})\mathbf{E}(g|\mathbf{Y}_K)(\mathbf{E}(g|\mathbf{Y}_{K+1})-\mathbf{E}(g|\mathbf{Y}_K)) \ll \eta_4^3.$$

Now observe that $\mathbf{E}(g|\mathbf{Y}_{K+1})-\mathbf{E}(g|\mathbf{Y}_K)$ is orthogonal to all $\mathbf{Y}_K$-measurable functions, and in particular

$$\int_X (1-1_{\Omega_K})\mathbf{E}(g|\mathbf{Y}_K)(\mathbf{E}(g|\mathbf{Y}_{K+1})-\mathbf{E}(g|\mathbf{Y}_K)) = 0.$$

Thus, it suffices to show that

$$\int_X (1_{\Omega_{K+1}}-1_{\Omega_K})\mathbf{E}(g|\mathbf{Y}_K)(\mathbf{E}(g|\mathbf{Y}_{K+1})-\mathbf{E}(g|\mathbf{Y}_K)) \ll \eta_4^3.$$

Since everything here is $\mathbf{Y}_{K+1}$-measurable, we may replace $\mathbf{E}(g|\mathbf{Y}_{K+1})$ by g. Using (62), it then suffices to show that

$$\int_X (1_{\Omega_{K+1}}-1_{\Omega_K})\big|g-\mathbf{E}(g|\mathbf{Y}_K)\big| \ll \eta_4^3.$$

But this follows from the pointwise bound $0\leqslant g\leqslant\nu$, from (62) and (59). This concludes the proof of Proposition 7.7, which in turn implies Theorem 7.1 and thus Theorem 4.7.

8. A pseudorandom measure which majorizes the primes

In the remainder of the paper we prove Theorem 3.18, which constructs the pseudorandom measure ν which will pointwise dominate the function f defined in (11). As in all

previous sections, we are using the notation from §2 to define quantities such as W, R, M and b.

The measure ν can in fact be described explicitly, following [36], [20] and [19]. Let $\chi\colon \mathbf{R}\to\mathbf{R}$ be a fixed smooth even function which vanishes outside of the interval $[-1,1]$ and obeys the normalization

$$\int_0^1 |\chi'(t)|^2\ dt = 1, \tag{71}$$

but is otherwise arbitrary.([33]) We then define ν by the formula

$$\nu(x) = \nu_\chi(x) := \frac{\phi(W)}{W}\log R\Big(\sum_{m|Wx+b}\mu(m)\chi\Big(\frac{\log m}{\log R}\Big)\Big)^2 \tag{72}$$

for $x\in[N]$, where the sum is over all positive integers m which divide $Wx+b$, and $\mu(m)$ is the *Möbius function* of m, defined as $(-1)^k$ when m is the product of k distinct primes for some $k\geqslant 0$, and zero otherwise (i.e. zero when m is divisible by a non-trivial square).

Remark 8.1. The definition of ν may seem rather complicated, but its behavior is in fact rather easily controlled, at least at "coarse-scales" (averaging x over intervals of length greater than a large power of R), by sieve theory techniques, and in particular by a method of Goldston and Yıldırım [14], though in the paper here we exploit the smoothness of the cutoff χ (as in [20], [32] and [19]) to avoid the need for multiple contour integration, relying on the somewhat simpler Fourier integral expansion instead. For instance, at such scales it is known from these methods that the average value of ν is $1+o(1)$ (see e.g. [20] and [32]), and more generally a large family of linear correlations of ν with itself are also $1+o(1)$ (see [18] and [19]). Thus one can view ν as being close to 1 in a weak (averaged) sense, though of course in a pointwise sense ν will fluctuate tremendously.

It is easy to verify the pointwise bound $f(x)\leqslant\nu(x)$. Indeed, from (11) and (72), it suffices to verify that

$$\sum_{m|Wx+b}\mu(m)\chi\Big(\frac{\log m}{\log R}\Big) = 1$$

whenever $x\in\big[\frac{1}{2}N\big]$ and $Wx+b\in A$. But this is clear, since $Wx+b$ is prime and greater than R. It is also easy to verify the bound (16), using the elementary result that the number of divisors of an integer n is $O_\varepsilon(n^\varepsilon)$ for any $\varepsilon>0$.

([33]) This differs slightly from the majorant introduced by Goldston and Yıldırım in [14] and used in [18]; in our notation, the majorants from those papers corresponds to the case $\chi(t):=\max(1-|t|,0)$. It turns out that choosing χ to be smooth allows for some technical simplifications, at the (acceptable) cost of lowering $\eta_2=\log R/\log N$ slightly.

The remaining task is to verify that ν obeys both the polynomial forms condition (17) and the polynomial correlation condition (18) (note that (15) follows from (17)). We can of course take N to be large compared with the parameters $\eta_0, ..., \eta_7$ and with the parameters D', D'', K and ε (in the case of (18)), as the claim is trivial otherwise.

We begin with a minor reduction designed to eliminate the "wraparound" effects caused by working in the cyclic group $X = \mathbf{Z}/N\mathbf{Z}$ rather than the interval $[N]$. Let us define the truncated domain X' to be the interval $X' := \{x \in \mathbf{Z} : \sqrt{N} \leqslant x \leqslant N - \sqrt{N}\}$ (say). From (16), we can replace the average in X by the average in X' in both (17) and (18) while only incurring an error of $o(1)$ or $o_{D',D'',K}(1)$ at worst. The point of restricting to X' is that all the shifts which occur in (17) and (18) have size at most $O(M^{O(1/\eta_1)})$ or $O_{D',D'',K}(M^{O(1/\eta_1)})$, because of the hypotheses on the degree and coefficients of the polynomials and because all convex bodies are contained in a ball $B(0, M^2)$. By choice of M, these shifts are thus less than $\sqrt{N}$ and so we do not encounter any wraparound issues. Thus (17) is now equivalent to

$$\mathbf{E}_{\vec{h} \in \Omega \cap \mathbf{Z}^d} \mathbf{E}_{x \in X'} \prod_{j \in [J]} \nu(x + Q_1(\vec{h})) = 1 + o_\varepsilon(1), \tag{73}$$

and (18) is similarly equivalent to

$$\begin{aligned} \mathbf{E}_{\vec{n} \in \Omega' \cap \mathbf{Z}^{D'}} \mathbf{E}_{\vec{h} \in \Omega'' \cap \mathbf{Z}^{D''}} \mathbf{E}_{x \in X'} \Big(\prod_{k \in [K]} \mathbf{E}_{\vec{m} \in \Omega \cap \mathbf{Z}^D} \prod_{j \in [J]} \nu(x + \vec{P}_j(\vec{h}) \cdot \vec{m} + \vec{Q}_{j,k}(\vec{h}) \cdot \vec{n}) \Big) \\ \times \prod_{l \in [L]} \nu(x + \vec{S}_l(\vec{h}) \cdot \vec{n}) = 1 + o_{D',D'',K,\varepsilon}(1), \end{aligned} \tag{74}$$

where ν is now viewed as a function on the integers rather than on $X = \mathbf{Z}/N\mathbf{Z}$, defined by (72).

We shall prove (73) and (74) in §11 and §12, respectively. Before doing so, let us first discuss what would happen if we tried to generalize these averages by considering the more general expression

$$\mathbf{E}_{\vec{x} \in \Omega \cap \mathbf{Z}^D} \prod_{j \in [J]} \nu(P_j(\vec{x})), \tag{75}$$

where $D, J \geqslant 0$ are integers, Ω is a convex body in $\mathbf{R}^D$, and $P_1, ..., P_J \in \mathbf{Z}[\mathbf{x}_1, ..., \mathbf{x}_D]$ are polynomials of bounded degree and whose coefficients are not too large (say of size $O(W^{O(1)})$). In light of the linear correlation theory, one would generally expect these polynomial correlations to also be $1 + o(1)$ as long as the polynomials $P_1, ..., P_J$ were suitably "distinct" and that the range Ω is suitably large.

There will however be some technical issues in establishing such a statement. For sake of exposition let us just discuss the case $J = 1$, so that we are averaging a single

factor $\nu(P(\vec{x}))$ for some polynomial P of D variables $\vec{x}=(x_1,...,x_D)$. Even in this simple case, two basic problems arise.

The first problem is that ν is not perfectly uniformly distributed modulo p for all primes p. The "W-trick" of using $Wx+b$ instead of x in (72) (and renormalizing by $\phi(W)/W$ to compensate) does guarantee a satisfactory uniform distribution of ν modulo p for small primes $p<w$. However, for larger primes $p\geqslant w$, it turns out that ν will generally avoid the residue class $\{x:Wx+b=0 \mod p\}$, and instead distribute itself uniformly among the other $p-1$ residue classes. This corresponds to the basic fact that primes (and almost primes) are mostly coprime to any given modulus p. Because of this, the expected value of an expression such as $\nu(P(x))$ will increase from 1 to roughly $(1-1/p)^{-1}$ if we know that $WP(x)+b$ is coprime to p, and conversely it will drop to essentially zero if we know that $WP(x)+b$ is divisible by p. These two effects will essentially balance each other out, provided that the algebraic variety $\{x\in F_p^D:WP(x)+b=0\}$ has the expected density of $1/p+O(1/p^{3/2})$ (say) over the finite field affine space F_p^D. The famous result of Deligne [7], [8], in which the Weil conjectures were proved, establishes this when $WP+b$ is non-constant and is absolutely irreducible modulo p (i.e. irreducible over the algebraic closure of F_p). However, there can be some "bad" primes $p\geqslant w$ for which this irreducibility fails; a particularly "terrible" case arises when p divides the polynomial $WP+b$, in which case the variety has density 1 in F_p^D and the expected value of (75) drops to zero. This reflects the intuitive fact that $WP(x)+b$ is much less likely to be prime or almost prime if $WP+b$ itself is divisible by some prime p. The other bad primes p do not cause such a severe change in the expectation (75), but can modify the expected answer of $1+o(1)$ by a factor of $1+O(1/p)=\exp(O(1/p))$, leading to a final value which is something like $\exp(O(\sum_{p \text{ bad}} 1/p)+o(1))$. In most cases, this expression will be in fact very close to 1, because of the restriction $p\geqslant w$. However, the (very slow) divergence of the sum $\sum_p 1/p$ means that there are some exceptional cases in which averages such as (75) are unpleasantly large. For instance, for any fixed $h\neq 0$, the average value of $\nu(x)\nu(x+h)$ over sufficiently coarse-scales turns out to be $\exp(O(\sum_{p\geqslant w:p|h} 1/p)+o(1))$, which can be arbitrarily large in the (very rare) case when h contains many prime factors larger than w, the basic problem being that the algebraic variety $\{x\in F_p:(Wx+b)(W(x+h)+b)=0\}$, which is normally empty, becomes unexpectedly large when $p\geqslant w$ and p divides h. This phenomenon was already present in [18], leading in particular to the rather technical "correlation condition" for ν.

The second problem, which is a new feature in the polynomial case compared with the previous linear theory, is that we will not necessarily be able to average *all* of the parameters $x_1,...,x_D$ over coarse-scales (e.g. at scales $O(M)$, $O\big(\sqrt{M}\,\big)$ or $O(M^{1/4})$). Instead, some of the parameters will be only averaged over fine scales such as $O(H)$. At

these scales, the elementary sieve theory methods we are employing cannot estimate the expression (75) directly; indeed, the problem then becomes analogous to that of understanding the distribution of primes in short intervals, which is notoriously difficult. Fortunately, we can proceed by first fixing the fine-scale parameters and using the sieve theory methods to compute the averages over the coarse-scale parameters rather precisely, leading to certain tractable divisor sums over "locally bad primes" which can then be averaged over fine scales. Here we will rely on a basic heuristic from algebraic geometry, which asserts that a "generic" slice of an algebraic variety by a linear subspace will have the same codimension as the original variety. In our context, this means that a prime which is "globally good" with respect to many parameters, will also be "locally good" when freezing one or more parameters, for "most" choices of such parameters. We will phrase the precise versions of these statements as a kind of "combinatorial Nullstellensatz" (cf. [1]) in Appendix D. This effect lets us deal with the previous difficulty that the sum of $1/p$ over bad primes can occasionally be very large.

We have already mentioned the need to control the density of varieties such as $\{x \in F_p^D : WP(x)+b=0\}$, which in general requires the Weil conjectures as proven by Deligne. Fortunately, for the application to polynomial progressions, the polynomials P involved will always be linear in at least one of the coarse-scale variables. This makes the density of the algebraic variety much easier to compute, provided that the coefficients in this linear representation do not degenerate (either by the linear coefficient vanishing, or by the linear and constant coefficients sharing a common factor). Thus we are able to avoid using the Weil conjectures. In fact we will be able to proceed by rather elementary algebraic methods, without using modern tools from arithmetic geometry; see Appendix D.

8.2. Notation

We now set out some notation which will be used throughout the proof of (73) and (74). If p is a prime, we use F_p to denote the finite field with p elements.

If P and Q lie in some ring R, we use $P|Q$ to denote the statement that Q is a multiple of P. An element of a ring is a *unit* if it is invertible, and *irreducible*([34]) if it is not a unit, and cannot be written as the product of two non-units. A ring is a *unique factorization domain* if every element is uniquely expressible as a finite product of irreducibles, up to permutations and units. If $P_1, ..., P_J$ lie in a unique factorization domain, we say that $P_1, ..., P_J$ are *jointly coprime* (or just *coprime* if $J=2$) if there exists no irreducible which divides all the $P_1, ..., P_J$, and *pairwise coprime* if each pair P_j, P_k is coprime for $1 \leqslant j < k \leqslant J$; thus pairwise coprime implies jointly coprime, but not conversely.

([34]) We shall reserve the term *prime* for the rational primes $2, 3, 5, 7, ...$ to avoid confusion.

As observed by Hilbert, if R is a unique factorization domain, then so is $R[\mathbf{x}]$ (due to the Euclidean algorithm). In particular, $F_p[\mathbf{x}_1,...,\mathbf{x}_D]$ and $\mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_D]$ are unique factorization domains (with units $F_p\backslash\{0\}$ and $\{-1,+1\}$, respectively).

Every polynomial in $R[\mathbf{x}_1,...,\mathbf{x}_D]$ can of course be viewed as a function from R^D to R. If $P\in\mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_D]$ is a polynomial and $N\geqslant 1$, we write $P \bmod N$ for the associated polynomial in $\mathbf{Z}_N[\mathbf{x}_1,...,\mathbf{x}_D]$ formed by projecting all the coefficients onto the ring $\mathbf{Z}_N$, and thus $P \bmod N$ can be viewed as a function from $\mathbf{Z}_N^D$ to $\mathbf{Z}_N$. Note that this projection may alter the property of two or more polynomials being jointly or pairwise coprime; the precise analysis of when this occurs will in fact be a major focus of our arguments here.

It will be convenient to introduce the modified exponential function

$$\mathrm{Exp}(x) := \max(e^x - 1, 0).$$

Thus $\mathrm{Exp}(x)\sim x$ when x is non-negative and small, while $\mathrm{Exp}(X)\sim e^x$ for x large. Observe the elementary inequalities

$$\mathrm{Exp}(x+y) \leqslant \mathrm{Exp}(2x)+\mathrm{Exp}(2y) \quad \text{and} \quad \mathrm{Exp}(x)^K \ll_K \mathrm{Exp}(Kx) \tag{76}$$

for any $x, y\geqslant 0$ and $K\geqslant 1$.

