Algebra & Number Theory

msp.org/ant

Zeros of *L*-functions
outside the critical strip

Andrew R. Booker and Frank Thorne

For a wide class of Dirichlet series associated to automorphic forms, we show that those without Euler products must have zeros within the region of absolute convergence. For instance, we prove that if $f \in S_k(\Gamma_1(N))$ is a classical holomorphic modular form whose *L*-function does not vanish for $\Re(s) > (k+1)/2$, then f is a Hecke eigenform. Our proof adapts and extends work of Saias and Weingartner, who proved a similar result for degree-1 *L*-functions.

1. Introduction

Saias and Weingartner [2009] showed that if $L(s) = \sum_{m=1}^{\infty} \lambda(m)/m^s$ is a Dirichlet series with periodic coefficients, then either $L(s) = 0$ for some s with real part > 1, or $\lambda(m)$ is multiplicative at almost all primes (so that $L(s) = D(s)L(s, \chi)$ for some primitive Dirichlet character χ and finite Dirichlet series D). Earlier work of Davenport and Heilbronn [1936a; 1936b] established this result for the special case of the Hurwitz zeta-function $\zeta(s, \alpha)$ with rational parameter α, and proved an analogue for the degree-2 Epstein zeta-functions. Also in degree 2, Conrey and Ghosh [1994] showed that the *L*-function associated to the square of Ramanujan's Δ modular form has infinitely many zeros outside of its critical strip. In this paper, we generalize all of these results and study the extent to which, among all Dirichlet series associated to automorphic forms (appropriately defined), the existence of an Euler product is characterized by nonvanishing in the region of absolute convergence. For instance, for classical degree-2 *L*-functions, we prove the following:

Theorem 1.1. *Let $f \in S_k(\Gamma_1(N))$ be a holomorphic cuspform of arbitrary weight and level. If the associated complete L-function $\Lambda_f(s) = \int_0^{\infty} f(iy)y^{s-1}\, dy$ does not vanish for $\Re(s) > (k+1)/2$, then f is an eigenfunction of the Hecke operators T_p for all primes $p \nmid N$.*

Our method is sufficiently general to apply to L-functions of all degrees, and in fact we obtain Theorem 1.1 as a corollary of the following general result:

Theorem 1.2. *Fix a positive integer n. For $j = 1, \ldots, n$, let r_j be a positive integer and π_j a unitary cuspidal automorphic representation of $\mathrm{GL}_{r_j}(\mathbb{A}_{\mathbb{Q}})$ with L-series $L(s, \pi_j) = \sum_{m=1}^{\infty} \lambda_j(m)m^{-s}$. Assume that the π_j satisfy the generalized Ramanujan conjecture at all finite places (so that, in particular, $|\lambda_j(p)| \leq r_j$ for all primes p) and are pairwise nonisomorphic. Let*

$$R = \left\{ \sum_{m=1}^{M} \frac{a_m}{m^s} : M \in \mathbb{Z}_{\geq 0}, (a_1, \ldots, a_M) \in \mathbb{C}^M \right\}$$

denote the ring of finite Dirichlet series, and let $P \in R[x_1, \ldots, x_n]$ be a polynomial with coefficients in R. Then either $P(L(s, \pi_1), \ldots, L(s, \pi_n))$ has a zero with real part > 1 or $P = D(s)x_1^{d_1} \cdots x_n^{d_n}$ for some $D \in R, d_1, \ldots, d_n \in \mathbb{Z}_{\geq 0}$.

Remarks. (1) For π_j as in the statement of the theorem, it is known (see [Jacquet and Shalika 1976]) that $L(s, \pi_j)$ does not vanish for $\Re(s) \geq 1$. Thus if $P = D(s)x_1^{d_1} \cdots x_n^{d_n}$ is a monomial, the matter of whether $P(L(s, \pi_1), \ldots, L(s, \pi_n))$ vanishes for some s with $\Re(s) > 1$ is determined entirely by the finite Dirichlet series $D(s)$. Further, the grand Riemann hypothesis (GRH) predicts that each $L(s, \pi_j)$ does not vanish for $\Re(s) > \frac{1}{2}$. Theorem 1.2 demonstrates that the GRH, if it is true, is a very rigid phenomenon.

(2) By the almost-periodicity of Dirichlet series, if $P(L(s, \pi_1), \ldots, L(s, \pi_n))$ has at least one zero with real part > 1 then it must have infinitely many such zeros. In fact, our proof shows that there is some number $\eta = \eta(P; \pi_1, \ldots, \pi_n) > 0$ such that for any σ_1, σ_2 with $1 < \sigma_1 < \sigma_2 \leq 1 + \eta$, we have

$$\#\left\{s \in \mathbb{C} : \Re(s) \in [\sigma_1, \sigma_2], \Im(s) \in [-T, T], P(L(s, \pi_1), \ldots, L(s, \pi_n)) = 0\right\} \gg T \quad (1\text{-}1)$$

for T sufficiently large (where both the implied constant and the meaning of "sufficiently large" depend on σ_1, σ_2 as well as P and $\pi_1, \ldots, \pi_n$).

On the other hand, if we restrict to $\mathbb{C}$-linear combinations (that is, homogeneous degree-1 polynomials $P \in \mathbb{C}[x_1, \ldots, x_n]$) and $\pi_1, \ldots, \pi_n$ with a common conductor and archimedean component $\pi_{1,\infty} \cong \cdots \cong \pi_{n,\infty}$, Bombieri and Hejhal [1995] showed, subject to GRH and a weak form of the pair correlation conjecture for $L(s, \pi_j)$, that asymptotically 100% of the nontrivial zeros of $P(L(s, \pi_1), \ldots, L(s, \pi_n))$ have real part $\frac{1}{2}$.

(3) The assumption of the Ramanujan conjecture in Theorem 1.2 could be relaxed. For instance, it would suffice to have, for each fixed j:

(i) Some mild control over the coefficients of the logarithmic derivative

$$\frac{L'}{L}(s, \pi_j) = \sum_{m=1}^{\infty} c_j(m) m^{-s}$$

at prime powers, namely $\sum_p |c_j(p^k)|^2/p^k < \infty$ for any fixed $k \geq 2$ (cf. [Rudnick and Sarnak 1996, Hypothesis H]).

(ii) An average bound for $|\lambda_j(p)|^4$ over arithmetic progressions of primes, namely

$$\limsup_{x \to \infty} \frac{\displaystyle\sum_{\substack{p \leq x \\ p \equiv a \,(\text{mod } q)}} |\lambda_j(p)|^4}{\displaystyle\sum_{\substack{p \leq x \\ p \equiv a \,(\text{mod } q)}} 1} \leq C_j$$

for all coprime $a, q \in \mathbb{Z}_{>0}$, where $C_j > 0$ is independent of a, q.

Note that (i) is known to hold when $r_j \leq 4$ (see [Rudnick and Sarnak 1996; Kim 2006]). Further, both estimates follow from the Rankin–Selberg method if, for instance, the tensor square $\pi_j \otimes \pi_j$ is automorphic for each j. Since this is known when $r_j = 2$ (see [Gelbart and Jacquet 1978]), Theorem 1.2 could be extended to include the *L*-functions associated to Maass forms.

(4) The main tool used in the proof is the quasi-orthogonality of the coefficients $\lambda_j(p)$, i.e., asymptotic estimates for sums of the form $\sum_{p \leq x} \lambda_j(p)\overline{\lambda_k(p)}/p$ as $x \to \infty$. These follow from the Rankin–Selberg method, and were obtained in a precise form independently by Wu and Ye [2007, Theorem 3] and Avdispahić and Smajlović [2010, Theorem 2.2]. (We also make use of similar estimates for sums over p in an arithmetic progression — see Lemma 2.1 for the exact statement — though it is likely that this could be avoided at the expense of making the proof more complicated.)

Since quasi-orthogonality and the Ramanujan conjecture are essentially the only properties of automorphic *L*-functions that we require, one could instead take these as hypotheses and state the theorem for an axiomatically defined class of *L*-functions, such as the Selberg class. However, it has been conjectured that the Selberg class coincides with the class of automorphic *L*-functions, so this likely offers no greater generality.

(5) The conclusion of Theorem 1.2 is interesting even for $n = 1$. For instance, Nakamura and Pańkowski [2012] have shown for a wide class of *L*-functions $L(s)$ that if $P \in R[x]$ is not a monomial and $\delta > 0$, then $P(L(s))$ necessarily has zeros in the half-plane $\Re(s) > 1 - \delta$. Our result strengthens this to $\Re(s) > 1$. (On the other hand, their results also yield the estimate (1-1) for any $[\sigma_1, \sigma_2] \subseteq \left(\frac{1}{2}, 1\right)$, which does not follow from our method.)

(6) Our results are related to universality results for zeta and L-functions. Voronin [1975] proved, for any compact set K with connected complement contained within the strip $\Re(s) \in \left(\frac{1}{2}, 1\right)$ and any nonvanishing, continuous function $f : K \to \mathbb{C}$ holomorphic on the interior of K, that f can be uniformly approximated by vertical translates of the zeta function.

Voronin's results were extended by a number of authors. One result similar to ours, due to Laurinčikas and Matsumoto [2004], states that given m functions $f_1, \ldots, f_m$ as above, and L-functions $L_j(s, F)$ associated to twists of a Hecke newform F by pairwise inequivalent Dirichlet characters, that the f_j may be simultaneously approximated by a single vertical translate of the functions $L_j(s, F)$. This implies [loc. cit., Theorem 4] that nontrivial linear combinations of the $L_j(s, F)$ must contain zeros inside the critical strip with $\Re(s) > \frac{1}{2}$.

References to many more works on universality can be found in [loc. cit.].

Summary of the proof. Our proof closely follows Saias and Weingartner in broad outline, but becomes more technical in some places. The reader may wish to read [Saias and Weingartner 2009].

The technical heart of our paper is Proposition 3.1, an extension of their Lemma 2. Given n nonzero complex numbers $z_1, \ldots, z_n$, we would like to simultaneously solve the equations $L(s, \pi_j) = z_j$, leading to a quick proof of the main theorem. As a substitute, we solve equations of the form $\prod_{p>y} L(\sigma + it_p, \pi_{j,p}) = z_j$, where the ordinate of s is allowed to vary for each prime.

Given this, in Section 4 we prove our main theorem, following the proof of Theorem 2 in [Saias and Weingartner 2009]. As in that work, the main tools are Weyl's criterion, allowing us to simultaneously approximate all of the $p^{-\sigma - it_p}$ by $p^{-\sigma - it}$ for a single t, and Rouché's theorem, which states that actual zeros must exist near approximate zeros.

The proof of Proposition 3.1 follows those of Lemmas 1 and 2 of [Saias and Weingartner 2009]. However, in that work the Dirichlet coefficients $\lambda(m)$ are all periodic to some fixed modulus, and this fact, combined with the prime number theorem for arithmetic progressions, allows for easy control of various partial sums that need to be estimated. Here, we must do without this periodicity.

To prove Proposition 3.1, we choose (in Proposition 3.3) a partition of the set of primes $p > y$ into disjoint subsets S, and complex numbers $\epsilon_p \in S^1$ for each $p > y$, so that the vectors of partial sums $\sum_{p \in S} \epsilon_p \lambda_j(p) p^{-\sigma}$ are linearly independent in a precise quantitative sense. Our main tool is the Rankin–Selberg method (substituting for periodicity and orthogonality of Dirichlet characters); see Lemma 2.1.

We also rely on the rather technical Proposition 3.2, which says that for matrices $g_1, \ldots, g_m$, we can continuously solve equations of the form $\sum_{i=1}^{m} g_i f_i(z) = z$ for n-tuples of complex numbers $z = (z_1, \ldots, z_n)$. The g_i are constructed from the sums

over $p \in S$ considered in Proposition 3.3, but we are able to formulate Proposition 3.2 in a general manner, without reference to automorphic forms or primes.

The conclusion of Proposition 3.2 is guaranteed only for large m, so that the number of subsets S needed may be large. We choose these subsets to be arithmetic progressions, for which the Rankin–Selberg estimates presented in Lemma 2.1 are known to hold. If such estimates were unavailable, it seems likely that we could still obtain our result by constructing the S in a more ad hoc fashion instead. In any case, and in contrast to Saias–Weingartner, the modulus of the arithmetic progression has no particular arithmetic significance, and is chosen to be coprime to all the conductors of the π_j.

2. Preliminaries

Automorphic L-functions. Let π_j be as in the statement of Theorem 1.2. Each π_j can be written as a restricted tensor product $\pi_{j,\infty} \otimes \bigotimes_p \pi_{j,p}$ of local representations, where p runs through all prime numbers. Then we have

$$L(s, \pi_j) = \prod_p L(s, \pi_{j,p}) \quad \text{for } \Re(s) > 1. \tag{2-1}$$

Here each local factor $L(s, \pi_{j,p})$ is a rational function of p^{-s}, of the form

$$L(s, \pi_{j,p}) = \frac{1}{(1 - \alpha_{j,p,1} p^{-s}) \cdots (1 - \alpha_{j,p,r_j} p^{-s})} \tag{2-2}$$

for certain complex numbers $\alpha_{j,p,\ell}$. The generalized Ramanujan conjecture asserts that $|\alpha_{j,p,\ell}| \leq 1$, with equality holding for all $p \nmid \text{cond}(\pi_j)$, where $\text{cond}(\pi_j) \in \mathbb{Z}_{>0}$ is the conductor of π_j. In particular, $|\lambda_j(p)| = |\alpha_{j,p,1} + \cdots + \alpha_{j,p,r_j}| \leq r_j$.

Lemma 2.1. *Let a and q be positive integers satisfying $\left(q, a \prod_{j=1}^n \text{cond}(\pi_j)\right) = 1$. Then*

$$\sum_{\substack{p > y \\ p \equiv a \,(\text{mod } q)}} \frac{|u_1 \lambda_1(p) + \cdots + u_n \lambda_n(p)|^2}{p^\sigma} = \left(\frac{1}{\phi(q)} + O(\sigma - 1)\right) \sum_{p > y} p^{-\sigma}$$

for all $y > 0$, $\sigma \in (1, 2]$ and all unit vectors $(u_1, \ldots, u_n)$, where the implied constant depends only on $\pi_1, \ldots, \pi_n$ and q.

Proof. Let $\chi \,(\text{mod } q)$ be a Dirichlet character, not necessarily primitive. We consider the sum

$$E_{jk\chi}(x) = \sum_{p \leq x} \left(\lambda_j(p) \overline{\lambda_k(p)} \chi(p) - \delta_{jk\chi}\right) \frac{\log p}{p},$$

running over primes $p \leq x$, where $\delta_{jk\chi} = 1$ if $j = k$ and χ is the trivial character, and $\delta_{jk\chi} = 0$ otherwise. Applying [Avdispahić and Smajlović 2010, (2) and (3)]

with $(\pi, \pi') = (\pi_j \otimes \chi, \pi_k)$ and, if χ is imprimitive, subtracting any contribution from the terms with $p \mid q$, we obtain the bound $E_{jk\chi}(x) \ll_q 1$.

Next, for any nonintegral $y \geq \frac{3}{2}$ and any $\sigma \in (1, 2]$, we have

$$\sum_{p>y} \frac{\lambda_j(p)\overline{\lambda_k(p)}\chi(p) - \delta_{jk\chi}}{p^\sigma} = \int_y^\infty \frac{t^{1-\sigma}}{\log t}\, dE_{jk\chi}(t).$$

Integrating by parts and applying the above estimate for $E_{jk\chi}$, we see that this is $\ll_q y^{1-\sigma}/\log y$.

Now, expanding the square and using orthogonality of Dirichlet characters, we have

$$\sum_{\substack{p>y \\ p \equiv a \ (\mathrm{mod}\ q)}} \frac{|u_1\lambda_1(p) + \cdots + u_n\lambda_n(p)|^2}{p^\sigma}$$

$$= \frac{1}{\phi(q)} \sum_{j=1}^n \sum_{k=1}^n \sum_{\chi\ (\mathrm{mod}\ q)} u_j \overline{u_k} \chi(a) \sum_{p>y} \frac{\lambda_j(p)\overline{\lambda_k(p)}\chi(p)}{p^\sigma}$$

$$= O_q\left(\frac{y^{1-\sigma}}{\log y}\right) + \frac{1}{\phi(q)} \sum_{p>y} p^{-\sigma}.$$

Finally, by the prime number theorem we have $\sum_{p>y} p^{-\sigma} \gg y^{1-\sigma}/((\sigma - 1)\log y)$, uniformly for $y \geq \frac{3}{2}$ and $\sigma \in (1, 2]$. The lemma follows. $\square$

A few lemmas. In the remainder of this section we discuss the topology of $\mathrm{GL}_n(\mathbb{C})$ and prove some simple lemmas, to be used in the more technical propositions which follow.

Let $\mathrm{Mat}_{n\times n}(\mathbb{C})$ denote the set of $n \times n$ matrices with entries in $\mathbb{C}$. For $A = (a_{ij}) \in \mathrm{Mat}_{n\times n}(\mathbb{C})$, the *Frobenius norm* is defined by

$$\|A\| = \sqrt{\mathrm{tr}(\overline{A}^T A)} = \sqrt{\sum |a_{ij}|^2}.$$

Note that this agrees with the Euclidean norm under the identification of $\mathrm{Mat}_{n\times n}(\mathbb{C})$ with $\mathbb{C}^{n^2}$. By the Schwarz inequality, we have $|Av| \leq \|A\| \cdot |v|$ for any $A \in \mathrm{Mat}_{n\times n}(\mathbb{C})$ and $v \in \mathbb{C}^n$.

We endow $\mathrm{GL}_n(\mathbb{C}) = \{g \in \mathrm{Mat}_{n\times n}(\mathbb{C}) : \det g \neq 0\}$ with the subspace topology. In particular, it is easy to see that a set $K \subseteq \mathrm{GL}_n(\mathbb{C})$ is compact if and only if K is closed in $\mathrm{Mat}_{n\times n}(\mathbb{C})$ and there are positive real numbers c and C such that

$$\|g\| \leq C \quad \text{and} \quad |\det g| \geq c \quad \text{for all } g \in K.$$

Since g^{-1} can be expressed in terms of $1/\det g$ and the cofactor matrix of g, it follows that $\|g^{-1}\|$ is bounded on K (and indeed the map $g \mapsto g^{-1}$ is continuous, so that $\mathrm{GL}_n(\mathbb{C})$ is a topological group with this topology).

Lemma 2.2. *Suppose K is a compact subset of $\mathrm{GL}_n(\mathbb{C})$, $g \in K$, and $U \subseteq \mathbb{C}^n$ contains an open δ-neighborhood of some point. Then gU contains an ε-neighborhood, where $\varepsilon > 0$ depends only on δ and K.*

Proof. By linearity, we may assume without loss of generality that U contains the δ-neighborhood of the origin, N_δ. Since K is compact, there is a number $C > 0$ such that $\|g^{-1}\| \leq C$ for all $g \in K$. Put $\varepsilon = C^{-1}\delta$, and let N_ε be the ε-neighborhood of the origin. For any $v \in N_\varepsilon$ we have $|g^{-1}v| \leq \|g^{-1}\| \cdot |v| < C\varepsilon = \delta$, so that $v = g(g^{-1}v) \in gN_\delta$. Since v was arbitrary, $gN_\delta \supseteq N_\varepsilon$. $\qquad\square$

Lemma 2.3. *For any $v_0, \ldots, v_k \in \mathbb{C}^n$, there exist $\theta_0, \ldots, \theta_k \in [0, 1]$ such that*

$$\left| \sum_{j=0}^{k} e(\theta_k)v_j \right| \leq \sqrt{\sum_{j=0}^{k} |v_j|^2}.$$

Proof. We have

$$\int_{[0,1]^k} \left| \sum_{j=0}^{k} e(\theta_j)v_j \right|^2 d\theta_1 \cdots d\theta_k = \sum_{j=0}^{k} |v_j|^2.$$

Thus, the average choice of $(\theta_0, \ldots, \theta_k)$ satisfies the conclusion. $\qquad\square$

Lemma 2.4. *Let $P \in \mathbb{C}[x_1, \ldots, x_n]$. Suppose that every solution to the equation $P(x_1, \ldots, x_n) = 0$ satisfies $x_1 \cdots x_n = 0$. Then P is a monomial; i.e., $P = cx_1^{d_1} \ldots x_n^{d_n}$ for some $c \neq 0$ and nonnegative integers $d_1, \ldots, d_n$.*

Proof. Let $V = \{(x_1, \ldots, x_n) \in \mathbb{C}^n : P(x_1, \ldots, x_n) = 0\}$ be the vanishing set of P. By hypothesis, the polynomial $x_1 \cdots x_n$ vanishes on V. Thus, since $\mathbb{C}$ is algebraically closed, Hilbert's Nullstellensatz implies that there is some $d \in \mathbb{Z}_{\geq 0}$ such that $(x_1 \cdots x_n)^d$ is contained in the ideal generated by P; i.e., $P \mid (x_1 \cdots x_n)^d$. Since $\mathbb{C}[x_1, \ldots, x_n]$ is a unique factorization domain, this is only possible if P is a monomial. $\qquad\square$

Lemma 2.5. *Let $P \in \mathbb{C}[x_1, \ldots, x_n]$ and suppose that $y \in \mathbb{C}^n$ is a zero of P. Then for any $\varepsilon > 0$ there exists $\delta > 0$ such that any polynomial $Q \in \mathbb{C}[x_1, \ldots, x_n]$ obtained by changing any of the nonzero coefficients of P by at most δ each has a zero $z \in \mathbb{C}^n$ with $|y - z| < \varepsilon$.*

Proof. If P is identically 0 then so is Q, so we may take $z = y$. Otherwise, set

$$p(t) = P(y + tu) \quad \text{and} \quad q(t) = Q(y + tu)$$

for $t \in \mathbb{C}$, where u is any unit vector for which $p(t)$ does not vanish for all t; shrinking ε if necessary, assume that $p(t)$ does not vanish on $C_\varepsilon = \{t \in \mathbb{C} : |t| = \varepsilon\}$, and let $\gamma > 0$ be the minimum of $|p(t)|$ on C_ε. For $t \in C_\varepsilon$ we have

$$|q(t) - p(t)| < \delta N (1 + \varepsilon + |y|)^{\deg P},$$

where N is the number of nonzero coefficients of P. Choosing δ so that the right side of this expression is bounded by γ, we have $|q(t) - p(t)| < |p(t)|$ for $t \in C_\varepsilon$. By Rouché's theorem $q(t)$ has a zero t_0 of modulus $|t_0| < \varepsilon$, and taking $z = y + t_0 u$ completes the proof. $\square$

3. Simultaneous representations of n-tuples of complex numbers

The technical heart of our work is the following analogue of Lemma 2 of [Saias and Weingartner 2009]:

Proposition 3.1. *For any real numbers* y, $R > 1$ *there exists* $\eta > 0$ *such that, for all* $\sigma \in (1, 1 + \eta]$, *we have*

$$\left\{ \left(\prod_{p>y} L(\sigma + it_p, \pi_{j,p}) \right)_{j=1,\dots,n} : t_p \in \mathbb{R} \text{ for each prime } p > y \right\}$$

$$\supseteq \{ (z_1, \dots, z_n) \in \mathbb{C}^n : R^{-1} \leq |z_j| \leq R \text{ for all } j \}.$$

Loosely speaking, after simultaneously approximating the t_p by a common t, it will follow that we can make the $L(s, \pi_j)$ independently approach any desired n-tuple of nonzero complex numbers, and this will allow us to find zeros in linear or polynomial combinations.

The proof relies on an analogue of Lemma 1 of [Saias and Weingartner 2009], whose adaptation is not especially straightforward. We carry out this work by proving two technical propositions; the first establishes the existence of solutions to a certain equation involving matrices in a fixed compact subset of $\mathrm{GL}_n(\mathbb{C})$.

Proposition 3.2. *Let*

$$T = \{ (z_1, \dots, z_n) \in \mathbb{C}^n : |z_1| = \cdots = |z_n| = 1 \},$$

$$D = \{ (z_1, \dots, z_n) \in \mathbb{C}^n : |z_1|, \dots, |z_n| \leq 1 \},$$

and fix a compact set $K \subseteq \mathrm{GL}_n(\mathbb{C})$. *Then there is a number* $m_0 > 0$ *such that for every* $m \geq m_0$ *and all* $(g_1, \dots, g_m) \in K^m$, *there are continuous functions* $f_1, \dots, f_m : D \to T$ *such that* $\sum_{i=1}^m g_i f_i(z) = z$ *for all* $z \in D$.

We will carry out the proof in three steps:

(1) We first show that there exist $\varepsilon > 0$ and m_1 such that for all $m \geq m_1$ and all $(g_1, \dots, g_m) \in K^m$, the set $\left\{ \sum_{i=1}^m g_i t_i : t_1, \dots, t_m \in T \right\}$ contains an open ε-neighborhood of a point in $\mathbb{C}^n$.

(2) "Fattening" the neighborhood constructed in the first step, we will then show that there exists m_2 such that for $m \geq m_2$ and all $(g_1, \dots, g_m) \in K^m$, $\left\{ \sum_{i=1}^m g_i t_i : t_1, \dots, t_m \in T \right\}$ contains $\{ (z_1, \dots, z_n) : |z_1|^2 + \cdots + |z_n|^2 \leq 4 \}$, the closed ball of radius 2.

(3) Although the previous step yields a parametrization of a large closed set, it is not obviously continuous. By repeating the construction from step (1) using the added knowledge of step (2), we show that one can achieve a continuous parametrization of D.

Proof. We begin by showing (1). By compactness, there is an m_1 such that for any $m \geq m_1$ and any m-tuple $(g_1, \ldots, g_m)$, there is a distinct pair of indices i, j such that $\|g_i^{-1} g_j - I\| < 1/(3\sqrt{n})$. Assume, without loss of generality, that $(i, j) = (1, 2)$, and put $\Delta = g_1^{-1} g_2 - I$. Then for any choice of $t_1, t_2 \in T$, we have

$$g_1 t_1 + g_2 t_2 = g_1(t_1 + (I + \Delta)t_2),$$

where $\|\Delta\| < 1/(3\sqrt{n})$.

We introduce some notation. First, define $A = \left\{ z \in \mathbb{C} : |z - 1| \leq \frac{1}{3} \right\}$ and $B = \left\{ z \in \mathbb{C} : |z - 1| \leq \frac{2}{3} \right\}$. Next, let $s_1, s_2 : B \to \mathbb{C}$ be the unique continuous functions satisfying $z = s_1(z) + s_2(z)$, $|s_1(z)| = |s_2(z)| = 1$ and $\Im(s_1(z)/s_2(z)) > 0$ for all $z \in B$. For $j = 1, 2$, let $t_j : B^n \to T$ be defined by $t_j(z_1, \ldots, z_n) = (s_j(z_1), \ldots, s_j(z_n))$.

Given an arbitrary element $w \in A^n$, we define a continuous function $h_w : B^n \to \mathbb{C}^n$ by $h_w(z) = w - \Delta t_2(z)$. Since $|t_2(z)| = \sqrt{n}$ and $\|\Delta\| < 1/(3\sqrt{n})$, we have $|\Delta t_2(z)| < \frac{1}{3}$. In particular, each entry of $\Delta t_2(z)$ is bounded in magnitude by $\frac{1}{3}$, so by the triangle inequality, the image of h_w is contained in B^n. By the Brouwer fixed point theorem, there exists $z \in B^n$ with $h_w(z) = z$, so that

$$t_1(z) + (I + \Delta)t_2(z) = z + \Delta t_2(z) = z + w - h_w(z) = w.$$

Therefore, all of A^n is in the image of the map $z \mapsto t_1(z) + (I + \Delta)t_2(z)$, so that in particular

$$A^n \subseteq \{t_1 + g_1^{-1} g_2 t_2 : t_1, t_2 \in T\}.$$

Applying Lemma 2.2 with $\delta = \frac{1}{3}$, there is an $\varepsilon > 0$ depending only on K such that $\{g_1 t_1 + g_2 t_2 : t_1, t_2 \in T\}$ contains an ε-neighborhood of some point in $\mathbb{C}^n$. We conclude the same of the set $\{g_1 t_1 + \cdots + g_m t_m : t_1, \ldots, t_m \in T\}$ by choosing arbitrary fixed $t_3, \ldots, t_m \in T$.

Proceeding to step (2), let k_1 be a large integer to be determined later, set $m_2 = m_1 k_1$, and for any $m \geq m_2$ write $m = km_1 + l$ with $k \geq k_1$ and $0 \leq l < m_1$.