9. Local estimates

Before we give correlation estimates for ν on the integers, we first need to consider the analogous problem modulo p. To formalize this problem, we introduce the following definition.

Definition 9.1. (Local factor) Let $P_1,...,P_J\in\mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_D]$ be polynomials with integer coefficients. For any prime p, we define the (principal) *local factor*

$$c_p(P_1,...,P_J) := \mathbf{E}_{x\in F_p^D} \prod_{j\in[J]} 1_{P_j(x)=0 \bmod p}.$$

We also define the *complementary local factor*

$$\bar{c}_p(P_1,...,P_J) := \mathbf{E}_{x\in F_p^D} \prod_{j\in[J]} 1_{P_j(x)\neq 0 \bmod p}.$$

Examples 9.2. If $P_1,...,P_J$ are homogeneous linear forms on F_p^D, with total rank r, then $c_p(P_1,...,P_J)=p^{-r}$. If the forms are independent (and thus $J=r$), then

$$\bar{c}_p(P_1,...,P_J) = \left(1-\frac{1}{p}\right)^J.$$

If $D=1$, then the local factor $c_p(\mathbf{x}^2+1)$ equals $2/p$ when $p=1 \bmod 4$ and equals 0 when $p=3 \bmod 4$, by quadratic reciprocity. (When $p=2$, it is equal to $1/p$.) More generally, the Artin reciprocity law [22] relates Artin characters to certain local factors. Deligne's celebrated proof [7], [8] of the Weil conjectures implies (as a very special case) that $c_p(P)=1/p+O_{k,D}(1/p^{3/2})$ whenever $P \in \mathbf{Z}[\mathbf{x}_1, ..., \mathbf{x}_D]$ determines a non-singular projective algebraic variety over F_p. For instance, if $P=\mathbf{x}_2^2-\mathbf{x}_1^3-a\mathbf{x}_1-b$, so that P determines an elliptic curve, with discriminant $\Delta=-16(4a^3+27b^2)$ coprime to p, then $c_p(P)=1/p+O(1/p^{3/2})$ (a classical result of Hasse). The Birch and Swinnerton–Dyer conjectures, if true, would provide more precise information (though not of upper bound type) on the error term in this case.

Remark 9.3. The factor c_p denotes the proportion of points on F_p^D which lie on the algebraic variety determined by the polynomials $P_1, ..., P_J$, while the complementary factor $\bar{c}_p$ is the proportion of points in F_p^D for which all the $P_1, ..., P_J$ are coprime to p. Clearly these factors lie between 0 and 1; for instance, when $J=0$ we have $c_p=1$ and $\bar{c}_p=0$. Our interest is to estimate c_p for higher values of J. This will be of importance when we come to the "global" estimates for $\prod_{j\in[J]} \nu(P_j(x))$ over various subsets of $\mathbf{Z}^d$; heuristically, the average value of this expression should be approximately the product of the complementary factors $\bar{c}_p$ as p ranges over the primes.

From the inclusion-exclusion principle we have the identity

$$\bar{c}_p(P_1, ..., P_J) = \sum_{S\subseteq[J]} (-1)^{|S|} c_p(\{P_j\}_{j\in S}) \tag{77}$$

and so we can estimate the complementary local factors using the principal local factors.

As mentioned earlier, the precise estimation of $c_p(P_1, ..., P_J)$ for general $P_1, ..., P_J$ is intimately connected to a number of deep results in arithmetic geometry such as the Weil conjectures and the Artin reciprocity law. Fortunately, for our applications, we will only need to know the $1/p$ coefficient of $c_p(P_1, ..., P_J)$ and can neglect lower order terms. Also, we will be working in the case where each of the polynomials P_j are linear in at least one of the coordinates $x_1, ..., x_D$ of x and are "non-degenerate" in the other coordinates. In such a simplified context, we will be able to control c_p quite satisfactorily using only arguments from elementary algebra. To state the results, we first need the notion of a prime p being good, bad or terrible with respect to a collection of polynomials.

Definition 9.4. (Good prime) Let $P_1, ..., P_J \in \mathbf{Z}[\mathbf{x}_1, ..., \mathbf{x}_D]$ be a collection of polynomials. We say that a prime p is *good* with respect to $P_1, ..., P_J$ if the following hold:

- the polynomials $P_1 \bmod p, ..., P_J \bmod p$ are pairwise coprime;

• for each $j \in [J]$, there exists a coordinate $1 \leqslant k_j \leqslant D$ for which we have the linear behavior

$$P_j(\mathbf{x}_1, ..., \mathbf{x}_D) = P_{j,1}(\mathbf{x}_1, ..., \mathbf{x}_{k_j-1}, \mathbf{x}_{k_j+1}, ..., \mathbf{x}_D)\mathbf{x}_{k_j} + P_{j,0}(\mathbf{x}_1, ..., \mathbf{x}_{k_j-1}, \mathbf{x}_{k_j+1}, ..., \mathbf{x}_D) \bmod p,$$

where $P_{j,1}, P_{j,0} \in F_p[\mathbf{x}_1, ..., \mathbf{x}_{k_j-1}, \mathbf{x}_{k_j+1}, ..., \mathbf{x}_D]$ are such that $P_{j,1}$ is non-zero and coprime to $P_{j,0}$.

We say that a prime is *bad* if it is not good. We say that a prime is *terrible* if at least one of the P_j's vanish identically modulo p (i.e. all the coefficients are divisible by p). Note that terrible primes are automatically bad.

Our main estimate on the local factors is then as follows.

LEMMA 9.5. (Local estimates) *Let $P_1, ..., P_J \in \mathbf{Z}[\mathbf{x}_1, ..., \mathbf{x}_D]$ have degree at most d, let p be a prime and let $S \subset [J]$.*

(a) *If $|S|=0$, then $c_p(\{P_j\}_{j\in S})=1$.*
(b) *If $|S| \geqslant 1$ and p is not terrible, then $c_p(\{P_j\}_{j\in S})=O_{d,D,J}(1/p)$.*
(c) *If $|S|=1$ and p is good, then $c_p(\{P_j\}_{j\in S})=1/p+O_{d,D,J}(1/p^2)$.*
(d) *If $|S|>1$ and p is good, then $c_p(\{P_j\}_{j\in S})=O_{d,D,J}(1/p^2)$.*
(e) *If p is terrible, then $\bar{c}_p(P_1, ..., P_J)=0$.*
(f) *If p is not terrible, then $\bar{c}_p(P_1, ..., P_J)=1+O_{d,D,J}(1/p)$.*

The proof of this lemma involves only elementary algebra, but we defer it to Appendix D so as not to disrupt the flow of the argument.

Remark 9.6. From (77) and Lemma 9.5 (a), (c), (d), we also have

$$\bar{c}_p(P_1, ..., P_J) = 1 - \frac{J}{p} + O_{d,D,J}\left(\frac{1}{p^2}\right)$$

when p is good. In practice we shall need a more sophisticated version of this fact, when certain complex weights $p^{-\sum_{j\in S} z_j}$ are inserted into the right-hand side of (77); see Lemma 10.4.

10. The initial correlation estimate

To prove (73) and (74) we shall need the following initial estimate which handles general polynomial averages of ν over large scales, but with an error term that can get large if there are many "bad" primes present. More precisely, this section is devoted to proving the following result.

PROPOSITION 10.1. (Correlation estimate) *Let $P_1, ..., P_J \in \mathbf{Z}[\mathbf{x}_1, ..., \mathbf{x}_D]$ have degree at most d for some $J, D, d \geqslant 0$. Let Ω be a convex body in $\mathbf{R}^D$ with inradius at least R^{4J+1}. Let $\mathbf{P}_b$ be the set of primes $w \leqslant p \leqslant R^{\log R}$ bad with respect to $WP_1+b, ..., WP_J+b$, and let $\mathbf{P}_t \subset \mathbf{P}_b$ be the set of primes $w \leqslant p \leqslant R^{\log R}$ which are terrible (as defined in Definition 9.4). Then*

$$\mathbf{E}_{x \in \Omega \cap \mathbf{Z}^D} \prod_{j \in [J]} \nu(P_j(x)) = 1_{\mathbf{P}_t = \varnothing} + o_{D,J,d}(1) + O_{D,J,d}\Big(\mathrm{Exp}\Big(O_{D,J,d}\Big(\sum_{p \in \mathbf{P}_b} \frac{1}{p}\Big)\Big)\Big). \quad (78)$$

Remark 10.2. We only expect this estimate to be useful when the number of bad primes is finite. This is equivalent to requiring that the polynomials $P_1, ..., P_J$ are coprime, and each one is linear in at least one variable. Because the sum $\sum_p 1/p$ is (very slowly) divergent (see (110)), the last error term can be unpleasantly large on occasion, but in practice we will be able to introduce averaging over additional parameters which will make the effect of the error small on average, the point being that the sets $\mathbf{P}_t$ and $\mathbf{P}_b$ are generically rather small. The radius R^{4J+1} is not best possible, but to lower it too much would require some deep analytical number theory estimates such as the Bombieri–Vinogradov inequality, which we shall avoid using here. The upper bound $R^{\log R}$ (which was not present in earlier work) can also be lowered, but for our purposes any bound which is subexponential in R will suffice.

Remark 10.3. All the primes $p < w$ will be bad (but not terrible); however, their contribution will be almost exactly canceled by the $\phi(W)/W$ term present in ν, and we do not need to include them into $\mathbf{P}_b$. Even a single terrible prime will cause the main term $1_{\mathbf{P}_t = \varnothing}$ to vanish (basically because one of the $P_j(x)$ will now be inherently composite and so will be unlikely to have a large value of ν), which will make asymptotics difficult; however, terrible primes are no worse than merely bad primes for the purposes of *upper* bounds.

Proof of Proposition 10.1. Throughout this proof we fix D, J and d, and allow the implicit constants in the $O(\cdot)$ and $o(\cdot)$ notation to depend on these parameters. We will also always assume R to be sufficiently large depending on D, J and d.

We expand the left-hand side using (72) as

$$\Big(\frac{\phi(W)}{W} \log R\Big)^J \sum_{m_1, m'_1, ..., m_J, m'_J \geqslant 1} \Big(\prod_{j \in [J]} \mu(m_j)\mu(m'_j) \chi\Big(\frac{\log m_j}{\log R}\Big) \chi\Big(\frac{\log m'_j}{\log R}\Big)\Big) \times \mathbf{E}_{x \in \Omega \cap \mathbf{Z}^D} \prod_{j=1}^{J} 1_{\mathrm{lcm}(m_j, m'_j) | WP_j(x)+b}. \quad (79)$$

Here of course $\mathrm{lcm}(\cdot)$ denotes least common multiple. Note that the presence of the μ and χ factors allows us to restrict $m_1, ..., m'_J$ to be square-free and at most R.

The first task is to eliminate the role of the convex body Ω, taking advantage of the large inradius assumption. Let $M := \mathrm{lcm}(m_1, m'_1, ..., m_j, m'_j)$. Thus M is square-free and at most R^{2J}. The function $x \mapsto 1_{\mathrm{lcm}(m_j,m'_j)|WP_j(x)+b}$ is periodic with respect to the lattice $M \cdot \mathbf{Z}^D$, and thus can be meaningfully defined on the group $\mathbf{Z}^D_M$. Applying Corollary C.3 (recalling that Ω is assumed to have inradius at least R^{4J+1}), we thus have

$$\mathbf{E}_{x \in \Omega \cap \mathbf{Z}^D} \prod_{j=1}^{J} 1_{\mathrm{lcm}(m_j,m'_j)|WP_j(x)+b} = \left(1 + O\left(\frac{1}{R^{2J+1}}\right)\right) \mathbf{E}_{y \in \mathbf{Z}^D_M} \prod_{j=1}^{J} 1_{\mathrm{lcm}(m_j,m'_j)|WP_j(y)+b}.$$

Let us first dispose of the error term $O(1/R^{2J+1})$. The contribution of this term to (79) can be crudely bounded by $O(R^{-2J-1})$, and so the contribution of this term to (79) can be crudely bounded by

$$O\left(\left(\frac{\phi(W)}{W} \log R\right)^J \sum_{1 \leqslant m_1, m'_1, ..., m_J, m'_J \leqslant R} R^{-2J-1}\right) \ll \frac{\log^J R}{R} = o(1).$$

Thus we may discard this error, and reduce to showing that

$$\begin{aligned}\left(\frac{\phi(W)}{W} \log R\right)^J \sum_{m_1, m'_1, ..., m_J, m'_J \geqslant 1} \Bigg(\prod_{j \in [J]} \mu(m_j)\mu(m'_j) \chi\left(\frac{\log m_j}{\log R}\right) \chi\left(\frac{\log m'_j}{\log R}\right)\Bigg) \\ \times \alpha_{\mathrm{lcm}(m_1,m'_1),...,\mathrm{lcm}(m_J,m'_J)} \qquad (80) \\ = 1_{\mathbf{P}_t = \varnothing} + o(1) + O\left(\mathrm{Exp}\left(O\left(\sum_{p \in \mathbf{P}_b} \frac{1}{p}\right)\right)\right),\end{aligned}$$

where $\alpha_{\mathrm{lcm}(m_1,m'_1),...,\mathrm{lcm}(m_J,m'_J)}$ is the local factor

$$\alpha_{\mathrm{lcm}(m_1,m'_1),...,\mathrm{lcm}(m_J,m'_J)} := \mathbf{E}_{y \in \mathbf{Z}^D_M} \prod_{j=1}^{J} 1_{\mathrm{lcm}(m_j,m'_j)|WP_j(y)+b}.$$

Observe, from the Chinese remainder theorem, that α is multiplicative, so that if

$$\mathrm{lcm}(m_j, m'_j) = \prod_p p^{r_j}$$

then

$$\alpha_{\mathrm{lcm}(m_1,m'_1),...,\mathrm{lcm}(m_J,m'_J)} = \prod_p \alpha_{p^{r_1},...,p^{r_J}}$$

(note that all but finitely many of the terms in the product are 1). If the $m_1, ..., m'_J$ are squarefree, then the r_j's are either 0 or 1, and we simplify further to

$$\alpha_{\mathrm{lcm}(m_1,m'_1),...,\mathrm{lcm}(m_J,m'_J)} = \prod_p c_p((WP_j+b)_{r_j=1}),$$

where the local factors c_p are defined in Definition 9.1, and the dummy variable j is ranging over all the indices for which $r_j=1$. Also note that the m_j and m'_j are bounded by R, and thus we may certainly restrict the primes p to be less than $R^{\log R}$ without difficulty.

The next step is to replace the χ factors by terms which are multiplicative in the m_j and m'_j. Since χ is smooth and compactly supported, we have the Fourier expansion

$$e^x\chi(x)=\int_{-\infty}^{\infty}\varphi(\xi)e^{-ix\xi}\,d\xi \tag{81}$$

for some smooth, rapidly decreasing function $\varphi(\xi)$ (so in particular $\varphi(\xi)=O_A((1+|\xi|)^{-A})$ for any $A>0$). For future reference, we observe that (81) and the hypotheses on χ will imply the identity

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(1+it)(1+it')}{2+it+it'}\varphi(t)\varphi(t')\,dt\,dt'=1 \tag{82}$$

(see [19, Lemma D.2], or the proof of [32, Proposition 2.2]).