For each j with $0 \leq j < k$, applying step (1) to $(g_{jm_1+1}, \ldots, g_{jm_1+m_1})$, we obtain an ε-neighborhood centered at some $v_j \in \mathbb{C}^n$. Further, we put $v_k = g_{km_1+1}\vec{1} + \cdots + g_{km_1+l}\vec{1}$, where $\vec{1} = (1, \ldots, 1) \in T$. Since m_1 is fixed and K is compact, we have $|v_j| \leq C$ for $0 \leq j \leq k$, for some C independent of the individual g_i.

Let $N_\varepsilon = \{(z_1, \ldots, z_n) \in \mathbb{C}^n : |z_1|^2 + \cdots + |z_n|^2 < \varepsilon^2\}$ be the ε-neighborhood of the origin in $\mathbb{C}^n$. Then by the above observations, for any $\theta_0, \ldots, \theta_k \in [0, 1]$,

$\left\{\sum_{i=1}^{m} g_i t_i : t_1, \ldots, t_m \in T\right\}$ contains the set

$$\sum_{j=0}^{k-1} e(\theta_j)(v_j + N_\varepsilon) + e(\theta_k)v_k = \sum_{j=0}^{k} e(\theta_j)v_j + kN_\varepsilon.$$

By Lemma 2.3, there is a choice of $\theta_0, \ldots, \theta_k$ for which $\left|\sum_{j=0}^{k} e(\theta_j)v_j\right| \leq C\sqrt{k+1}$. Now let k_1 be the smallest positive integer satisfying $k_1\varepsilon > C\sqrt{k_1+1}+2$. Then for $k \geq k_1$, we have shown that $\left\{\sum_{i=1}^{m} g_i t_i : t_1, \ldots, t_m \in T\right\}$ contains the closed ball of radius 2.

Proceeding to step (3), we put $m_0 = 3nm_2$. Suppose that $m \geq m_0$ and $(g_1, \ldots, g_m)$ are given, and choose a partition of $\{1, \ldots, m\}$ into $3n$ sets $I_{j,\ell}$ (for $1 \leq j \leq n$, $1 \leq \ell \leq 3$), each of size at least m_2. For each j with $1 \leq j \leq n$, write

$$v_j = v_{j,1} = v_{j,2} = v_{j,3} = (0, \ldots, 0, 2, 0, \ldots, 0),$$

where the 2 is in the j-th position. For each j and ℓ we use step (2) to express $v_{j,\ell}$ in the form

$$v_{j,\ell} = \sum_{i \in I_{j,\ell}} g_i t_i \tag{3-1}$$

for some $t_i \in T$.

Next, note that the set $\left\{2[(1, \ldots, 1)+\alpha+\beta]:\alpha, \beta \in T\right\}$ contains D. As in step (1), we can choose continuous functions $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\beta = (\beta_1, \ldots, \beta_n) : D \to T$ such that $z_j = 2[1 + \alpha_j(z) + \beta_j(z)]$ for every $z = (z_1, \ldots, z_n) \in D$. Thus,

$$z = \sum_{j=1}^{n}[1 + \alpha_j(z) + \beta_j(z)]v_j = \sum_{j=1}^{n}[v_{j,1} + \alpha_j(z)v_{j,2} + \beta_j(z)v_{j,3}].$$

Finally, we use (3-1) to rewrite this as

$$z = \sum_{j=1}^{n}\left(\sum_{i \in I_{j,1}} g_i t_i + \sum_{i \in I_{j,2}} g_i\left(t_i \alpha_j(z)\right) + \sum_{i \in I_{j,3}} g_i\left(t_i \beta_j(z)\right)\right),$$

which is a decomposition of the type required. $\square$

Next, we use the quasi-orthogonality of the coefficients $\lambda_j(p)$ (Lemma 2.1) to show that, by choosing an arbitrary "twist" $\epsilon_p \in S^1$ for each large prime p, we can make sums of the $\epsilon_p \lambda_j(p)$ line up in linearly independent directions, as quantified in the following proposition.

Given a real parameter $y > 0$, we write

$$S(y) = \{p \text{ prime} : p > y\} \quad \text{and} \quad s(y, \sigma) = \sum_{p \in S(y)} p^{-\sigma}.$$

Proposition 3.3. *There is a compact set $K \subseteq \mathrm{GL}_n(\mathbb{C})$, explicitly defined in (3-5), depending only on the degrees $r_1, \ldots, r_n$, with the following property:*

Let m be a positive integer. Then there is a real number $\delta > 0$ (depending on the π_j and m) such that for any $y > 0$ and any $\sigma \in (1, 1 + \delta]$, there exists a partition of $S(y)$ into mn pairwise disjoint subsets $S_{ik}(y)$ $(i = 1, \ldots, m, k = 1, \ldots, n)$ and a choice of $\epsilon_p \in S^1$ for each $p \in S(y)$ such that the m-tuple of matrices $(g_1, \ldots, g_m)$ defined by

$$g_i = \left(\frac{mn}{s(y, \sigma)} \sum_{p \in S_{ik}(y)} \frac{\epsilon_p \lambda_j(p)}{p^\sigma} \right)_{1 \le j, k \le n}, \quad i = 1, \ldots, m \tag{3-2}$$

lies in K^m.

Proof. Let q be the smallest prime number satisfying $q \equiv 1 \pmod{mn}$ and $q \nmid \prod_{j=1}^n \mathrm{cond}(\pi_j)$. We put $t = (q - 1)/mn$ and define $S_{ik}^\circ(y)$ to be the union of residue classes

$$S_{ik}^\circ(y) = \bigcup_{\ell=1}^t \{p \in S(y) : p \equiv tn(i - 1) + t(k - 1) + \ell \pmod{q}\},$$

and

$$S_{ik}(y) = \begin{cases} S_{ik}^\circ(y) \cup \{q\} & \text{if } i = k = 1 \text{ and } y < q, \\ S_{ik}^\circ(y) & \text{otherwise.} \end{cases}$$

Then the $S_{ik}(y)$ are pairwise disjoint and cover $S(y)$.

For a fixed choice of i, let v_k denote the k-th column of g_i, as defined in (3-2), with the ϵ_p yet to be chosen. We will show by induction that there is a choice of the ϵ_p such that

$$|v_\ell - \mathrm{proj}_{\mathrm{span}\{v_1, \ldots, v_{\ell-1}\}} v_\ell| \ge \frac{1}{2r} \tag{3-3}$$

holds for every $\ell = 1, \ldots, n$, where $r = \sqrt{r_1^2 + \cdots + r_n^2}$. To that end, let k be given, and assume that (3-3) has been established for $\ell = 1, \ldots, k - 1$. Choose a unit vector $u = (u_1, \ldots, u_n)$ orthogonal to $v_1, \ldots, v_{k-1}$. By the Schwarz inequality and the Ramanujan bound $|\lambda_j(p)| \le r_j$, for each prime p we have $|\bar{u}_1 \lambda_1(p) + \cdots + \bar{u}_n \lambda_n(p)| \le r$. Therefore

$$\frac{mn}{s(y, \sigma)} \sum_{p \in S_{ik}(y)} \frac{|\bar{u}_1 \lambda_1(p) + \cdots + \bar{u}_n \lambda_n(p)|}{p^\sigma}$$

$$\ge \frac{mn}{rs(y, \sigma)} \sum_{p \in S_{ik}^\circ(y)} \frac{|\bar{u}_1 \lambda_1(p) + \cdots + \bar{u}_n \lambda_n(p)|^2}{p^\sigma}$$

$$= \frac{1 + O_{m,n}(\sigma - 1)}{r}, \tag{3-4}$$

the latter equality following by Lemma 2.1. We choose δ so that the O term above is bounded in modulus by $\frac{1}{2}$, and for each $p \in S_{ik}(y)$ we choose ϵ_p such that $\epsilon_p(\bar{u}_1\lambda_1(p) + \cdots + \bar{u}_n\lambda_n(p))$ is real and nonnegative. Then the left side of (3-4) equals

$$\langle u, v_k \rangle = \langle u, v_k - \text{proj}_{\text{span}\{v_1,\ldots,v_{k-1}\}} v_k \rangle \leq |v_k - \text{proj}_{\text{span}\{v_1,\ldots,v_{k-1}\}} v_k|,$$

so that (3-3) follows for $\ell = k$.

Applying Gram–Schmidt orthogonalization to $v_1, \ldots, v_n$, it follows from (3-3) that $|\det g_i| \geq (2r)^{-n}$. Moreover, by the Schwarz inequality and Lemma 2.1 again, each entry of g_i is bounded above by $1 + O_{m,n}(\sigma - 1)$, so that $\|g_i\| \leq 2n$ for a suitable choice of δ. Thus,

$$K = \{g \in \text{GL}_n(\mathbb{C}) : \|g\| \leq 2n, |\det g| \geq (2r)^{-n}\} \tag{3-5}$$

has the desired properties. $\square$

We are now ready to prove Proposition 3.1, largely following [Saias and Weingartner 2009].

Proof of Proposition 3.1. We use Propositions 3.3 and 3.2 to determine a compact set $K \subseteq \text{GL}_n(\mathbb{C})$, a positive integer m_0, and a real number $\delta > 0$ with the properties described there. Taking $m = m_0$, the aforementioned propositions yield, for any $\sigma \in (1, 1 + \delta]$, an m-tuple of matrices $(g_1, \ldots, g_m) \in K^m$, elements $\epsilon_p \in S^1$ for each prime $p > y$, and continuous functions $f_1, \ldots, f_m : D \to T$ such that

$$\sum_{i=1}^{m} g_i f_i(z) = z \quad \text{for all } z \in D. \tag{3-6}$$

Now, let $\mu = s(y, \sigma)/(mn)$. For each prime $p > y$, we define a continuous function $t_p : \mu D \to \mathbb{R}$ satisfying

$$p^{-it_p(z)} = \epsilon_p f_i(\mu^{-1}z)_k, \tag{3-7}$$

where (i, k) is the unique pair of indices for which $p \in S_{ik}(y)$, and $f_i(\mu^{-1}z)_k$ denotes the k-th component of $f_i(\mu^{-1}z)$. (Note that the lift from S^1 to $\mathbb{R}$ is possible since D is simply connected.)

Define an error term $E(z) = (E_1(z), \ldots, E_n(z))$ by writing, for each $j = 1, \ldots, n$,

$$E_j(z) = \sum_{p>y} \left(\log L(\sigma + it_p(z), \pi_{j,p}) - \lambda_j(p)p^{-(\sigma+it_p(z))} \right).$$

By the Ramanujan bound, we have

$$\log L(s, \pi_{j,p}) - \lambda_j(p)p^{-s} = O(p^{-2})$$

uniformly for $\Re(s) \geq 1$. Since $\sum_p p^{-2}$ converges, the continuity of E follows from

that of the individual t_p. Moreover, each component $E_j(z)$ is bounded by a number $C > 0$, independent of j, z, y, or σ.

Set $R' = \sqrt{\pi^2 + \log^2 R}$. We take $\eta \in (0, \delta]$ small enough that the condition $\sigma \in (1, 1+\eta]$ ensures that $\mu \geq C + R'$. By (3-6), (3-7), and Proposition 3.3 we have

$$\sum_{p>y} \lambda_j(p) p^{-(\sigma+it_p(z))} = \sum_{i=1}^{m} \sum_{k=1}^{n} \sum_{p \in S_{ik}(y)} \frac{\lambda_j(p) \epsilon_p f_i(\mu^{-1}z)_k}{p^\sigma} = z_j$$

for any $z = (z_1, \ldots, z_n) \in \mu D$. Now, fix $w \in R'D$ and define a function $F_w : (C + R')D \to \mathbb{C}$ by $F_w(z) = w - E(z)$. By the estimate for $E_j(z)$ above, the image of F_w is contained in $(C + R')D$. Thus, by the Brouwer fixed point theorem, there exists $z \in (C + R')D$ with $F_w(z) = z$, so that

$$\left(\sum_{p>y} \log L(\sigma + it_p(z), \pi_{j,p}) \right)_{j=1,\ldots,n} = z + E(z) = z + w - F_w(z) = w.$$

Taking exponentials yields the proposition. $\qquad\square$

4. Proof of Theorem 1.2

The proof will be carried out in two steps:

(1) Applying our previous results, we show that unless P is a monomial (as described in Theorem 1.2), for every $\sigma > 1$ sufficiently close to 1 there are real numbers t_p (for each prime p) and t_0 such that $P|_{s=\sigma+it_0}$ vanishes at $(\prod_p L(\sigma + it_p, \pi_{1,p}), \ldots, \prod_p L(\sigma + it_p, \pi_{n,p}))$.

(2) Simultaneously, approximating the p^{-it_p} by p^{-it} for a common value of t, we use Rouché's theorem to find a zero of $P(L(s, \pi_1), \ldots, L(s, \pi_n))$ close to $\sigma + it$.

Note that the second step is standard and is applied in [Saias and Weingartner 2009] in much the same way.

We begin with a polynomial P whose coefficients are finite Dirichlet series $D(s) = \sum_{m=1}^{M} a_m m^{-s}$, and let y be the largest value of M occurring in any of these coefficients. We rewrite each $L(s, \pi_j)$ as $L_{\leq y}(s, \pi_j) L_{>y}(s, \pi_j)$, splitting each Euler product into products over primes $p \leq y$ and $p > y$ respectively. Setting

$$Q(x_1, \ldots, x_n) = P(L_{\leq y}(s, \pi_1)x_1, \ldots, L_{\leq y}(s, \pi_n)x_n),$$

we have $P(L(s, \pi_1), \ldots, L(s, \pi_n)) = Q(L_{>y}(s, \pi_1), \ldots, L_{>y}(s, \pi_n))$.

The coefficients of Q are rational functions of the p^{-s} for $p \leq y$. More precisely, for any monomial term $D(s)x_1^{d_1} \cdots x_n^{d_n}$ in the expansion of P, the corresponding term of Q is

$$D(s)L_{\leq y}(s, \pi_1)^{d_1} \cdots L_{\leq y}(s, \pi_n)^{d_n} x_1^{d_1} \cdots x_n^{d_n}.$$

Since the finite Euler products $L_{\leq y}(s, \pi_j)$ are nonvanishing holomorphic functions on $\{s \in \mathbb{C} : \Re(s) \geq 1\}$, the corresponding terms of P and Q have the same zeros there.

Let $D_1(s), \ldots, D_m(s)$ run through the coefficients of P which do not vanish identically, and consider their product $f(s) = D_1(s) \cdots D_m(s)$. Then f is itself a finite Dirichlet series which does not vanish identically. By complex analysis, f cannot vanish at $1 + it$ for every $t \in \mathbb{R}$, so there is some t_0 for which $D_1(1 + it_0), \ldots, D_m(1 + it_0)$ are all nonzero, and the same holds for the corresponding terms of Q.

Next, we specialize the coefficients of Q to a fixed value of s, obtaining a polynomial $h_s \in \mathbb{C}[x_1, \ldots, x_n]$. Considering $s = 1 + it_0$, Lemma 2.4 implies either that $h_{1+it_0} = c x_1^{d_1} \cdots x_n^{d_n}$ for some $c \in \mathbb{C}$ and $d_1, \ldots, d_n \in \mathbb{Z}_{\geq 0}$, or that there are $y_1, \ldots, y_n \in \mathbb{C}$, none zero, for which $h_{1+it_0}(y_1, \ldots, y_n) = 0$. In the former case, it follows from our choice of t_0 that $P = D(s) x_1^{d_1} \cdots x_n^{d_n}$ is a monomial, as allowed in the conclusion of Theorem 1.2. Henceforth, we assume that we are in the latter case, and aim to show that $P(L(s, \pi_1), \ldots, L(s, \pi_n))$ has a zero with $\Re(s) > 1$.

We choose $R > 1$ so that $R^{-1/2} \leq |y_j| \leq R^{1/2}$ for every j. By Lemma 2.5, there is a number $\varepsilon > 0$ such that for every $\sigma \in (1, 1+\varepsilon]$, there exists $(z_1(\sigma), \ldots, z_n(\sigma)) \in \mathbb{C}^n$ satisfying $h_{\sigma+it_0}(z_1(\sigma), \ldots, z_n(\sigma)) = 0$ and $R^{-1} \leq |z_j(\sigma)| \leq R$ for every j. We use Proposition 3.1 to determine η in terms of y and R, and assume that $\eta \leq \varepsilon$ by shrinking η if necessary. Proposition 3.1 then guarantees that, for every $\sigma \in (1, 1+\eta]$, we can solve the simultaneous system of equations

$$\prod_{p > y} L(\sigma + it_p, \pi_{j,p}) = z_j(\sigma), \quad j = 1, \ldots, n,$$

in the t_p for $p > y$. For $p \leq y$ we set $t_p = t_0$, thereby completing step (1).

Turning to step (2), let $\sigma_1, \sigma_2 \in \mathbb{R}$ with $1 < \sigma_1 < \sigma_2 \leq 1+\eta$, and put $\sigma = (\sigma_1+\sigma_2)/2$. With the t_0 and t_p resulting from step (1) for this choice of σ, let P_{it_0} denote the polynomial obtained from P by replacing s by $s+it_0$ in all of its coefficients, and define

$$F(s) = P_{it_0}\left(\prod_p L(s + it_p, \pi_{1,p}), \ldots, \prod_p L(s + it_p, \pi_{n,p})\right). \quad (4\text{-}1)$$

Then F is holomorphic for $|s - \sigma| < \sigma - 1$ and satisfies $F(\sigma) = 0$ by construction. It follows that there is a number $\rho \in (0, (\sigma_2 - \sigma_1)/2]$ such that $F(s) \neq 0$ for all $s \in C_\rho = \{s \in \mathbb{C} : |s - \sigma| = \rho\}$. Write γ for the minimum of $|F(s)|$ on C_ρ.

Next, by abuse of notation, we write $P(s)$ to denote $P(L(s, \pi_1), \ldots, L(s, \pi_n))$. As $P(s) = \sum_{m=1}^{\infty} a_m m^{-s}$ converges absolutely as a Dirichlet series for $\Re(s) > 1$, there is an integer $M > 0$ with $\sum_{m=M}^{\infty} |a_m| m^{-\sigma_1} \leq \gamma/3$. By (4-1) we have $F(s) = \sum_{m=1}^{\infty} b_m m^{-s}$, where $b_m = a_m \prod_{p|m} p^{-it_p \operatorname{ord}_p(m)}$, and by the joint uniform

distribution of p^{it} for primes $p < M$, it follows that the set of $t \in \mathbb{R}$ satisfying

$$\sum_{m=1}^{M-1} \frac{|a_m m^{-it} - b_m|}{m^{\sigma_1}} < \frac{\gamma}{3}$$

has positive lower density. The triangle inequality yields $|P(s+it) - F(s)| < \gamma$ for any such t and for all s with $\Re(s) \geq \sigma_1$, and in particular for all $s \in C_\rho$. By Rouché's theorem, it follows that $P(s+it)$ has a zero s with $|s - \sigma| < \rho$. Thus, $P(s)$ has zeros with real part in $[\sigma_1, \sigma_2]$, and indeed we have

$$\#\{s \in \mathbb{C} : \Re(s) \in [\sigma_1, \sigma_2], \Im(s) \in [-T, T], P(s) = 0\} \gg_{\sigma_1, \sigma_2} T$$

for all $T \geq T_0(\sigma_1, \sigma_2)$.

Acknowledgements

References

[Avdispahić and Smajlović 2010] M. Avdispahić and L. Smajlović, "On the Selberg orthogonality for automorphic *L*-functions", *Arch. Math.* (*Basel*) **94**:2 (2010), 147–154. MR 2011b:11068 Zbl 1221.11120

[Bombieri and Hejhal 1995] E. Bombieri and D. A. Hejhal, "On the distribution of zeros of linear combinations of Euler products", *Duke Math. J.* **80**:3 (1995), 821–862. MR 96m:11071 Zbl 0853.11074

[Conrey and Ghosh 1994] J. B. Conrey and A. Ghosh, "Turán inequalities and zeros of Dirichlet series associated with certain cusp forms", *Trans. Amer. Math. Soc.* **342**:1 (1994), 407–419. MR 94e:11056 Zbl 0796.11021

[Davenport and Heilbronn 1936a] H. Davenport and H. Heilbronn, "On the Zeros of Certain Dirichlet Series", *J. London Math. Soc.* **S1-11**:3 (1936), 181–185. MR 1574345 Zbl 0014.21601

[Davenport and Heilbronn 1936b] H. Davenport and H. Heilbronn, "On the Zeros of Certain Dirichlet Series", *J. London Math. Soc.* **S1-11**:4 (1936), 307–312. MR 1574931 Zbl 0015.19802

[Gelbart and Jacquet 1978] S. Gelbart and H. Jacquet, "A relation between automorphic representations of GL(2) and GL(3)", *Ann. Sci. École Norm. Sup.* (4) **11**:4 (1978), 471–542. MR 81e:10025 Zbl 0406.10022

[Jacquet and Shalika 1976] H. Jacquet and J. A. Shalika, "A non-vanishing theorem for zeta functions of GL$_n$", *Invent. Math.* **38**:1 (1976), 1–16. MR 55 #5583 Zbl 0349.12006

[Kim 2006] H. H. Kim, "A note on Fourier coefficients of cusp forms on GL$_n$", *Forum Math.* **18**:1 (2006), 115–119. MR 2007a:11058 Zbl 1108.11041

[Laurinčikas and Matsumoto 2004] A. Laurinčikas and K. Matsumoto, "The joint universality of twisted automorphic *L*-functions", *J. Math. Soc. Japan* **56**:3 (2004), 923–939. MR 2005h:11100 Zbl 1142.11032

[Nakamura and Pánkowski 2012] T. Nakamura and Ł. Pánkowski, "Any non-monomial polynomial of the Riemann zeta-function has complex zeros off the critical line", preprint, 2012. arXiv 1212.5890

[Rudnick and Sarnak 1996] Z. Rudnick and P. Sarnak, "Zeros of principal *L*-functions and random matrix theory", *Duke Math. J.* **81**:2 (1996), 269–322. MR 97f:11074 Zbl 0866.11050

[Saias and Weingartner 2009] E. Saias and A. Weingartner, "Zeros of Dirichlet series with periodic coefficients", *Acta Arith.* **140**:4 (2009), 335–344. MR 2010m:11107 Zbl 1205.11101

[Voronin 1975] S. M. Voronin, "A theorem on the "universality" of the Riemann zeta-function", *Izv. Akad. Nauk SSSR Ser. Mat.* **39**:3 (1975), 475–486, 703. In Russian; translated in *Math. UUSR-Izv.* **9**:3 (1975), 443–453. MR 57 #12419 Zbl 0315.10037

[Wu and Ye 2007] J. Wu and Y. Ye, "Hypothesis H and the prime number theorem for automorphic representations", *Funct. Approx. Comment. Math.* **37**:part 2 (2007), 461–471. MR 2009g:11063 Zbl 1230.11065

andrew.booker@bristol.ac.uk   *School of Mathematics, University of Bristol,*
                                   University Walk, Bristol, BS8 1TW, United Kingdom

thorne@math.sc.edu   *Department of Mathematics, University of South Carolina,*
                                   1523 Greene Street, Columbia, SC 29208, United States

Tropical independence I: Shapes of divisors and a proof of the Gieseker–Petri theorem

David Jensen and Sam Payne

We develop a framework to apply tropical and nonarchimedean analytic methods to multiplication maps for linear series on algebraic curves, studying degenerations of these multiplications maps when the special fiber is not of compact type. As an application, we give a new proof of the Gieseker–Petri theorem, including an explicit tropical criterion for a curve over a valued field to be Gieseker–Petri general.

1. Introduction

Classical Brill–Noether theory studies the schemes $\mathcal{G}^r_d(X)$ parametrizing linear series of degree d and rank r on a smooth curve X of genus g. The Brill–Noether number $\rho(g, r, d) = g - (r+1)(g-d+r)$ is a naive dimension estimate for $\mathcal{G}^r_d(X)$, and the following two fundamental results give the local structure of these schemes when the curve is general in its moduli space.

Brill–Noether Theorem [Griffiths and Harris 1980]. *Let X be a general curve of genus g. Then $\mathcal{G}^r_d(X)$ has pure dimension $\rho(g, r, d)$, if this is nonnegative, and is empty otherwise.*

Gieseker–Petri Theorem [Gieseker 1982]. *Let X be a general curve of genus g. Then $\mathcal{G}^r_d(X)$ is smooth.*

The Zariski tangent space to $\mathcal{G}^r_d(X)$ at a linear series $W \subset \mathcal{L}(D_X)$ has dimension $\rho(g, r, d) + \dim \ker \mu_W$, where

$$\mu_W : W \otimes \mathcal{L}(K_X - D_X) \to \mathcal{L}(K_X)$$

is the adjoint multiplication map. In particular, $\mathcal{G}^r_d(X)$ is smooth of dimension $\rho(g, r, d)$ at a linear series W if and only if the multiplication map μ_W is injective [Arbarello et al. 1985, §IV.4].
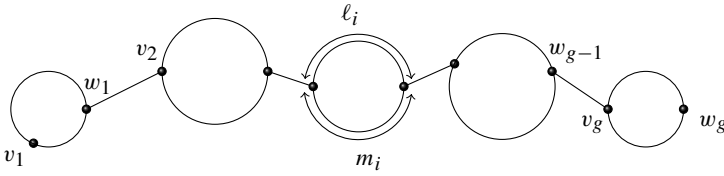
Figure 1. The graph Γ.

Gieseker's original proof that μ_W is injective for all W when X is general involves a subtle degeneration argument. Eisenbud and Harris [1983; 1986] developed a more systematic method for studying limits of linear series for one-parameter degenerations of curves in which the special fiber has compact type, and applied this theory to give a simpler proof of the Gieseker–Petri theorem. Lazarsfeld [1986] gave another proof, without degenerations, using vector bundles on K3 surfaces.

Here, we give a new proof of the Gieseker–Petri theorem, using a different class of degenerations, where the special fiber is not of compact type. Our arguments are based in tropical geometry and Berkovich's theory of nonarchimedean analytic curves and their skeletons.

Let Γ be a chain of g loops connected by bridges, with generic edge lengths.

The genericity condition on edge lengths on the loops is the same as in [Cools et al. 2012]; we require that ℓ_i/m_i is not equal to the ratio of two positive integers whose sum is less than or equal to $2g - 2$.

Theorem 1.1. *Let X be a smooth projective curve of genus g over a complete nonarchimedean field such that the minimal skeleton of the Berkovich analytic space X^{an} is isometric to Γ. Then the multiplication map*

$$\mu_W : W \otimes \mathscr{L}(K_X - D_X) \to \mathscr{L}(K_X)$$

is injective for all linear series $W \subset \mathscr{L}(D_X)$ on X.

There do exist such curves over valued fields of arbitrary pure or mixed characteristic. This follows from the fact that the moduli space of tropical curves is the skeleton of the Deligne–Mumford compactification of the moduli space of curves [Abramovich et al. 2012], and can also be proved by deformation theory, as in [Baker 2008, Appendix B]. The existence of Gieseker–Petri general curves over an arbitrary algebraically closed field then follows by standard arguments from scheme theory, using the fact that the coarse moduli space of curves is defined over Spec $\mathbb{Z}$, as in [Cools et al. 2012, Section 3]. In particular, the Gieseker–Petri theorem follows from Theorem 1.1, by standard arguments.

The proof of Theorem 1.1 is essentially independent of the tropical proof of the Brill–Noether theorem and does not involve the combinatorial classification of special divisors on a chain of loops from [Cools et al. 2012]. (In Section 6, we give

a simplified proof in the special case where $\rho(g, r, d)$ is zero, which does use this classification; see Remark 1.5.) Our approach involves not only the distribution of degrees over components of the special fiber, but also algebraic geometry over the residue field. In particular, we use Thuillier's nonarchimedean analytic Poincaré–Lelong formula [Thuillier 2005; Baker et al. 2011], which relates orders of vanishing at nodes in the special fiber of a semistable model to slopes of piecewise linear functions on the skeleton. The resulting interplay between tropical geometry and algebraic linear series is close in spirit to the important recent work [Amini and Baker 2014] on linear series on metrized complexes of curves, which was a source of inspiration.

Remark 1.2. The graph Γ differs from the chain of loops studied in [Cools et al. 2012] only by the addition of bridges between the loops. The tropical Jacobians of two graphs that differ by the addition or deletion of bridges are canonically isomorphic, and these isomorphisms respect the images of the Abel–Jacobi maps, so the Brill–Noether theory of Γ is the same as that of the chain of loops. See [Lim et al. 2012; Len 2014] for the basics of tropical Brill–Noether theory.

We do not need to introduce bridges for the case where $\rho(g, r, d)$ is zero; the arguments in Section 6 work equally well for a chain of loops without bridges. However, when $\rho(g, r, d)$ is positive we need to relate the slopes of piecewise linear functions along the bridge edges to orders of vanishing at nodes in the special fiber, through the nonarchimedean Poincaré–Lelong formula, in order to produce bases for the algebraic linear series $\mathcal{L}(D_X)$ with the required properties. In particular, we do not know whether the conclusion of Theorem 1.1 holds for chains of loops without bridges when $\rho(g, r, d)$ is positive.

On the way to proving Theorem 1.1, we introduce some new techniques for working with tropical linear series and relating them to algebraic linear series. In Section 3A, we present a notion of *tropical independence*, which gives a sufficient condition for linear independence of rational functions on an algebraic curve X in terms of the associated piecewise linear functions on the Berkovich skeleton of the analytic curve X^{an}. The key to applying such an independence condition is to produce well-understood piecewise linear functions on the skeleton that are not only in the tropical linear series, but are in fact tropicalizations of rational functions in a given algebraic linear series. In the case where $\rho(g, r, d)$ is zero, the necessary piecewise linear functions come from tropicalizing a basis for the linear series and a basis for the adjoint linear series. In this case, the piecewise linear functions are explicit and uniquely determined by the graph, and the proof that they all come from the algebraic linear series is essentially combinatorial. (See Proposition 6.3.) When ρ is positive, we have much less control over which tropical functions come from a given algebraic linear series. In the general case, we work one loop at a

time on the metric graph and use an existence argument from algebraic geometry, inspired by [Eisenbud and Harris 1983, Lemma 1.2]. (See Lemma 7.2.)

One new insight on the tropical side is the importance of *shapes* of effective divisors, expressed in terms of connected subsets that do or do not meet the divisor. When the metric graph is a chain of loops, a typical connected subset to consider would be a loop minus a single point. See Sections 3B and 4B, along with the proofs of Theorems 6.6 and 1.1 at the ends of Sections 6 and 7, respectively.

We also use a new *patching* construction, gluing together tropicalizations of different rational functions in a fixed algebraic linear series on different parts of the graph, to arrive at a piecewise linear function in the corresponding tropical linear series that may or may not come from any linear combination of the original rational functions. See the construction of θ at the beginning of the proof of Theorem 1.1. The most delicate step in this construction is to ensure that no poles are introduced at the gluing points.

We now briefly sketch relations between the approach developed here, the classical theory of limit linear series, and the tropical theory of divisors on graphs.

Suppose X is defined over a discretely valued field with valuation ring R, and let L be a line bundle on X. Consider a regular model $\mathcal{X}$ over Spec R with general fiber X, in which the special fiber $\overline{\mathcal{X}}$ is semistable with smooth components $\overline{\mathcal{X}}_i$. (By the semistable reduction theorem, such a model exists after a finite, totally ramified extension of the valued field.) The special fiber of this model has compact type, meaning that its Jacobian is compact, if and only if its dual graph is a tree. In this case, for each component $\overline{\mathcal{X}}_i$ there is a unique extension $\mathcal{L}_i$ of the line bundle L such that

$$\deg\left(\mathcal{L}_i|_{\overline{\mathcal{X}}_j}\right) = \begin{cases} d & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Given a linear subspace $W \subset H^0(X, L)$ of degree d and dimension $r + 1$, the R-submodule $\mathcal{W}_i \subset W$ consisting of sections that extend to $\mathcal{L}_i$ is free of rank $r + 1$, and restricts to a linear series of degree d and dimension r on $\overline{\mathcal{X}}_i$. The theory of limit linear series studies these distinguished linear series on the components of the special fiber, with special attention to their vanishing sequences at the nodes of $\overline{\mathcal{X}}$.

In contrast, if $\overline{\mathcal{X}}$ is not of compact type, then its dual graph is not a tree, and there is an obstruction to extending L to a line bundle $\mathcal{L}_i$ with degrees as above on the components of the special fiber. This obstruction is given by an element in the component group of the Néron model of the Jacobian of X.

The theory of divisors on graphs follows a deep analogy between divisors on algebraic curves and the distributions of degrees of specializations of L over the components of the special fiber. In this framework, one considers the dual graph whose vertices v_i correspond to components $\overline{\mathcal{X}}_i$ and whose edges correspond to

nodes of $\overline{\mathscr{X}}$. Then an extension $\mathscr{L}$ of L to $\mathscr{X}$ gives rise to a formal sum

$$D_{\mathscr{L}} = \sum_i \deg\big(\mathscr{L}|_{\overline{\mathscr{X}}_i}\big) v_i,$$

which is considered as a divisor on the graph. Since the divisors arising from different specializations of L differ by a sequence of chip-firing moves, one studies the tropical Picard group parametrizing equivalence classes of divisors on the graph modulo the relation generated by chip-firing. The tropical Jacobian, the degree zero part of this tropical Picard group, is canonically identified with the component group of the Néron model of the Jacobian of X.

Baker's specialization lemma [2008] says that a line bundle whose complete linear series has dimension r can be specialized so that all degrees are nonnegative and the distribution of degrees dominates any given divisor of degree r on the dual graph. In other words, it has rank at least r in the sense of [Baker and Norine 2007]. Therefore, the specialization of any line bundle whose complete linear series has dimension at least r lies in the tropical Brill–Noether locus parametrizing divisor classes of degree d with rank at least r. In [Cools et al. 2012], a careful analysis of the Brill–Noether loci of the chain of loops shows that if a curve X has a regular semistable model whose special fiber has this dual graph, then the curve must be Brill–Noether general, meaning that $\mathscr{G}_d^r(X)$ has dimension $\rho(g, r, d)$ if this is nonnegative, and is empty otherwise. In particular, we get not only a new proof of the Brill–Noether theorem, but an explicit and computationally verifiable sufficient condition for a curve to be Brill–Noether general, the existence of a regular semistable model whose special fiber has a particular dual graph.

Remark 1.3. This tropical proof of the Brill–Noether theorem can be reframed in the language of Berkovich's nonarchimedean analytic geometry to show that any curve of genus g over a valued field whose skeleton is a chain of g loops with generic edge lengths must be Brill–Noether general. Here, we follow this more general approach, with skeletons of analytifications in place of dual graphs of regular semistable models. Similar arguments, combined with the basepoint-free pencil trick, lead to a proof of the Gieseker–Petri theorem in the special case where $r = 1$ [Baratham et al. 2014].

Remark 1.4. In some ways, the tropical geometry of divisors on a chain of loops with generic edge lengths appears similar to the geometry of limit linear series on a chain of elliptic curves with generic attaching points. As is well-known to experts in Brill–Noether theory, the theory of limit linear series on such curves gives a characteristic-free proof of the Brill–Noether and Gieseker–Petri theorems [Osserman 2011; Castorena et al. 2012], and some steps in our approach, including Lemma 7.2 and Proposition 7.4, can be viewed as tropical analogues of such arguments from classical algebraic geometry.

Other steps seem more difficult to translate. In the limit linear series proofs of Gieseker–Petri, both [Eisenbud and Harris 1983] and [Castorena et al. 2012] assume the multiplication map is not injective and use a degeneration argument to construct a divisor in $|K_X|$ of *impossible degree*. We assume the multiplication map is not injective and reach a contradiction by constructing an impossible divisor in $|K_\Gamma|$, but it is not the degree of this divisor that creates the contradiction. Our argument relies on Proposition 3.5 and Lemma 4.4 to show that the divisor has *impossible shape*.

The relations to the geometry of the Deligne–Mumford compactification of $\mathcal{M}_g$ are also different. Limit linear series arguments produce stable curves corresponding to points in the boundary of $\overline{\mathcal{M}}_g$ that are not in the closure of the Gieseker–Petri special locus, whereas the special fibers of our models are semistable, but necessarily unstable, and their stabilizations are always in the closure of the hyperelliptic locus. (Limit linear series arguments may also involve semistable curves that are not stable, but the configurations of rational curves collapsed by stabilization tend to play an incidental role. In sharp contrast, the precise combinatorial configurations of collapsed curves are essential to our arguments.)

It may still be tempting to try to interpret the tropical approach as a rephrasing or retranslation of classical degeneration arguments, at least in broad strokes, but there are fundamental obstacles to overcome. As explained above, the data in our tropical arguments are in some sense strictly complementary to the data involved in limit linear series. We work primarily in the component group of the Néron model of the Jacobian (or its analytic counterpart, the tropical Jacobian) whereas classical limit linear series are defined only in the case where this component group is trivial. On the other hand, the limit linear series approach depends on computations in the compact part of the Jacobian of the special fiber, which is trivial in the cases we consider.

Finally, we note that even the tropical Riemann–Roch theorem has not been reinterpreted or reproved using classical algebraic geometry, despite multiple attempts. Our proof of Gieseker–Petri uses this result in a crucial way, to control the shapes of effective canonical divisors (Lemma 4.4), so any satisfying interpretation of our argument in terms of classical degeneration methods should explain tropical Riemann–Roch as well.

Remark 1.5. In Section 6, we give a simplified proof of Theorem 1.1 in the special case where $\rho(g, r, d)$ is zero. The simplified argument in this special case is essentially combinatorial, and relies on the classification of special divisors on a chain of loops in terms of rectangular tableaux [Cools et al. 2012] and the interpretation of adjunction in terms of transposition [Agrawal et al. 2013]. It does not involve algebraic geometry over the residue field or the Poincaré–Lelong formula.

Although the guts of the argument are different, the overall structure of the proof by contradiction is the same as in the general case. We assume that the multiplication

map has nonzero kernel, deduce that certain carefully constructed collections of piecewise linear functions are tropically dependent, and use this dependence to produce a canonical divisor of impossible shape. Although this section is not logically necessary, we believe that most readers will find it helpful to work through this special case first, as we did, before proceeding to the proof of Theorem 1.1.

2. Background

We briefly review the theory of divisors and divisor classes on metric graphs, along with relations to the classical theory of algebraic curves via Berkovich analytification and specialization to skeletons. For further details and references, see [Baker and Norine 2007; Baker 2008; Baker et al. 2011; Amini and Baker 2014].

2A. *Divisors on graphs and Riemann–Roch.* Let Γ be a metric graph. A *divisor* on Γ is a finite formal sum

$$D = a_1 v_1 + \cdots + a_s v_s,$$

where the v_i are points in Γ and the coefficients a_i are integers. The *degree* of a divisor is the sum of its coefficients

$$\deg(D) = a_1 + \cdots + a_s,$$

and a divisor is *effective* if all of its coefficients are nonnegative. We say that an effective divisor *contains* a point v_i if its coefficient a_i is strictly positive. We will frequently consider questions about whether a given effective divisor D contains at least one point in a connected subset $\Gamma' \subset \Gamma$. See, for instance, Section 3B.

Let $\mathrm{PL}(\Gamma)$ be the additive group of continuous piecewise linear functions ψ with integer slopes on Γ. (Throughout, all of the piecewise linear functions that we consider have integer slopes.) The *order* of such a piecewise linear function ψ at a point v is the sum of its incoming slopes along edges containing v, and is denoted $\mathrm{ord}_v(\psi)$. Note that $\mathrm{ord}_v(\psi)$ is zero for all but finitely many points v in Γ, so

$$\mathrm{div}(\psi) = \sum_{v \in \Gamma} \mathrm{ord}_v(\psi)\, v$$

is a divisor. A divisor is *principal* if it is equal to $\mathrm{div}(\psi)$ for some piecewise linear function ψ, and two divisors D and D' are *equivalent* if $D - D'$ is principal. Note that every principal divisor has degree zero, so the group $\mathrm{Pic}(\Gamma)$ of equivalence classes of divisors is graded by degree.

Let D be a divisor on Γ. The *complete linear series* $|D|$ is the set of effective divisors on Γ that are equivalent to D, and

$$R(D) = \{\psi \in \mathrm{PL}(\Gamma) \mid D + \mathrm{div}(\psi) \text{ is effective}\}.$$

These objects are closely analogous to the complete linear series of a divisor on an algebraic curve and the vector space of rational functions with poles bounded by that divisor. There is a natural surjective map from $R(D)$ to $|D|$ taking a piecewise linear function ψ to $\mathrm{div}(\psi) + D$, and two functions ψ and ψ' have the same image in $|D|$ if and only if $\psi - \psi'$ is constant. The vector space structure on rational functions with bounded poles is analogous to the *tropical module* structure on $R(D)$. Addition in this tropical module is given by the pointwise minimum; if $\psi_0, \ldots, \psi_r$ are in $R(D)$ and $b_0, \ldots, b_r$ are real numbers, then the function θ given by

$$\theta(v) = \min_j \big\{ \psi_j(v) + b_j \big\},$$

is also in $R(D)$ [Haase et al. 2012].

The *rank $r(D)$* is the largest integer r such that $D - E$ is equivalent to an effective divisor for every effective divisor E of degree r. In other words, a divisor D has rank at least r if and only if its linear series contains divisors that dominate any effective divisor of degree r. This invariant satisfies the following Riemann–Roch theorem with respect to the *canonical divisor* $K_\Gamma = \sum_{v \in \Gamma} (\deg(v) - 2) \, v$:

Tropical Riemann–Roch Theorem [Baker and Norine 2007; Gathmann and Kerber 2008; Mikhalkin and Zharkov 2008]. *Let D be a divisor on a metric graph Γ with first Betti number g. Then*

$$r(D) - r(K_\Gamma - D) = \deg(D) - g + 1.$$

Remark 2.1. Although it is closely analogous to the classical Riemann–Roch theorem for curves, the tropical Riemann–Roch theorem has no known proof via algebraic geometry. Indeed, neither of these results is known to imply the other.

2B. *Specialization of divisors from curves to graphs.* Throughout, we work over a fixed algebraically closed field K that is complete with respect to a nontrivial valuation

$$\mathrm{val} : K^* \to \mathbb{R}.$$

Let $R \subset K$ be the valuation ring, and let κ be the residue field.

Let X be an algebraic curve over K. The underlying set of the Berkovich analytic space X^{an} consists of the closed points $X(K)$ together with the set of valuations on the function field $K(X)$ that extend the given valuation on K. We write

$$\mathrm{val}_y : K(X) \to \mathbb{R} \cup \{+\infty\}$$

for the valuation corresponding to a point y in $X^{\mathrm{an}} \setminus X(K)$.

Remark 2.2. We treat the points in $X(K)$ differently, because they do not correspond to valuations on the function field $K(X)$. Nevertheless, one can still study

the closed points in terms of generalized valuations on rings, as follows. If $U \subset X$ is any affine open neighborhood of a closed point $x \in X(K)$, then the map

$$\mathrm{val}_x : \mathbb{O}_X(U) \to \mathbb{R} \cup \{+\infty\}$$

is a ring valuation. Note that val_x, unlike a valuation on a field, may take a nonzero element to $+\infty$.

The topology on X^{an} is the weakest containing U^{an} for every Zariski-open U in X and such that, for any $f \in \mathbb{O}_X(U)$, the function taking $x \in U^{\mathrm{an}}$ to $\mathrm{val}_x(f)$ is continuous.

The points in $X(K)$ are called type-1 points of X^{an}, and the remaining points in $X^{\mathrm{an}} \setminus X(K)$ are classified into three more types according to the algebraic properties of the corresponding valuation on $K(X)$. For our purposes, the most relevant points are type-2 points, the points y such that the residue field of $K(X)$ with respect to val_y has transcendence degree 1 over κ. We write X_y for the smooth projective curve over the residue field of K with this function field.

Remark 2.3. By passing to a spherically complete extension field whose valuation surjects onto $\mathbb{R}$, one could assume that all points in $X^{\mathrm{an}} \setminus X(K)$ are of type-2.

Suppose X is smooth and projective. Then X has a *semistable vertex set*, a finite set of type-2 points whose complement is a disjoint union of a finite number of open annuli and an infinite number of open balls. Each semistable vertex set $V \subset X^{\mathrm{an}}$ corresponds to a semistable model $\mathscr{X}_V$ of X. The normalized irreducible components of the special fiber $\overline{\mathscr{X}}_V$ are naturally identified with the curves X_y, for $y \in V$, and the preimages of the nodes in $\overline{\mathscr{X}}_V$ under specialization are the annuli in $X^{\mathrm{an}} \setminus V$. The annulus corresponding to a node where X_y meets $X_{y'}$ contains a unique embedded open segment with endpoints y and y', whose length is the logarithmic modulus of the annulus. The union of these open segments together with V is a closed connected metric graph embedded in $X^{\mathrm{an}} \setminus X(K)$ with a natural metric. We write Γ_V for this metric graph, and call it the *skeleton* of the semistable model $\mathscr{X}_V$. If X has genus at least 2, which we may assume since the Gieseker–Petri theorem is trivial for curves of genus 0 and 1, there is a unique minimal semistable vertex set in X^{an}. We write Γ for the skeleton of this minimal semistable vertex set, and call it simply the *skeleton* of X^{an}.

Each connected component of $X^{\mathrm{an}} \setminus \Gamma$ has a unique boundary point in Γ, and there is a canonical retraction to the skeleton

$$X^{\mathrm{an}} \to \Gamma$$

taking a connected component of $X^{\mathrm{an}} \setminus \Gamma$ to its boundary point. Restricting to $X(K)$ and extending linearly gives the tropicalization map on divisors

$$\mathrm{Trop} : \mathrm{Div}(X) \to \mathrm{Div}(\Gamma).$$

This map respects rational equivalence of divisors, as follows.

Let $f \in K(X)$ be a rational function. We write $\mathrm{trop}(f)$ for the real-valued function on the skeleton Γ given by $y \mapsto \mathrm{val}_y(f)$. The function $\mathrm{trop}(f)$ is piecewise linear with integer slopes. Furthermore, if y is a type-2 point and $\mathrm{trop}(f)(y) = 0$, then the residue $\bar{f}_y$ is a nonzero rational function on X_y whose slope along an edge incident to y is the order of vanishing of $\bar{f}_y$ at the corresponding node. This is the nonarchimedean Poincaré–Lelong formula, due to Thuillier; see [Thuillier 2005] and [Baker et al. 2011, §5]. One immediate consequence of this formula is that the tropical specialization map for rational functions

$$\mathrm{trop} : K(X)^* \to \mathrm{PL}(\Gamma)$$

is compatible with passing to principal divisors. More precisely, for any nonzero rational function $f \in K(X)$, we have

$$\mathrm{Trop}(\mathrm{div}(f)) = \mathrm{div}(\mathrm{trop}(f)).$$

Therefore, the tropicalization map on divisors respects equivalences and descends to a natural map on Picard groups

$$\mathrm{Trop} : \mathrm{Pic}(X) \to \mathrm{Pic}(\Gamma).$$

Furthermore, since tropicalizations of effective divisors are effective, if D_X is a divisor on X and f is a rational function in $\mathcal{L}(D_X)$, then $\mathrm{trop}(f)$ is in $R(\mathrm{Trop}(D_X))$. This leads to the following version of Baker's specialization lemma:

Lemma 2.4. *Let D_X be a divisor on X. Then $r(\mathrm{Trop}(D_X)) \geq r(D_X)$.*

Here, the rank $r(D_X)$ is the dimension of the complete linear series of D_X on X.

Remark 2.5. The specialization lemma and Riemann–Roch theorem together imply that $\mathrm{Trop}(K_X) = K_\Gamma$, and hence tropicalization respects adjunction. In other words, $\mathrm{Trop}(K_X - D_X) = K_\Gamma - \mathrm{Trop}(D_X)$.

Remark 2.6. Note that $\mathrm{trop}(\mathcal{L}(D_X))$ is often much smaller than $R(\mathrm{Trop}(D_X))$. It is difficult in general to determine which piecewise linear functions in $R(\mathrm{Trop}(D_X))$ are tropicalizations of rational functions in $\mathcal{L}(D_X)$.

3. Tropical multiplication maps

We now introduce a basic tropical lemma for studying linear dependence of rational functions and ranks of multiplication maps on linear series.

3A. *Tropical independence.* Let $f_0, \ldots, f_r$ be rational functions on X. Suppose $\{f_0, \ldots, f_r\}$ is linearly dependent, so there are constants $c_0, \ldots, c_r$ in K, not all zero, such that

$$c_0 f_0 + \cdots + c_r f_r = 0.$$

Then, for any point $v \in X^{\mathrm{an}}$, the minimum of the valuations

$$\{\mathrm{val}_v(c_0 f_0), \ldots, \mathrm{val}_v(c_r f_r)\}$$

must occur at least twice. In particular, if $f_0, \ldots, f_r$ are linearly dependent in $K(X)$ then there are real numbers $b_0, \ldots, b_r$ such that the minimum of the piecewise linear functions $\{\mathrm{trop}(f_0) + b_0, \ldots, \mathrm{trop}(f_r) + b_r\}$ occurs at least twice at every point of the skeleton Γ. Here, take $b_j = \mathrm{val}(c_j)$ if c_j is nonzero, and otherwise make b_j sufficiently large such that $\psi_j + b_j$ is never minimal.

Definition 3.1. A set of piecewise linear functions $\{\psi_0, \ldots, \psi_r\}$ is *tropically dependent* if there are real numbers $b_0, \ldots, b_r$ such that the minimum

$$\min\{\psi_0(v) + b_0, \ldots, \psi_r(v) + b_r\}$$

occurs at least twice at every point v in Γ.

If there are no such real numbers $b_0, \ldots, b_r$ then we say $\{\psi_0, \ldots, \psi_r\}$ is *tropically independent*.

Lemma 3.2. *Let D_X and E_X be divisors on X, with $\{f_0, \ldots, f_r\}$ and $\{g_0, \ldots, g_s\}$ bases for $\mathscr{L}(D_X)$ and $\mathscr{L}(E_X)$, respectively. If $\{\mathrm{trop}(f_i) + \mathrm{trop}(g_j)\}_{ij}$ is tropically independent, then the multiplication map*

$$\mu : \mathscr{L}(D_X) \otimes \mathscr{L}(E_X) \to \mathscr{L}(D_X + E_X)$$

is injective.

Proof. The elementary tensors $f_i \otimes g_j$ form a basis for $\mathscr{L}(D_X) \otimes \mathscr{L}(E_X)$. The image of $f_i \otimes g_j$ under μ is the rational function $f_i g_j$, and these are linearly independent, since their tropicalizations are tropically independent. $\square$

Remark 3.3. The main difficulty in applying this lemma is that one must prove the existence of rational functions in the algebraic linear series whose tropicalizations have the appropriate independence property. Finding such piecewise linear functions in the tropical linear series is not enough.

3B. *Shapes of equivalent divisors.* Here we prove a technical proposition about how the tropical module structure on $R(D)$ is reflected in the shapes of divisors in $|D|$. The proposition will be particularly useful when combined with our notion of tropical dependence of piecewise linear functions.

Lemma 3.4. *Let D be a divisor on a metric graph Γ, with $\psi_0, \ldots, \psi_r$ piecewise linear functions in $R(D)$, and let*

$$\theta = \min\{\psi_0, \ldots, \psi_r\}.$$

Let $\Gamma_j \subset \Gamma$ be the closed set where $\theta = \psi_j$. Then $\mathrm{div}(\theta) + D$ contains a point $v \in \Gamma_j$ if and only if v is in either

(1) *the divisor* $\operatorname{div}(\psi_j) + D$, *or*

(2) *the boundary of* Γ_j.

Proof. If ψ_j agrees with θ on some open neighborhood of v, then $\operatorname{ord}_v(\theta) = \operatorname{ord}_v(\psi_j)$, and hence $\operatorname{div}(\theta) + D$ contains v if and only if $\operatorname{div}(\psi_j) + D$ does. On the other hand, if v is in the boundary of Γ_j then there is an edge containing v such that the incoming slope of θ along this edge is strictly greater than that of ψ_j, and the incoming slope of θ along any other edge containing v must be at least as large as that of ψ_j. By summing over all edges containing v, we find that $\operatorname{ord}_v(\theta)$ is strictly greater than $\operatorname{ord}_v(\psi_j)$. Since $\operatorname{div} \psi_j + D$ is effective, by hypothesis, it follows that the coefficient of v in $\operatorname{div}(\theta) + D$ is strictly positive, as required. $\qquad\square$

Proposition 3.5. *Let D be a divisor on a metric graph Γ, with $\psi_0, \ldots, \psi_r$ in $R(D)$, and*

$$\theta = \min\{\psi_0, \ldots, \psi_r\}.$$

Let $\Gamma' \subset \Gamma$ be a connected subset, and suppose that $\operatorname{div}(\psi_j) + D$ contains a point in Γ' for all j. Then $\operatorname{div}(\theta) + D$ also contains a point in Γ'.

Proof. Pick j such that θ is equal to ψ_j at some point in Γ', and let

$$\Gamma'_j = \{v \in \Gamma' \mid \theta(v) = \psi_j(v)\}.$$

If Γ'_j is properly contained in Γ', then its boundary is nonempty, since Γ' is connected, and each of the boundary points is contained in $\operatorname{div}(\theta) + D$, by Lemma 3.4.

Otherwise, if θ agrees with ψ_j on all of Γ', then $\operatorname{div}(\theta) + D$ contains the points of $\operatorname{div}(\psi_j) + D$ in Γ', and the proposition follows. $\qquad\square$

4. The chain of loops with bridges

We now restrict attention to the specific graph Γ shown in Figure 1, consisting of a chain of g loops separated by bridges. Throughout, we assume that the loops of Γ have generic edge lengths in the same sense as in [Cools et al. 2012], meaning that ℓ_i / m_i is never equal to the ratio of two positive integers whose sum is less than or equal to $2g - 2$.

4A. *Reduced divisors.* Fix a point $v \in \Gamma$. Recall that an effective divisor D is v-reduced if the multiset of distances from v to points in D is lexicographically minimal among all effective divisors equivalent to D. Every effective divisor is equivalent to a unique v-reduced divisor, and the rank of a v-reduced divisor is bounded above by the coefficient of v. In particular, if D is a v-reduced divisor that does not contain v, then $r(D)$ is zero. See [Luo 2011, Proposition 2.1].

It is relatively straightforward to classify v-reduced divisors on Γ. We will only need the special case of w_g-reduced divisors. For each i, let γ_i be the i-th loop
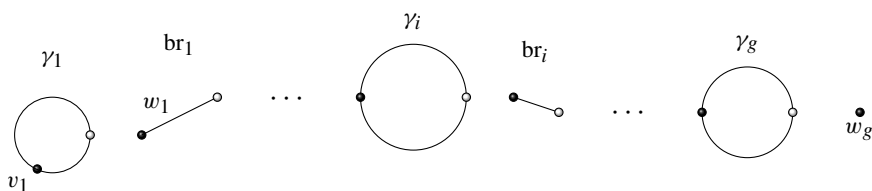
Figure 2. A decomposition of Γ.

minus w_i, the union of the two half-open edges $[v_i, w_i)$, and let br_i be the half-open bridge $[w_i, v_{i+1})$. Note that Γ decomposes as a disjoint union

$$\Gamma = \gamma_1 \sqcup \mathrm{br}_1 \sqcup \cdots \sqcup \gamma_g \sqcup \{w_g\},$$

as shown in Figure 2.

Proposition 4.1. *An effective divisor D is w_g-reduced if and only if it contains*

(1) *no points in the bridges* $\mathrm{br}_1, \ldots, \mathrm{br}_{g-1}$, *and*

(2) *at most one point in each cell* $\gamma_1, \ldots, \gamma_g$.

Proof. This is a straightforward application of Dhar's burning algorithm, as in [Cools et al. 2012, Example 2.6]. $\qquad\square$

4B. *The shape of a canonical divisor.* As mentioned in the introduction, our strategy is a proof by contradiction; we assume that a multiplication map has nonzero kernel and use Proposition 3.5 to construct a canonical divisor of *impossible shape*.