We follow the arguments in [36], [20], [32] and [19], except that for technical reasons (having to do with the terrible primes) we will be unable to truncate the ξ variables. From (81) we have

$$\chi\biggl(\frac{\log m_j}{\log R}\biggr)=\int_{-\infty}^{\infty}m_j^{-z_j}\,d\xi_j\quad\text{and}\quad\chi\biggl(\frac{\log m'_j}{\log R}\biggr)=\int_{-\infty}^{\infty}(m'_j)^{-z'_j}\,d\xi'_j,$$

where we adopt the notational conventions

$$z_j:=\frac{1+\xi_j}{\log R}\quad\text{and}\quad z'_j:=\frac{1+\xi'_j}{\log R}.$$

Our task is thus to show that

$$\begin{aligned}&\biggl(\frac{\phi(W)}{W}\log R\biggr)^J\sum_{m_1,m'_1,\dots,m_J,m'_J\geqslant 1}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\\&\biggl(\prod_{j\in[J]}\mu(m_j)\mu(m'_j)m_j^{-z_j}(m'_j)^{-z'_j}\varphi(\xi_j)\varphi(\xi'_j)\,d\xi_j\,d\xi'_j\biggr)\prod_{p\leqslant R^{\log R}}c_p((WP_j+b)_{r_j=1})\\&\quad=1+o(1)+O\biggl(\operatorname{Exp}\biggl(O\Bigl(\sum_{p\in\mathbf{P}_b}\frac{1}{p}\Bigr)\biggr)\biggr).\end{aligned}\tag{83}$$

The left-hand side can be factorized as

$$\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\biggl(\frac{\phi(W)}{W}\log R\biggr)^J\prod_{p\leqslant R^{\log R}}E_p\prod_{j\in[J]}\varphi(\xi_j)\varphi(\xi'_j)\,d\xi_j\,d\xi'_j,$$

where $E_p = E_p(z_1, ..., z'_J)$ is the Euler factor

$$E_p := \sum_{m_1,...,m'_J \in \{1,p\}} \prod_{j \in [J]} \mu(m_j)\mu(m'_j) m_j^{-z_j} (m'_j)^{-z'_j} c_p((WP_j+b)_{r_j=1}).$$

Note that if the z_j and z'_j were zero, then this would just be the complementary factor $\bar{c}_p(WP_1+b, ..., WP_J+b)$ defined in Definition 9.1; see (77). Of course, z_j and z'_j are non-zero. To approximate E_p in this case, we introduce the Euler factor

$$E'_p := \prod_{j \in [J]} \frac{(1-1/p^{1+z_j})(1-1/p^{1+z'_j})}{1-1/p^{1+z_j+z'_j}}.$$

Note that E'_p never vanishes.

LEMMA 10.4. (Euler product estimate) *We have*

$$\prod_{p \leqslant R^{\log R}} \frac{E_p}{E'_p} = \Big(1_{\mathbf{P}_t=\varnothing} + o(1) + O\Big(\mathrm{Exp}\Big(O\Big(\sum_{p \in \mathbf{P}_b} \frac{1}{p}\Big)\Big)\Big)\Big)\Big(\frac{W}{\phi(W)}\Big)^J.$$

Proof. For $p<w$, we directly compute (since w is slowly growing compared to R, and WP_j+b is equal modulo p to b, which is coprime to p) that

$$E_p = 1+o(1) \quad \text{and} \quad E'_p = \Big(1-\frac{1}{p}\Big)^J + o(1),$$

and hence (again because w is slowly growing)

$$\prod_{p<w} \frac{E_p}{E'_p} = (1+o(1))\Big(\frac{W}{\phi(W)}\Big)^J.$$

Thus it will suffice to show that

$$\prod_{w \leqslant p \leqslant R^{\log R}} \frac{E_p}{E'_p} = 1_{\mathbf{P}_t=\varnothing} + o(1) + O\Big(\mathrm{Exp}\Big(O\Big(\sum_{p \in \mathbf{P}_b} \frac{1}{p}\Big)\Big)\Big).$$

For p terrible, Lemma 9.5 gives the estimate

$$E_p \ll \frac{1}{p} \ll \frac{1}{p} E'_p,$$

and so it will suffice to show that

$$\prod_{\substack{w \leqslant p \leqslant R^{\log R} \\ p \text{ not terrible}}} \frac{E_p}{E'_p} = 1 + o(1) + O\Big(\mathrm{Exp}\Big(O\Big(\sum_{p \in \mathbf{P}_b} \frac{1}{p}\Big)\Big)\Big).$$

For p bad but not terrible, Lemma 9.5 gives the crude estimate

$$E_p = 1+O\Big(\frac{1}{p}\Big) = \exp\Big(O\Big(\frac{1}{p}\Big)\Big)E'_p,$$

and thus

$$\prod_{\substack{w\leqslant p\leqslant R^{\log R}\\ p \text{ bad but not terrible}}} \frac{E_p}{E'_p} = 1+O\Big(\mathrm{Exp}\Big(O\Big(\sum_{p\in\mathbf{P}_b}\frac{1}{p}\Big)\Big)\Big).$$

Thus it suffices to show that

$$\prod_{\substack{w\leqslant p\leqslant R^{\log R}\\ p \text{ good}}} \frac{E_p}{E'_p} = 1+o(1).$$

Since the product $\prod_p(1+O(1/p^2))$ is convergent, and w goes to infinity, it in fact suffices to show that

$$E_p = \Big(1+O\Big(\frac{1}{p^2}\Big)\Big)E'_p$$

for all good primes larger than w. But this easily follows from Lemma 9.5 and Taylor expansion (recall that the real parts of z_j and z'_j are $1/\log R>0$). □

Now we use the theory of the Riemann zeta function. From (113) we have that

$$\prod_{p\leqslant R^{\log R}} E'_p = \prod_{j\in[J]}(1+o(1))\frac{\zeta(1+z_j+z'_j)}{\zeta(1+z_j)\zeta(1+z'_j)}.$$

On the other hand, from (108) we have

$$\frac{1}{\zeta(1+(1+i\xi)/\log R)} = (1+o((1+|\xi|)^2))\frac{1+i\xi}{\log R}$$

and

$$\zeta(1+(1+i\xi)/\log R) = (1+o((1+|\xi|)^2))\frac{\log R}{1+i\xi}$$

for any real ξ, and hence

$$\prod_{p\leqslant R^{\log R}} E'_p = \prod_{j\in[J]}(1+o((1+|\xi_j|+|\xi'_j|)^6))\frac{1}{\log R}\frac{(1+i\xi_j)(1+i\xi'_j)}{2+i\xi_j+i\xi'_j}.$$

Applying Lemma 10.4, we conclude that

$$\begin{aligned}&\Big(\frac{\phi(W)}{W}\log R\Big)^J \prod_{p\leqslant R^{\log R}} E_p\\ &= \Big(1_{\mathbf{P}_t=\varnothing}+o\Big(\prod_{j\in[J]}(1+|\xi_j|+|\xi'_j|)^6\Big)+O\Big(\mathrm{Exp}\Big(O\Big(\sum_{p\in\mathbf{P}_b}\frac{1}{p}\Big)\Big)\Big)\prod_{j\in[J]}\frac{(1+i\xi_j)(1+i\xi'_j)}{2+i\xi_j+i\xi'_j}.\end{aligned}$$

Thus, by the triangle inequality, to show (83) it will suffice to show that

$$\int_{\mathbf{R}} \dots \int_{\mathbf{R}} \prod_{j\in[J]} \varphi(\xi_j)\varphi(\xi'_j) \frac{(1+i\xi_j)(1+i\xi'_j)}{2+i\xi_j+i\xi'_j}\, d\xi_j\, d\xi'_j = 1$$

and

$$\int_I \dots \int_I \prod_{j\in[J]} |\varphi(\xi_j)||\varphi(\xi'_j)|(1+|\xi_j|+|\xi'_j|)^6 \frac{|1+i\xi_j||1+i\xi'_j|}{|2+i\xi_j+i\xi'_j|}\, d\xi_j\, d\xi'_j = O(1).$$

But the first estimate follows from (82), while the second estimate follows from the rapid decrease of φ. This proves Proposition 10.1. □

To illustrate the above proposition, let us specialize to the case of monic linear polynomials of one variable (this case was essentially treated in [18] and [14]).

COROLLARY 10.5. (Correlation condition) *Let $h_1, \dots, h_J$ be integers, and let $I \subset \mathbf{R}$ be an interval of length at least R^{4J+1}. Then*

$$\mathbf{E}_{x\in I\cap\mathbf{Z}} \prod_{j\in[J]} \nu(x+h_j) = 1 + o_{D,J,d}(1) + O_{D,J,d}\Big(\mathrm{Exp}\Big(O_{D,J,d}\Big(\sum_{p\in\mathbf{P}_b} \frac{1}{p}\Big)\Big)\Big),$$

where

$$\mathbf{P}_b := \{w \leqslant p \leqslant R^{\log R} : p \,|\, h_j - h_{j'} \text{ for some } 1 \leqslant j < j' \leqslant J\}.$$

Proof. Apply Proposition 10.1 with $P_j(x) := x+h_j$ and $\Omega := I$. Then there are no terrible primes, and the only bad primes larger than w are those which divide $h_j - h_{j'}$ for some $1 \leqslant j < j' \leqslant J$. □

This can already be used to derive the "correlation condition" in [18]; a similar aplication of Proposition 10.1 also gives the "linear forms condition" from that paper. We will also need the following variant of the above estimate.

COROLLARY 10.6. (Correlation condition on progressions) *Let $h_1, \dots, h_J$ be integers, $q \geqslant 1$, $a \in \mathbf{Z}_q$ and $I \subset \mathbf{R}$ be an interval of length at least qR^{4J+1}. Then*

$$\mathbf{E}_{x\in I\cap\mathbf{Z}, x=a \bmod q} \prod_{j\in[J]} \nu(x+h_j) = O_{D,J,d}\Big(\exp\Big(O_{D,J,d}\Big(\sum_{p\in\mathbf{P}_b} \frac{1}{p}\Big)\Big)\Big),$$

where

$$\mathbf{P}_b := \{w \leqslant p \leqslant R^{\log R} : p \,|\, h_j - h_{j'} \text{ for some } 1 \leqslant j < j' \leqslant J\} \cup \{p \geqslant w : p \,|\, q\}. \tag{84}$$

Proof. Apply Proposition 10.1 with $P_j(x) := (qx+a)+h_j$ and with

$$\Omega := \{x \in \mathbf{R} : qx+a \in I\}.$$

Then the bad primes are those which divide $h_j - h_{j'}$ or which divide q. (There are terrible primes if a and q are not coprime, but this will not affect the upper bound. One can get more precise estimates as in Corollary 10.5, but we will not need them here.) □

This in turn implies the following result.

COROLLARY 10.7. (Correlation condition with periodic weight) *Let $h_1, ..., h_J$ be integers, $q \geqslant 1$, $I \subset \mathbf{R}$ be an interval of length at least qR^{4J+1} and $f : \mathbf{Z} \to \mathbf{R}^+$ be periodic modulo q (and thus definable on $\mathbf{Z}_q$). Then*

$$\mathbf{E}_{x \in I \cap \mathbf{Z}} f(x) \prod_{j \in [J]} \nu(x+h_j) = O_{D,J,d}\Big((\mathbf{E}_{y \in \mathbf{Z}_q} f(y)) \exp\Big(O_{D,J,d}\Big(\sum_{p \in \mathbf{P}_b} \frac{1}{p} \Big)\Big)\Big),$$

where $\mathbf{P}_b$ was defined in (84).

Proof. The left-hand side can be bounded by

$$\mathbf{E}_{y \in \mathbf{Z}_q} f(y) \mathbf{E}_{x \in I \cap \mathbf{Z}, x = a \bmod q} \prod_{j \in [J]} \nu(x+h_j),$$

simply because the set $\{x \in I \cap \mathbf{Z} : x = a \bmod q\}$ has cardinality roughly $|I \cap \mathbf{Z}|/q$ (by the hypotheses on the length of I). The claim then follows from Corollary 10.6. □

11. The polynomial forms condition

In this section we use the above correlation estimates to prove the polynomial forms condition (73). We begin with a preliminary bound in this direction.

THEOREM 11.1. (Polynomial forms condition) *Let $M, D, d, J \geqslant 0$ and $\varepsilon > 0$. Let $P_1, ..., P_J \in \mathbf{Z}[\mathbf{m}_1, ..., \mathbf{m}_D]$ be polynomials of degree d with all coefficients of size at most W. Let $I \subset \mathbf{R}$ be any interval of length at least R^{4J+1}, and let $\Omega \subset \mathbf{R}^D$ be any convex body with inradius at least R^{ε}. Then*

$$\mathbf{E}_{x \in I \cap \mathbf{Z}, \vec{m} \in \Omega \cap \mathbf{Z}^D} \prod_{j \in [J]} \nu(x+P_j(\vec{m})) = 1 + o_{D,d,\varepsilon,J}\Big(\exp\Big(O_{D,d,\varepsilon,J}\Big(\sum_{p \in \mathbf{P}_b} \frac{1}{p} \Big)\Big)\Big),$$

where $\mathbf{P}_b$ denote the set of all $w \leqslant p \leqslant \mathbf{R}^{\log R}$ which are "globally bad" in the sense that $p | P_j - P_{j'}$ for some $1 \leqslant j < j' \leqslant J$.

Proof. Let us fix D, d, J and ε, and allow all implicit constants to depend on these quantities. From Corollary 10.5 we have

$$\mathbf{E}_{x\in I\cap\mathbf{Z}}\prod_{j\in[J]}\nu(x+P_j(\vec{m}))=1+o(1)+O\Big(\mathrm{Exp}\Big(O\Big(\sum_{p\in\mathbf{P}_{\vec{m}}}\frac{1}{p}\Big)\Big)\Big)$$

for all $\vec{m}\in\Omega'\cap\mathbf{Z}^D$, where $\mathbf{P}_{\vec{m}}$ is the set of primes $w\leqslant p\leqslant R^{\log R}$ such that $p|P_j(\vec{m})-P_{j'}(\vec{m})$ for some $1\leqslant j<j'\leqslant d$. Thus it suffices to show that

$$\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}\mathrm{Exp}\Big(O\Big(\sum_{p\in\mathbf{P}_{\vec{m}}}\frac{1}{p}\Big)\Big)=o\Big(\exp\Big(O\Big(\sum_{p\in\mathbf{P}_b}\frac{1}{p}\Big)\Big)\Big).$$

Applying (76), we reduce to showing that

$$\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}\mathrm{Exp}\Big(O\Big(\sum_{p\in\mathbf{P}_{\vec{m}}\backslash\mathbf{P}_b}\frac{1}{p}\Big)\Big)=o(1).$$

Applying Lemma E.1, it suffices to show that

$$\sum_{\substack{w\leqslant p\leqslant R^{\log R}\\ p\notin\mathbf{P}_b}}\frac{\log^{O(1)}p}{p}\,\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}1_{p\in\mathbf{P}_{\vec{m}}}=o(1).$$

From (111) and (112), it will suffice to establish the bounds

$$\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}1_{p\in\mathbf{P}_{\vec{m}}}=O\Big(\frac{1}{p}\Big)+O\Big(\frac{1}{R^\varepsilon}\Big)$$

for any $w\leqslant p\leqslant R^{\log R}$ with $p\notin\mathbf{P}_b$ (note that $\log(R^{\log R})^{O(1)}=o(R^\varepsilon)$). By the triangle inequality, it suffices to show that

$$\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}1_{p|P_j(\vec{m})-P_{j'}(\vec{m})}=O\Big(\frac{1}{p}\Big)+O\Big(\frac{1}{R^\varepsilon}\Big)$$

for all $1\leqslant j<j'\leqslant J$.