The following basic lemma, which we state and prove but do not use, restricts the possibilities for the shape of a canonical divisor on an arbitrary graph:

Lemma 4.2. *Let Γ' be a metric graph of genus g, let $e_1, \ldots, e_g$ be disjoint open edges of Γ' whose complement is a tree, and let D be an effective divisor equivalent to $K_{\Gamma'}$. Then at least one of the open edges $e_1, \ldots, e_g$ contains no point of D.*

Proof. Suppose that each open edge $e_1, \ldots, e_g$ contains a point of D, let p_i be a point in e_i, and let $D' = p_1 + \cdots + p_g$. Since $K_{\Gamma'} - D'$ is effective, by construction, the tropical Riemann–Roch theorem says that $r(D')$ is at least 1. However, Dhar's burning algorithm [1990] shows that D' is v-reduced for any point v in the complement of $e_1 \cup \cdots \cup e_g$. Since D' does not contain v, it follows that $r(D')$ is zero. $\qquad\square$

Remark 4.3. Lemma 4.2 also follows from the rigidity of effective representatives for classes in the relative interiors of top-dimensional cells in the natural subdivision of $\mathrm{Pic}_g(\Gamma)$ into parallelotopes studied by An, Baker, Kuperberg, and Shokrieh [An et al. 2014, Lemma 3.5].

On the chain of loops with bridges, we can use the classification of w_g-reduced divisors to refine the preceding lemma as follows:

Lemma 4.4. *Let D be an effective divisor equivalent to K_Γ. Then D contains no point in at least one of the cells $\gamma_1, \ldots, \gamma_g$.*

Proof. Suppose each cell $\gamma_1, \ldots, \gamma_g$ contains a point of D. Let p_i be a point of D in γ_i, and let $D' = p_1 + \cdots + p_g$. Then $K_\Gamma - D'$ is equivalent to an effective divisor, by construction, so the tropical Riemann–Roch theorem says that $r(D')$ is at least 1. However, D' is w_g-reduced by Proposition 4.1 and does not contain w_g, so $r(D')$ is zero. □

Remark 4.5. Note that the point p_i in the proof of Lemma 4.4 may be equal to v_i for some $2 \leq i \leq g$, in which case the complement of $\{p_1, \ldots, p_g\}$ is not a tree. For this reason, the lemma does not follow from Lemma 4.2. We use Lemma 4.4 to obtain contradictions and prove our main results at the end of Sections 6 and 7.

5. Preliminaries for the proof of injectivity

Let X be a curve over K with skeleton Γ, and let D_X be a divisor of degree d and rank r on X. To prove that X is Gieseker–Petri general we must show that the multiplication map μ_W is injective for every linear subspace $W \subset \mathcal{L}(D_X)$. It clearly suffices to consider the case where $W = \mathcal{L}(D_X)$. In other words, we must show that

$$\mu : \mathcal{L}(D_X) \otimes \mathcal{L}(K_X - D_X) \to \mathcal{L}(K_X)$$

is injective.

Given Lemma 3.2, a natural strategy is to show that there are bases $\{f_i\}$ and $\{g_j\}$ for $\mathcal{L}(D_X)$ and $\mathcal{L}(K_X - D_X)$, respectively, such that the set of piecewise linear functions

$$\{\text{trop}(f_i) + \text{trop}(g_j)\}_{ij}$$

is tropically independent. We prove the existence of such a basis when the Brill–Noether number $\rho(g, r, d)$ is zero. The following section, which treats this special case, is not logically necessary for the proof of Theorem 1.1. However, the basic strategy that we use is the same as in the general case, only the details are simpler.

Remark 5.1. When $\rho(g, r, d)$ is positive, we do not know whether there are bases $\{f_i\}, \{g_j\}$ for $\mathcal{L}(D_X)$ and $\mathcal{L}(K_X - D_X)$, respectively, such that $\{\text{trop}(f_i) + \text{trop}(g_j)\}$ is tropically independent.

6. A special case: Brill–Noether number zero

The results of this sections are not used in the proof of Theorem 1.1, but working through this special case where $\rho(g, r, d)$ is zero before proceeding to the proof of the general case should be helpful for most readers. An overview of the argument is as follows.

We start by assuming that the multiplication map has a kernel, and therefore the tropicalization of the image under μ of any basis for $\mathscr{L}(D_X) \otimes \mathscr{L}(K_X - D_X)$ is tropically dependent. We use this tropical dependence together with Proposition 3.5 to construct a divisor in $|K_\Gamma|$ that violates Lemma 4.4, i.e., a canonical divisor of impossible shape. When the Brill–Noether number is zero, the bases for $\mathscr{L}(D_X)$ and $\mathscr{L}(K_X - D_X)$ are explicit and canonically determined, and we only need to choose one basis for each.

Additional subtleties in the general case include the choice of g different bases for $\mathscr{L}(D_X)$ and $\mathscr{L}(K_X - D_X)$, one for each loop in Γ, and the application of Poincaré–Lelong to control the slopes of tropicalizations along the bridges. Furthermore, the bases are not explicit in the general case, but Lemma 7.2 gives the existence of bases with the required properties.

Remark 6.1. For a completely different tropical proof of the Gieseker–Petri theorem in the case $\rho(g, r, d) = 0$, using lifting arguments instead of tropical independence, see [Cartwright et al. 2014, Proposition 1.6].

Suppose D_X is a divisor of degree d and rank r on X, with $\rho(g, r, d) = 0$, and let D be the v_1-reduced divisor equivalent to $\mathrm{Trop}(D_X)$. There are only finitely many v_1-reduced divisors of degree d and rank r on Γ, and they are explicitly classified in [Cools et al. 2012]. These divisors correspond naturally and bijectively to the rectangular standard tableau with $(g - d + r)$ rows and $(r + 1)$ columns. Note that, since $\rho(g, r, d) = 0$, the genus g factors as

$$g = (r + 1)(g - d + r).$$

In particular, the entries in the tableau corresponding to D are the integers $1, \dots, g$.

Fix the tableau corresponding to D. We label the columns from 0 to r and the rows from 0 to $g - d + r - 1$. The tableau determines a Dyck path, consisting of a series of points $p_0, \dots, p_g$ in $\mathbb{Z}^r$, as follows. We write $e_0, \dots, e_{r-1}$ for the standard basis vectors on $\mathbb{Z}^r$. The starting and ending point of the Dyck path is

$$p_0 = p_g = (r, \dots, 1),$$

and the i-th step $p_i - p_{i-1}$ is equal to

- the standard basis vector e_j if i appears in the j-th column of the tableau, for $0 \le j < r$, or
- the vector $(-1, \dots, -1)$ if i appears in the last column.

The tableau properties exactly ensure that each p_i lies in the open Weyl chamber $x_0 > \cdots > x_{r-1} > 0$. We write $p_i(j)$ for the j-th coordinate of p_i.

The divisor D can be recovered from the Dyck path as follows. The coefficient of v_1 is r. If i appears in the j-th column of the tableau, for $0 \le j < r$, then D contains

the point on the i-th loop at distance $p_{i-1}(j)m_i$ modulo $(\ell_i + m_i)$ counterclockwise from w_i with coefficient 1. If i appears in the last column of the tableau, then D contains no point in the i-th loop.

Remark 6.2. In this bijection, adjunction of divisors corresponds to transposition of tableaux [Agrawal et al. 2013, Theorem 39]. Therefore, the v_1-reduced divisor E equivalent to $\mathrm{Trop}(K_X - D_X)$ is exactly the divisor corresponding to the transpose of the tableau for D.

Proposition 6.3. *For each integer* $0 \leq j \leq r$, *there is a unique divisor* D_j *equivalent to* D *such that* $D_j - j v_1 - (r - j)w_g$ *is effective. Moreover,* γ_i *contains no point of* D_j *if and only if* i *appears in the* j-th column of the tableau corresponding to D.

Proof. The divisor D_r is exactly D. The remaining divisors D_j are constructed in the proof of Proposition 4.10 in [Cools et al. 2012] by an explicit chip-firing procedure. One takes a pile of $r - j$ chips from v_1 and moves it to the right. The pile of chips changes size as it moves, and has $p_i(j)$ chips when it reaches v_i. As the pile moves across the i-th loop, there is a single chip left behind in the interior of one of the edges unless i appears in the j-th loop, in which case the i-th loop is left empty. When the pile reaches w_g, it has $p_g(j) = r - j$ chips. Since j chips were left at v_1 at the start of the procedure, $D_j - j v_1 - (r - j)w_g$ is effective. To see that D_j is the unique divisor equivalent to D with this property, note that $D_j - j v_1 - (r - j)w_g$ does not move; it is effective and contains no points on the bridges or at the vertices, and hence is v-reduced for every v in Γ. $\square$

Similarly, for $0 \leq k \leq g - d + r - 1$ there is a unique divisor E_k equivalent to the v_1-reduced adjoint divisor E such that $E_k - k v_1 - (g - d + r - 1 - k)w_g$ is effective, and γ_i contains no point of E_k if and only if i appears in the k-th row of the tableau.

It follows that the g divisors $D_j + E_k$ are distinct and correspond to the loops of Γ, as follows.

Corollary 6.4. *The connected subset* $\gamma_i \subset \Gamma$ *contains no point of* $D_j + E_k$ *if and only if* i *appears in the* j-th column and k-th row of the tableau corresponding to D.

Proposition 6.5. *There is a basis* $f_0, \ldots, f_r$ *for* $\mathscr{L}(D_X)$ *such that*

$$\mathrm{Trop}(D_X + \mathrm{div}(f_j)) = D_j.$$

Proof. Let x and y be points in $X(K)$ specializing to v_1 and w_g, respectively. Since D_X has rank r, there is a rational function $f_j \in \mathscr{L}(D_X)$ such that $D_X + \mathrm{div}(f_j)$ contains x with coefficient at least j and y with coefficient at least $r - j$. Then $\mathrm{Trop}(D_X + \mathrm{div}(f_j))$ is an effective divisor and contains v_1 and w_g with coefficient at least j and $r - j$, respectively. By Proposition 6.3, $\mathrm{Trop}(D_X + \mathrm{div}(f_j))$ must be equal to D_j. $\square$

Similarly, there is a basis $\{g_0, \ldots, g_{g-d+r-1}\}$ for $\mathcal{L}(K_X - D_X)$ such that

$$\text{Trop}(K_X - D_X + \text{div}(g_k)) = E_k.$$

We proceed to study the piecewise linear functions

$$\phi_j = \text{trop}(f_j) \quad \text{and} \quad \psi_k = \text{trop}(g_k).$$

Note that $D + \text{div}(\phi_j) = D_j$ and $E + \text{div}(\psi_k) = E_k$, and this determines each ϕ_j and ψ_k up to an additive constant.

Theorem 6.6. *The set of g piecewise linear functions $\{\phi_j + \psi_k\}_{jk}$ is tropically independent.*

Proof. Suppose that $\{\phi_j + \psi_k\}_{jk}$ is tropically dependent. Then there exist real numbers b_{jk} such that the minimum

$$\theta = \min_{j,k}\{\phi_j + \psi_k + b_{jk}\}$$

occurs at least twice at every point in Γ. Note that $D + E + \text{div}(\theta)$ is an effective canonical divisor, since $R(D + E)$ is a tropical module and D and E are adjoint.

We claim that $D + E + \text{div}\,\theta$ contains a point in each γ_i. Choose j_0 and k_0 such that i appears in the j_0-th column and k_0-th row of the tableau corresponding to D. Then Corollary 6.4 says that $D + E + \text{div}(\phi_j + \psi_k + b_{jk})$ contains a point in γ_i for $(j, k) \neq (j_0, k_0)$. Also, since the minimum of $\{\phi_j + \psi_k + b_{jk}\}$ occurs at least twice at every point of Γ, we have

$$\theta = \min_{(j,k)\neq(j_0,k_0)}\{\phi_j + \psi_k + b_{jk}\}.$$

Therefore, by Proposition 3.5, the divisor $D + E + \text{div}(\theta)$ contains a point in γ_i, as claimed.

We have shown that $D + E + \text{div}(\theta)$ is an effective canonical divisor that contains a point in each of $\gamma_1, \ldots, \gamma_g$. But this is impossible, by Lemma 4.4. $\square$

7. Proof of Theorem 1.1

As in the previous two sections, let X be a smooth projective curve of genus g over K with skeleton Γ. Since skeletons are invariant under base change with respect to extensions of algebraically closed valued fields, we can and do assume that K is spherically complete.

Remark 7.1. Spherical completeness is equivalent to completeness for discretely valued fields, but stronger in general. We use spherical completeness only in the proof of Lemma 7.2, to ensure that normed K-vector spaces have orthogonal bases.
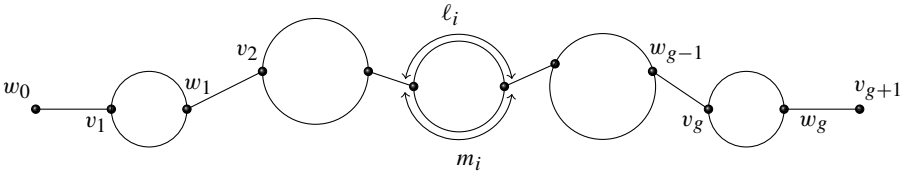
Figure 3. The skeleton Γ_V.

Let D_X be an effective divisor on X. We must show that the multiplication map

$$\mu : \mathcal{L}(D_X) \otimes \mathcal{L}(K_X - D_X) \to \mathcal{L}(D_X)$$

is injective. This is trivial if $\mathcal{L}(K_X - D_X)$ is zero, so we assume there is an effective divisor E_X equivalent to $K_X - D_X$. We may also assume v_1 and w_g are type-2 points, and choose type-2 points w_0 and v_{g+1} in the connected components of $X^{\mathrm{an}} \setminus \Gamma$ with boundary points v_1 and w_g, respectively. Then

$$V = \{v_1, \ldots, v_{g+1}, w_0, \ldots, w_g\}$$

is a semistable vertex set, with skeleton $\Gamma_V \supset \Gamma$ as shown in [Figure 3](#).

Let $\mathcal{X}_V$ be the semistable model of X associated to V, with X_i the component of the special fiber $\overline{\mathcal{X}}_V$ corresponding to v_i, and $x_i \in X_i$ the node corresponding to the edge $e_i = [w_{i-1}, v_i]$, for $1 \leq i \leq g + 1$.

Recall that the reduction of f in $\kappa(X_i)^*$ is the residue of af with respect to the valuation val_{v_i} on $K(X)$, where $a \in K^*$ is chosen such that $\mathrm{val}_{v_i}(af) = 0$ [Amini and Baker 2014]. This reduction is defined only up to multiplication by elements of κ^*, but its order of vanishing at x_i is independent of all choices. Similarly, if $f_0, \ldots, f_r$ are rational functions in $K(X)^*$, then the κ-span of their reductions in $\kappa(X_i)$ is independent of all choices. In particular, it makes sense to talk about whether these reductions are linearly independent.

Lemma 7.2. *Let D_X be a divisor of rank r on X. For each $1 \leq i \leq g$, there is a basis $f_0, \ldots, f_r$ for $\mathcal{L}(D)$ such that*

(1) *the reductions of $f_0, \ldots, f_r$ in $\kappa(X_i)$ have distinct orders of vanishing at x_i, and*

(2) *the reductions of $f_0, \ldots, f_r$ in $\kappa(X_{i+1})$ are linearly independent.*

Proof. We consider $\mathcal{L}(D_X)$ as a normed vector space over K, with respect to the norms $| \ |_i$ and $| \ |_{i+1}$ whose logarithms are $-\mathrm{val}(v_i)$ and $-\mathrm{val}(v_{i+1})$, respectively, and use the basic properties of nonarchimedean normed vector spaces developed in [Bosch et al. 1984, Chapter 2]. Since K is spherically complete, the vector space $\mathcal{L}(D_X)$ is K-cartesian [Bosch et al. 1984, 2.4.4.2], and since v_i and v_{i+1} are type-2 points, the image of $\mathcal{L}(D_X)$ under each of these norms is equal to the image of

K under its given norm. Therefore, $\mathcal{L}(D_X)$ is strictly K-cartesian [Bosch et al. 1984, 2.5.1.2], which means that all of its subspaces have orthonormal bases. So, first choose an orthonormal basis for $\mathcal{L}(D_X)$ with respect to $|\ |_i$. The reductions of these basis elements are linearly independent [Bosch et al. 1984, 2.5.1.3], so we can take suitable combinations with coefficients in R^* to ensure that they have distinct orders of vanishing at x_i.

Let $f_0, \ldots, f_r$ be a basis for $\mathcal{L}(D_X)$ whose reductions in $\kappa(X_i)$ have strictly decreasing order of vanishing at x_i. Then, for each j, we can replace f_j by a suitable linear combination of $f_0, \ldots, f_j$ that is orthogonal to the span of $f_0, \ldots, f_{j-1}$ with respect to $|\ |_{i+1}$. This does not change the order of vanishing at x_i of the reduction in $\kappa(X_i)$, but ensures that the reductions in $\kappa(X_{i+1})$ are linearly independent. $\square$

This lemma, closely analogous to [Eisenbud and Harris 1983, Lemma 1.2], will be especially useful in combination with the following identity relating orders of vanishing of reductions of rational functions to the slopes of their tropicalizations. For any piecewise linear function ψ on Γ_V, we write $s_i(\psi)$ for the incoming slope of ψ at v_i along e_i. Suppose $\psi = \text{trop}(f)$ for some rational function f in $K(X)^*$. Then Thuillier's nonarchimedean analytic Poincaré–Lelong formula [Thuillier 2005; Baker et al. 2011] says that $s_i(\text{trop}(f))$ is the order of vanishing at x_i of the reduction of f in $\kappa(X_i)$.

Fix a basis $f_0, \ldots, f_r$ for $\mathcal{L}(D_X)$ whose reductions in $\kappa(X_i)$ have distinct orders of vanishing at x_i and whose reductions at X_{i+1} are linearly independent. Let $a_0, \ldots, a_r$ be constants in K. Define

$$\psi = \text{trop}(a_0 f_0 + \cdots + a_r f_r),$$
$$\psi' = \min\{\text{trop}(f_0) + \text{val}(a_0), \ldots, \text{trop}(f_r) + \text{val}(a_r)\}.$$

Note that $\psi(v) \geq \psi'(v)$ for all v, with equality when v is equal to v_i or v_{i+1}. This is because the reductions of the $a_j f_j$ in both $\kappa(X_i)$ and $\kappa(X_{i+1})$ are linearly independent.

Proposition 7.3. *The piecewise linear functions ψ and ψ' are equal on some nonempty interval $(v, v_i) \subset e_i$.*

Proof. The two functions ψ and ψ' agree at any point v where the minimum of $\{\text{trop}(f_0)(v) + \text{val}(a_0), \ldots, \text{trop}(f_r)(v) + \text{val}(a_r)\}$ occurs only once. By construction, the reductions of $f_0, \ldots, f_r$ in $\kappa(X_i)$ have distinct orders of vanishing at x_i, so the Poincaré–Lelong formula says that $\text{trop}(f_0), \ldots, \text{trop}(f_r)$ have distinct incoming slopes at v_i along e_i. It follows that the minimum occurs only once on some open interval (v, v_i), and ψ and ψ' agree on this interval. $\square$

The final ingredient in our proof of Theorem 1.1 is the following proposition relating slopes along bridges to shapes of divisors in a linear series on Γ_V:

Proposition 7.4. *Let D be an effective divisor of degree at most $2g - 2$ on Γ_V, and let $\psi_0, \ldots, \psi_r \in R(D)$ be piecewise linear functions with distinct incoming slopes at v_i along e_i, for some $1 \leq i \leq g$. Then at most one of the divisors $D + \mathrm{div}(\psi_0), \ldots, D + \mathrm{div}(\psi_r)$ contains no point in γ_i.*

Proof. Let Γ' be the union of the i-th loop together with a small closed subsegment of $[v, v_i] \subset [w_{i-1}, v_i]$ along which $\psi_0, \ldots, \psi_r$ all have constant slope. We may choose v sufficiently close to v_i so that D contains no points in $[v, v_i)$. Let $D' = D|_{\Gamma'}$ and $\psi'_j = \psi_j|_{\Gamma'}$. Note that the coefficient of v in $\mathrm{div}(\psi'_j)$ is $-s_i(\psi_j)$, and $D' + \mathrm{div}(\psi'_j)$ agrees with $D + \mathrm{div}(\psi_j)$ on γ_i. We now show that at most one of the divisors $D' + \mathrm{div}(\psi'_j)$ contains no point in γ_i.

Suppose $D' + \mathrm{div}(\psi'_j)$ and $D' + \mathrm{div}(\psi'_k)$ both contain no point in γ_i. Then both of these divisors are supported at v and w_i. Subtracting one from the other, we find an equivalence of divisors

$$(s_i(\psi_j) - s_i(\psi_k)) \, v \; \sim \; (s_i(\psi_j) - s_i(\psi_k)) \, w_i$$

on Γ'. Note that $s_i(\psi_j)$ is bounded above by the sum of the coefficients of D at points to the left of v_i and bounded below by minus the sum of its coefficients at v_i and to the right. Similarly, $-s_i(\psi_k)$ is bounded above by the sum of the coefficients of D at v_i and to the right, and bounded below by minus the sum of its coefficients at points to the left of v_i. Therefore, $|s_i(\psi_j) - s_i(\psi_k)|$ is bounded by the degree of D. The equivalence above then implies that ℓ_i / m_i is a ratio of two positive integers whose sum is less than or equal to the degree of D, contradicting the genericity hypothesis on the edge lengths. $\square$

Proof of Theorem 1.1. Suppose the multiplication map

$$\mu : \mathcal{L}(D_X) \otimes \mathcal{L}(E_X) \to \mathcal{L}(K_X)$$

has nonzero kernel. For $1 \leq i \leq g$, let $\{f_0^i, \ldots, f_r^i\}$ be a basis for $\mathcal{L}(D_X)$ consisting of rational functions whose reductions in $\kappa(X_i)$ have distinct orders of vanishing at x_i and whose reductions in $\kappa(X_{i+1})$ are linearly independent. Similarly, let $\{g_0^i, \ldots, g_{g-d+r-1}^i\}$ be a basis for $\mathcal{L}(E_X)$ consisting of rational functions satisfying the same conditions.

Fix an element in the kernel of μ. Then, for each i, we can express this element uniquely as a sum of elementary tensors

$$\sum_{j,k} a_{j,k}^i f_j^i \otimes g_k^i.$$

Define a piecewise linear function

$$\theta_i = \min_{j,k} \{\mathrm{trop}(f_j^i) + \mathrm{trop}(g_k^i) + \mathrm{val}(a_{j,k}^i)\},$$

and note that the minimum must occur at least twice at every point in Γ_V.

Replacing $\{f_0^i, \ldots, f_r^i\}$ by $\{af_0^i, \ldots, af_r^i\}$ for some $a \in K^*$, we may assume that $\theta_i(v_{i+1}) = \theta_{i+1}(v_{i+1})$ for $1 \le i < g$, and proceed by *patching* these piecewise linear functions together.

Let θ be the unique continuous piecewise linear function on Γ_V that agrees with θ_i between v_i and v_{i+1} for $1 \le i \le g$. A priori, it is not clear whether θ is in the tropical linear series $R(D + E)$, where

$$D = \mathrm{Trop}(D_X) \quad \text{and} \quad E = \mathrm{Trop}(E_X).$$

Nevertheless, we claim not only that $D + E + \mathrm{div}(\theta)$ is effective but also that it contains a point in γ_i, for $1 \le i \le g$. (Note that θ may or may not be the tropicalization of a rational function in $\mathscr{L}(D_X + E_X)$.)

First we show that $D + E + \mathrm{div}(\theta)$ is effective. In the open subgraph between v_i and v_{i+1}, the divisor $D + E + \mathrm{div}(\theta)$ agrees with $D + E + \mathrm{div}(\theta_i)$, which is effective because $R(D + E)$ is a tropical module that contains $\mathrm{trop}(f_j^i) + \mathrm{trop}(g_k^i)$ for all j and k. It remains to check that the coefficient of v_i is nonnegative. Since $D + E + \mathrm{div}(\theta_i)$ is effective, it will suffice to show

$$s_i(\theta_{i-1}) \ge s_i(\theta_i).$$

We prove this by changing coordinates in two steps, first replacing the basis $\{f_j^i\}_j$ with $\{f_j^{i-1}\}_j$ and then replacing the basis $\{g_k^i\}_k$ with $\{g_k^{i-1}\}_k$.

Fix k, write

$$\sum_j a_{j,k}^i f_j^i = \sum_j b_{j,k} f_j^{i-1},$$

and define

$$\theta' = \min_{j,k}\{\mathrm{trop}(f_j^{i-1}) + \mathrm{trop}(g_j^i) + \mathrm{val}(b_{j,k})\}.$$

Note that

$$\min_j\{\mathrm{trop}(f_j^{i-1})(v_i) + \mathrm{val}(b_{j,k})\} = \min_j\{\mathrm{trop}(f_j^i)(v_i) + \mathrm{val}(a_{j,k}^i)\},$$

since the reductions of both $\{f_j^i\}_j$ and $\{f_j^{i-1}\}_j$ in $\kappa(X_i)$ are linearly independent. By adding the constant $g_k^i(v_i)$ and taking the minimum over all k, we see that

$$\theta'(v_i) = \theta(v_i).$$

We now examine the slopes $s_i(\theta)$ and $s_i(\theta')$. At any point v on the edge between w_{i-1} and v_i, we have

$$\mathrm{trop}\left(\sum_j b_{j,k} f_j^{i-1}\right)(v) \ge \min_j\{\mathrm{trop}(b_{j,k}) + \mathrm{trop}(f_j^{i-1})\}(v).$$

Since this inequality holds with equality at v_i, it follows that

$$s_i\left(\text{trop}\left(\sum_j b_{j,k} f_j^{i-1}\right)\right) \le s_i\left(\min_j\{\text{trop}(b_{j,k}) + \text{trop}(f_j^{i-1})\}\right).$$

Now Proposition 7.3 tells us that, on some nonempty interval $(v, v_i) \subset e_i$,

$$\text{trop}\left(\sum_j b_{j,k} f_j^{i-1}\right) = \min_j\{\text{trop}(a_{j,k}^i) + \text{trop}(f_j^i)\}.$$

Taking the minimum over all k with $\min_j\{\text{trop}(a_{j,k}^i) + \text{trop}(f_j^i)\}(v_i) + \text{trop}(g_k^i)(v_i) = \theta(v_i)$, we see that

$$s_i(\theta_i) \le s_i(\theta').$$

A similar argument, fixing j and replacing the basis $\{g_k^i\}$ with $\{g_k^{i-1}\}$, shows that $s_i(\theta') \le s_i(\theta_{i-1})$, as required. This proves that $D + E + \text{div}(\theta)$ is effective. It remains to show that $D + E + \text{div}(\theta)$ contains a point in each cell $\gamma_1, \ldots, \gamma_g$.

We now show that $D + E + \text{div}(\theta)$ contains a point in γ_i. By Proposition 7.4, there is at most one index j such that $D + \text{div}(\text{trop}(f_j^i))$ contains no point in γ_i. Similarly, there is at most one index k such that $E + \text{div}(\text{trop}(g_k^i))$ contains no point in γ_i. Call these indices j_0 and k_0, respectively, if they exist. Note that, for $(j, k) \ne (j_0, k_0)$, the divisor $D + E + \text{div}(\text{trop}(f_j^i)) + \text{div}(\text{trop}(g_k^i))$ contains a point in γ_i.

The minimum of the piecewise linear functions $\text{trop}(f_j^i) + \text{div}(\text{trop}(g_k^i)) + \text{val}(a_{j,k}^i)$ occurs at least twice at every point, by hypothesis. Thus

$$\theta_i = \min_{(j,k) \ne (j_0, k_0)}\left\{\text{trop}(f_j^i) + \text{div}(\text{trop}(g_k^i)) + \text{val}(a_{j,k}^i)\right\}.$$

Then Proposition 3.5 says that $D + E + \text{div}(\theta_i)$ contains a point in γ_i. Now, $D + E + \text{div}(\theta)$ agrees with $D + E + \text{div}(\theta_i)$ on $\gamma_i \setminus \{v_i\}$. Furthermore, since $s_i(\theta_i) \le s_i(\theta_{i-1})$, the coefficient of v_i in $D + E + \text{div}(\theta)$ is greater than or equal to the coefficient of v_i in $D + E + \text{div}(\theta_i)$. It follows that $D + E + \text{div}(\theta)$ also contains a point in γ_i, as claimed.