Fix j and j'. Observe that the property $p|P_j(\vec{m})-P_{j'}(\vec{m})$ is periodic in each component of $\vec{m}$ of period p, and can thus meaningfully be defined for $\vec{m}\in F_p^D$. Applying Corollary C.3 (for $p\ll R^\varepsilon$) or Lemma C.4 (for $p\gg R^\varepsilon$), it will thus suffice to show the bound

$$\mathbf{E}_{m_1\in A_1,\dots,m_D\in A_D}1_{p|P_j(m_1,\dots,m_D)-P_{j'}(m_1,\dots,m_D)}=O\Big(\frac{1}{M}\Big)$$

for all subsets $A_1,\dots,A_D$ in F_p of size at least $M\geqslant 1$ for some M. But since $p\notin\mathbf{P}_b$, the polynomial $P_j-P_{j'}$ does not vanish modulo p, and the claim follows from Lemma D.3. □

We can improve the error term if the coefficients of the polynomials are not too large.

COROLLARY 11.2. (Polynomial forms condition, again) *Let $M, D, d, J \geqslant 0$ and $\varepsilon > 0$. Let $P_1, ..., P_J \in \mathbf{Z}[\mathbf{m}_1, ..., \mathbf{m}_D]$ be distinct polynomials of degree d with all coefficients of size at most W^M. Let $I \subset \mathbf{R}$ be any interval of length at least R^{4J+1}, and let $\Omega \subset \mathbf{R}^D$ be any convex body with inradius at least R^ε. Then*

$$\mathbf{E}_{x \in I \cap \mathbf{Z}, \vec{m} \in \Omega \cap \mathbf{Z}^D} \prod_{j \in [J]} \nu(x + P_j(\vec{m})) = 1 + o_{D,d,\varepsilon,J,M}(1).$$

Proof. Let $\mathbf{P}_b$ denote the set of all primes $w \leqslant p \leqslant R^{\log R}$ such that $p | P_j - P_{j'}$ for some $1 \leqslant j < j' \leqslant J$. Since $P_j - P_{j'}$ is non-zero, this p must then divide a non-zero difference of two of the coefficients of the P_j's, which is $O(W^M)$. Thus the total product of all such p is at most $O(W^{O(1)})$, and hence, by Lemma E.3, we have $\sum_{p \in \mathbf{P}_b} 1/p = o(1)$. The claim now follows from Theorem 11.1. □

From Corollary 11.2, the desired estimate (73) quickly follows.

12. The polynomial correlation condition

Now we use the estimates from §10 to prove the polynomial correlation condition (74). It will suffice to prove the following estimate.

THEOREM 12.1. (Polynomial correlation condition) *Let $B, D, D', D'', d, J, K, L \geqslant 0$ and $\varepsilon > 0$. For any $j \in [J]$, $k \in [K]$ and $l \in [L]$, let*

$$\vec{P}_j \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_{D''}]^D \quad \text{and} \quad Q_{j,k}, S_l \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_{D''}]^{D'}$$

be polynomials obeying the following conditions:

- *for any $1 \leqslant j < j' \leqslant d$ and $k \in [K]$, the vector-valued polynomials*

$$(\vec{P}_j, \vec{Q}_{j,k}) \quad \text{and} \quad (\vec{P}_{j'}, \vec{Q}_{j',k})$$

are not parallel;

- *the coefficients of $P_{j,d}$ and $S_{l,d'}$ are bounded in magnitude by W^B*;
- *the vector-valued polynomials $\vec{S}_l$ are distinct as l varies in $[L]$.*

Let $I \subset \mathbf{R}$ be any interval of length at least R^{4L+1}, let $\Omega \subset \mathbf{R}^D$ be a bounded convex body with inradius at least R^{8J+2}, and let $\Omega' \subset \mathbf{R}^{D'}$ and $\Omega'' \subset \mathbf{R}^{D''}$ have inradii at least R^ε. Suppose also that Ω'' is contained in the ball $B(0, R^B)$. Then

$$\begin{aligned} \mathbf{E}_{x \in I, \vec{n} \in \Omega' \cap \mathbf{Z}^{D'}, \vec{h} \in \Omega'' \cap \mathbf{Z}^{D''}} \Big(\prod_{k \in [K]} \mathbf{E}_{\vec{m} \in \Omega \cap \mathbf{Z}^D} \prod_{j \in [J]} \nu(x + \vec{P}_j(\vec{h}) \cdot \vec{m} + \vec{Q}_{j,k}(\vec{h}) \cdot \vec{n}) \Big) \\ \times \prod_{l \in [L]} \nu(x + \vec{S}_l(\vec{h}) \cdot \vec{n}) = 1 + o_{B,D,D',D'',d,J,K,L,\varepsilon}(1). \end{aligned} \tag{85}$$

Proof. We repeat the same strategy of proof as in the preceding section. We fix B, D, D', D'', d, J, K, L and ε, and allow implicit constants to depend on these parameters. Thus, for instance, the right-hand side of (85) is now simply $1+o(1)$.

We begin by fixing k, x, $\vec{h}$ and $\vec{n}$, and considering a single average

$$\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}\prod_{j\in[J]}\nu(x+\vec{P}_j(\vec{h})\cdot\vec{m}+\vec{Q}_{j,k}(\vec{h})\cdot\vec{n}).$$

By Proposition 10.1, this average is

$$1_{\mathbf{P}_t[k,x,\vec{h},\vec{n}]=\varnothing}+o(1)+O\Big(\mathrm{Exp}\Big(O\Big(\sum_{p\in\mathbf{P}_b[k,x,\vec{h},\vec{n}]}\frac{1}{p}\Big)\Big)\Big), \tag{86}$$

where $\mathbf{P}_t[k,x,\vec{h},\vec{n}]$ is the collection of primes $w\leqslant p\leqslant R^{\log R}$ which are terrible with respect to the linear polynomials

$$W\times[x+\vec{P}_j(\vec{h})\cdot\vec{\mathbf{m}}+\vec{Q}_{j,k}(\vec{h})\cdot\vec{n}]+b\in\mathbf{Z}[\mathbf{m}_1,...,\mathbf{m}_D],\quad j\in[J], \tag{87}$$

and $\mathbf{P}_b[k,x,\vec{h},\vec{n}]$ is the collection of primes which are bad. We can thus express

$$\prod_{k\in[K]}\mathbf{E}_{\vec{m}\in\Omega\cap\mathbf{Z}^D}\prod_{j\in[J]}\nu(x+\vec{P}_j(\vec{h})\cdot\vec{m}+\vec{Q}_{j,k}(\vec{h})\cdot\vec{n})$$

using (76) as

$$1_{\mathbf{P}_t[k,x,\vec{h},\vec{n}]=\varnothing\text{ for all }k}+o(1)+O\Big(\mathrm{Exp}\Big(O\Big(\sum_{p\in\bigcup_{k\in[K]}\mathbf{P}_b[k,x,\vec{h},\vec{n}]}\frac{1}{p}\Big)\Big)\Big),$$

which we estimate crudely by

$$1+o(1)+O\Big(\sum_{k\in[K]}\sum_{p\in\mathbf{P}_t[k,x,\vec{h},\vec{n}]}1\Big)+O\Big(\mathrm{Exp}\Big(O\Big(\sum_{p\in\bigcup_{k\in[K]}\mathbf{P}_b[k,x,\vec{h},\vec{n}]\backslash\mathbf{P}_t[k,x,\vec{h},\vec{n}]}\frac{1}{p}\Big)\Big)\Big).$$

Observe that if $p\geqslant w$ is terrible for (87), then $p|\vec{P}_j(\vec{h})$ and $p|x+\vec{Q}_{j,k}(\vec{h})\cdot\vec{n}$ for some $j\in[J]$, while if $p\geqslant w$ is bad but not terrible, then

$$p\,|\,(\vec{P}_j(\vec{h}),\vec{Q}_{j,k}(\vec{h})\cdot\vec{n})\wedge(\vec{P}_{j'}(\vec{h}),\vec{Q}_{j',k}(\vec{h})\cdot\vec{n})$$

for some $1\leqslant j<j'\leqslant J$, where $\wedge$ denotes the wedge product on the $(D+1)$-dimensional space. Thus we may estimate the preceding sum (using (76)) by

$$\begin{aligned}1+o(1)+\sum_{j\in[J]}\sum_{k\in[K]}O\Big(\sum_{\substack{w\leqslant p\leqslant R^{\log R}\\ p|\vec{P}_j(\vec{h}),x+\vec{Q}_{j,k}(\vec{h})\cdot\vec{n}}}1\Big)^{1/2}\\ +\sum_{1\leqslant j<j'\leqslant J}O\Big(\mathrm{Exp}\Big(O\Big(\sum_{\substack{w\leqslant p\leqslant R^{\log R}\\ p|(\vec{P}_j(\vec{h}),\vec{Q}_{j,k}(\vec{h})\cdot\vec{n})\wedge(\vec{P}_{j'}(\vec{h}),\vec{Q}_{j',k}(\vec{h})\cdot\vec{n})}}\frac{1}{p}\Big)\Big)\Big)^{1/2}.\end{aligned}$$

At this point we pause to remove some "globally bad" primes. Let $\mathbf{P}_b$ denote the primes $w\leqslant p\leqslant R^{\log R}$ which divide $(\vec{P}_j,\vec{Q}_{j,k})\wedge(\vec{P}_{j'},\vec{Q}_{j'})$ for some $1\leqslant j<j'\leqslant J$ and $k\in[K]$ (note that this is now the wedge product in $D+D'$ dimensions). Because the wedge products $(\vec{P}_j,\vec{Q}_{j,k})\wedge(\vec{P}_{j'},\vec{Q}_{j'})$ are non-zero and have coefficients $O(W^{O(1)})$, the product of all these primes is $O(W^{O(1)})$, and hence, by Lemma E.3, we have $\sum_{p\in\mathbf{P}_b}1/p=o(1)$. Thus we may safely delete these primes from the expression inside the $\operatorname{Exp}(\cdot)$. If we then apply Lemma E.1, we can bound the above sum as

$$1+o(1)+\sum_{j\in[J]}O\Bigg(\sum_{\substack{w\leqslant p\leqslant R^{\log R}\\ p|\vec{P}_j(\vec{h}),x+\vec{Q}_{j,k}(\vec{h})\cdot\vec{n}}}1\Bigg)^{1/2}$$
$$+\sum_{1\leqslant j<j'\leqslant J}O\Bigg(\sum_{\substack{w\leqslant p\leqslant R^{\log R}\\ p\notin\mathbf{P}_b}}\frac{\log^{O(1)}p}{p}1_{p|(\vec{P}_j(\vec{h}),\vec{Q}_{j,k}(\vec{h})\cdot\vec{n})\wedge(\vec{P}_{j'}(\vec{h}),\vec{Q}_{j',k}(\vec{h})\cdot\vec{n})}\Bigg)^{1/2}.$$

Inserting this bound into (85) and using the Cauchy–Schwarz inequality, we reduce to showing the bounds

$$\mathbf{E}_{x\in I,\vec{n}\in\Omega',\vec{h}\in\Omega''}\prod_{l\in[L]}\nu(x+\vec{S}_l(\vec{h})\cdot\vec{n})=1+o(1),\tag{88}$$

$$\mathbf{E}_{x\in I,\vec{n}\in\Omega',\vec{h}\in\Omega''}\sum_{\substack{w\leqslant p\leqslant R^{\log R}\\ p|\vec{P}_j(\vec{h}),x+\vec{Q}_{j,k}(\vec{h})\cdot\vec{n}}}\prod_{l\in[L]}\nu(x+\vec{S}_l(\vec{h})\cdot\vec{n})=o(1),\tag{89}$$

$$\mathbf{E}_{x\in I,\vec{n}\in\Omega',\vec{h}\in\Omega''}\sum_{\substack{w\leqslant p\leqslant R^{\log R}\\ p\notin\mathbf{P}_b}}\frac{\log^{O(1)}p}{p}1_{p|(\vec{P}_j(\vec{h}),\vec{Q}_{j,k}(\vec{h})\cdot\vec{n})\wedge(\vec{P}_{j'}(\vec{h}),\vec{Q}_{j',k}(\vec{h})\cdot\vec{n})}\times\prod_{l\in[L]}\nu(x+\vec{S}_l(\vec{h})\cdot\vec{n})=o(1)\tag{90}$$

for all $1\leqslant j<j'\leqslant J$ and $k\in[K]$.

The bound (88) already follows from Corollary 11.2 and the hypotheses on $S_{l,d'}$. We now turn to (89). We rewrite the left-hand side as

$$\sum_{w\leqslant p\leqslant R^{\log R}}\mathbf{E}_{\vec{n}\in\Omega',\vec{h}\in\Omega''}\Bigg(1_{p|\vec{P}_j(\vec{h})}\mathbf{E}_{x\in I}\,1_{x=-\vec{Q}_{j,k}(\vec{h})\cdot\vec{n}\bmod p}\prod_{l\in[L]}\nu(x+\vec{S}_l(\vec{h})\cdot\vec{n})\Bigg).\tag{91}$$

Let us first consider the contributions of the primes p which are larger than R^{4L+1}. In this case, we bound ν extremely crudely by $O(R^2\log R)$ (taking absolute values in (72)), to bound the inner expectation of (91) by

$$O\Bigg(\frac{(R^2\log R)^L}{R^{4L+1}}\Bigg)=o(R^{-1/2})$$

(say), and to show that

$$\mathbf{E}_{\vec{h}\in\Omega''}\sum_{R^{4L}\leqslant p\leqslant R^{\log R}} 1_{p|\vec{P}_j(\vec{h})} \ll R^{1/2}.$$

But from the bounds on Ω'' and $\vec{h}$, we see that $\vec{P}_j(\vec{h})=O(R^{O(1)})$, and so at most $O(1)$ primes p can contribute to the sum for each $\vec{h}$. The claim follows.

Now we consider the contributions of the primes p between w and R^{4L+1}. We can then apply Corollary 10.7 and estimate the inner expectation of (91) by

$$\frac{1}{p}O\Big(\exp\Big(O\Big(\sum_{1\leqslant l<l'\leqslant L}\sum_{\substack{w\leqslant p'\leqslant R^{\log R}\\ p'|\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h})}}\frac{1}{p'}\Big)\Big)\Big),$$

which, by Lemma E.1, can be bounded by

$$O\Big(\frac{1}{p}\Big)+\sum_{1\leqslant l<l'\leqslant L}\sum_{w\leqslant p'\leqslant R^{\log R}}\frac{\log^{O(1)}p'}{pp'}O(1_{p'|\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h})}).$$

The contribution to (91) can thus be bounded by the sum of

$$\sum_{w\leqslant p\leqslant R^{\log R}}\frac{1}{p}O(\mathbf{E}_{\vec{h}\in\Omega''}1_{p|\vec{P}_j(\vec{h})})$$

and

$$\sum_{1\leqslant l<l'\leqslant L}\sum_{w\leqslant p,p'\leqslant R^{\log R}}\frac{\log^{O(1)}p'}{pp'}\mathbf{E}_{\vec{h}\in\Omega''}1_{p|\vec{P}_j(\vec{h})}1_{p'|\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h})}.$$

Now, by hypothesis, the vector-valued polynomials $\vec{P}_j$ and $\vec{S}_l-\vec{S}_{l'}$ are non-zero. Thus, by Lemma D.3,

$$\mathbf{E}_{\vec{h}\in A_1\times\ldots\times A_{D''}}1_{p'|\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h})}\ll\frac{1}{M}$$

whenever $A_1,\ldots,A_{D''}\subset F_{p'}$ have cardinality at least M. Applying Corollary C.3 and Lemma C.4, we conclude that

$$\mathbf{E}_{\vec{h}\in\Omega''}1_{p'|\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h})}\ll\frac{1}{p'}+\frac{1}{R^{\varepsilon}}.$$

A similar argument gives

$$\mathbf{E}_{\vec{h}\in\Omega''}1_{p|\vec{P}_j(\vec{h})}\ll\frac{1}{p}+\frac{1}{R^{\varepsilon}},$$

and hence, by the Cauchy–Schwarz inequality,

$$\mathbf{E}_{\vec{h}\in\Omega''}1_{p|\vec{P}_j(\vec{h})}1_{p'|\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h})}\ll\frac{1}{(pp')^{1/2}}+\frac{1}{R^{\varepsilon/2}}.$$

Applying all of these bounds, we can thus bound the total contribution of this case to (91) by

$$\sum_{w\leqslant p\leqslant R^{\log R}} \frac{1}{p}\Big(O\Big(\frac{1}{p}\Big)+O\Big(\frac{1}{R^{\varepsilon}}\Big)\Big)+\sum_{w\leqslant p,p'\leqslant R^{\log R}} \frac{\log^{O(1)} p'}{pp'}\Big(O\Big(\frac{1}{(pp')^{1/2}}\Big)+O\Big(\frac{1}{R^{\varepsilon/2}}\Big)\Big),$$

which is $o(1)$, by (111) and (112).

Finally, we consider (90). We first apply Theorem 11.1 to bound

$$\mathbf{E}_{x\in I}\prod_{l\in[L]}\nu(x+\vec{S}_l(\vec{h})\cdot\vec{n})=O\Big(\exp\Big(O\Big(\sum_{1\leqslant l<l'\leqslant L}\ \sum_{\substack{w\leqslant p'\leqslant R^{\log R}\\ p'|(\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h}))\cdot\vec{n}}}\frac{1}{p'}\Big)\Big)\Big).$$

Once again, we must extract the "globally bad" primes. Let Π'_b denote all the primes $w\leqslant p'\leqslant R^{\log R}$ which divide $\vec{S}_l-\vec{S}_{l'}$ for some $1\leqslant l<l'\leqslant L$. Since these polynomials are non-zero and have coefficients $O(W^{O(1)})$, the product of all the primes in Π'_b is $O(W^{O(1)})$, and hence, by Lemma E.3, as before these primes contribute only $o(1)$ and can be discarded. If we then apply Lemma E.1, we can bound the preceding expression by

$$O(1)+\sum_{1\leqslant l<l'\leqslant L}\ \sum_{\substack{w\leqslant p'\leqslant R^{\log R}\\ p'\notin\Pi'_b}}\frac{\log^{O(1)} p'}{p'}\,O(1_{p'|(\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h}))\cdot\vec{n}}).$$

Thus, to prove (90), it suffices to show the estimates

$$\sum_{\substack{w\leqslant p\leqslant R^{\log R}\\ p\notin\mathbf{P}_b}}\frac{\log^{O(1)} p}{p}\,\mathbf{E}_{\vec{n},\vec{h}}1_{p|(\vec{P}_j(\vec{h}),\vec{Q}_{j,k}(\vec{h})\cdot\vec{n})\wedge(\vec{P}_{j'}(\vec{h}),\vec{Q}_{j',k}(\vec{h})\cdot\vec{n})}=o(1)$$

and

$$\sum_{\substack{w\leqslant p,p'\leqslant R^{\log R}\\ p\notin\mathbf{P}_b\\ p'\notin\Pi'_b}}\frac{\log^{O(1)} p\log^{O(1)} p'}{pp'}\times\mathbf{E}_{\vec{n},\vec{h}}1_{p|(\vec{P}_j(\vec{h}),\vec{Q}_{j,k}(\vec{h})\cdot\vec{n})\wedge(\vec{P}_{j'}(\vec{h}),\vec{Q}_{j',k}(\vec{h})\cdot\vec{n})}1_{p'|(\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h}))\cdot\vec{n}}=o(1)$$

for all $1\leqslant j<j'\leqslant J$ and all $1\leqslant l<l'\leqslant L$, where $\vec{n}$ and $\vec{h}$ are averaged over $\Omega'\cap\mathbf{Z}^{D'}$ and $\Omega''\cap\mathbf{Z}^{D''}$, respectively. Applying the Cauchy–Schwarz inequality to the expectation in the latter estimate and then factorizing the double sum, we see that it will suffice to show that

$$\sum_{\substack{w\leqslant p\leqslant R^{\log R}\\ p\notin\mathbf{P}_b}}\frac{\log^{O(1)} p}{p}(\mathbf{E}_{\vec{n},\vec{h}}1_{p|(\vec{P}_j(\vec{h}),\vec{Q}_{j,k}(\vec{h})\cdot\vec{n})\wedge(\vec{P}_{j'}(\vec{h}),\vec{Q}_{j',k}(\vec{h})\cdot\vec{n})})^{1/2}=o(1)\qquad(92)$$

and

$$\sum_{\substack{w\leqslant p'\leqslant R^{\log R}\\ p\notin\prod_{b'}}} \frac{\log^{O(1)} p'}{p'}(\mathbf{E}_{\vec{n},\vec{h}} 1_{p|(\vec{S}_l(\vec{h})-\vec{S}_{l'}(\vec{h}))\cdot\vec{n}})^{1/2} = o(1). \tag{93}$$

Since $p\notin\mathbf{P}_b$, we observe from Lemma D.3, Corollary C.3 and Lemma C.4 that

$$\mathbf{E}_{\vec{h}} 1_{p|(\vec{P}_j(\vec{h}),\vec{Q}_{j,k}(\vec{h})\cdot\vec{n})\wedge(\vec{P}_{j'}(\vec{h}),\vec{Q}_{j',k}(\vec{h})\cdot\vec{n})} \ll \frac{1}{p}+\frac{1}{R^\varepsilon},$$

and the claim (92) now follows from (111) and (112). The estimate (93) is proven similarly. This (finally!) completes the proof of Theorem 12.1. □

Appendix A. Local Gowers uniformity norms

In this appendix we shall collect a number of elementary inequalities based on the Cauchy–Schwarz inequality, including several related to Gowers-type uniformity norms.

The formulation of the Cauchy–Schwarz inequality which we shall rely on is

$$|\mathbf{E}_{a\in A,b\in B} f(a)g(a,b)|^2 \leqslant (\mathbf{E}_{a\in A}F(a))(\mathbf{E}_{a\in A}F(a)|\mathbf{E}_{b\in B}g(a,b)|^2) \tag{94}$$

whenever $f\colon A\to\mathbf{R}$, $F\colon A\to\mathbf{R}^+$ and $g\colon A\times B\to\mathbf{R}$ are functions on non-empty finite sets A and B with the pointwise bound $|f|\leqslant F$.

A well-known consequence of the Cauchy–Schwarz inequality is the *van der Corput lemma*, which allows one to estimate a coarse-scale average of a function f by coarse-scale averages of "derivatives" of f over short scales. Here is the precise formulation we need.

LEMMA A.1. (van der Corput) *Let N, M and H be as in §2. Let $\{x_m\}_{m\in\mathbf{Z}}$ be a sequence of real numbers obeying the bound*

$$x_m \ll_\varepsilon N^\varepsilon \tag{95}$$

for any $\varepsilon>0$ and $m\in\mathbf{Z}$. Then

$$\mathbf{E}_{m\in[M]}x_m = \mathbf{E}_{h\in[H]}\mathbf{E}_{m\in[M]}x_{m+h}+o(1) \tag{96}$$

and

$$|\mathbf{E}_{m\in[M]}x_m|^2 \ll \mathbf{E}_{h,h'\in[H]}\mathbf{E}_{m\in[M]}x_{m+h}x_{m+h'}+o(1). \tag{97}$$

Proof. From (95) we see that

$$\mathbf{E}_{m\in[M]}x_m = \mathbf{E}_{m\in[M]}x_{m+h}+o(1)$$

for all $h\in[H]$; averaging over all h and rearranging, we obtain (96). Applying (94), we conclude that

$$|\mathbf{E}_{m\in[M]}x_m| \ll \mathbf{E}_{m\in[M]}|\mathbf{E}_{h\in[H]}x_{m+h}|^2+o(1)$$

and (97) follows. □

We will use Lemma A.1 only in one place, namely in Proposition 5.14, which is the key inductive step needed to estimate a polynomial average by a collection of linear averages.

Next, we recall some Cauchy–Schwarz–Gowers inequalities, which can be found for instance in [19, Appendix B]. Let X be a finite non-empty set. If A is a finite set and $f: X^A \to \mathbf{R}$, define the *Gowers box norm* $\|f\|_{\square^A}$ as

$$\|f\|_{\square^A} := \Big(\mathbf{E}_{m^{(0)},m^{(1)}\in X^A} \prod_{\omega\in\{0,1\}^A} f((m_\alpha^{(\omega_\alpha)})_{\alpha\in A})\Big)^{1/2^{|A|}}, \tag{98}$$

where $\omega=(\omega_\alpha)_{\alpha\in A}$ and $m^{(j)}=(m_\alpha^{(j)})_{\alpha\in A}$ for $j=0,1$. This is indeed a norm([35]) for $|A|\geqslant 2$. It obeys the *Cauchy–Schwarz–Gowers inequality*

$$\Big|\mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in X^A} \prod_{\omega\in\{0,1\}^A} f_\omega((m_\alpha^{(\omega_\alpha)})_{\alpha\in A})\Big| \leqslant \prod_{\omega\in\{0,1\}^A} \big|\|f_\omega\|_{\square^A}\big|. \tag{99}$$

We shall also need a weighted variant of this inequality.

PROPOSITION A.2. (Weighted generalized von Neumann inequality) *Let A be a non-empty finite set, and let $f: X^A\to\mathbf{R}$ be a function. For every $\alpha\in A$, let $f_\alpha: X^{A\setminus\{\alpha\}}\to\mathbf{R}$ and $\nu_\alpha: X^{A\setminus\{\alpha\}}\to\mathbf{R}^+$ be functions with the pointwise bound $|f_\alpha|\leqslant\nu_\alpha$. Then we have*

$$\Big|\mathbf{E}_{\vec{m}\in X^A} f(\vec{m}) \prod_{\alpha\in A} f_\alpha(\vec{m}|_{A\setminus\{\alpha\}})\Big| \leqslant \big|\|f\|_{\square^A(\nu)}\big| \prod_{\alpha\in A} \|\nu_\alpha\|_{\square^{A\setminus\{\alpha\}}}^{1/2},$$

where $\vec{m}|_{A\setminus\{\alpha\}}$ is the restriction of $\vec{m}\in X^A$ to $X^{A\setminus\{\alpha\}}$, and $\|f\|_{\square^A(\nu)}$ is the weighted Gowers box norm of f, defined by the formula

$$\|f\|_{\square^A(\nu)}^{2^{|A|}} := \mathbf{E}_{\vec{m}^{(0)},\vec{m}^{(1)}\in X^A}\Big(\prod_{\omega\in\{0,1\}^A} f((m_\alpha^{(\omega_\alpha)})_{\alpha\in A})\Big) \times \prod_{\alpha\in A}\prod_{\omega^{(\alpha)}\in\{0,1\}^{A\setminus\{\alpha\}}} \nu_\alpha((m_\beta^{(\omega_\beta^{(\alpha)})})_{\beta\in A\setminus\{\alpha\}}).$$

Proof. This is a special case of [19, Corollary B.4], in which all functions f_B and ν_B associated with subsets B of A of cardinality $|A|-2$ or less are set equal to 1. □

The local Gowers norm $U_{\sqrt{M}}^{a_1,\dots,a_d}$ defined in (22) is related to the above Gowers box norms by the obvious identity

$$\|f\|_{U_{\sqrt{M}}^{a_1,\dots,a_d}} = (\mathbf{E}_{x\in X}\|F_x\|_{\square^d}^{2^d})^{1/2^d} \tag{100}$$

([35]) If $|A|=0$ then $\|f\|_{\square^A}=f(\varnothing)$, while if $|A|=1$ then $\|f\|_{\square^A}=|\mathbf{E}f|$.

for any $f: X \to \mathbf{R}$, where for each $x \in X$ the function $F_x: [\sqrt{M}]^d \to \mathbf{R}$ is defined by

$$F_x(m_1, ..., m_d) := f(x + a_1 m_1 + ... + a_d m_d).$$

In particular, since $\square^d$ is a norm for $d \geqslant 2$, we easily verify from Minkowski's inequality that $U^{a_1,...,a_d}_{\sqrt{M}}$ is a norm also when $d \geqslant 2$. This in turn implies that the averaged local Gowers norms $U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}$ are also indeed norms.

Now we introduce the concept of concatenation of two or more averaged local Gowers norms. If $\vec{Q} \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{W}]^d$ and $\vec{Q}' \in \mathbf{Z}[\mathbf{h}'_1, ..., \mathbf{h}'_{t'}, \mathbf{W}]^{d'}$ are a d-tuple and d'-tuple of polynomials, respectively, we define the *concatenation*

$$\vec{Q} \oplus \vec{Q}' \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{h}'_1, ..., \mathbf{h}'_t, \mathbf{W}]^{d+d'}$$

to be the $(d+d')$-tuple of polynomials whose first d components are those of $\vec{Q}$ (using the obvious embedding of $\mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{W}]$ into $\mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{h}'_1, ..., \mathbf{h}'_{t'}, \mathbf{W}]$) and the last d' components are those of $\vec{Q}'$ (using the obvious embedding of $\mathbf{Z}[\mathbf{h}'_1, ..., \mathbf{h}'_{t'}, \mathbf{W}]$ into $\mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{h}'_1, ..., \mathbf{h}'_t, \mathbf{W}]$). One can similarly define the concatenation of more than two tuples of polynomials in the obvious manner.

The key lemma concerning concatenation is as follows.