Pushing forward the divisor $D + E + \text{div}(\theta)$ under the natural contraction $\Gamma_V \to \Gamma$ gives an effective canonical divisor that contains a point in each cell $\gamma_1, \ldots, \gamma_g$. But this is impossible, by Lemma 4.4. $\qquad\square$

Acknowledgements

References

[Abramovich et al. 2012] D. Abramovich, L. Caporaso, and S. Payne, "The tropicalization of the moduli space of curves", preprint, 2012. To appear in *Ann. Sci. Éc. Norm. Sup.* arXiv 1212.0373

[Agrawal et al. 2013] R. Agrawal, G. Musiker, V. Sotirov, and F. Wei, "Involutions on standard Young tableaux and divisors on metric graphs", *Electron. J. Combin.* **20**:3 (2013), Paper 33, 23. MR 3104531 Zbl 06330305

[Amini and Baker 2014] O. Amini and M. Baker, "Linear series on metrized complexes of algebraic curves", *Math. Ann.* (online publication October 2014).

[An et al. 2014] Y. An, M. Baker, G. Kuperberg, and F. Shokrieh, "Canonical representatives for divisor classes on tropical curves and the matrix–tree theorem", *Forum of Math., Sigma* **2** (2014), e24. MR 3264262

[Arbarello et al. 1985] E. Arbarello, M. Cornalba, P. A. Griffiths, and J. Harris, *Geometry of algebraic curves, Vol. I*, Grundlehren der Mathematischen Wissenschaften **267**, Springer, New York, 1985. MR 86h:14019 Zbl 0559.14017

[Baker 2008] M. Baker, "Specialization of linear systems from curves to graphs", *Algebra Number Theory* **2**:6 (2008), 613–653. MR 2010a:14012 Zbl 1162.14018

[Baker and Norine 2007] M. Baker and S. Norine, "Riemann–Roch and Abel–Jacobi theory on a finite graph", *Adv. Math.* **215**:2 (2007), 766–788. MR 2008m:05167 Zbl 1124.05049

[Baker et al. 2011] M. Baker, S. Payne, and J. Rabinoff, "Nonarchimedean geometry, tropicalization, and metrics on curves", preprint, 2011. arXiv 1104.0320v1

[Baratham et al. 2014] V. Baratham, D. Jensen, C. Mata, D. Nguyen, and S. Parekh, "Towards a tropical proof of the Gieseker–Petri theorem", *Collect. Math.* **65**:1 (2014), 17–27. MR 3147766

[Bosch et al. 1984] S. Bosch, U. Güntzer, and R. Remmert, *Non-Archimedean analysis*, Grundlehren der Mathematischen Wissenschaften **261**, Springer, Berlin, 1984. MR 86b:32031 Zbl 0539.14017

[Cartwright et al. 2014] D. Cartwright, D. Jensen, and S. Payne, "Lifting divisors on a generic chain of loops", preprint, 2014. To appear in *Canad. Math. Bull.* arXiv 1404.4001

[Castorena et al. 2012] A. Castorena, A. L. Martín, and M. T. i Bigas, "Petri map for vector bundles near good bundles", preprint, 2012. arXiv 1203.0983

[Cools et al. 2012] F. Cools, J. Draisma, S. Payne, and E. Robeva, "A tropical proof of the Brill–Noether theorem", *Adv. Math.* **230**:2 (2012), 759–776. MR 2914965 Zbl 06040347

[Dhar 1990] D. Dhar, "Self-organized critical state of sandpile automaton models", *Phys. Rev. Lett.* **64**:14 (1990), 1613–1616. MR 90m:82053 Zbl 0943.82553

[Eisenbud and Harris 1983] D. Eisenbud and J. Harris, "A simpler proof of the Gieseker–Petri theorem on special divisors", *Invent. Math.* **74**:2 (1983), 269–280. MR 85e:14039 Zbl 0533.14012

[Eisenbud and Harris 1986] D. Eisenbud and J. Harris, "Limit linear series: basic theory", *Invent. Math.* **85**:2 (1986), 337–371. MR 87k:14024 Zbl 0598.14003

[Gathmann and Kerber 2008] A. Gathmann and M. Kerber, "A Riemann–Roch theorem in tropical geometry", *Math. Z.* **259**:1 (2008), 217–230. MR 2009a:14014 Zbl 1187.14066

[Gieseker 1982] D. Gieseker, "Stable curves and special divisors: Petri's conjecture", *Invent. Math.* **66**:2 (1982), 251–275. MR 83i:14024 Zbl 0522.14015

[Griffiths and Harris 1980] P. Griffiths and J. Harris, "On the variety of special linear systems on a general algebraic curve", *Duke Math. J.* **47**:1 (1980), 233–272. MR 81e:14033 Zbl 0446.14011

[Haase et al. 2012] C. Haase, G. Musiker, and J. Yu, "Linear systems on tropical curves", *Math. Z.* **270**:3-4 (2012), 1111–1140. MR 2892941 Zbl 06031800

[Lazarsfeld 1986] R. Lazarsfeld, "Brill–Noether–Petri without degenerations", *J. Differential Geom.* **23**:3 (1986), 299–307. MR 88b:14019 Zbl 0608.14026

[Len 2014] Y. Len, "The Brill–Noether rank of a tropical curve", *J. Algebraic Combin.* **40**:3 (2014), 841–860. MR 3265236

[Lim et al. 2012] C. M. Lim, S. Payne, and N. Potashnik, "A note on Brill–Noether theory and rank-determining sets for metric graphs", *Int. Math. Res. Not.* **2012**:23 (2012), 5484–5504. MR 2999150 Zbl 06126085

[Luo 2011] Y. Luo, "Rank-determining sets of metric graphs", *J. Combin. Theory Ser. A* **118**:6 (2011), 1775–1793. MR 2012d:05122 Zbl 1227.05133

[Mikhalkin and Zharkov 2008] G. Mikhalkin and I. Zharkov, "Tropical curves, their Jacobians and theta functions", pp. 203–230 in *Curves and abelian varieties*, edited by C. H. C. Valery Alexeev, Arnaud Beauville and E. Izadi, Contemp. Math. **465**, Amer. Math. Soc., Providence, RI, 2008. MR 2011c:14163 Zbl 1152.14028

[Osserman 2011] B. Osserman, "A simple characteristic-free proof of the Brill–Noether theorem", preprint, 2011. To appear in *Bull. Braz. Math. Soc.* arXiv 1108.4967v1

[Thuillier 2005] A. Thuillier, *Théorie du potentiel sur ler courbes en géométrie analytique non archimédienne. Applications à la thééorie d'Arakelov*, Ph.D. thesis, Université Rennes, 2005, Available at https://tel.archives-ouvertes.fr/tel-00010990/document.

dave.h.jensen@gmail.com   *Department of Mathematics, University of Kentucky, 719 Patterson Office Tower, Lexington, KY 40506, United States*

sam.payne@yale.edu   *Department of Mathematics, Yale University, 10 Hillhouse Avenue, New Haven, CT 06511, United States*

New equidistribution estimates of Zhang type

D. H. J. Polymath

We prove distribution estimates for primes in arithmetic progressions to large smooth squarefree moduli, with respect to congruence classes obeying Chinese remainder theorem conditions, obtaining an exponent of distribution $\frac{1}{2} + \frac{7}{300}$.

1. Introduction

In May 2013, Y. Zhang [2014] proved the existence of infinitely many pairs of primes with bounded gaps. In particular, he showed that there exists at least one $h \geq 2$ such that the set

$$\{p \text{ prime} : p + h \text{ is prime}\}$$

is infinite. (In fact, he showed this for some even h between 2 and 7×10^7, although the precise value of h could not be extracted from his method.)

Zhang's work started from the method of Goldston, Pintz and Yıldırım [Goldston et al. 2009], who had earlier proved the bounded gap property, conditionally on

distribution estimates concerning primes in arithmetic progressions to *large moduli*, i.e., beyond the reach of the Bombieri–Vinogradov theorem.

Based on work of Fouvry and Iwaniec [1985; 1980; 1983; 1992] and Bombieri, Friedlander and Iwaniec [Bombieri et al. 1986; 1987; 1989], distribution estimates going beyond the Bombieri–Vinogradov range for arithmetic functions such as the von Mangoldt function were already known. However, they involved restrictions concerning the residue classes which were incompatible with the method of Goldston, Pintz and Yıldırım.

Zhang's resolution of this difficulty proceeded in two stages. First, he isolated a weaker distribution estimate that sufficed to obtain the bounded gap property (still involving the crucial feature of going beyond the range accessible to the Bombieri–Vinogradov technique), where (roughly speaking) only smooth (i.e, friable) moduli were involved and the residue classes had to obey strong multiplicative constraints (the possibility of such a weakening had been already noticed by Motohashi and Pintz [2008]). Secondly, and more significantly, Zhang then proved such a distribution estimate.

This revolutionary achievement led to a flurry of activity. In particular, the POLYMATH8 project was initiated by T. Tao with the goal first of understanding, and then of improving and streamlining, where possible, the argument of Zhang. This was highly successful, and through the efforts of a number of people, reached a conclusion in October 2013, when the first version of this paper [Polymath 2014a] established the bounded gap property in the form

$$\liminf(p_{n+1} - p_n) \leq 4680,$$

where p_n denotes the n-th prime number.

However, at that time, J. Maynard [2013] obtained another conceptual breakthrough, by showing how a modification of the structure and of the main-term analysis of the method of Goldston, Pintz and Yıldırım was able to establish not just the bounded gap property using only the Bombieri–Vinogradov theorem (in fact the bound

$$\liminf(p_{n+1} - p_n) \leq 600$$

obtained was significantly better than the one obtained by POLYMATH8), but also the bounds

$$\liminf(p_{n+k} - p_n) < +\infty$$

for any fixed $k \geq 1$ (in a quantitative way), something which was out of reach of the earlier methods, even for $k = 2$. (Similar results were obtained independently in unpublished work of Tao.)

Because of this development, a part of the POLYMATH8 paper became essentially obsolete. Nevertheless, the distribution estimate for primes in arithmetic progressions are not superseded by the new method, and they have considerable interest for analytic number theory. Indeed, it is the best known result concerning primes in arithmetic progressions to large moduli without fixing the residue class. (When the class is fixed, the best results remain those of Bombieri, Friedlander and Iwaniec [Bombieri et al. 1986], improving on those of [Fouvry and Iwaniec 1983].) The results here are also needed to obtain the best known bounds on $\liminf(p_{n+k} - p_n)$ for large values of k; see [Polymath 2014b].

The present version of the work of POLYMATH8 therefore contains only the statement and proof of these estimates. We note however that some of the earlier version is incorporated in our subsequent paper [Polymath 2014b], which builds on Maynard's method to further improve many bounds concerning gaps between primes, both conditional and unconditional. Furthermore, the original version of this paper, and the history of its elaboration, remain available online [Polymath 2014a].

Our main theorem is:

Theorem 1.1. *Let $\theta = \frac{1}{2} + \frac{7}{300}$. Let $\varepsilon > 0$ and $A \geq 1$ be fixed real numbers. For all primes p, let a_p be a fixed invertible residue class modulo p, and for $q \geq 1$ squarefree, denote by a_q the unique invertible residue class modulo q such that $a_q \equiv a_p$ modulo all primes p dividing q.*

There exists $\delta > 0$, depending only on ε, such that for $x \geq 1$, we have

$$\sum_{\substack{q \leq x^{\theta-\varepsilon} \\ qx^{\delta}\text{-smooth, squarefree}}} \left| \psi(x; q, a_q) - \frac{x}{\varphi(q)} \right| \ll \frac{x}{(\log x)^A},$$

where the implied constant depends only on A, ε and δ, and in particular is independent of the residue classes (a_p).

In this statement, we have, as usual, defined

$$\psi(x; q, a) = \sum_{\substack{n \leq x \\ n = a\,(q)}} \Lambda(n),$$

where Λ is the von Mangoldt function. Zhang [2014] established a weaker form of Theorem 1.1, with $\theta = \frac{1}{2} + \frac{1}{584}$, and with the a_q required to be roots of a polynomial P of the form $P(n) := \prod_{1 \leq j \leq k;\, j \neq i} (n + h_j - h_i)$ for a fixed admissible tuple $(h_1, \ldots, h_k)$ and $i = 1, \ldots, k$.

In fact, we will prove a number of variants of this bound. These involve either weaker restrictions on the moduli ("dense-divisibility", instead of smoothness, which may be useful in some applications), or smaller values of $\theta > \frac{1}{2}$, but with significantly simpler proofs. In particular, although the full strength of Theorem 1.1

depends in crucial ways on applications of Deligne's deepest form of the Riemann hypothesis over finite fields, we show that, for a smaller value of $\theta > \frac{1}{2}$, it is possible to obtain the same estimate by means of Weil's theory of exponential sums in one variable over finite fields.

The outline of this paper is as follows: in the next section, we briefly outline the strategy, starting from the work of Bombieri, Fouvry, Friedlander, and Iwaniec (in chronological order, [Fouvry and Iwaniec 1980; 1983; Friedlander and Iwaniec 1985; Bombieri et al. 1986; 1987; 1989; Fouvry and Iwaniec 1992]), and explain Zhang's innovations. These involve different types of estimates of bilinear or trilinear nature, which we present in turn. All involve estimates for exponential sums over finite fields. We therefore survey the relevant theory, separating that part depending only on one-variable character sums of Weil type (Section 4), and the much deeper one which depends on Deligne's form of the Riemann hypothesis (Section 6). In both cases, we present the formalism in sometimes greater generality than strictly needed, as these results are of independent interest and might be useful for other applications.

1A. *Overview of proof.* We begin with a brief and informal overview of the methods used in this paper.

Important work of Fouvry and Iwaniec [1980; 1983] and of Bombieri, Friedlander and Iwaniec [Bombieri et al. 1986; 1987; 1989] had succeeded, in some cases, in establishing distribution results similar to Theorem 1.1, in fact with θ as large as $\frac{1}{2} + \frac{1}{14}$, but with the restriction that the residue classes a_p are obtained by reduction modulo p of a fixed integer $a \geq 1$.

Following the techniques of Bombieri, Fouvry, Friedlander and Iwaniec, Zhang used the Heath-Brown identity [1982] to reduce the proof of (his version of) Theorem 1.1 to the verification of three families of estimates, which he called "Type I", "Type II", and "Type III". These estimates were then reduced to exponential sum estimates, using techniques such as Linnik's dispersion method, completion of sums, and Weyl differencing. Ultimately, the exponential sum estimates were established by applications of the Riemann hypothesis over finite fields, in analogy with all previous works of this type. The final part of Zhang's argument is closely related to the study of the distribution of the ternary divisor function in arithmetic progressions by Friedlander and Iwaniec [1985], and indeed the final exponential sum estimate that Zhang uses already appears in their work (this estimate was proved by Birch and Bombieri in [Friedlander and Iwaniec 1985, Appendix]). An important point is that by using techniques that are closer to those of [Fouvry and Iwaniec 1980], Zhang avoids using the spectral theory of automorphic forms, which is a key ingredient in [Fouvry and Iwaniec 1983] and [Bombieri et al. 1986], and one of the sources of the limitation to a fixed residue in these works.

Our proof of Theorem 1.1 follows the same general strategy as Zhang's, with improvements and refinements.

First, we apply the Heath-Brown identity [1982] in Section 3, with little change compared with Zhang's argument, reducing to the "bilinear" (Types I/II) and "trilinear" (Type III) estimates.

For the Type I and Type II estimates, we follow the arguments of Zhang to reduce to the task of bounding incomplete exponential sums similar to

$$\sum_{N < n \leq 2N} e\left(\frac{c_1\bar{n} + c_2\overline{n+l}}{q}\right),$$

(where $e(z) = e^{2i\pi z}$, and $\bar{x}$ denotes the inverse of x modulo q) for various parameters N, c_1, c_2, l, q. We obtain significant improvements of Zhang's numerology at this stage, by exploiting the smooth (or at least densely divisible) nature of q, using the q-van der Corput A-process of [Heath-Brown 1978] and [Graham and Ringrose 1990], combined with the Riemann hypothesis for curves over finite fields. Additional gains are obtained by optimizing the parametrizations of sums prior to application of the Cauchy–Schwarz inequality. In our strongest Type I estimate, we also exploit additional averaging over the modulus by means of higher-dimensional exponential sum estimates, which now do depend on the deep results of Deligne. We refer to Sections 4, 5 and 8 for details of these parts of the arguments.

Finally, for the Type III sums, Zhang's delicate argument [2014] adapts and improves the work of Friedlander and Iwaniec [1985] on the ternary divisor function in arithmetic progressions. As we said, it ultimately relies on a three-variable exponential sum estimate that was proved by Birch and Bombieri in [Friedlander and Iwaniec 1985, Appendix]. Here, we proceed slightly differently, inspired by the streamlined approach of Fouvry, Kowalski, and Michel [Fouvry et al. 2014b]. Namely, in Section 7 we show how our task can be reduced to obtaining certain correlation bounds on hyper-Kloosterman sums. These bounds are established in Section 6, by fully exploiting the formalism of "trace functions" over finite fields (which relies on Deligne's second, more general proof of the Riemann hypothesis over finite fields [1980]). The very general techniques presented in Section 6 are also used in the proof of the strongest Type I estimate in Section 8, and we present them in considerable detail in order to make them more accessible to analytic number theorists.

1B. *Basic notation.* We use $|E|$ to denote the cardinality of a finite set E, and $\mathbf{1}_E$ to denote the indicator function of a set E; thus $\mathbf{1}_E(n) = 1$ when $n \in E$ and $\mathbf{1}_E(n) = 0$ otherwise.

All sums and products will be over the natural numbers $\mathbb{N} := \{1, 2, 3, \dots\}$ unless otherwise specified, with the exceptions of sums and products over the variable p, which will be understood to be over primes.

The following important asymptotic notation will be in use throughout most of the paper; when it is not (as in Section 6), we will mention this explicitly.

Definition 1.2 (asymptotic notation). We use x to denote a large real parameter, which one should think of as going off to infinity; in particular, we will implicitly assume that it is larger than any specified fixed constant. Some mathematical objects will be independent of x and referred to as *fixed*; but unless otherwise specified we allow all mathematical objects under consideration to depend on x (or to vary within a range that depends on x, e.g., the summation parameter n in the sum $\sum_{x \leq n \leq 2x} f(n)$). If X and Y are two quantities depending on x, we say that $X = O(Y)$ or $X \ll Y$ if one has $|X| \leq CY$ for some fixed C (which we refer to as the *implied constant*), and $X = o(Y)$ if one has $|X| \leq c(x)Y$ for some function $c(x)$ of x (and of any fixed parameters present) that goes to zero as $x \to \infty$ (for each choice of fixed parameters). We use $X \lll Y$ to denote the estimate $|X| \leq x^{o(1)}Y$, $X \asymp Y$ to denote the estimate $Y \ll X \ll Y$, and $X \approx Y$ to denote the estimate $Y \lll X \lll Y$. Finally, we say that a quantity n is of *polynomial size* if one has $n = O(x^{O(1)})$.

If asymptotic notation such as $O(\ )$ or $\lll$ appears on the left-hand side of a statement, this means that the assertion holds true for any specific interpretation of that notation. For instance, the assertion $\sum_{n=O(N)} |\alpha(n)| \lll N$ means that for each fixed constant $C > 0$, one has $\sum_{|n| \leq CN} |\alpha(n)| \lll N$.

If q and a are integers, we write $a \mid q$ if a divides q.

If q is a natural number and $a \in \mathbb{Z}$, we use a (q) to denote the congruence class

$$a\ (q) := \{a + nq : n \in \mathbb{Z}\},$$

and we denote by $\mathbb{Z}/q\mathbb{Z}$ the ring of all such congruence classes. The notation $b = a\ (q)$ is synonymous to $b \in a\ (q)$. We use (a, q) to denote the greatest common divisor of a and q, and $[a, q]$ to denote the least common multiple.[1] More generally, we let $(q_1, \ldots, q_k)$ denote the greatest simultaneous common divisor of $q_1, \ldots, q_k$. We note in particular that $(0, q) = q$ for any natural number q. Note that $a \mapsto (a, q)$ is periodic with period q, and so we may also define (a, q) for $a \in \mathbb{Z}/q\mathbb{Z}$ without ambiguity. We also let

$$(\mathbb{Z}/q\mathbb{Z})^\times := \{a\ (q) : (a, q) = 1\}$$

denote the primitive congruence classes of $\mathbb{Z}/q\mathbb{Z}$. More generally, for any commutative ring R (with unity) we use $R^\times$ to denote the multiplicative group of units. If $a \in (\mathbb{Z}/q\mathbb{Z})^\times$, we use $\bar{a}$ to denote the inverse of a in $\mathbb{Z}/q\mathbb{Z}$.

---

[1]When a, b are real numbers, we will also need to use (a, b) and $[a, b]$ to denote the open and closed intervals respectively with endpoints a, b. Unfortunately, this notation conflicts with the notation given above, but it should be clear from the context which notation is in use. Similarly for the notation $\bar{a}$ for $a \in \mathbb{Z}/q\mathbb{Z}$, and the notation $\bar{z}$ to denote the complex conjugate of a complex number z.

For any real number x, we write $e(x) := e^{2\pi i x}$. We set $e_q(a) := e(a/q) = e^{2\pi i a/q}$ (see also the conventions concerning this additive character in Section 4A).

We use the following standard arithmetic functions:

(i) $\varphi(q) := |(\mathbb{Z}/q\mathbb{Z})^\times|$ denotes the Euler totient function of q.

(ii) $\tau(q) := \sum_{d|q} 1$ denotes the divisor function of q.

(iii) $\Lambda(q)$ denotes the von Mangoldt function of q, thus $\Lambda(q) = \log p$ if q is a power of a prime p and $\Lambda(q) = 0$ otherwise.

(iv) $\theta(q)$ is defined to be equal to $\log q$ when q is a prime and to be 0 otherwise.

(v) $\mu(q)$ denotes the Möbius function of q, thus $\mu(q) = (-1)^k$ if q is the product of k distinct primes for some $k \geq 0$ and $\mu(q) = 0$ otherwise.

(vi) $\Omega(q)$ denotes the number of prime factors of q (counting multiplicity).

The *Dirichlet convolution* $\alpha \star \beta : \mathbb{N} \to \mathbb{C}$ of two arithmetic functions $\alpha, \beta : \mathbb{N} \to \mathbb{C}$ is defined in the usual fashion as

$$\alpha \star \beta(n) := \sum_{d|n} \alpha(d)\beta\left(\frac{n}{d}\right) = \sum_{ab=n} \alpha(a)\beta(b).$$

Many of the key ideas in Zhang's work (as well as in the present article) concern the uniform distribution of arithmetic functions in arithmetic progressions. For any function $\alpha : \mathbb{N} \to \mathbb{C}$ with finite support (that is, α is nonzero only on a finite set) and any primitive congruence class $a\ (q)$, we define the (signed) *discrepancy* $\Delta(\alpha; a\ (q))$ to be the quantity

$$\Delta(\alpha; a\ (q)) := \sum_{n=a\ (q)} \alpha(n) - \frac{1}{\varphi(q)} \sum_{(n,q)=1} \alpha(n). \tag{1-1}$$

There are some additional concepts and terminology that will be used in multiple sections of this paper. These are listed in Table 1.

We will often use the following simple estimates for the divisor function τ and its powers.

Lemma 1.3 (crude bounds on τ).

(i) (*divisor bound*) *One has*

$$\tau(d) \ll 1 \tag{1-2}$$

whenever d is of polynomial size. In particular, d has $o(\log x)$ distinct prime factors.

(ii) *One has*

$$\sum_{d \leq y} \tau^C(d) \ll y \log^{O(1)} x \tag{1-3}$$

for any fixed $C > 0$ and any $y > 1$ of polynomial size.

ϖ	level of distribution	Section 2
δ	smoothness/dense divisibility parameter	Section 2
i	multiplicity of dense divisibility	Definition 2.1
σ	Type I/III boundary parameter	Definition 2.6
$\mathrm{MPZ}^{(i)}[\varpi,\delta]$	MPZ conjecture for densely divisible moduli	Claim 2.3
$\mathrm{Type}_{\mathrm{I}}^{(i)}[\varpi,\delta,\sigma]$	Type I estimate	Definition 2.6
$\mathrm{Type}_{\mathrm{II}}^{(i)}[\varpi,\delta]$	Type II estimate	Definition 2.6
$\mathrm{Type}_{\mathrm{III}}^{(i)}[\varpi,\delta,\sigma]$	Type III estimate	Definition 2.6
$\mathscr{S}_I$	squarefree products of primes in I	Definition 2.2
P_I	product of all primes in I	Definition 2.2
$\mathscr{D}^{(i)}(y)$	i-tuply y-densely divisible integers	Definition 2.1
$\mathrm{FT}_q(f)$	normalized Fourier transform of f	(4-11)
	coefficient sequence at scale N	Definition 2.5
	Siegel–Walfisz theorem	Definition 2.5
	(shifted) smooth sequence at scale N	Definition 2.5

Table 1. Notation and terminology.

(iii) *More generally*, *one has*

$$\sum_{\substack{d \leq y \\ d \equiv a\,(q)}} \tau^C(d) \ll \frac{y}{q}\tau^{O(1)}(q)\log^{O(1)}x + x^{o(1)} \tag{1-4}$$

for any fixed $C > 0$, any residue class a (q) (not necessarily primitive), and any $y > 1$ of polynomial size.

Proof. For the divisor bound (1-2), see for example [Montgomery and Vaughan 2007, Theorem 2.11]. For the bound (1-3), see Corollary 2.15 of the same book. Finally, to prove the bound (1-4), observe using (1-2) that we may factor out any common factor of a and q, so that a (q) is primitive. Next, we may assume that $q \leq y$, since the case $q > y$ is trivial by (1-2). The claim now follows from the Brun–Titchmarsh inequality for multiplicative functions (see [Shiu 1980] or [Barban and Vehov 1969]).   □

Note that we have similar bounds for the higher divisor functions

$$\tau_k(n) := \sum_{d_1,\ldots,d_k:\, d_1\cdots d_k = n} 1$$

for any fixed $k \geq 2$, thanks to the crude upper bound $\tau_k(n) \leq \tau(n)^{k-1}$.

The following elementary consequence of the divisor bound will also be useful:

Lemma 1.4. *Let $q \geq 1$ be an integer. Then for any $K \geq 1$ we have*

$$\sum_{1 \leq k \leq K} (k, q) \leq K\tau(q).$$

In particular, if q is of polynomial size, then we have

$$\sum_{a \in \mathbb{Z}/q\mathbb{Z}} (a, q) \ll q,$$

and we also have

$$\sum_{|k| \leq K} (k, q) \ll Kq^{\varepsilon} + q$$

for any fixed $\varepsilon > 0$ and arbitrary q (not necessarily of polynomial size).

Proof. We have

$$(k, q) \leq \sum_{d \mid (q, k)} d$$

and hence

$$\sum_{1 \leq k \leq K} (k, q) \leq \sum_{d \mid q} \sum_{\substack{1 \leq k \leq K \\ d \mid k}} d \leq K\tau(q). \qquad \square$$

2. Preliminaries

2A. *Statements of results.* In this section we will give the most general statements that we prove, and in particular define the concept of "dense divisibility", which weakens the smoothness requirement of Theorem 1.1.

Definition 2.1 (multiple dense divisibility). Let $y \geq 1$. For each natural number $i \geq 0$, we define a notion of *i-tuply y-dense divisibility* recursively as follows:

(i) Every natural number n is 0-tuply y-densely divisible.

(ii) If $i \geq 1$ and n is a natural number, we say that n is i-tuply y-densely divisible if, whenever $j, k \geq 0$ are natural numbers with $j + k = i - 1$, and $1 \leq R \leq yn$, one can find a factorization

$$n = qr \quad \text{with } y^{-1}R \leq r \leq R \tag{2-1}$$

such that q is j-tuply y-densely divisible and r is k-tuply y-densely divisible.

We let $\mathcal{D}^{(i)}(y)$ denote the set of i-tuply y-densely divisible numbers. We abbreviate "1-tuply densely divisible" as "densely divisible", "2-tuply densely divisible" as "doubly densely divisible", and so forth; we also abbreviate $\mathcal{D}^{(1)}(y)$ as $\mathcal{D}(y)$, and since we will often consider squarefree densely divisible integers with prime factors in an interval I, we will set

$$\mathcal{D}_I^{(j)}(y) = \mathcal{S}_I \cap \mathcal{D}^{(j)}(y). \tag{2-2}$$

A number of basic properties of this notion will be proved at the beginning of Section 2C, but the intent is that we want to have integers which can always be factored, in such a way that we can control the location of the divisors. For instance, the following fact is quite easy to check: any y-smooth integer is also i-tuply y-densely divisible, for any $i \geq 0$ (see Lemma 2.10(iii) for details).