LEMMA A.3. (Domination lemma) *Let $k \geqslant 1$. For each $j \in [k]$, let $t_j \geqslant 0$ be an integer and $\vec{Q}_j \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_{t_j}, \mathbf{W}]^{d_j}$ be a polynomial. Let $t := t_1 + ... + t_k$, $d := d_1 + ... + d_k$, and let $\vec{Q} \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, \mathbf{W}]^d$ be the concatenation of all the $\vec{Q}_j$. Then we have*

$$\|g\|_{U^{\vec{Q}_j([H]^{t_j},W)}_{\sqrt{M}}} \leqslant \|g\|_{U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}}$$

for all $j \in [k]$ and all $g: X \to \mathbf{R}$.

Proof. By induction we may take $k = 2$. By symmetry it thus suffices to show that

$$\|g\|_{U^{\vec{Q}([H]^t,W)}_{\sqrt{M}}} \leqslant \|g\|_{U^{\vec{Q} \oplus \vec{Q}'([H]^{t+t'},W)}_{\sqrt{M}}}$$

for any $g: X \to \mathbf{R}$ and any $\vec{Q} \in \mathbf{Z}[\mathbf{h}_1, ..., \mathbf{h}_t, W]^d$ and $\vec{Q}' \in \mathbf{Z}[\mathbf{h}'_1, ..., \mathbf{h}'_{t'}, W]^{d'}$. We may take $d' \geqslant 1$ as the case $d' = 0$ is trivial. From (105) and Hölder's inequality, it suffices to prove the estimate

$$\|g\|_{U^{a_1,...,a_d}_{\sqrt{M}}} \leqslant \|g\|_{U^{a_1,...,a_d,a'_1,...,a'_{d'}}_{\sqrt{M}}}$$

for all $a_1, ..., a_d, a'_1, ..., a'_d \in \mathbf{Z}$. Applying (100), we see that it suffices to prove the monotonicity formula([36])

$$\|f\|_{\square^d([\sqrt{M}])} \leqslant \|f\|_{\square^{d+d'}([\sqrt{M}])}$$

([36]) This is of course closely connected with the monotonicity of the Gowers U^d norms, noted for instance in [18].

for any $f\colon [\sqrt{M}]^d \to \mathbf{R}$, where we extend f to $[\sqrt{M}]^{d+d'}$ by adding d' dummy variables. Thus

$$f(m_1, ..., m_d, m_{d+1}, ..., m_{d+d'}) := f(m_1, ..., m_d).$$

But this easily follows by raising both sides to the power 2^d and using the Cauchy–Schwarz–Gowers inequality (99) for the $\square^{d+d'}$ norm (setting 2^d factors equal to f, and the other $2^{d+d'} - 2^d$ factors equal to 1). □

Appendix B. The uniform polynomial Szemerédi theorem

In this appendix we use the Furstenberg correspondence principle and the Bergelson–Leibman theorem [6] to prove the quantitative polynomial Szemerédi theorem, Theorem 3.2. The arguments here are reminiscent of those in [5]; see also [35] for another argument in a similar spirit.

Firstly, observe that to prove Theorem 3.2 it certainly suffices to do so in the case when g is an indicator function 1_E, since in the general case one can obtain a lower bound $g \geqslant \frac{1}{2}\delta 1_E$, where $E := \{x \in X : g(x) \geqslant \frac{1}{2}\delta\}$, which must have measure at least $\frac{1}{2}\delta - o(1)$.

Fix $P_1, ..., P_k$ and δ, and suppose by contradiction that Theorem 3.2 failed. Then (by the axiom of choice([37])) we can find a sequence of N going to infinity, and a sequence of indicator functions $1_{E_N}\colon \mathbf{Z}/N\mathbf{Z} \to \mathbf{R}$ of density

$$\frac{|E_N|}{N} \geqslant \delta - o(1) \tag{101}$$

such that

$$\lim_{N\to\infty} \mathbf{E}_{m\in[M]} \int_{\mathbf{Z}/N\mathbf{Z}} \prod_{j\in[k]} T_N^{P_j(Wm)/W} 1_{E_N} = 0,$$

where T_N is the shift on $\mathbf{Z}/N\mathbf{Z}$ and N is always understood to lie along the sequence (recall that M and W both depend on N).

The next step is to use an averaging argument (dating back to Varnavides [38]) to deal with the fact that M is growing rather rapidly in N. Let $B \geqslant 1$ be an integer, then

([37]) It is not difficult to rephrase this argument so that the axiom of choice is not used; we leave the details to the interested reader. The weak sequential compactness of probability measures which we need later in this section can also be established by an Arzelà–Ascoli type diagonalization argument which avoids the axiom of choice. On the other hand, the only known proof of the multi-dimensional Bergelson–Leibman theorem does use the axiom of choice, and so the main result of this paper also currently requires this axiom. However, it is expected that the Bergelson–Leibman theorem (and hence our result also) will eventually be provable by means which avoid using this axiom. For instance, for the 1-dimensional Bergelson–Leibman theorem one can use the Gowers–Host–Kra seminorm characteristic factors as in [9], which do not require choice.

for N sufficiently large we have

$$\mathbf{E}_{m\in[M]}\int_{\mathbf{Z}/N\mathbf{Z}}\prod_{j\in[k]}T_N^{P_j(Wm)/W}1_{E_N}\leqslant\frac{1}{B^3}$$

and hence

$$\mathbf{E}_{m\in[M/B]}\int_{\mathbf{Z}/N\mathbf{Z}}\prod_{j\in[k]}T_N^{P_j(Wbm)/W}1_{E_N}\ll\frac{1}{B^2}$$

for all $b\in[B]$. In particular

$$\mathbf{E}_{m\in[M/B]}\sum_{b=1}^{B}\int_{\mathbf{Z}/N\mathbf{Z}}\prod_{j\in[k]}T_N^{P_j(Wbm)/W}1_{E_N}\ll\frac{1}{B},$$

and hence by the pigeonhole principle (and the axiom of choice) we can find $m_N\in[M/B]$ for all sufficiently large N such that

$$\sum_{b=1}^{B}\int_{\mathbf{Z}/N\mathbf{Z}}\prod_{j\in[k]}T_N^{P_j(Wbm_N)/W}1_{E_N}\ll\frac{1}{B}$$

and hence

$$\lim_{N\to\infty}\int_{\mathbf{Z}/N\mathbf{Z}}\prod_{j\in[k]}T_N^{P_j(Wbm_N)/W}1_{E_N}=0$$

for each $b\geqslant 1$.

We now eliminate the W and m_N dependences by "lifting" the 1-dimensional shift to several dimensions. Let d be the maximum degree of the $P_1,...,P_k$, then we may write

$$\frac{P_j(Wbm_N)}{W}=\sum_{h\in[d]}W^{h-1}m_N^h b^h c_{j,h}$$

for some integer constants $c_{j,h}$. Thus, if we set $T_{N,h}:=T_N^{W^{h-1}m_N^h}$, we have

$$\lim_{N\to\infty}\mathbf{E}_{m\in[M]}\int_{\mathbf{Z}/N\mathbf{Z}}\prod_{j\in[k]}\Big(\prod_{h\in[d]}T_{N,h}^{c_{j,h}b^h}\Big)1_{E_N}=0\quad\text{for all }\ b\geqslant 1.\tag{102}$$

Now we use the Furstenberg correspondence principle to take a limit. Let Ω be the product space $\Omega:=\{0,1\}^{\mathbf{Z}^d}$, endowed with the usual product σ-algebra and with the standard commuting shifts $T_1,...,T_d$ defined by

$$T_h((\omega_n)_{n\in\mathbf{Z}^d}):=(\omega_{n-e_h})_{n\in\mathbf{Z}^d}\quad\text{for }\ h\in[d],$$

where $e_1, ..., e_d$ is the standard basis for $\mathbf{Z}_d$. We can define a probability measure μ_N on this space by $\mu_N := \mathbf{E}_{x \in \mathbf{Z}/N\mathbf{Z}} \mu_{N,x}$, where $\mu_{N,x}$ is the Dirac measure at the point

$$(1_{T_1^{n_1} \dots T_d^{n_d} x \in E_N})_{n \in \mathbf{Z}^d}.$$

One easily verifies that μ_N is invariant under the commuting shifts $T_1, ..., T_d$. Also if we let $A \subset \Omega$ be the cylinder set

$$A := \{(\omega_n)_{n \in \mathbf{Z}^d} : \omega_0 = 1\},$$

then we see from (101) and (102) that $\mu_N(A) \geqslant \delta - o(1)$ and

$$\lim_{N \to \infty} \mu_N\Big(\bigcap_{j \in [k]} \Big(\prod_{h \in [d]} T_h^{c_{j,h} b^h} \Big) A \Big) = 0$$

for all $b \geqslant 1$. By weak sequential compactness, we may after passing to a subsequence assume that the measures μ_N converge weakly to another probability measure μ on Ω, which is thus translation-invariant and obeys the bounds

$$\mu(A) \geqslant \delta > 0$$

and

$$\mu\Big(\bigcap_{j \in [k]} \Big(\prod_{h \in [d]} T_h^{c_{j,h} b^h} \Big) A \Big) = 0 \quad \text{for all } b \geqslant 1.$$

But this contradicts the multidimensional Bergelson–Leibman recurrence theorem [6, Theorem A_0]. This contradiction concludes the proof of Theorem 3.2.

Appendix C. Elementary convex geometry

In this paper we shall frequently be averaging over sets of the form $\Omega \cap \mathbf{Z}^D$, where $\Omega \subset \mathbf{R}^D$ is a convex body. It is thus of interest to estimate the size of such sets. Fortunately we will be able to do this using only very crude estimates (we only need the main term in the asymptotics, and do not need the deeper theory of error estimates). We shall bound the geometry of Ω using the inradius $r(\Omega)$; this is more or less dual to the approach in [19], which uses instead the circumradius.

Observe that the cardinality of $\Omega \cap \mathbf{Z}^D$ equals the Lebesgue measure of the Minkowski sum $(\Omega \cap \mathbf{Z}^D) + \big[-\frac{1}{2}, \frac{1}{2}\big]^D$ of $\Omega \cap \mathbf{Z}^D$ with the unit cube $\big[-\frac{1}{2}, \frac{1}{2}\big]^D$. The latter set differs from Ω only on the $\sqrt{D}/2$-neighborhood $\mathcal{N}_{\sqrt{D}/2}(\partial\Omega)$ of the boundary $\partial\Omega$. We thus have the *Gauss bound*

$$|\Omega \cap \mathbf{Z}^D| = \mathrm{mes}(\Omega) + O(\mathrm{mes}(\mathcal{N}_{\sqrt{D}/2}(\partial\Omega))),$$

where $\mathrm{mes}(\cdot)$ denotes Lebesgue measure. By dilation and translation, we thus have

$$|\Omega\cap(m\cdot\mathbf{Z}^D+a)| = m^{-D}[\mathrm{mes}(\Omega)+O(\mathrm{mes}(\mathcal{N}_{m\sqrt{D}/2}(\partial\Omega)))] \tag{103}$$

for any $m>0$ and $a\in\mathbf{R}^D$.

Now we estimate the boundary term in terms of the inradius $r(\Omega)$ of Ω.

LEMMA C.1. (Gauss bound) *Suppose that $\Omega\subset\mathbf{R}^D$ is a convex body. Then for any $0<r<r(\Omega)$ we have*

$$\mathrm{mes}(\mathcal{N}_r(\partial\Omega)) \ll_D \frac{r}{r(\Omega)}\mathrm{mes}(\Omega).$$

Proof. We may rescale $r(\Omega)=1$ (so $0<r<1$), and translate so that Ω contains the open unit ball $B(0,1)$. Elementary convex geometry then shows that for a sufficiently large constant $C_D>0$, we have

$$B(x,r)\subset\Omega \quad \text{whenever } x\in(1-C_Dr)\cdot\Omega$$

and

$$B(x,r)\cap\Omega=\varnothing \quad \text{whenever } x\notin(1+C_Dr)\cdot\Omega.$$

This shows that

$$\mathcal{N}_r(\partial\Omega)\subset[(1+C_Dr)\cdot\Omega]\setminus[(1-C_Dr)\cdot\Omega]$$

and the claim follows. □

From Lemma C.1 and (103) we conclude that

$$|\Omega\cap(m\cdot\mathbf{Z}^D+a)| = \left(1+O_D\left(\frac{m}{r(\Omega)}\right)\right)m^{-D}\mathrm{mes}(\Omega), \tag{104}$$

whenever $0<m\leqslant r(\Omega)$ and $a\in\mathbf{Z}^D$. As a corollary we obtain the following result.

COROLLARY C.2. (Equidistribution of residue classes) *Let $m\geqslant 1$ be an integer, $a\in\mathbf{Z}^D_m$ and $\Omega\subset\mathbf{R}^D$ be a convex body. If $r(\Omega)\geqslant C_Dm$ for some sufficiently large constant $C_D>0$, then we have*

$$\mathbf{E}_{x\in\Omega\cap\mathbf{Z}^D}1_{x\in m\cdot\mathbf{Z}^D+a} = \left(1+O_D\left(\frac{m}{r(\Omega)}\right)\right)m^{-D}.$$

This lets us average m-periodic functions on convex bodies as follows.

COROLLARY C.3. (Averaging lemma) *Let $m\geqslant 1$ be an integer, and let $f:\mathbf{Z}^D\to\mathbf{R}^+$ be a non-negative m-periodic function (and thus f can also be identified with a function on $\mathbf{Z}^D_m$). Let $\Omega\subset\mathbf{R}^D$ be a convex body. If $r(\Omega)\geqslant C_Dm$ for some sufficiently large constant $C_D>0$, then we have*

$$\mathbf{E}_{x\in\Omega\cap\mathbf{Z}^D}f(x) = \left(1+O_D\left(\frac{m}{r(\Omega)}\right)\right)\mathbf{E}_{y\in\mathbf{Z}^D_m}f.$$

Proof. We expand the left-hand side as

$$\mathbf{E}_{x\in\Omega\cap\mathbf{Z}^D} f(x) = \sum_{y\in\mathbf{Z}_m^D} f(y)\mathbf{E}_{x\in\Omega\cap\mathbf{Z}^D} 1_{x\in m\cdot\mathbf{Z}^D+y}$$

and apply Corollary C.2. □

Corollary C.3 is no longer useful when the period m is large compared to the inradius $r(\Omega)$. In such cases we shall need to rely instead on the following cruder estimate.