Definition 2.2. For any set $I \subset \mathbb{R}$ (possibly depending on x), let $\mathcal{S}_I$ denote the set of all squarefree natural numbers whose prime factors lie in I. If I is also a bounded set (with the bound allowed to depend on x), we let P_I denote the product of all the primes in I; thus in this case $\mathcal{S}_I$ is the set of divisors of P_I.

For every fixed $0 < \varpi < \frac{1}{4}$ and $0 < \delta < \frac{1}{4} + \varpi$ and every natural number i, we let MPZ$^{(i)}[\varpi, \delta]$ denote the following claim:

Claim 2.3 (modified Motohashi–Pintz–Zhang estimate, MPZ$^{(i)}[\varpi, \delta]$). *Let $I \subset \mathbb{R}$ be a bounded set, which may vary with x, and let $Q \ll x^{1/2+2\varpi}$. If a is an integer coprime to P_I and $A \geq 1$ is fixed, then*

$$\sum_{\substack{q \leq Q \\ q \in \mathcal{D}_I^{(i)}(x^\delta)}} |\Delta(\Lambda \mathbf{1}_{[x,2x]}; a\,(q))| \ll x \log^{-A} x. \tag{2-3}$$

We will prove the following cases of these estimates:

Theorem 2.4 (Motohashi–Pintz–Zhang-type estimates).

  (i) *We have* MPZ$^{(4)}[\varpi, \delta]$ *for any fixed* $\varpi, \delta > 0$ *such that* $600\varpi + 180\delta < 7$.

 (ii) *We can prove* MPZ$^{(2)}[\varpi, \delta]$ *for any fixed* $\varpi, \delta > 0$ *such that* $168\varpi + 48\delta < 1$, *without invoking any of Deligne's results* [1974; 1980] *on the Riemann hypothesis over finite fields.*

The statement MPZ$^{(i)}[\varpi, \delta]$ is easier to establish as i increases. If true for some $i \geq 1$, it implies that

$$\sum_{\substack{q \leq x^{1/2+2\varpi-\varepsilon} \\ q\ x^\delta\text{-smooth, squarefree}}} |\Delta(\Lambda \mathbf{1}_{[x,2x]}; a\,(q))| \ll x \log^{-A} x$$

for any $A \geq 1$ and $\varepsilon > 0$. Using a dyadic decomposition and the Chinese remainder theorem, this shows that Theorem 2.4(i) implies Theorem 1.1.

2B. *Bilinear and trilinear estimates.* As explained, we will reduce Theorem 2.4 to bilinear or trilinear estimates. In order to state these precisely, we introduction some further notation.

Definition 2.5 (coefficient sequences). A *coefficient sequence* is a finitely supported sequence $\alpha : \mathbb{N} \to \mathbb{R}$ (which may depend on x) that obeys the bounds

$$|\alpha(n)| \ll \tau^{O(1)}(n) \log^{O(1)}(x) \tag{2-4}$$

for all n (recall that τ is the divisor function).

(i) A coefficient sequence α is said to be *located at scale N* for some $N \geq 1$ if it is supported on an interval of the form $[cN, CN]$ for some $1 \ll c < C \ll 1$.

(ii) A coefficient sequence α located at scale N for some $N \geq 1$ is said to *obey the Siegel–Walfisz theorem*, or to *have the Siegel–Walfisz property*, if one has

$$|\Delta(\alpha \mathbf{1}_{(\cdot,r)=1}; a\ (q))| \ll \tau(qr)^{O(1)} N \log^{-A} x \tag{2-5}$$

for any $q, r \geq 1$, any fixed A, and any primitive residue class $a\ (q)$.

(iii) A coefficient sequence α is said to be *shifted smooth at scale N* for some $N \geq 1$ if it has the form $\alpha(n) = \psi((n - x_0)/N)$ for some smooth function $\psi : \mathbb{R} \to \mathbb{C}$ supported on an interval $[c, C]$ for some fixed $0 < c < C$ and some real number x_0, with ψ obeying the derivative bounds

$$|\psi^{(j)}(x)| \ll \log^{O(1)} x \tag{2-6}$$

for all fixed $j \geq 0$, where the implied constant may depend on j, and where $\psi^{(j)}$ denotes the j-th derivative of ψ. If we can take $x_0 = 0$, we call α *smooth at scale N*; note that such sequences are also located at scale N.

Note that for a coefficient sequence α at scale N, an integer $q \geq 1$ and a primitive residue class $a\ (q)$, we have the trivial estimate

$$\Delta(\alpha; a\ (q)) \ll \frac{N}{\varphi(q)} (\log x)^{O(1)}. \tag{2-7}$$

In particular, we see that the Siegel–Walfisz property amounts to a requirement that the sequence α be uniformly equidistributed in arithmetic progressions to moduli $q \ll (\log x)^A$ for any A. In the most important arithmetic cases, it is established using methods from the classical theory of L-functions.

Definition 2.6 (Type I, II, III estimates). Let $0 < \varpi < \frac{1}{4}$, $0 < \delta < \frac{1}{4} + \varpi$, and $0 < \sigma < \frac{1}{2}$ be fixed quantities, and let $i \geq 1$ be a fixed natural number. We let I be an arbitrary bounded subset of $\mathbb{R}$ and define $P_I = \prod_{p \in I} p$ as before. Let $a\ (P_I)$ be a primitive congruence class.

(i) We say that $\text{Type}_I^{(i)}[\varpi, \delta, \sigma]$ holds if, for any I and $a\ (P_I)$ as above, any quantities $M, N \gg 1$ with

$$MN \asymp x \tag{2-8}$$

and

$$x^{1/2-\sigma} \ll N \ll x^{1/2-2\varpi-c} \tag{2-9}$$

for some fixed $c > 0$, any $Q \ll x^{1/2+2\varpi}$, and any coefficient sequences α, β located at scales M, N respectively, with β having the Siegel–Walfisz property, we have

$$\sum_{\substack{q \leq Q \\ q \in \mathcal{D}_I^{(i)}(x^\delta)}} |\Delta(\alpha \star \beta; a\,(q))| \ll x \log^{-A} x \tag{2-10}$$

for any fixed $A > 0$. (Recall the definition (2-2) of the set $\mathcal{D}_I^{(i)}(x^\delta)$.)

(ii) We say that $\text{Type}_{\text{II}}^{(i)}[\varpi, \delta]$ holds if, for any I and $a\,(P_I)$ as above, any quantities $M, N \gg 1$ obeying (2-8) and

$$x^{1/2-2\varpi-c} \ll N \ll x^{1/2} \tag{2-11}$$

for some sufficiently small fixed $c > 0$, any $Q \ll x^{1/2+2\varpi}$, and any coefficient sequences α, β located at scales M, N respectively, with β having the Siegel–Walfisz property, we have (2-10) for any fixed $A > 0$.

(iii) We say that $\text{Type}_{\text{III}}^{(i)}[\varpi, \delta, \sigma]$ holds if, for any I and $a\,(P_I)$ as above, for any quantities $M, N_1, N_2, N_3 \gg 1$ which satisfy the conditions

$$M N_1 N_2 N_3 \asymp x,$$
$$N_1 N_2, N_1 N_3, N_2 N_3 \gg x^{1/2+\sigma}, \tag{2-12}$$
$$x^{2\sigma} \ll N_1, N_2, N_3 \ll x^{1/2-\sigma}, \tag{2-13}$$

for any coefficient sequences $\alpha, \psi_1, \psi_2, \psi_3$ located at scales M, N_1, N_2, N_3, respectively, with ψ_1, ψ_2, ψ_3 smooth, and finally for any $Q \ll x^{1/2+2\varpi}$, we have

$$\sum_{\substack{q \leq Q \\ q \in \mathcal{D}_I^{(i)}(x^\delta)}} |\Delta(\alpha \star \psi_1 \star \psi_2 \star \psi_3; a\,(q))| \ll x \log^{-A} x \tag{2-14}$$

for any fixed $A > 0$.

Roughly speaking, Type I estimates control the distribution of Dirichlet convolutions $\alpha \star \beta$ where α, β are rough coefficient sequences at moderately different scales, Type II estimates control the distribution of Dirichlet convolutions $\alpha \star \beta$ where α, β are rough coefficient sequences at almost the same scale, and Type III estimates control the distribution of Dirichlet convolutions $\alpha \star \psi_1 \star \psi_2 \star \psi_3$ where ψ_1, ψ_2, ψ_3 are smooth and α is rough but supported at a fairly small scale.

In Section 3, we will use the Heath-Brown identity to reduce $\text{MPZ}^{(i)}[\varpi, \delta]$ to a combination of $\text{Type}_{\text{I}}^{(i)}[\varpi, \delta, \sigma]$, $\text{Type}_{\text{II}}^{(i)}[\varpi, \delta]$, and $\text{Type}_{\text{III}}^{(i)}[\varpi, \delta, \sigma]$:

Lemma 2.7 (combinatorial lemma). *Let $i \geq 1$ be a fixed integer, and let $0 < \varpi < \frac{1}{4}$, $0 < \delta < \frac{1}{4} + \varpi$, and $\frac{1}{10} < \sigma < \frac{1}{2}$ be fixed quantities with $\sigma > 2\varpi$, such that the estimates $\mathrm{Type}_{\mathrm{I}}^{(i)}[\varpi, \delta, \sigma]$, $\mathrm{Type}_{\mathrm{II}}^{(i)}[\varpi, \delta]$, and $\mathrm{Type}_{\mathrm{III}}^{(i)}[\varpi, \delta, \sigma]$ all hold. Then $\mathrm{MPZ}^{(i)}[\varpi, \delta]$ holds.*

Furthermore, if $\sigma > \frac{1}{6}$, then the hypothesis $\mathrm{Type}_{\mathrm{III}}^{(i)}[\varpi, \delta, \sigma]$ may be omitted.

As stated earlier, this lemma is a simple consequence of the Heath-Brown identity, a dyadic decomposition (or more precisely, a finer-than-dyadic decomposition), some standard analytic number theory estimates (in particular, the Siegel–Walfisz theorem) and some elementary combinatorial arguments.

In [Zhang 2014], the claims $\mathrm{Type}_{\mathrm{I}}[\varpi, \delta, \sigma]$, $\mathrm{Type}_{\mathrm{II}}[\varpi, \delta]$, $\mathrm{Type}_{\mathrm{III}}[\varpi, \delta, \sigma]$ are (implicitly) proven with $\varpi = \delta = \frac{1}{1168}$ and $\sigma = \frac{1}{8} - 8\varpi$. In fact, if one optimizes the numerology in his arguments, one can derive $\mathrm{Type}_{\mathrm{I}}[\varpi, \delta, \sigma]$ whenever $44\varpi + 12\delta + 8\sigma < 1$, $\mathrm{Type}_{\mathrm{II}}[\varpi, \delta]$ whenever $116\varpi + 20\delta < 1$, and $\mathrm{Type}_{\mathrm{III}}[\varpi, \delta, \sigma]$ whenever $\sigma > \frac{3}{26} + \frac{32}{13}\varpi + \frac{2}{13}\delta$ (see [Pintz 2013] for details). We will obtain the following improvements to these estimates, where the dependency with respect to σ is particularly important:

Theorem 2.8 (new Type I, II, III estimates). *Let $\varpi, \delta, \sigma > 0$ be fixed quantities.*

 (i) *If $54\varpi + 15\delta + 5\sigma < 1$, then $\mathrm{Type}_{\mathrm{I}}^{(1)}[\varpi, \delta, \sigma]$ holds.*

 (ii) *If $56\varpi + 16\delta + 4\sigma < 1$, then $\mathrm{Type}_{\mathrm{I}}^{(2)}[\varpi, \delta, \sigma]$ holds.*

 (iii) *If $\frac{160}{3}\varpi + 16\delta + \frac{34}{9}\sigma < 1$ and $64\varpi + 18\delta + 2\sigma < 1$, then $\mathrm{Type}_{\mathrm{I}}^{(4)}[\varpi, \delta, \sigma]$ holds.*

 (iv) *If $68\varpi + 14\delta < 1$, then $\mathrm{Type}_{\mathrm{II}}^{(1)}[\varpi, \delta]$ holds.*

 (v) *If $\sigma > \frac{1}{18} + \frac{28}{9}\varpi + \frac{2}{9}\delta$ and $\varpi < \frac{1}{12}$, then $\mathrm{Type}_{\mathrm{III}}^{(1)}[\varpi, \delta, \sigma]$ holds.*

The proofs of the claims in (iii) and (v) require Deligne's work on the Riemann hypothesis over finite fields, but the claims in (i), (ii) and (iv) do not.

In proving these estimates, we will rely on the following general "bilinear" form of the Bombieri–Vinogradov theorem (the principle of which is due to Gallagher [1968] and Motohashi [1976]).

Theorem 2.9 (Bombieri–Vinogradov theorem). *Let $N, M \gg 1$ be such that $NM \asymp x$ and $N \geq x^{\varepsilon}$ for some fixed $\varepsilon > 0$. Let α, β be coefficient sequences at scales M, N respectively such that β has the Siegel–Walfisz property. Then for any fixed $A > 0$ there exists a fixed $B > 0$ such that*

$$\sum_{q \leq x^{1/2} \log^{-B} x} \sup_{a \in (\mathbb{Z}/q\mathbb{Z})^{\times}} |\Delta(\alpha \star \beta; a\,(q))| \ll x \log^{-A} x.$$

See [Bombieri et al. 1986, Theorem 0] for the proof. Besides the assumption of the Siegel–Walfisz property, the other main ingredient used to establish Theorem 2.9 is the large sieve inequality for Dirichlet characters, from which the critical limitation to moduli less than $x^{1/2}$ arises.

The Type I and Type II estimates in Theorem 2.8 will be proven in Section 5, with the exception of the more difficult Type I estimate (iii), which is proven in Section 8. The Type III estimate is established in Section 7. In practice, the estimate in Theorem 2.8(i) gives inferior results to that in Theorem 2.8(ii), but we include it here because it has a slightly simpler proof.

The proofs of these estimates involve essentially all the methods that have been developed or exploited for the study of the distribution of arithmetic functions in arithmetic progressions to large moduli, for instance the dispersion method, completion of sums, the Weyl differencing technique, and the q-van der Corput A process. All rely ultimately on some estimates of (incomplete) exponential sums over finite fields, either one-dimensional or higher-dimensional. These final estimates are derived from forms of the Riemann hypothesis over finite fields, either in the (easier) form due to Weil [1948], or in the much more general form due to Deligne [1980].

2C. *Properties of dense divisibility.* We present the most important properties of the notion of multiple dense divisibility, as defined in Definition 2.1. Roughly speaking, dense divisibility is a weaker form of smoothness which guarantees a plentiful supply of divisors of the given number in any reasonable range, and multiple dense divisibility is a hereditary version of this property which also partially extends to some factors of the original number.

Lemma 2.10 (properties of dense divisibility). *Let $i \geq 0$ and $y \geq 1$.*

(0) *If n is i-tuply y-densely divisible, and $y_1 \geq y$, then n is i-tuply y_1-densely divisible. Furthermore, if $0 \leq j \leq i$, then n is j-tuply y-densely divisible.*

(i) *If n is i-tuply y-densely divisible, and m is a divisor of n, then m is i-tuply $y(n/m)$-densely divisible. Similarly, if l is a multiple of n, then l is i-tuply $y(l/n)$-densely divisible.*

(ii) *If m, n are y-densely divisible, then $[m, n]$ is also y-densely divisible.*

(iii) *Any y-smooth number is i-tuply y-densely divisible.*

(iv) *If n is z-smooth and squarefree for some $z \geq y$, and*

$$\prod_{\substack{p|n \\ p \leq y}} p \geq \frac{z^i}{y}, \tag{2-15}$$

then n is i-tuply y-densely divisible.

Proof. We abbreviate "i-tuply y-densely divisible" in this proof by the shorthand "(i, y)-d.d."

The monotony properties of (0) are immediate from the definition.

Before we prove the other properties, we make the following remark: in checking that an integer n is (i, y)-d.d., it suffices to consider parameters R with $1 \leq R \leq n$ when looking for factorizations of the form (2-1): indeed, if $n < R \leq yn$, the factorization $n = qr$ with $r = n$ and $q = 1$ satisfies the condition $y^{-1}R \leq r \leq R$, and $r = n$ is (j, y)-d.d. (or $q = 1$ is (k, y)-d.d.) whenever $j + k = i - 1$. We will use this reduction in (i), (ii), (iii), (iv) below.

We prove the first part of (i) by induction on i. For $i = 0$, the statement is obvious since every integer is $(0, y)$-d.d. for every $y \geq 1$. Now assume the property holds for j-tuply dense divisibility for $j < i$, let n be (i, y)-d.d., and let $m \mid n$ be a divisor of n. We proceed to prove that m is (i, ym_1)-d.d.

We write $n = mm_1$. Let R be such that $1 \leq R \leq m$, and let $j, k \geq 0$ be integers with $j + k = i - 1$. Since $R \leq n$, and n is (i, y)-d.d., there exists by definition a factorization $n = qr$, where q is (j, y)-d.d., r is (k, y)-d.d., and $y/R \leq r \leq y$. Now we write $m_1 = n_1 n_1'$, where $n_1 = (r, m_1)$ is the gcd of r and m_1. We have then a factorization $m = q_1 r_1$, where

$$q_1 = \frac{q}{n_1'}, \quad r_1 = \frac{r}{n_1},$$

and we check that this factorization satisfies the condition required for checking that m is (i, ym_1)-d.d. First, we have

$$\frac{R}{ym_1} \leq \frac{r}{m_1} \leq \frac{r}{n_1} = r_1 \leq R,$$

so the divisor r_1 is well-located. Next, by induction applied to the divisor $r_1 = r/n_1$ of the (k, y)-d.d. integer r, this integer is (k, yn_1)-d.d., and hence by (0), it is also (k, ym_1)-d.d. Similarly, q_1 is (j, yn_1')-d.d., and hence also (j, ym_1)-d.d. This finishes the proof that m is (i, ym_1)-d.d.

The second part of (i) is similar and left to the reader.

To prove (ii), recall that y-densely divisible means $(1, y)$-densely divisible. We may assume that $m \leq n$. Let $a = [m, n]n^{-1}$. Now let R be such that $1 \leq R \leq [m, n]$. If $R \leq n$, then a factorization $n = qr$ with $Ry^{-1} \leq r \leq R$, which exists since n is y-d.d., gives the factorization $[m, n] = aqr$, which has the well-located divisor r. If $n < R \leq [m, n]$, we get

$$1 \leq \frac{n}{a} \leq \frac{R}{a} \leq n,$$

and therefore there exists a factorization $n = qr$ with $R(ay)^{-1} \leq r \leq Ra^{-1}$. Then $[m, n] = q(ar)$ with $Ry^{-1} \leq ar \leq R$. Thus we see that $[m, n]$ is y-d.d.

We now prove (iii) by induction on i. The case $i = 0$ is again obvious, so we assume that (iii) holds for j-tuply dense divisibility for $j < i$. Let n be a y-smooth integer, let $j, k \geq 0$ satisfy $j + k = i - 1$, and let $1 \leq R \leq n$ be given. Let r be the largest divisor of n which is $\leq R$, and let $q = n/r$. Since all prime divisors of n are $\leq y$, we have

$$Ry^{-1} \leq r \leq R,$$

and furthermore both q and r are y-smooth. By the induction hypothesis, q is (j, y)-d.d. and r is (k, y)-d.d., hence it follows that n is (i, y)-d.d.

We now turn to (iv). The claim is again obvious for $i = 0$. Assume then that $i = 1$. Let R be such that $1 \leq R \leq n$. Let

$$s_1 = \prod_{\substack{p|n \\ p \leq y}} p, \quad r_1 = \prod_{\substack{p|n \\ p > y}} p.$$

Assume first that $r_1 \leq R$. Since $n/r_1 = s_1$ is y-smooth, it is 1-d.d., and since $1 \leq Rr_1^{-1} \leq s_1$, we can factor s_1 into $q_2 r_2$ with $R(r_1 y)^{-1} \leq r_2 \leq Rr_1^{-1}$. Then $n = q_2(r_1 r_2)$ with

$$Ry^{-1} \leq r_1 r_2 \leq R.$$

So assume that $r_1 > R$. Since n and hence r_1 are z-smooth, we can factor r_1 into $r_2 q_2$ with $Rz^{-1} \leq r_2 \leq R$. Let r_3 be the smallest divisor of s_1 such that $r_3 r_2 \geq Ry^{-1}$, which exists because $s_1 r_2 \geq zy^{-1} r_2 \geq Ry^{-1}$ by the assumption (2-15). Since s_1 is y-smooth, we have $r_3 r_2 \leq R$ (since otherwise we must have $r_3 \neq 1$, hence r_3 is divisible by a prime $p \leq y$, and $r_3 p^{-1}$ is a smaller divisor with the required property $r_3 p^{-1} r_2 > Ry^{-1}$, contradicting the minimality of r_3). Therefore $n = q(r_3 r_2)$ with

$$\frac{R}{y} \leq r_3 r_2 \leq R,$$

as desired.

Finally we consider the $i > 1$ case. We assume, by induction, that (iv) holds for integers $j < i$. Let $j, k \geq 0$ be such that $j + k = i - 1$. By assumption, using the notation r_1, s_1 as above, we have

$$s_1 \geq z^i y^{-1} = z^j \cdot z^k \cdot \frac{z}{y}.$$

We can therefore write $s_1 = n_1 n_2 n_3$, where

$$\begin{aligned} z^j y^{-1} &\leq n_1 \leq z^j, \\ z^k y^{-1} &\leq n_2 \leq z^k, \end{aligned} \tag{2-16}$$

and thus

$$n_3 \geq \frac{z}{y}.$$

Now we divide into several cases in order to find a suitable factorization of n. Suppose first that $n_1 \leq R \leq n/n_2$. Then

$$1 \leq \frac{R}{n_1} \leq \frac{n}{n_1 n_2}$$

and the integer $n/(n_1 n_2) = r_1 n_3$ satisfies the assumptions of (iv) for $i = 1$. Thus, by the previous case, we can find a factorization $r_1 n_3 = q'r'$ with $y^{-1}(R/n_1) \leq r' \leq R/n_1$. We set $r = n_1 r'$ and $q = n_2 q'$, and observe that by (2-16), r and q satisfy the assumption of (iv) for $i = j$ and $i = k$ respectively. By induction, the factorization $n = qr$ has the required property.

Next, we assume that $R < n_1$. Since n_1 is y-smooth, we can find a divisor r of n_1 such that $y^{-1}R \leq r \leq R$. Then $q = n/r$ is a multiple of n_2, and therefore it satisfies

$$\prod_{\substack{p|q \\ p \leq y}} p \geq n_2 \geq z^k y^{-1}.$$

By induction, it follows that q is (k, y)-d.d. Since r is y-smooth, q is also (j, y)-d.d. by (iii), and hence the factorization $n = qr$ is suitable in this case.

Finally, suppose that $R > n/n_2$, i.e., that $nR^{-1} < n_2$. We then find a factor q of the y-smooth integer n_2 such that $n(Ry)^{-1} \leq q \leq nR^{-1}$. Then the complementary factor $r = n/q$ is a multiple of n_1, and therefore it satisfies

$$\prod_{\substack{p|r \\ p \leq y}} p \geq z^j y^{-1},$$

so that r is (j, y)-d.d. by induction, and since q is also (j, y)-d.d. by (iii), we also have the required factorization in this case. $\qquad\square$

3. Applying the Heath-Brown identity

The goal of this and the next sections is to prove the assumption $MPZ^{(i)}[\varpi, \delta]$ (Claim 2.3) for as wide a range of ϖ and δ as possible, following the outline in Section 1A. The first step, which we implement in this section, is the proof of Lemma 2.7. We follow standard arguments, particularly those in [Zhang 2014]. The main tool is the Heath-Brown identity, which is combined with a purely combinatorial result about finite sets of nonnegative numbers. We begin with the latter statement:

Lemma 3.1. *Let $\frac{1}{10} < \sigma < \frac{1}{2}$, and let $t_1, \ldots, t_n$ be nonnegative real numbers such that $t_1 + \cdots + t_n = 1$. Then at least one of the following three statements holds*:

(Type 0) *There is a t_i with $t_i \geq \frac{1}{2} + \sigma$.*

(Type I/II) *There is a partition $\{1, \ldots, n\} = S \cup T$ such that*

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

(Type III) *There exist distinct i, j, k with $2\sigma \leq t_i \leq t_j \leq t_k \leq \frac{1}{2} - \sigma$ and*

$$t_i + t_j, \ t_i + t_k, \ t_j + t_k \geq \frac{1}{2} + \sigma. \tag{3-1}$$

Furthermore, if $\sigma > \frac{1}{6}$, then the Type III alternative cannot occur.

Proof. We dispense with the final claim first: if $\sigma > \frac{1}{6}$, then $2\sigma > \frac{1}{2} - \sigma$, and so the inequalities $2\sigma \leq t_i \leq t_j \leq t_k \leq \frac{1}{2} - \sigma$ of the Type III alternative are inconsistent.

Now we prove the main claim. Let σ and $(t_1, \ldots, t_n)$ be as in the statement. We assume that the Type 0 and Type I/II statements are false, and will deduce that the Type III statement holds.

From the failure of the Type 0 conclusion, we know that

$$t_i < \frac{1}{2} + \sigma \tag{3-2}$$

for all $i = 1, \ldots, n$. From the failure of the Type I/II conclusion, we also know that, for any $S \subset \{1, \ldots, n\}$, we have

$$\sum_{i \in S} t_i \notin \left(\frac{1}{2} - \sigma, \frac{1}{2} + \sigma\right),$$

since otherwise we would obtain the conclusion of Type I/II by taking T to be the complement of S, possibly after swapping the roles of S and T.

We say that a set $S \subset \{1, \ldots, n\}$ is *large* if $\sum_{i \in S} t_i \geq \frac{1}{2} + \sigma$, and that it is *small* if $\sum_{i \in S} t_i \leq \frac{1}{2} - \sigma$. Thus, the previous observation shows that every set $S \subset \{1, \ldots, n\}$ is either large or small, and also (from (3-2)) that singletons are small, as is the empty set. Also, it is immediate that the complement of a large set is small, and that the converse holds (since $t_1 + \cdots + t_n = 1$).

Further, we say that an element $i \in \{1, \ldots, n\}$ is *powerful* if there exists a small set $S \subset \{1, \ldots, n\} \setminus \{i\}$ such that $S \cup \{i\}$ is large, i.e., if i can be used to turn a small set into a large set. Then we say that an element i is *powerless* if it is not powerful. Thus, adding or removing a powerless element from a set S cannot alter its smallness or largeness, and in particular, the union of a small set and a set of powerless elements is small.

We claim that there exist exactly three powerful elements. First, there must be at least two, because if P is the set of powerless elements, then it is small, and

hence its complement is large, and thus contains at least two elements, which are powerful. But picking one of these powerful i, the set $\{i\} \cup P$ is small, and therefore its complement also has at least two elements, which together with i are three powerful elements.

Now, we observe that if i is powerful, then $t_i \geq 2\sigma$, since the gap between a large sum $\sum_{j \in S \cup \{i\}} t_j$ and a small sum $\sum_{j \in S} t_j$ is at least 2σ. In particular, if $i \neq j$ are two powerful numbers, then

$$t_i + t_j \geq 4\sigma > \tfrac{1}{2} - \sigma,$$

where the second inequality holds because of the assumption $\sigma > \frac{1}{10}$. Thus the set $\{i, j\}$ is not small, and is therefore large. But then if $\{i, j, k, l\}$ was a set of four powerful elements, it would follow that

$$1 = t_1 + \cdots + t_n \geq (t_i + t_j) + (t_k + t_l) \geq 2\left(\tfrac{1}{2} + \sigma\right) > 1,$$

a contradiction.