LEMMA C.4. (Covering inequality) *Let $\Omega\subset\mathbf{R}^D$ be a convex body with $r(\Omega)>C_D$ for some large constant $C_D>1$, and let $f:\mathbf{Z}^D\to\mathbf{R}^+$ be an arbitrary function. Then*

$$\mathbf{E}_{x\in\Omega\cap\mathbf{Z}^D} f(x) \ll_D \sup_{y\in\mathbf{R}^D} \mathbf{E}_{x\in y+[-r(\Omega),r(\Omega)]^D\cap\mathbf{Z}^D} f(x).$$

Proof. From (104) we have that $|\Omega\cap\mathbf{Z}^D|\sim\mathrm{mes}(\Omega)$, so it suffices to show that

$$\sum_{x\in\Omega\in\mathbf{Z}^D} f(x) \ll_D \mathrm{mes}(\Omega) \sup_{y\in\mathbf{R}^D} \mathbf{E}_{x\in y+[-r(\Omega),r(\Omega)]^D\cap\mathbf{Z}^D} f(x).$$

This will follow if we can cover Ω by $O_D(\mathrm{mes}(\Omega)/r(\Omega)^D)$ translates of the cube

$$[-r(\Omega),r(\Omega)]^D.$$

By rescaling and translating, we reduce to verifying the following fact: if Ω is a convex body containing $B(0,1)$, then Ω can be covered by $O_D(\mathrm{mes}(\Omega))$ translates of $[-1,1]^D$. To see this, we use a covering argument of Ruzsa [27]. First observe that because the cube $\left[-\frac{1}{2},\frac{1}{2}\right]^D$ is contained in a dilate of $B(0,1)$ (and hence Ω) by $O_D(1)$, the Minkowski sum $\Omega+\left[-\frac{1}{2},\frac{1}{2}\right]^D$ is also contained in an $O_D(1)$-dilate of Ω and thus has volume $O_D(\mathrm{mes}(\Omega))$. Now let $x_1+\left[-\frac{1}{2},\frac{1}{2}\right]^D,...,x_N+\left[-\frac{1}{2},\frac{1}{2}\right]^D$ be a maximal collection of disjoint shifted cubes with $x_1,...,x_N\in\Omega$, then by the previous volume bound we have $N\ll_D\mathrm{mes}(\Omega)$. But by maximality we see that the cubes $x_1+[-1,1]^D,...,x_N+[-1,1]^D$ must cover Ω, and the claim follows. □

Appendix D. Counting points of varieties over F_p

Let R be an arbitrary ring, and let $P_1,...,P_J\in R[\mathbf{x}_1,...,\mathbf{x}_D]$ be polynomials. Our interest here is to control the "density" of the (affine) algebraic variety

$$\{(x_1,...,x_D)\in R^D : P_j(x_1,...,x_D)=0 \text{ for all } j\in[J]\},$$

and more precisely to estimate quantities such as

$$\mathbf{E}_{x_1\in A_1,\dots,x_D\in A_D}\prod_{j\in[J]}1_{P_j(x_1,\dots,x_D)=0} \tag{105}$$

for certain finite non-empty subsets $A_1,...,A_D\subset R$ (typically $A_1,...,A_D$ will either be all of R, or some arithmetic progression). We are particularly interested in the case when R is the finite field F_p, but in order to also encompass the case of the integers $\mathbf{Z}$ (and of polynomial rings over F_p or $\mathbf{Z}$), we shall start by working in the more general context of a unique factorization domain.

Of course, the proper way to do this would be to use the tools of modern algebraic geometry, for instance using the concepts of generic point and algebraic dimension of varieties. Indeed, the results in this appendix are "morally trivial" if one uses the fact that the codimension of an algebraic variety is preserved under restriction to generic subspaces. However, to keep the exposition simple we have chosen a very classical, pedestrian and elementary approach, to emphasize that the facts from algebraic geometry which we will need are not very advanced.

From the factor theorem (which is valid over any unique factorization domain) we have the following result.

LEMMA D.1. (Generic points of a one-dimensional polynomial) *Let $P\in R[\mathbf{x}]$ be a polynomial of one variable of degree at most d over a ring R. If $P\neq 0$, then $P(x)\neq 0$ for all but at most d values of $x\in R$.*

As a corollary, we obtain the following result.

COROLLARY D.2. (Generic points of a multi-dimensional polynomial) *Let*

$$P\in R[\mathbf{x}_1,...,\mathbf{x}_D]$$

be a polynomial in D variables of degree at most d over a ring R. If $P\neq 0$, then $P(\cdot,x_D)\neq 0$ for all but at most d values of $x_D\in R$, where $P(\cdot,x_D)\in F[\mathbf{x}_1,...,\mathbf{x}_{D-1}]$ is the polynomial in $D-1$ variables formed from P by replacing $\mathbf{x}_D$ by x_D.

Proof. View P as a 1-dimensional polynomial of $\mathbf{x}_D$ with coefficients in the ring $R[\mathbf{x}_1,...,\mathbf{x}_{D-1}]$ (which contains R), and apply Lemma D.1. □

As a consequence, we obtain a "baby combinatorial Nullstellensatz" (cf. [1]).

LEMMA D.3. (Baby Nullstellensatz) *Let $P\in R[\mathbf{x}_1,...,\mathbf{x}_D]$ be a polynomial of D variables of degree at most d over a ring R. Let $A_1,...,A_D$ be finite subsets of R with $|A_1|,...,|A_D|\geqslant M$ for some $M>0$. If $P\neq 0$, then*

$$\mathbf{E}_{x_1\in A_1,\dots,x_D\in A_D}1_{P(x_1,\dots,x_D)=0}\leqslant\frac{Dd}{M}\ll_{D,d}\frac{1}{M}.$$

Proof. We use induction over D. The case $D=0$ is vacuous. Now suppose that $D\geqslant 1$ and that the claim has already been proven for $D-1$. By Corollary D.2, we have $P(\cdot, x_D)\neq 0$ for all but at most d values of $x_D\in A_D$. The exceptional values of x_D can contribute at most d/M, while the remaining values of x_D will contribute at most $(D-1)d/M$ by the induction hypothesis. This completes the induction. □

This gives us a reasonable upper bound on the quantity (105) in the case $J=1$, which then trivially implies the same bound for $J>1$. However, we expect to do better than $1/M$ type bounds for higher J when the polynomials $P_1, ..., P_J$ are jointly coprime. To exploit the property of being coprime we recall the classical *resultant* of two polynomials.

Definition D.4. (Resultant) Let R be a ring, $d, d'\geqslant 1$, and

$$P = a_0 + a_1\mathbf{x} + ... + a_d\mathbf{x}^d \quad \text{and} \quad Q = b_0 + b_1 + ... + b_{d'}\mathbf{x}^{d'}$$

be two polynomials in $R[\mathbf{x}]$ of degree at most d and d', respectively. Then the *resultant* $\mathrm{Res}_{d,d'}(P,Q)\in R$ is defined to be the determinant of the $(d+d')\times(d+d')$ matrix whose rows are the coefficients in R of the polynomials $P,\ \mathbf{x}P, ..., \mathbf{x}^{d'-1}P,\ Q, \mathbf{x}Q, ..., \mathbf{x}^{d-1}Q$, with respect to the basis $1, \mathbf{x}, ..., \mathbf{x}^{d+d'-1}$.

More generally, if $D\geqslant 1,\ j\in[D],\ d_j, d'_j\geqslant 1$ and P and Q are two polynomials in $R[\mathbf{x}_1, ..., \mathbf{x}_D]$ with $\deg_{\mathbf{x}_j}(P)\leqslant d_j$ and $\deg_{\mathbf{x}_j}(Q)\leqslant d'_j$, then we define the resultant

$$\mathrm{Res}_{d_j,d'_j,\mathbf{x}_j}(P,Q)\in R[\mathbf{x}_1, ..., \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, ..., \mathbf{x}_D]$$

by viewing P and Q as 1-dimensional polynomials of $\mathbf{x}_j$ over the ring

$$R[\mathbf{x}_1, ..., \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, ..., \mathbf{x}_D]$$

and using the 1-dimensional resultant defined earlier.

Example D.5. If $d=d'=1$, then the resultant of $a+b\mathbf{x}$ and $c+d\mathbf{x}$ is $ad-bc$, and the resultant of $a(\mathbf{x}_1)+b(\mathbf{x}_1)\mathbf{x}_2$ and $c(\mathbf{x}_1)+d(\mathbf{x}_1)\mathbf{x}_2$ in the $\mathbf{x}_2$ variable is

$$a(\mathbf{x}_1)d(\mathbf{x}_1) - b(\mathbf{x}_1)c(\mathbf{x}_1).$$

Let $P, Q\in R[\mathbf{x}]$ have degrees d and d', respectively, for some $d, d'\geqslant 1$, where R is a unique factorization domain. By splitting the determinant into a matrix and its adjugate, we obtain an identity

$$\mathrm{Res}_{d,d'}(P,Q) = AP + BQ \tag{106}$$

for some polynomials $A, B\in R[\mathbf{x}]$ of degree at most $d'-1$ and $d-1$, respectively. Thus, if P and Q are irreducible and coprime, then the resultant cannot vanish by unique factorization. The same extends to higher dimensions, as we show next.

LEMMA D.6. *Let R be a unique factorization domain, let $j \in [D]$, and suppose that $P, Q \in \mathbf{R}[\mathbf{x}_1, ..., \mathbf{x}_D]$ are such that $\deg_{\mathbf{x}_j}(P) = d_j \geqslant 1$ and $\deg_{\mathbf{x}_j}(Q) = d'_j \geqslant 1$. If P and Q are irreducible and coprime, then $\mathrm{Res}_{d,d',\mathbf{x}_j}(P,Q) \neq 0$.*

Proof. View P and Q as 1-dimensional polynomials over the unique factorization domain $R[\mathbf{x}_1, ..., \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, ..., \mathbf{x}_D]$ and apply the preceding argument. □

LEMMA D.7. (Generic points of multiple polynomials) *Let $P_1, ..., P_J \in R[\mathbf{x}_1, ..., \mathbf{x}_D]$ have degrees at most d over a unique factorization domain R, and suppose that all the $P_1, ..., P_J$ are non-zero and jointly coprime. Then $P_1(\cdot, x_D), ..., P_J(\cdot, x_D)$ are non-zero and jointly coprime for all but at most $O_{J,d}(1)$ values of $x_D \in R$.*

Proof. By splitting each of the P_j's into factors, we may assume that all the P_j's are irreducible. By eliminating any two polynomials which are scalar multiples of each other, we may then assume that the P_j's are pairwise coprime. The claim is vacuous for $k < 2$, so it will suffice to verify the claim for $k = 2$.

Suppose first that P_1 is constant in $\mathbf{x}_D$. Then $P_1(\cdot, x_D) = P_1$ is irreducible, and the only way it can fail to be coprime to $P_2(\cdot, x_D)$ is if $P_2(\cdot, x_D)$ is a multiple of P_1. But we know that P_2 itself is not a multiple of P_1; viewing P_2 modulo P_1 as a polynomial of degree at most d in $\mathbf{x}_D$ over the ring $R[\mathbf{x}_1, ..., \mathbf{x}_{D-1}]/(P_1)$, we see from Lemma D.1 that the number of exceptional x_D is at most d.

A similar argument works if P_2 is constant in $\mathbf{x}_D$. So now we may assume that $\deg_{\mathbf{x}_D}(P_1) = d_1$ and $\deg_{\mathbf{x}_D}(P_2) = d_2$ for some $d_1, d_2 \geqslant 1$, which allows us to compute the resultant $\mathrm{Res}_{d_1,d_2,\mathbf{x}_D}(P_1, P_2) \in F[\mathbf{x}_1, ..., \mathbf{x}_{D-1}]$. By Lemma D.6, this resultant is non-zero; also, by definition, we see that the resultant has degree $O_d(1)$.

From (106) we see that if $P_1(\cdot, x_D)$ and $P_2(\cdot, x_D)$ have any common factor in $R[\mathbf{x}_1, ..., \mathbf{x}_{D-1}]$ (which we may assume to be irreducible), then this factor must also divide $\mathrm{Res}_{d_1,d_2,\mathbf{x}_D}(P_1, P_2)$. From degree considerations we see that there are at most $O_d(1)$ such factors. Let $Q \in R[\mathbf{x}_1, ..., \mathbf{x}_{D-1}]$ be one such possible factor. Since P_1 and P_2 are coprime, we know that Q cannot divide both P_1 and P_2; say it does not divide P_2. Then, by viewing P_2 modulo Q as a polynomial of y_D over $R[x_1, ..., x_{D-1}]/Q$, as before we see that there are at most d values of x_D for which Q divides $P_2(\cdot, x_D)$. Putting this all together, we obtain the claim. □

This gives us a variant of Lemma D.3.

LEMMA D.8. (Second baby Nullstellensatz) *Let $P_1, ..., P_J \in R[\mathbf{x}_1, ..., \mathbf{x}_D]$ be polynomials in D variables of degree at most d over a unique factorization domain R. Let $A_1, ..., A_D$ be finite subsets of F with $|A_1|, ..., |A_D| \geqslant M$ for some $M > 0$. If all the*

$P_1, ..., P_J$ are non-zero and jointly coprime, then

$$\mathbf{E}_{x_1 \in A_1, ..., x_D \in A_D} \prod_{j \in [J]} 1_{P_j(x_1, ..., x_D) \text{ non-zero, jointly coprime}} = 1 - O_{D,d,J}\left(\frac{1}{M}\right)$$

and

$$\mathbf{E}_{x_1 \in A_1, ..., x_D \in A_D} \prod_{j \in [J]} 1_{P_j(x_1, ..., x_D) = 0} \ll_{D,d,J} \frac{1}{M^2}.$$

Remark D.9. One could obtain sharper results by using Bezout's lemma, but the result here will suffice for our applications.

Proof. The first claim follows by repeating the proof of Lemma D.3 (replacing Corollary D.2 by Lemma D.7) and we leave it to the reader. To prove the second claim, we again use induction over D. The base case $D=0$ is again trivial, so assume $D \geqslant 1$ and that the claim has already been proven for $D-1$.

By Lemma D.7, we know that for all but $O_{J,d}(1)$ values of x_D the polynomials $P_1(\cdot, x_D), ..., P_J(\cdot, x_D)$ are all non-zero and jointly coprime. This case will contribute $O_{D,d,J}(1/M^2)$ by the induction hypothesis. Now consider one of the $O_{J,d}(1)$ exceptional values of x_D. For each such x_D, at least one of the polynomials $P_j(\cdot, x_D)$ has to be non-zero, otherwise $\mathbf{x}_D - x_D$ would be a common factor of $P_1, ..., P_J$, a contradiction. Applying Lemma D.3, we see that the contribution of each such x_D is thus $O_{D,d,J}(1/M^2)$, and the claim follows. □

We now specialize the above discussion to compute the local factors c_p and $\bar{c}_p$ defined in Definition 9.1. We first observe the following easy upper bounds.

LEMMA D.10. (Crude local bounds) *Let $P_1, ..., P_J \in \mathbf{Z}[\mathbf{x}_1, ..., \mathbf{x}_D]$ be polynomials of degree at most d, and let p be a prime.*

(i) *If all the $P_1, ..., P_J$ vanish identically modulo p, then*

$$c_p(P_1, ..., P_J) = 1.$$

(ii) *If at least one of $P_1, ..., P_J$ vanish identically modulo p, then*

$$\bar{c}_p(P_1, ..., P_J) = 0.$$

(iii) *If at least one of $P_1, ..., P_J$ is a non-zero constant modulo p, then*

$$c_p(P_1, ..., P_J) = 0.$$

(iv) *If at least one of $P_1, ..., P_J$ is non-constant modulo p, then*

$$c_p(P_1, ..., P_J) \ll_{d,D} \frac{1}{p}.$$

(v) *If the $P_1,...,P_J$ are jointly coprime modulo p, then*

$$c_p(P_1,...,P_J) \ll_{d,D} \frac{1}{p^2}.$$

Proof. (i), (ii) and (iii) are trivial, while (iv) and (v) follow from Lemmas D.3 and D.8, respectively (setting $A_1=...=A_D=R=F_p$). □

Now we can refine the bound for a single polynomial P in the case when P is linear in one variable, with linear and constant coefficients coprime.