Let therefore i, j, k be the three powerful elements. We may order them so that $t_i \leq t_j \leq t_k$. We have

$$2\sigma \leq t_i \leq t_j \leq t_k \leq \tfrac{1}{2} - \sigma$$

by (3-2) and the previous argument, which also shows that $\{i, j\}$, $\{i, k\}$ and $\{j, k\}$ are large, which is (3-1). □

Remark 3.2. For $\frac{1}{10} < \sigma \leq \frac{1}{6}$, the Type III case can indeed occur, as can be seen by considering the examples $(t_1, t_2, t_3) = \left(2\sigma, \frac{1}{2} - \sigma, \frac{1}{2} - \sigma\right)$. The lemma may be extended to the range $\frac{1}{14} < \sigma < \frac{1}{2}$, but at the cost of adding two additional cases (corresponding to the case of four or five powerful elements respectively):

(Type IV) There exist distinct i, j, k, l with $2\sigma \leq t_i \leq t_j \leq t_k \leq t_l \leq \frac{1}{2} - \sigma$ and $t_i + t_l \geq \frac{1}{2} + \sigma$.

(Type V) There exist distinct i, j, k, l, m with $2\sigma \leq t_i \leq t_j \leq t_k \leq t_l \leq t_m \leq \frac{1}{2} - \sigma$ and $t_i + t_j + t_k \geq \frac{1}{2} + \sigma$.

We leave the verification of this extension to the reader. Again, for $\frac{1}{14} < \sigma \leq \frac{1}{10}$, the Type IV and Type V cases can indeed occur, as can be seen by considering the examples $(t_1, t_2, t_3, t_4) = \left(2\sigma, 2\sigma, \frac{1}{2} - 3\sigma, \frac{1}{2} - \sigma\right)$ and $(t_1, t_2, t_3, t_4, t_5) = (2\sigma, 2\sigma, 2\sigma, 2\sigma, 1 - 8\sigma)$. With this extension, it is possible to extend Lemma 2.7 to the regime $\frac{1}{14} < \sigma < \frac{1}{2}$, but at the cost of requiring additional "Type IV" and "Type V" estimates as hypotheses. Unfortunately, while the methods in this paper do seem to be able to establish some Type IV estimates, they do not seem to give enough Type V estimates to make it profitable to try to take σ below $\frac{1}{10}$.

To apply Lemma 3.1 to distribution theorems concerning the von Mangoldt function Λ, we recall the Heath-Brown identity (see [Heath-Brown 1982] or [Iwaniec and Kowalski 2004, Proposition 13.3]).

Lemma 3.3 (Heath-Brown identity). *For any $K \geq 1$, we have the identity*

$$\Lambda = \sum_{j=1}^{K} (-1)^{j-1} \binom{K}{j} \mu_{\leq}^{\star j} \star \mathbf{1}^{\star(j-1)} \star L \tag{3-3}$$

on the interval $[x, 2x]$, where $\mathbf{1}$ is the constant function $\mathbf{1}(n) := 1$, L is the logarithm function $L(n) := \log n$, $\mu_{\leq}$ is the truncated Möbius function

$$\mu_{\leq}(n) := \mu(n) \mathbf{1}_{n \leq (2x)^{1/K}},$$

and where we denote by $f^{\star j} = f \star \cdots \star f$ the j-fold Dirichlet convolution of an arithmetic function f, i.e.,

$$f^{\star j}(n) := \sum_{a_1 \cdots a_j = n} \cdots \sum f(a_1) \cdots f(a_j).$$

Proof. Write $\mu = \mu_{\leq} + \mu_{>}$, where $\mu_{>}(n) := \mu(n) \mathbf{1}_{n > (2x)^{1/K}}$. Clearly the convolution

$$\mu_{>}^{\star K} \star \mathbf{1}^{\star K-1} \star L$$

vanishes on $[1, 2x]$. Expanding out $\mu_{>} = \mu - \mu_{\leq}$ and using the binomial formula, we conclude that

$$0 = \sum_{j=0}^{K} (-1)^j \binom{K}{j} \mu^{\star(K-j)} \star \mu_{\leq}^{\star j} \star \mathbf{1}^{\star(K-1)} \star L \tag{3-4}$$

on $[x, 2x]$. Since Dirichlet convolution is associative, the standard identities $\Lambda = \mu \star L$ and $\delta = \mu \star \mathbf{1}$ (where the Kronecker delta function $\delta(n) := \mathbf{1}_{n=1}$ is the unit for Dirichlet convolution) show that the $j = 0$ term of (3-4) is

$$\mu^{\star K} \star \mathbf{1}^{\star(K-1)} \star L = \mu \star L = \Lambda.$$

For all the other terms, we can use commutativity of Dirichlet convolution and (again) $\mu \star \mathbf{1} = \delta$ to write

$$\mu^{\star K-j} \star \mu_{\leq}^{\star j} \star \mathbf{1}^{\star K-1} \star L = \mu_{\leq}^{\star j} \star \mathbf{1}^{\star(j-1)} \star L,$$

so that we get (3-3). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

We will now prove Lemma 2.7, which the reader is invited to review. Let $i, \varpi, \delta, \sigma$ satisfy the hypotheses of that lemma, and let $A_0 > 0$ be fixed. By the definition of $\mathrm{MPZ}^{(i)}(\varpi, \delta)$, which is the conclusion of the lemma, it suffices to show that for any $Q \ll x^{1/2+2\varpi}$, any bounded set $I \subset (0, +\infty)$ and any residue class $a\ (P_I)$, we have

$$\sum_{q \in \mathcal{Q}} |\Delta(\Lambda \mathbf{1}_{[x,2x]}; a\ (q))| \ll x \log^{-A_0+O(1)} x, \tag{3-5}$$

where

$$\mathcal{Q} := \{q \leq Q : q \in \mathscr{D}_I^{(i)}(x^\delta)\} \tag{3-6}$$

(recalling the definition (2-2)) and the $O(1)$ term in the exponent is independent of A_0.

Let K be any fixed integer with

$$\frac{1}{K} < 2\sigma \tag{3-7}$$

(e.g., one can take $K = 10$). We apply Lemma 3.3 with this value of K. By the triangle inequality, it suffices to show that

$$\sum_{q \in \mathcal{Q}} |\Delta((\mu_{\leq}^{\star j} \star 1^{\star j-1} \star L)\mathbf{1}_{[x,2x]}; a\ (q))| \ll x \log^{-A_0/2+O(1)} x \tag{3-8}$$

for each $1 \leq j \leq K$, which we now fix.

The next step is a finer-than-dyadic decomposition (a standard idea going back at least to [Fouvry 1984] and [Fouvry and Iwaniec 1983]). We define $\Theta := 1 + \log^{-A_0} x$. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a smooth function supported on $[-\Theta, \Theta]$ that is equal to 1 on $[-1, 1]$ and obeys the derivative estimates

$$|\psi^{(m)}(x)| \ll \log^{mA_0} x$$

for $x \in \mathbb{R}$ and any fixed $m \geq 0$, where the implied constant depends only on m. We then have a smooth partition of unity

$$1 = \sum_{N \in \mathscr{D}} \psi_N(n)$$

indexed by the multiplicative semigroup

$$\mathscr{D} := \{\Theta^m : m \in \mathbb{N} \cup \{0\}\}$$

for any natural number n, where

$$\psi_N(n) := \psi\left(\frac{n}{N}\right) - \psi\left(\frac{\Theta n}{N}\right)$$

is supported in $[\Theta^{-1}N, \Theta N]$. We thus have decompositions

$$1 = \sum_{N \in \mathcal{D}} \psi_N, \quad \mu_{\le} = \sum_{N \in \mathcal{D}} \psi_N \mu_{\le}, \quad L = \sum_{N \in \mathcal{D}} \psi_N L.$$

For $1 \le j \le K$, we have

$$(\mu_{\le}^{\star j} \star 1^{\star(j-1)} \star L)\mathbf{1}_{[x,2x]}$$

$$= \sum_{N_1,\dots,N_{2j} \in \mathcal{D}} \cdots \sum \{(\psi_{N_1}\mu_{\le}) \star \cdots \star (\psi_{N_j}\mu_{\le}) \star \psi_{N_{j+1}} \star \cdots \star \psi_{N_{2j-1}} \star \psi_{N_{2j}} L\}\mathbf{1}_{[x,2x]}$$

$$= \sum_{N_1,\dots,N_{2j} \in \mathcal{D}} \cdots \sum \log(N_{2j})\{(\psi_{N_1}\mu_{\le}) \star \cdots \star (\psi_{N_j}\mu_{\le}) \star \psi_{N_{j+1}} \star \cdots \star \psi_{N_{2j-1}} \star \psi'_{N_{2j}}\}\mathbf{1}_{[x,2x]},$$

where $\psi'_N := \psi_N(L/\log N)$ is a simple variant of ψ_N.

For each $N_1, \dots, N_{2j}$, the summand in this formula vanishes unless

$$N_1, \dots, N_j \ll x^{1/K} \tag{3-9}$$

and

$$\frac{x}{\Theta^{2K}} \le N_1 \cdots N_{2j} \le 2x\Theta^{2K}.$$

In particular, it vanishes unless

$$x\left(1 - O\left(\frac{1}{\log^{A_0} x}\right)\right) \le N_1 \cdots N_{2j} \le 2x\left(1 + O\left(\frac{1}{\log^{A_0} x}\right)\right). \tag{3-10}$$

We conclude that there are at most

$$\ll \log^{2j(A_0+1)} x \tag{3-11}$$

tuples $(N_1, \dots, N_{2j}) \in \mathcal{D}^{2j}$ for which the summand is nonzero. Let $\mathcal{E}$ be the set of these tuples. We then consider the arithmetic function

$$\alpha = \sum_{(N_1,\dots,N_{2j}) \in \mathcal{E}} \cdots \sum \log(N_{2j})\{(\psi_{N_1}\mu_{\le}) \star \cdots \star (\psi_{N_j}\mu_{\le}) \star \psi_{N_{j+1}} \star \cdots \star \psi_{N_{2j-1}} \star \psi'_{N_{2j}}\}$$

$$- (\mu_{\le}^{\star j} \star 1^{\star j-1} \star L)\mathbf{1}_{[x,2x]}. \tag{3-12}$$

Note that the cutoff $\mathbf{1}_{[x,2x]}$ is only placed on the second term in the definition of α, and is not present in the first term.

By the previous remarks, this arithmetic function is supported on

$$[x(1 - O(\log^{-A_0} x)), x] \cup [2x, 2x(1 + O(\log^{-A_0} x))],$$

and using the divisor bound and trivial estimates, it satisfies

$$\alpha(n) \ll \tau(n)^{O(1)}(\log n)^{O(1)},$$

where the exponents are bounded independently of A_0. In particular, we deduce from Lemma 1.3 that

$$\Delta(\alpha; a\ (q)) \ll x \log^{-A_0 + O(1)} x$$

for all $q \geq 1$. Using the estimate (3-11) for the number of summands in $\mathcal{E}$, we see that, in order to prove (3-8), it suffices to show that

$$\sum_{q \in \mathcal{Q}} |\Delta(\alpha_1 \star \cdots \star \alpha_{2j}; a\ (q))| \ll x \log^{-A} x \tag{3-13}$$

for $A > 0$ arbitrary, where each α_i is an arithmetic function of the form $\psi_{N_i} \mu_\leq$, ψ_{N_i} or ψ'_{N_i}, where $(N_1, \ldots, N_{2j})$ satisfies (3-9) and (3-10).

We now establish some basic properties of the arithmetic functions α_k that may arise. For a subset $S \subset \{1, \ldots, 2j\}$, we will denote by

$$\alpha_S := \bigstar_{k \in S} \alpha_k$$

the convolution of the α_k for $k \in S$.

Lemma 3.4. *Let $1 \leq k \leq 2j$ and $S \subset \{1, \ldots, 2j\}$. The following facts hold:*

(i) *Each α_k is a coefficient sequence located at scale N_k, and more generally, the convolution α_S is a coefficient sequence located at scale $\prod_{k \in S} N_k$.*

(ii) *If $N_k \gg x^{2\sigma}$, then α_k is smooth at scale N_k.*

(iii) *If $N_k \gg x^\varepsilon$ for some fixed $\varepsilon > 0$, then α_k satisfies the Siegel–Walfisz property. More generally, α_S satisfies the Siegel–Walfisz property if $\prod_{k \in S} N_k \gg x^\varepsilon$ for some fixed $\varepsilon > 0$.*

(iv) *$N_1 \cdots N_{2j} \asymp x$.*

Proof. The first part of (i) is clear from construction. For the second part of (i), we use the easily verified fact that if α, β are coefficient sequences located at scales N, M respectively, then $\alpha \star \beta$ is a coefficient sequence located at scale NM.

For (ii), we observe that since $2\sigma > K^{-1}$, the condition $N_k \gg x^{2\sigma}$ can only occur for $k > j$ in view of (3-9), so that α_k takes the form ψ_{N_k} or ψ'_{N_k}, and the smoothness then follows directly from the definitions.

For (iii), the Siegel–Walfisz property for α_k when $k \leq j$ follows from the Siegel–Walfisz theorem for the Möbius function and for Dirichlet characters (see, e.g., [Siebert 1971, Satz 4] or [Iwaniec and Kowalski 2004, Theorem 5.29]), using summation by parts to handle the smooth cutoff, and we omit the details. For $k > j$, α_k is smooth, and the Siegel–Walfisz property for α_k follows from the Poisson summation formula (and the rapid decay of the Fourier transform of smooth, compactly supported functions; compare with the arguments at the end of this section for the Type 0 case).

To handle the general case, it therefore suffices to check that if α, β are coefficient sequences located at scales N, M, respectively, with $x^\varepsilon \ll M \ll x^C$ for some fixed $\varepsilon, C > 0$, and β satisfies the Siegel–Walfisz property, then so does $\alpha \star \beta$. This is again relatively standard, but we give the proof for completeness.

By Definition 2.5, our task is to show that

$$|\Delta((\alpha \star \beta)\mathbf{1}_{(\cdot,q)=1}; a\,(r))| \ll \tau(qr)^{O(1)} N \log^{-A} x$$

for any $q, r \geq 1$, any fixed A, and any primitive residue class $a\,(r)$. We replace α, β by their restriction to integers coprime to qr (without indicating this in the notation), which allows us to remove the constraint $\mathbf{1}_{(n,q)=1}$. We may also assume that $r = O(\log^{A+O(1)} x)$, since the desired estimate follows from the trivial estimate (2-7) for the discrepancy otherwise.

For any integer n, we have

$$\sum_{n=a\,(r)} (\alpha \star \beta)(n) = \sum_{b \in (\mathbb{Z}/r\mathbb{Z})^\times} \left(\sum_{d=b\,(r)} \alpha(d) \right) \left(\sum_{m=\bar{b}a\,(r)} \beta(m) \right)$$

and

$$\sum_n (\alpha \star \beta)(n) = \left(\sum_d \alpha(d) \right) \left(\sum_m \beta(m) \right) = \sum_{b \in (\mathbb{Z}/r\mathbb{Z})^\times} \left(\sum_{d=b\,(r)} \alpha(d) \right) \left(\sum_m \beta(m) \right)$$

so that

$$|\Delta(\alpha \star \beta, a\,(r))| \leq \sum_{b \in (\mathbb{Z}/r\mathbb{Z})^\times} \left| \sum_{d=b\,(r)} \alpha(d) \right| |\Delta(\beta; \bar{b}a\,(r))|.$$

From (1-4) (and Definition 2.5), we have

$$\sum_{d=b\,(r)} \alpha(d) \ll \frac{N}{r} \tau(r)^{O(1)} \log^{O(1)} x + N^{o(1)}$$

for any $b\,(r)$, and since β has the Siegel–Walfisz property, we have

$$|\Delta(\beta; \bar{b}a\,(r))| \ll \tau(r)^{O(1)} M \log^{-B} x$$

for any $b\,(r)$ and any fixed $B > 0$. Thus

$$\begin{aligned}
|\Delta(\alpha \star \beta, a\,(r))| &\ll \tau(r)^{O(1)} \varphi(r) \left(\frac{N}{r} + N^{o(1)} \right) M \log^{-B+O(1)} x \\
&\ll \tau(r)^{O(1)} M N \log^{-B+O(1)} x,
\end{aligned}$$

by the assumption concerning the size of r.

Finally, claim (iv) follows from (3-10). $\square$

We now conclude this section by showing how the assumptions $\text{Type}_{\text{I}}^{(i)}[\varpi, \delta, \sigma]$, $\text{Type}_{\text{II}}^{(i)}[\varpi, \delta]$ and $\text{Type}_{\text{III}}^{(i)}[\varpi, \delta, \sigma]$ of Lemma 2.7 imply the estimates (3-13).

Let therefore $(\alpha_1, \ldots, \alpha_{2j})$ be given satisfying the condition after (3-13). By Lemma 3.4(iv), we can write $N_k \asymp x^{t_k}$ for $k = 1, \ldots, 2j$, where the t_k are nonnegative reals (not necessarily fixed) that sum to 1. By Lemma 3.1, the t_i satisfy one of the three conclusions (Type 0), (Type I/II), (Type III) of that lemma. We deal with each in turn. The first case can be dealt with directly, while the others require one of the assumptions of Lemma 2.7, and we begin with these.

Suppose that we are in the Type I/II case, with the partition $\{1, \ldots, 2j\} = S \cup T$ given by the combinatorial lemma. We have

$$\alpha_1 \star \cdots \star \alpha_{2j} = \alpha_S \star \alpha_T.$$

By Lemma 3.4, α_S, α_T are coefficient sequences located at scales N_S, N_T respectively, where

$$N_S N_T \asymp x,$$

and (by (iii)) α_S and α_T satisfy the Siegel–Walfisz property. By Lemma 3.1, we also have

$$x^{1/2-\sigma} \ll N_S \ll N_T \ll x^{1/2+\sigma}.$$

Thus, directly from Definition 2.6 and (3-6), the required estimate (3-13) follows either from the hypothesis $\text{Type}_{\text{I}}^{(i)}[\varpi, \delta, \sigma]$ (if one has $N_S \leq x^{1/2-2\varpi-c}$ for some sufficiently small fixed $c > 0$) or from $\text{Type}_{\text{II}}^{(i)}[\varpi, \delta]$ (if $N_S > x^{1/2-2\varpi-c}$, for the same value of c).

Similarly, in the Type III case, comparing Lemmas 3.4 and 3.1 with Definition 2.6 and (3-6) shows that (3-8) is a direct translation of $\text{Type}_{\text{III}}^{(i)}[\varpi, \delta, \sigma]$.

It remains to prove (3-8) in the Type 0 case, and we can do this directly. In this case, there exists some $k \in \{1, \ldots, 2j\}$ such that $t_k \geq \frac{1}{2} + \sigma > 2\sigma$. Intuitively, this means that α_k is smooth (by Lemma 3.4(ii)) and has a long support, so that it is very well-distributed in arithmetic progressions to relatively large moduli, and we can just treat the remaining α_j trivially.

Precisely, we write

$$\alpha_1 \star \cdots \star \alpha_{2j} = \alpha_k \star \alpha_S,$$

where $S = \{1, \ldots, 2j\}\backslash\{k\}$. By Lemma 3.4, α_k is a coefficient sequence which is smooth at a scale $N_k \gg x^{1/2+\sigma}$, and α_S is a coefficient sequence which is located at a scale N_S with $N_k N_S \asymp x$. We argue as in Lemma 3.4(iii): we have

$$\Delta(\alpha_k \star \alpha_S; a\,(q)) = \sum_{m \in (\mathbb{Z}/q\mathbb{Z})^\times} \sum_{\ell = m\,(q)} \alpha_S(\ell) \Delta(\alpha_k; \bar{m}a\,(q)),$$

and since

$$\sum_m |\alpha_S(m)| \ll N_S$$

(by (1-3) and Definition 2.5), we get

$$\sum_{q \in \mathcal{D}} |\Delta(\alpha_1 \star \cdots \star \alpha_{2j}; a\,(q))| \ll N_S \sum_{q \le Q} \sup_{b \in (\mathbb{Z}/q\mathbb{Z})^\times} |\Delta(\alpha_k; b\,(q))|. \qquad (3\text{-}14)$$

Since α_k is smooth at scale N_k, we can write

$$\alpha_k(n) = \psi(n/N_k)$$

for some smooth function $\psi : \mathbb{R} \to \mathbb{R}$ supported on an interval of size $\ll 1$ which satisfies the estimates

$$|\psi^{(j)}(t)| \ll 1$$

for all t and all fixed $j \ge 0$. By the Poisson summation formula, we have

$$\sum_{n=b\,(q)} \alpha_k(n) = \frac{N_k}{q} \sum_{m \in \mathbb{Z}} e_q(mb)\hat{\psi}\left(\frac{mN_k}{q}\right) = \frac{N_k}{q}\hat{\psi}(0) + \frac{N_k}{q}\sum_{m \ne 0} e_q(mb)\hat{\psi}\left(\frac{mN_k}{q}\right)$$

for $q \ge 1$ and $b\,(q)$, where

$$\hat{\psi}(s) := \int_{\mathbb{R}} \psi(t)e(-ts)\,dt$$

is the Fourier transform of ψ. From the smoothness and support of ψ, we get the bound

$$\left|\hat{\psi}\left(\frac{mN_k}{q}\right)\right| \ll \left(\frac{mN_k}{q}\right)^{-2}$$

for $m \ne 0$ and $q \le Q$, and thus we derive that

$$\sum_{n=b\,(q)} \alpha_k(n) = \frac{N_k}{q}\hat{\psi}(0) + O\left(\frac{N_k}{q}(N_k/q)^{-2}\right).$$

Since by definition

$$\Delta(\alpha_k; b\,(q)) = \sum_{n=b\,(q)} \alpha_k(n) - \frac{1}{\varphi(q)}\sum_{c \in (\mathbb{Z}/q\mathbb{Z})^\times}\sum_{n=c\,(q)} \alpha_k(n),$$

we get

$$|\Delta(\alpha_k; b\,(q))| \ll \frac{N_k}{q}(N_k/q)^{-2}.$$

Therefore, from (3-14), we have

$$\sum_{q \in \mathcal{Q}} |\Delta(\alpha_1 \star \cdots \star \alpha_{2j}; a \ (q))| \ll N_S N_k \left(\frac{Q}{N_k} \right)^2 \ll x^{1-2\sigma+4\varpi},$$

and since $\sigma > 2\varpi$ (by assumption in Lemma 2.7), this implies (3-13), which concludes the proof of Lemma 2.7.

Remark 3.5. In the case $\sigma > \frac{1}{6}$, one can replace the Heath-Brown identity of Lemma 3.3 with other decompositions of the von Mangoldt function Λ, and in particular with the well-known *Vaughan identity* [1977]

$$\Lambda_{\geq} = \mu_< \star L - \mu_< \star \Lambda_< \star 1 + \mu_{\geq} \star \Lambda_{\geq} \star 1,$$

where

$$\Lambda_{\geq}(n) := \Lambda(n) \mathbf{1}_{n \geq V}, \qquad \Lambda_<(n) := \Lambda(n) \mathbf{1}_{n < V}, \qquad (3\text{-}15)$$

$$\mu_{\geq}(n) := \mu(n) \mathbf{1}_{n \geq U}, \qquad \mu_<(n) := \mu(n) \mathbf{1}_{n < U}, \qquad (3\text{-}16)$$

where $U, V > 1$ are arbitrary parameters. Setting $U = V = x^{1/3}$, we then see that to show (3-5), it suffices to establish the bounds

$$\sum_{q \in \mathcal{Q}} |\Delta((\mu_< \star L) \mathbf{1}_{[x,2x]}; a \ (q))| \ll x \log^{-A_0/2+O(1)} x, \qquad (3\text{-}17)$$

$$\sum_{q \in \mathcal{Q}} |\Delta((\mu_< \star \Lambda_< \star 1) \mathbf{1}_{[x,2x]}; a \ (q))| \ll x \log^{-A_0/2+O(1)} x, \qquad (3\text{-}18)$$

$$\sum_{q \in \mathcal{Q}} |\Delta((\mu_{\geq} \star \Lambda_{\geq} \star 1) \mathbf{1}_{[x,2x]}; a \ (q))| \ll x \log^{-A_0/2+O(1)} x. \qquad (3\text{-}19)$$

To prove (3-17), we may perform dyadic decomposition on $\mu_<$ and L, much as in the previous arguments. The components of L which give a nontrivial contribution to (3-17) will be located at scales $\gg x^{2/3}$. One can then use the results of the Type 0 analysis above. In order to prove (3-19), we similarly decompose the $\mu_{\geq}$, $\Lambda_{\geq}$, and 1 factors and observe that the resulting components of $\mu_{\geq}$ and $\Lambda_{\geq} \star 1$ that give a nontrivial contribution to (3-19) will be located at scales M, N with $x^{1/3} \ll M, N \ll x^{2/3}$ and $MN \asymp x$, and one can then argue using Type I and Type II estimates as before since $\sigma > \frac{1}{6}$. Finally, for (3-18), we decompose $\mu_< \star \Lambda_<$ and 1 into components at scales M, N, respectively, with $M \ll x^{2/3}$ and $MN \asymp x$, so $N \gg x^{1/3}$. If $N \gg x^{2/3}$, then the Type 0 analysis applies again, and otherwise we may use the Type I and Type II estimates with $\sigma > \frac{1}{6}$.

Remark 3.6. An inspection of the arguments shows that the interval $[x, 2x]$ used in Lemma 2.7 may be replaced by a more general interval $[x_1, x_2]$ for any $x \leq x_1 \leq x_2 \leq 2x$, leading to a slight generalization of the conclusion $\mathrm{MPZ}^{(i)}[\varpi, \delta]$.

By telescoping series, one may then generalize the intervals $[x_1, x_2]$ further, to the range $1 \leq x_1 \leq x_2 \leq 2x$.

In the next sections, we will turn our attention to the task of proving distribution estimates of Type I, II and III. All three turn out to be intimately related to estimates for exponential sums over $\mathbb{Z}/q\mathbb{Z}$, either "complete" sums over all of $\mathbb{Z}/q\mathbb{Z}$ or "incomplete" sums over suitable subsets, such as reductions modulo q of intervals or arithmetic progressions (this link goes back to the earliest works in proving distribution estimates beyond the range of the large sieve). In the next section, we consider the basic theory of the simplest of those sums, where the essential results go back to Weil's theory of exponential sums in one variable over finite fields. These are enough to handle basic Type I and II estimates, which we consider next. On the other hand, for Type III estimates and the most refined Type I estimates, we require the much deeper results and insights of Deligne's second proof of the Riemann hypothesis for algebraic varieties over finite fields.

4. One-dimensional exponential sums

The results of this section are very general and are applicable to many problems in analytic number theory. Since the account we provide might well be useful as a general reference beyond the applications to the main results of this paper, we will not use the asymptotic convention of Definition 1.2, but provide explicit estimates that can easily be quoted in other contexts. (In particular, we will sometimes introduce variables named x in our notation.)

4A. *Preliminaries.* We begin by setting up some notation and conventions. We recall from Section 1B that we defined $e_q(a) = e^{2i\pi a/q}$ for $a \in \mathbb{Z}$ and $q \geq 1$. This is a group homomorphism $\mathbb{Z} \to \mathbb{C}^\times$, and since $q\mathbb{Z} \subset \ker e_q$, it naturally induces a homomorphism, which we also denote by e_q, from $\mathbb{Z}/q\mathbb{Z}$ to $\mathbb{C}^\times$. In fact, for any multiple qr of q, we can also view e_q as a homomorphism $\mathbb{Z}/qr\mathbb{Z} \to \mathbb{C}^\times$.

It is convenient for us (and compatible with the more algebraic theory for multivariable exponential sums discussed in Section 6) to extend further e_q to the projective line $\mathbb{P}^1(\mathbb{Z}/q\mathbb{Z})$ by extending it by zero to the point(s) at infinity. Precisely, recall that $\mathbb{P}^1(\mathbb{Z}/q\mathbb{Z})$ is the quotient of

$$X_q = \{(a, b) \in (\mathbb{Z}/q\mathbb{Z})^2 : a \text{ and } b \text{ have no common factor}\}$$

(where a common factor of a and b is a prime $p \mid q$ such that a and b are zero modulo p) by the equivalence relation

$$(a, b) = (ax, bx)$$

for all $x \in (\mathbb{Z}/q\mathbb{Z})^{\times}$. We identify $\mathbb{Z}/q\mathbb{Z}$ with a subset of $\mathbb{P}^1(\mathbb{Z}/q\mathbb{Z})$ by sending x to the class of $(x, 1)$. We note that

$$|\mathbb{P}^1(\mathbb{Z}/q\mathbb{Z})| = q \prod_{p|q} \left(1 + \frac{1}{p}\right),$$

and that a point $(a, b) \in \mathbb{P}^1(\mathbb{Z}/q\mathbb{Z})$ belongs to $\mathbb{Z}/q\mathbb{Z}$ if and only if $b \in (\mathbb{Z}/q\mathbb{Z})^{\times}$, in which case $(a, b) = (ab^{-1}, 1)$.