LEMMA D.11. (Linear case) *Let $P \in \mathbf{Z}[\mathbf{x}_1,...,\mathbf{x}_D]$ have degree at most d, and let p be a prime. Suppose that $P \bmod p$ is linear in the $\mathbf{x}_j$ variable for some $j \in [d]$, and thus*

$$P(\mathbf{x}_1,...,\mathbf{x}_D) = P_1(\mathbf{x}_1,...,\mathbf{x}_{j-1},\mathbf{x}_{j+1},...,\mathbf{x}_D)\mathbf{x}_j + P_0(\mathbf{x}_1,...,\mathbf{x}_{j-1},\mathbf{x}_{j+1},...,\mathbf{x}_D) \bmod p$$

for some polynomials $P_0, P_1 \in F_p[\mathbf{x}_1,...,\mathbf{x}_{j-1},\mathbf{x}_{j+1},...,\mathbf{x}_D]$. Suppose also that the linear coefficient P_1 is non-zero and coprime to the constant coefficient P_0. Then

$$c_p(P) = \frac{1}{p} + O_{d,D}\left(\frac{1}{p^2}\right).$$

Proof. Let us split $F_p^{D-1} = A \cup B \cup C$, where A is the subset of F_p^{D-1} where $P_1 \neq 0$, B is the subset of F_p^{D-1} where $P_1=0$ and $P_2 \neq 0$, and C is the subset of F_p^{D-1} where $P_1=P_2=0$. Then an elementary counting argument shows that

$$c_p(P) = \frac{|A|+|C|p}{p^D} = \frac{1}{p} - \frac{|B|+|C|}{p^D} + \frac{|C|}{p^{D-1}}.$$

As P_1 is not zero, by Lemma D.10 (iv) we have $|B|+|C| \ll_d p^{D-2}$. Since P_1 and P_2 are coprime modulo p, by Lemma D.10 (v) we have $|C| \ll_d p^{D-2}$. The claim follows. □

We can now quickly prove Lemma 9.5.

Proof of Lemma 9.5. The claims (a), (b) and (d) follow from Lemma D.10 and Definition 9.4, while (c) follows from Lemma D.11 and Definition 9.4. The claim (e) is trivial, and the claim (f) follows from (a), (b) and (77). □

Appendix E. The distribution of primes

In this section we recall some classical results about the distribution of primes.

For $\mathrm{Re}\, s > 1$, define the *Riemann zeta function*

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

Our argument will be elementary enough that we will not need the meromorphic continuation of ζ to the region $\mathrm{Re}\, s \leqslant 1$. From the unique factorization of the positive integers, we have the *Euler product formula*

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}. \tag{107}$$

We also have the bounds

$$\zeta(s) = \frac{1}{s-1} + O(\log(2+|\mathrm{Im}\, s|)) \quad \text{and} \quad \frac{1}{\zeta(s)} = O(\log(2+|\mathrm{Im}\, s|)) \tag{108}$$

whenever $1 < \mathrm{Re}\, s < 10$ (see e.g. [37, Chapter 3]).

From the prime number theorem

$$\sum_{p<x} 1 = (1+o(1))\frac{x}{\log x}, \qquad \text{as } x \to \infty$$

(which, incidentally, can be deduced readily from (108)), and summation by parts, we easily obtain the estimates

$$\sum_{p<x} \log p = x + o(x), \quad \text{as } x \to \infty, \tag{109}$$

$$\sum_{p<x} \frac{1}{p} = \log\log(10+x) + O(1) \quad \text{for } x > 0, \tag{110}$$

$$\sum_{p<x} \frac{\log^K p}{p} \ll_K \log^K(10+x) \quad \text{for } K > 0 \text{ and } x > 0, \tag{111}$$

$$\left|\sum_{p>x} \frac{\log^K p}{p^s}\right| \ll_{K,s} \frac{\log^{K-1} x}{x^{\mathrm{Re}\, s - 1}} \quad \text{for } K \geqslant 0,\ x > 1 \text{ and } \mathrm{Re}\, s > 1. \tag{112}$$

In a similar spirit we have, whenever $1 < \mathrm{Re}\, s < 2$ and $x > 2$,

$$\begin{aligned}\left|\sum_{p>x} \log\left(1 - \frac{1}{p^s}\right)\right| &\ll \sum_{p>x} \frac{1}{p^{\mathrm{Re}\, s}} \ll \sum_{n=0}^{\infty} \sum_{2^n x < p \leqslant 2^{n+1} x} \frac{1}{2^{n \mathrm{Re}\, s} x^{\mathrm{Re}\, s}} \\ &\ll \sum_{n=0}^{\infty} \frac{1}{2^{n(\mathrm{Re}\, s - 1)} x^{\mathrm{Re}\, s - 1} \log x} \ll \frac{1}{(\mathrm{Re}\, s - 1) x^{\mathrm{Re}\, s - 1} \log x}.\end{aligned}$$

In particular, when $\mathrm{Re}\, s = 1 + 1/\log R$ and $x = R^{\log R}$, we have

$$\sum_{p > R^{\log R}} \log\left(1 - \frac{1}{p^s}\right) = o(1),$$

and hence, from (107),

$$\prod_{p \leqslant R^{\log R}} \left(1 - \frac{1}{p^s}\right)^{-1} = (1+o(1))\zeta(s) \quad \text{whenever } \operatorname{Re} s = 1 + \frac{1}{\log R}. \tag{113}$$

We will frequently encounter expressions of the form $\exp\bigl(K \sum_{p \in \mathbf{P}} 1/p\bigr)$, where p ranges over some set of primes (typically finite). Such sums can eventually be somewhat large, due to (110). Fortunately the very slow nature of the divergence of $\sum_p 1/p$ lets us estimate this exponential by a slowly divergent sum over primes, conceding only a few logarithmic factors.

LEMMA E.1. (Exponentials can be replaced by logarithms) *Let* $\mathbf{P}$ *be any set of primes, and let* $K \geqslant 1$. *Then*

$$\exp\left(K \sum_{p \in \mathbf{P}} \frac{1}{p}\right) \leqslant 1 + O_K\left(\sum_{p \in \mathbf{P}} \frac{\log^K p}{p}\right)$$

or equivalently

$$\operatorname{Exp}\left(K \sum_{p \in \mathbf{P}} \frac{1}{p}\right) \ll_K \sum_{p \in \mathbf{P}} \frac{\log^K p}{p}.$$

Remark E.2. Note that the sums are only over primes in $\mathbf{P}$, rather than products of primes in $\mathbf{P}$, which would have been the case if we had written

$$\exp\left(K \sum_{p \in \mathbf{P}} \frac{1}{p}\right) = \prod_{p \in \mathbf{P}} \exp\left(\frac{K}{p}\right) = \prod_{p \in \mathbf{P}} \left(1 + O_K\left(\frac{1}{p}\right)\right).$$

The fact that we keep the sum over primes is useful for applications, as it allows us to work over fields F_p rather than mere rings $\mathbf{Z}_N$ when performing certain local counting estimates. This lets us avoid certain technical issues involving zero divisors which would otherwise complicate the argument. The additional logarithmic powers of p are sometimes dangerous, but in several cases we will be able to acquire an additional factor of $1/p$ from an averaging argument, which will make the summation on the right-hand side safely convergent regardless of how many logarithms are present, due to (112).

Proof. Let us fix K and suppress the dependence of the $O(\cdot)$ notation on K. By a limiting argument we may take $\mathbf{P}$ to be finite. We expand the left-hand side as a power series

$$1 + \sum_{n=1}^{\infty} \frac{K^n}{n!} \sum_{p_1, \ldots, p_n \in \mathbf{P}} \frac{1}{p_1 \ldots p_n}.$$

By paying a factor of n, we may assume that p_n is greater than or equal to the other primes, thus bounding the previous expression by

$$1+\sum_{n=1}^{\infty}\frac{K^n}{(n-1)!}\sum_{p_n\in\mathbf{P}}\sum_{\substack{p_1,\dots,p_{n-1}\in\mathbf{P}\\ p_1,\dots,p_{n-1}\leqslant p_n}}\frac{1}{p_1\dots p_n}.$$

We rewrite p_n as p and rearrange this as

$$1+\sum_{p\in\mathbf{P}}\frac{1}{p}\sum_{n=1}^{\infty}\frac{K^n}{(n-1)!}\Big(\sum_{\substack{p'\in\mathbf{P}\\ p'\leqslant p}}\frac{1}{p'}\Big)^{n-1}.$$

From (110) we have

$$\sum_{\substack{p'\in\mathbf{P}\\ p'\leqslant p}}\frac{1}{p'}\leqslant\log\log(10+p)+O(1),$$

and so we can bound the previous expression by

$$1+\sum_{p\in\mathbf{P}}\frac{1}{p}\sum_{n=1}^{\infty}\frac{K(K\log\log(10+p)+O(1))^{n-1}}{(n-1)!}.$$

Summing the power series we obtain the result. □

Finally, we record a very simple lemma, using the quantities w and W defined in §8.2.

LEMMA E.3. (Divisor bound) *Let* $\mathbf{P}$ *be any collection of primes such that*

$$\prod_{p\in\mathbf{P}}p\leqslant MW^M$$

for some $M>0$. *Then*

$$\sum_{\substack{p\in\mathbf{P}\\ p\geqslant w}}\frac{1}{p}=o_M(1).$$

Proof. We trivially bound $1/p$ by $1/w$, and observe that the number of primes in $\mathbf{P}$ larger than w is at most $\log(MW^M)/\log w=o_M(\log W)$. But from (109) we have $\log W\ll w$. The claim follows. □

References

[1] Alon, N., Combinatorial Nullstellensatz. *Combin. Probab. Comput.*, 8 (1999), 7–29.

[2] Balog, A., Pelikán, J., Pintz, J. & Szemerédi, E., Difference sets without $\varkappa$th powers. *Acta Math. Hungar.*, 65 (1994), 165–187.

[3] Bateman, P. T. & Horn, R. A., A heuristic asymptotic formula concerning the distribution of prime numbers. *Math. Comp.*, 16 (1962), 363–367.

[4] Bergelson, V., Weakly mixing PET. *Ergodic Theory Dynam. Systems*, 7 (1987), 337–349.

[5] Bergelson, V., Host, B., McCutcheon, R. & Parreau, F., Aspects of uniformity in recurrence. *Colloq. Math.*, 84/85 (2000), 549–576.

[6] Bergelson, V. & Leibman, A., Polynomial extensions of van der Waerden's and Szemerédi's theorems. *J. Amer. Math. Soc.*, 9 (1996), 725–753.

[7] Deligne, P., La conjecture de Weil. I. *Inst. Hautes Études Sci. Publ. Math.*, 43 (1974), 273–307.

[8] — La conjecture de Weil. II. *Inst. Hautes Études Sci. Publ. Math.*, 52 (1980), 137–252.

[9] Frantzikinakis, N. & Kra, B., Polynomial averages converge to the product of integrals. *Israel J. Math.*, 148 (2005), 267–276.

[10] Furstenberg, H., Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J. Anal. Math.*, 31 (1977), 204–256.

[11] — *Recurrence in Ergodic Theory and Combinatorial Number Theory*. Princeton University Press, Princeton, NJ, 1981.

[12] Furstenberg, H. & Katznelson, Y., An ergodic Szemerédi theorem for commuting transformations. *J. Anal. Math.*, 34 (1978), 275–291 (1979).

[13] Goldston, D. A., Pints, J. & Yildirim, C. Y., Small gaps between primes. II. Preprint, 2008.

[14] Goldston, D. A. & Yildirim, C. Y., Higher correlations of divisor sums related to primes. I. Triple correlations. *Integers*, 3 (2003), A5, 66 pp.

[15] Gowers, W. T., A new proof of Szemerédi's theorem. *Geom. Funct. Anal.*, 11 (2001), 465–588.

[16] Green, B., On arithmetic structures in dense sets of integers. *Duke Math. J.*, 114 (2002), 215–238.

[17] Green, B. & Tao, T., An inverse theorem for the Gowers $U^3(G)$ norm. *Proc. Edinb. Math. Soc.*, 51 (2008), 73–153.

[18] — The primes contain arbitrarily long arithmetic progressions. *Ann. of Math.*, 167 (2008), 481–547.

[19] — Linear equations in primes. To appear in *Ann. of Math.*

[20] Host, B., Progressions arithmétiques dans les nombres premiers (d'après B. Green et T. Tao). *Astérisque*, 307 (2006), viii, 229–246.

[21] Host, B. & Kra, B., Convergence of polynomial ergodic averages. *Israel J. Math.*, 149 (2005), 1–19.

[22] Janusz, G. J., *Algebraic Number Fields*. Pure and Applied Mathematics, 55. Academic Press, New York–London, 1973.

[23] Leibman, A., Convergence of multiple ergodic averages along polynomials of several variables. *Israel J. Math.*, 146 (2005), 303–315.

[24] Pintz, J., Steiger, W. L. & Szemerédi, E., On sets of natural numbers whose difference set contains no squares. *J. London Math. Soc.*, 37 (1988), 219–231.

[25] Ramaré, O., On Shnirel′man's constant. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.*, 22 (1995), 645–706.

[26] Ramaré, O. & Ruzsa, I. Z., Additive properties of dense subsets of sifted sequences. *J. Théor. Nombres Bordeaux*, 13 (2001), 559–581.

[27] Ruzsa, I. Z., An analog of Freiman's theorem in groups. *Astérisque*, 258 (1999), xv, 323–326.
[28] Sárközy, A., On difference sets of sequences of integers. I. *Acta Math. Acad. Sci. Hungar.*, 31 (1978), 125–149.
[29] Slijepčević, S., A polynomial Sárközy–Furstenberg theorem with upper bounds. *Acta Math. Hungar.*, 98 (2003), 111–128.
[30] Szemerédi, E., On sets of integers containing no k elements in arithmetic progression. *Acta Arith.*, 27 (1975), 199–245.
[31] Tao, T., The Gaussian primes contain arbitrarily shaped constellations. *J. Anal. Math.*, 99 (2006), 109–176.
[32] — Obstructions to uniformity and arithmetic patterns in the primes. *Pure Appl. Math. Q.*, 2 (2006), 395–433.
[33] — A quantitative ergodic theory proof of Szemerédi's theorem. *Electron. J. Combin.*, 13 (2006), Research Paper 99, 49 pp.
[34] — A variant of the hypergraph removal lemma. *J. Combin. Theory Ser. A*, 113 (2006), 1257–1280.
[35] — An ergodic transference theorem. Unpublished notes. `http://www.math.ucla.edu/~tao/preprints/Expository/limiting.dvi`.
[36] — A remark on Goldston–Yıldırım correlation estimates. Preprint, 2007. `http://www.math.ucla.edu/~tao/preprints/Expository/gy-corr.dvi`.
[37] Titchmarsh, E. C., *The Theory of the Riemann Zeta-Function*. Oxford University Press, New York, 1986.
[38] Varnavides, P., On certain sets of positive density. *J. London Math. Soc.*, 34 (1959), 358–360.

Terence Tao
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90095-1555
U.S.A.
tao@math.ucla.edu

Tamar Ziegler
Department of Mathematics
Technion – Israel Institute of Technology
Haifa 32000
Israel
tamarzr@tx.technion.ac.il

Received October 10, 2006