Thus, we can extend e_q to $\mathbb{P}^1(\mathbb{Z}/q\mathbb{Z})$ by defining

$$e_q((a, b)) = e_q(ab^{-1})$$

if $b \in (\mathbb{Z}/q\mathbb{Z})^{\times}$, and $e_q((a, b)) = 0$ otherwise.

We have well-defined reduction maps $\mathbb{P}^1(\mathbb{Z}/qr\mathbb{Z}) \to \mathbb{P}^1(\mathbb{Z}/q\mathbb{Z})$ for all integers $r \geq 1$, as well as $\mathbb{P}^1(\mathbb{Q}) \to \mathbb{P}^1(\mathbb{Z}/q\mathbb{Z})$, and we can therefore also naturally define $e_q(x)$ for $x \in \mathbb{P}^1(\mathbb{Z}/qr\mathbb{Z})$ or for $x \in \mathbb{P}^1(\mathbb{Q})$ (for the map $\mathbb{P}^1(\mathbb{Q}) \to \mathbb{P}^1(\mathbb{Z}/q\mathbb{Z})$, we use the fact that any $x \in \mathbb{P}^1(\mathbb{Q})$ is the class of (a, b) where a and b are coprime integers, so that $(a\ (q), b\ (q)) \in X_q$).

We will use these extensions especially in the following context: let $P, Q \in \mathbb{Z}[X]$ be polynomials, with $Q \neq 0$, and consider the rational function $f = P/Q \in \mathbb{Q}(X)$. This defines a map $\mathbb{P}^1(\mathbb{Q}) \to \mathbb{P}^1(\mathbb{Q})$, and then, by reduction modulo q, a map

$$f\ (q) : \mathbb{P}^1(\mathbb{Z}/q\mathbb{Z}) \to \mathbb{P}^1(\mathbb{Z}/q\mathbb{Z}).$$

We can therefore consider the function $x \mapsto e_q(f(x))$ for $x \in \mathbb{Z}/q\mathbb{Z}$. If $x \in \mathbb{Z}$ is such that $Q(x)$ is coprime to q, then this is just $e_q(P(x)\overline{Q(x)})$. If $Q(x)$ is not coprime to q, on the other hand, one must be a bit careful. If q is prime, then one should write $f\ (q) = P_1/Q_1$ with $P_1, Q_1 \in (\mathbb{Z}/q\mathbb{Z})[X]$ coprime, and then $e_q(f(x)) = e_q(P_1(x)\overline{Q_1(x)})$ if $Q_1(x) \neq 0$, while $e_q(f(x)) = 0$ otherwise. If q is squarefree, one combines the prime components according to the Chinese remainder theorem, as we will recall later.

Example 4.1. Let $P = X$, $Q = X + 3$ and $q = 3$, and set $f := P/Q$. Then, although $P\ (q)$ and $Q\ (q)$ both take the value 0 at $x = 0 \in \mathbb{Z}/q\mathbb{Z}$, we have $e_q(f(0)) = 1$.

In rare cases (in particular the proof of Proposition 8.4 in Section 8D) we will use one more convention: quantities

$$e_p\left(\frac{a}{b}\right)$$

may arise, where a and b are integers that depend on other parameters, and with b allowed to be divisible by p. However, this will only happen when the formula is to be interpreted as

$$e_p\left(\frac{a}{b}\right) = \psi\left(\frac{1}{b}\right) = \psi(\infty),$$

where $\psi(x) = e_p(ax)$ defines an additive character of $\mathbb{F}_p$. Thus we use the convention

$$e_p\left(\frac{a}{b}\right) = \begin{cases} 0 & \text{if } a \neq 0\,(p),\, b = 0\,(p), \\ 1 & \text{if } a = 0\,(p),\, b = 0\,(p), \end{cases}$$

since in the second case we are evaluating the trivial character at ∞.

4B. *Complete exponential sums over a finite field.* As is well-known since early works of Davenport and Hasse in particular, the Riemann hypothesis for curves over finite fields (proved by Weil [1948]) implies bounds with "square root cancellation" for one-dimensional exponential sums over finite fields. A special case is the following general bound:

Lemma 4.2 (one-variable exponential sums with additive characters). *Let $P, Q \in \mathbb{Z}[X]$ be polynomials over $\mathbb{Z}$ in one indeterminate X. Let p be a prime number such that $Q\,(p) \in \mathbb{F}_p[X]$ is nonzero and such that there is no identity of the form*

$$\frac{P}{Q}\,(p) = g^p - g + c \tag{4-1}$$

in $\mathbb{F}_p(X)$ for some rational function $g = g(X) \in \mathbb{F}_p(X)$ and some $c \in \mathbb{F}_p$. Then we have

$$\left| \sum_{x \in \mathbb{F}_p} e_p\left(\frac{P(x)}{Q(x)}\right) \right| \ll \sqrt{p}, \tag{4-2}$$

where the implicit constant depends only on $\max(\deg P, \deg Q)$, and this dependency is linear.

Note that, by our definitions, we have

$$\sum_{x \in \mathbb{F}_p} e_p\left(\frac{P(x)}{Q(x)}\right) = \sum_{\substack{x \in \mathbb{F}_p \\ Q_1(x) \neq 0}} e_p(P_1(x)\overline{Q_1(x)}),$$

where $P/Q\,(p) = P_1/Q_1$ with $P_1, Q_1 \in \mathbb{F}_p[X]$ coprime polynomials.

As key examples of Lemma 4.2, we record Weil's bound for Kloosterman sums, namely,

$$\left| \sum_{x \in \mathbb{F}_p} e_p\left(ax + \frac{b}{x}\right) \right| \ll \sqrt{p} \tag{4-3}$$

when $a, b \in \mathbb{F}_p$ are not both zero, as well as the variant

$$\left| \sum_{x \in \mathbb{F}_p} e_p\left(ax + \frac{b}{x} + \frac{c}{x+l} + \frac{d}{x+m} + \frac{e}{x+l+m}\right) \right| \ll \sqrt{p} \tag{4-4}$$

for $a, b, c, d, e, l, m \in \mathbb{F}_p$ with $b, c, d, e, l, m, l + m$ nonzero. In fact, these two estimates are almost the only two cases of Lemma 4.2 that are needed in our arguments. In both cases, one can determine a suitable implied constant, e.g., the Kloosterman sum in (4-3) has modulus at most $2\sqrt{p}$.

We note also that the case (4-1) must be excluded, since $g^p(x) - g(x) + c = c$ for all $x \in \mathbb{F}_p$, and therefore the corresponding character sum has size equal to p.

Proof. This estimate follows from the Riemann hypothesis for the algebraic curve C over $\mathbb{F}_p$ defined by the Artin–Schreier equation

$$y^p - y = P(x)/Q(x).$$

This was first explicitly stated by Perel′muter [1969], although this was undoubtedly known to Weil; an elementary proof based on Stepanov's method may also be found in [Cochrane and Pinner 2006]. A full proof for all curves, using a minimal amount of the theory of algebraic curves, is found in [Bombieri 1974]. $\qquad\square$

Remark 4.3. For our purpose of establishing some nontrivial Type I and Type II estimates for a given choice of σ (and in particular for σ slightly above $\frac{1}{6}$) and for sufficiently small ϖ, δ, it is not necessary to have the full square root cancellation in (4-2), and any power savings of the form p^{1-c} for some fixed absolute constant $c > 0$ would suffice (with the same dependency on P and Q); indeed, assuming such a power savings, one obtains a nontrivial bound on the relevant short exponential sums arising in these estimates once one invokes the q-van der Corput method a sufficient number of times (depending on c and σ), by an appropriate modification of Proposition 4.12 below. The Type I and Type II estimates established in later sections need such a power savings to overcome a variety of inefficiencies in the remainder of the argument, but all of these losses are of the form $O(x^{O(\varpi+\delta)})$ (with the most serious loss coming from the use of completion of sums, which worsens the trivial bound by a factor of about H, where H is defined in (5-25)). The power savings of p^{-c} will be attenuated by a number of applications of the Cauchy–Schwarz inequality (each use of which, roughly speaking, halves the exponent c in the power savings); however, this inequality is only used a bounded number of times, and so any power savings in (4-2) will still lead to enough Type I and Type II estimates to obtain a nontrivial equidistribution estimate for sufficiently small ϖ, δ if one is willing to use the q-van der Corput method a sufficiently large number of times. (In fact, even just Type II estimates alone are sufficient for this task; see Remark 5.11.)

Such a power saving in (4-2) (with $c = \frac{1}{4}$) was obtained for the Kloosterman sum (4-3) by Kloosterman [1927] using an elementary dilation argument (see also [Mordell 1932] for a generalization), but this argument does not appear to be available for estimates such as (4-4).

In order to prove parts (i), (ii) and (iv) of Theorem 2.8, we need to extend the bounds of Lemma 4.2 in two ways: to sums over $\mathbb{Z}/q\mathbb{Z}$ for q squarefree instead of prime, and to incomplete sums over suitable subsets of $\mathbb{Z}/q\mathbb{Z}$ (the other two parts of the theorem also require exponential sum estimates, but these require the much deeper work of Deligne [1980], and will be considered in Section 6).

4C. *Complete exponential sums to squarefree moduli.* To extend Lemma 4.2 to squarefree moduli, we first need some preliminaries. We begin with a version of the Chinese remainder theorem.

Lemma 4.4 (Chinese remainder theorem). *If q_1, q_2 are coprime natural numbers, then for any integer a, or indeed for any $a \in \mathbb{P}^1(\mathbb{Q})$, we have*

$$e_{q_1 q_2}(a) = e_{q_1}\left(\frac{a}{q_2}\right) e_{q_2}\left(\frac{a}{q_1}\right). \tag{4-5}$$

More generally, if $q_1, \ldots, q_k$ are pairwise coprime natural numbers, then for any integer a or any $a \in \mathbb{P}^1(\mathbb{Q})$, we have

$$e_{q_1 \cdots q_k}(a) = \prod_{i=1}^{k} e_{q_i}\left(\frac{a}{\prod_{j \neq i} q_j}\right).$$

Proof. It suffices to prove the former claim for $a \in \mathbb{P}^1(\mathbb{Q})$, as the latter then follows by induction.

If a maps to a point at infinity in $\mathbb{P}^1(\mathbb{Z}/q_1 q_2 \mathbb{Z})$, then it must map to a point at infinity in $\mathbb{P}^1(\mathbb{Z}/q_1\mathbb{Z})$ or $\mathbb{P}^1(\mathbb{Z}/q_2\mathbb{Z})$, so that both sides of (4-5) are zero.

So we can assume that $a \in \mathbb{Z}/q_1 q_2 \mathbb{Z}$. Let $\overline{q_1}, \overline{q_2}$ be integers such that $q_1 \overline{q_1} = 1$ (q_2) and $q_2 \overline{q_2} = 1$ (q_1), respectively. Then we have $q_1 \overline{q_1} + q_2 \overline{q_2} = 1$ $(q_1 q_2)$, and hence

$$e_{q_1 q_2}(a) = e_{q_1 q_2}(a(q_1 \overline{q_1} + q_2 \overline{q_2})) = e_{q_1 q_2}(q_1 \overline{q_1} a) e_{q_1 q_2}(q_2 \overline{q_2} a).$$

Since $e_{q_1 q_2}(q_1 \overline{q_1} a) = e_{q_2}(a/q_1)$ and $e_{q_1 q_2}(q_2 \overline{q_2} a) = e_{q_1}(a/q_2)$, the claim follows. $\square$

If $q \in \mathbb{Z}$ is an integer, we say that q divides f, and write $q \mid f$, if q divides f in $\mathbb{Z}[X]$. We denote by (q, f) the largest factor of q that divides f (i.e., the positive generator of the ideal of $\mathbb{Z}$ consisting of integers dividing f). Thus for instance $(q, 0) = q$. We also write f $(q) \in (\mathbb{Z}/q\mathbb{Z})[X]$ for the reduction of f modulo q.

We need the following algebraic lemma, which can be viewed as a version of (a special case of) the fundamental theorem of calculus:

Lemma 4.5. *Let $f = P/Q \in \mathbb{Q}(X)$ with $P, Q \in \mathbb{Z}[X]$ coprime, and let q be a natural number such that Q (p) is a nonzero polynomial for all primes $p \mid q$ (automatic if Q is monic).*

  (i) *If $q \mid f'$ and all prime factors of q are sufficiently large depending on the degrees of P and Q, then there exists $c \in \mathbb{Z}/q\mathbb{Z}$ such that $q \mid f - c$.*

(ii) *If q is squarefree, if Q (p) has degree* $\deg(Q)$ *for all* $p \mid q$ *and*[2] $\deg(P) < \deg(Q)$, *and if all prime factors of q are sufficiently large depending on the degrees of P and Q, then* (q, f') *divides* (q, f). *In particular, if* $(q, f) = 1$ *then* $(q, f') = 1$.

Proof. We first prove (i). By the Chinese remainder theorem, we may assume that $q = p^j$ is the power of a prime. Write $f' = P_1/Q_1$, where P_1 and $Q_1 \in \mathbb{Z}[X]$ are coprime. By definition, the condition $q \mid f'$ implies that $P_1(x) = 0 \ (q)$ for all $x \in \mathbb{Z}/q\mathbb{Z}$. On the other hand, since $Q_1 \ (p)$ is nonzero in $\mathbb{Z}/p\mathbb{Z}[X]$, the rational function $f' \ (q)$ is well-defined at all $x \in \mathbb{Z}/q\mathbb{Z}$ except at most $\deg(Q)$ zeros of Q_1, and takes the value 0 at all these $\geq q - \deg(Q)$ values. If q is large enough in terms of $\deg(P)$ and $\deg(Q)$, this implies that $f' \ (q) = 0 \in \mathbb{Z}/q\mathbb{Z}[X]$, and therefore that $f \ (q) = c$ for some $c \in \mathbb{Z}/q\mathbb{Z}$, i.e., that $q \mid f - c$.

Now we prove (ii). If a prime p divides (q, f'), then by (i) there exists $c \in \mathbb{Z}/p\mathbb{Z}$ such that $p \mid f - c$. If $p \nmid (q, f)$, we must have $c \neq 0$. But then $p \mid P - cQ$, where $P - cQ \ (p) \in \mathbb{Z}/p\mathbb{Z}[X]$ is (by assumption) a polynomial of degree $\deg(Q) \geq 1$. For $p > \deg(Q)$, this is a contradiction, so that $p \mid (q, f)$. □

We use this to give an estimate for complete exponential sums, which combines the bounds for Ramanujan sums with those from the Riemann hypothesis for curves.

Proposition 4.6 (Ramanujan–Weil bounds). *Let q be a squarefree natural number, and let* $f = P/Q \in \mathbb{Q}(X)$, *where* $P, Q \in \mathbb{Z}[X]$ *are coprime polynomials with Q nonzero modulo p for every* $p \mid q$ *(for instance, with Q monic). Then we have*

$$\left| \sum_{n \in \mathbb{Z}/q\mathbb{Z}} e_q(f(n)) \right| \leq C^{\Omega(q)} q^{1/2} \frac{(f', q)}{(f'', q)^{1/2}}$$

for some constant $C \geq 1$ *depending only on* $\deg(P)$ *and* $\deg(Q)$.

Example 4.7. (1) Let $f(X) := b/X$ for some integer b. We get, after changing the summation variable, a slightly weaker version of the familiar Ramanujan sum bound

$$\left| \sum_{n \in \mathbb{Z}/q\mathbb{Z}} e(bn) \mathbf{1}_{(n,q)=1} \right| \leq (b, q) \tag{4-6}$$

since $(q, f') = (b, q)$ and $(q, f'') = c(b, q)$ in this case for some $c = 1, 2$.

(2) More generally, let $f := a/X + bX$ for some integers a, b. We get a weaker form of Weil's bound for Kloosterman sums:

$$\left| \sum_{n \in \mathbb{Z}/q\mathbb{Z}} e_q(a\bar{n} + bn) \mathbf{1}_{(n,q)=1} \right| \leq 2^{\Omega(q)} q^{1/2} \frac{(a, b, q)}{(a, q)^{1/2}},$$

which generalizes (4-3).

---

[2]We adopt the convention $\deg(0) = -\infty$.

Proof. By Lemma 4.4, we can factor the sum as a product of exponential sums over the prime divisors of q:

$$\sum_{n\in\mathbb{Z}/q\mathbb{Z}} e_q(f(n)) = \prod_{p\mid q} \sum_{n\in\mathbb{Z}/p\mathbb{Z}} e_p\left(\frac{f(n)}{(q/p)}\right).$$

Since, for each $p \mid q$, the constant q/p is an invertible element in $\mathbb{Z}/p\mathbb{Z}$, we see that it suffices to prove the estimates

$$\sum_{n\in\mathbb{Z}/p\mathbb{Z}} e_p(f(n)) \ll p \qquad \text{when } p \mid f' \text{ (which implies } p \mid f''\text{)}, \qquad (4\text{-}7)$$

$$\sum_{n\in\mathbb{Z}/p\mathbb{Z}} e_p(f(n)) \ll 1 \qquad \text{when } p \mid f'' \text{ but } p \nmid f', \qquad (4\text{-}8)$$

$$\sum_{n\in\mathbb{Z}/p\mathbb{Z}} e_p(f(n)) \ll \sqrt{p} \qquad \text{otherwise,} \qquad (4\text{-}9)$$

where the implied constants, in all three cases, depend only on $\deg(P)$ and $\deg(Q)$. Thus we may always assume that $p \mid q$ is large enough in terms of $\deg(P)$ and $\deg(Q)$, since otherwise the result is trivial.

The first bound is clear, with implied constant equal to 1. For (4-8), since $p \mid f''$, we conclude from Lemma 4.5 (since p is large enough) that there exists $c \in \mathbb{Z}/p\mathbb{Z}$ such that $p \mid f' - c$. Since $p \nmid f'$, we see that c must be nonzero. Then, since $f' - c = (f - ct)'$, another application of Lemma 4.5 shows that there exists $d \in \mathbb{Z}/p\mathbb{Z}$ such that $p \mid f - ct - d$. This implies that $f(n) = cn + d \ (p)$ whenever n is not a pole of $f \ (p)$. The denominator Q of f (which is nonzero modulo p by assumption) has at most $\deg(Q)$ zeroes, and therefore we see that $e_p(f(n)) = e_p(cn + d)$ for all but $\leq \deg(Q)$ values of $n \in \mathbb{Z}/p\mathbb{Z}$. Thus (by orthogonality of characters) we get

$$\left|\sum_{n\in\mathbb{Z}/p\mathbb{Z}} e_p(f(n))\right| = \left|\sum_{n\in\mathbb{Z}/p\mathbb{Z}} e_p(f(n)) - \sum_{n\in\mathbb{Z}/p\mathbb{Z}} e_p(cn+d)\right| \leq \deg(Q).$$

Now we prove (4-9). This estimate follows immediately from Lemma 4.2, except if the reduction $\tilde{f} \in \mathbb{F}_p(X)$ of f modulo p satisfies an identity

$$\tilde{f} = g^p - g + c \qquad (4\text{-}10)$$

for some $g \in \mathbb{F}_p(X)$ and $c \in \mathbb{F}_p$. We claim that if p is large enough, this can only happen if $p \mid f'$, which contradicts the assumption of (4-9) and therefore concludes the proof.

To prove the claim, we just observe that if (4-10) holds, then any pole of g would be a pole of $\tilde{f}$ of order p, and thus g must be a polynomial if p is large enough.

But then (4-10) implies that $\tilde{f} - c$ either vanishes or has degree at least p. If p is large enough, the latter conclusion is not possible, and thus $p \mid f'$. $\qquad\square$

We also need a variant of Proposition 4.6, which is a slight refinement of an estimate appearing in the proof of [Zhang 2014, Proposition 11]:

Lemma 4.8. *Let d_1, d_2 be squarefree integers, so that $[d_1, d_2]$ is squarefree, and let c_1, c_2, l_1, l_2 be integers. Then there exists $C \geq 1$ such that*

$$\left| \sum_{n \in \mathbb{Z}/[d_1, d_2]\mathbb{Z}} e_{d_1}\left(\frac{c_1}{n + l_1}\right) e_{d_2}\left(\frac{c_2}{n + l_2}\right) \right| \leq C^{\Omega([d_1, d_2])}(c_1, \delta_1)(c_2, \delta_2)(d_1, d_2),$$

where $\delta_i := d_i/(d_1, d_2)$ for $i = 1, 2$.

Proof. As in the proof of Proposition 4.6, we may apply Lemma 4.4 to reduce to the case where $[d_1, d_2] = p$ is a prime number. The bound is then trivial if (c_1, δ_1), (c_2, δ_2), or (d_1, d_2) is equal to p, so we may assume without loss of generality that $d_1 = p$, $d_2 = 1$, and that c_1 is coprime to p. We then need to prove that

$$\sum_{n \in \mathbb{Z}/p\mathbb{Z}} e_p\left(\frac{c_1}{n + l}\right) \ll 1,$$

but this is clear since after the change of variable $m = c_1/(n + l)$ this sum is just a Ramanujan sum. $\qquad\square$

4D. *Incomplete exponential sums.* The bounds in the previous section control "complete" additive exponential sums in one variable in $\mathbb{Z}/q\mathbb{Z}$, by which we mean sums where the variable n ranges over all of $\mathbb{Z}/q\mathbb{Z}$. For our applications, as well as for many others, one needs also to have good estimates for "incomplete" versions of the sums, in which the variable n ranges over an interval, or more generally over the integers weighted by a coefficient sequence which is (shifted) smooth at some scale N.

The most basic technique to obtain such estimates is the method of completion of sums, also called the Pólya–Vinogradov method. In essence, this is an elementary application of discrete Fourier analysis, but the importance of the results cannot be overestimated.

We begin with some facts about the discrete Fourier transform. Given a function

$$f : \mathbb{Z}/q\mathbb{Z} \to \mathbb{C},$$

we define its *normalized Fourier transform* $\mathrm{FT}_q(f)$ to be the function on $\mathbb{Z}/q\mathbb{Z}$ given by

$$\mathrm{FT}_q(f)(h) := \frac{1}{q^{1/2}} \sum_{x \in \mathbb{Z}/q\mathbb{Z}} f(x) e_q(hx). \tag{4-11}$$

The normalization factor $1/q^{1/2}$ is convenient because the resulting Fourier transform operator is then unitary with respect to the inner product

$$\langle f, g \rangle := \sum_{x \in \mathbb{Z}/q\mathbb{Z}} f(x)\overline{g(x)}$$

on the space of functions $\mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$. In other words, the Plancherel formula

$$\sum_{x \in \mathbb{Z}/q\mathbb{Z}} f(x)\overline{g(x)} = \sum_{h \in \mathbb{Z}/q\mathbb{Z}} \mathrm{FT}_q(f)(h)\overline{\mathrm{FT}_q(g)(h)}$$

holds for any functions $f, g : \mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$. Furthermore, by the orthogonality of additive characters, we have the discrete Fourier inversion formula

$$\mathrm{FT}_q(\mathrm{FT}_q(f))(x) = f(-x)$$

for all $x \in \mathbb{Z}/q\mathbb{Z}$.

Lemma 4.9 (completion of sums). *Let $M \geq 1$ be a real number and let ψ_M be a function on $\mathbb{R}$ defined by*

$$\psi_M(x) = \psi\left(\frac{x - x_0}{M}\right),$$

where $x_0 \in \mathbb{R}$ and ψ is a smooth function supported on $[c, C]$ satisfying

$$|\psi^{(j)}(x)| \ll \log^{O(1)} M$$

for all fixed $j \geq 0$, where the implied constant may depend on j. Let $q \geq 1$ be an integer, and let

$$M' := \sum_{m \geq 1} \psi_M(m) \ll M(\log M)^{O(1)}.$$

We have:

  (i) *If $f : \mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$ is a function, then*

$$\left|\sum_m \psi_M(m) f(m) - \frac{M'}{q} \sum_{m \in \mathbb{Z}/q\mathbb{Z}} f(m)\right|$$
$$\ll q^{1/2}(\log M)^{O(1)} \sup_{h \in \mathbb{Z}/q\mathbb{Z}\setminus\{0\}} |\mathrm{FT}_q(f)(h)|. \quad (4\text{-}12)$$

  In particular, if $M \ll q(\log M)^{O(1)}$, then

$$\left|\sum_m \psi_M(m) f(m)\right| \ll q^{1/2}(\log M)^{O(1)} \|\mathrm{FT}_q(f)\|_{\ell^\infty(\mathbb{Z}/q\mathbb{Z})}. \quad (4\text{-}13)$$

We also have the variant

$$\left| \sum_m \psi_M(m) f(m) - \frac{M'}{q} \sum_{m \in \mathbb{Z}/q\mathbb{Z}} f(m) \right|$$

$$\ll (\log M)^{O(1)} \frac{M}{q^{1/2}} \sum_{0 < |h| \le qM^{-1+\varepsilon}} |\mathrm{FT}_q(f)(h)| + M^{-A} \sum_{m \in \mathbb{Z}/q\mathbb{Z}} |f(m)| \quad (4\text{-}14)$$

for any fixed $A > 0$ and $\varepsilon > 0$, where the implied constant depends on ε and A.

(ii) *If I is a finite index set, and for each $i \in I$, c_i is a complex number and a_i (q) is a residue class, then for each fixed $A > 0$ and $\varepsilon > 0$, one has*

$$\left| \sum_{i \in I} c_i \sum_m \psi_M(m) \mathbf{1}_{m = a_i \ (q)} - \frac{M'}{q} \sum_{i \in I} c_i \right|$$

$$\ll (\log M)^{O(1)} \frac{M}{q} \sum_{0 < |h| \le qM^{-1+\varepsilon}} \left| \sum_{i \in I} c_i e_q(a_i h) \right| + M^{-A} \sum_{i \in I} |c_i|, \quad (4\text{-}15)$$

where the implied constant depends on ε and A.

Remark 4.10. One could relax the derivative bounds on ψ to $|\psi^{(j)}(x)| \ll M^{\varepsilon_j}$ for various small fixed $\varepsilon_j > 0$, at the cost of similarly worsening the various powers of $\log M$ in the conclusion of the lemma to small powers of M, and assuming the ε_j small enough depending on ε and A; however this variant of the lemma is a little tricky to state, and we will not have use for it here.

Proof. Define the function

$$\psi_{M,q}(x) = \sum_{n \in \mathbb{Z}} \psi_M(x + qn).$$

This is a smooth q-periodic function on $\mathbb{R}$. By periodization and the Plancherel formula, we have

$$\sum_m \psi_M(m) f(m) = \sum_{x \in \mathbb{Z}/q\mathbb{Z}} f(x) \psi_{M,q}(x)$$

$$= \sum_{h \in \mathbb{Z}/q\mathbb{Z}} \mathrm{FT}_q(f)(h) \mathrm{FT}_q(\psi_{M,q})(-h). \quad (4\text{-}16)$$

The contribution of the frequency $h = 0$ is given by

$$\mathrm{FT}_q(f)(0) \mathrm{FT}_q(\psi_{M,q})(0) = \frac{1}{q} \sum_{m \in \mathbb{Z}/q\mathbb{Z}} f(m) \sum_{m \in \mathbb{Z}/q\mathbb{Z}} \psi_{M,q}(m) = \frac{M'}{q} \sum_{m \in \mathbb{Z}/q\mathbb{Z}} f(m).$$

We now consider the contribution of the nonzero frequencies. For $h \in \mathbb{Z}/q\mathbb{Z}$, the definition of $\psi_{M,q}$ leads to

$$q^{1/2} \mathrm{FT}_q(\psi_{M,q})(-h) = \Psi\left(\frac{h}{q}\right),$$

where the function Ψ is defined on $\mathbb{R}/\mathbb{Z}$ by

$$\Psi(y) := \sum_m \psi_M(m)e(-my).$$

This is a smooth function $\Psi : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$. We then have

$$\left| \sum_{h \in \mathbb{Z}/q\mathbb{Z} \setminus \{0\}} \mathrm{FT}_q(f)(h)\mathrm{FT}_q(\psi_{M,q})(-h) \right|$$
$$\leq \sup_{h \in \mathbb{Z}/q\mathbb{Z} \setminus \{0\}} |\mathrm{FT}_q(f)(h)| q^{-1/2} \sum_{\substack{-q/2 < h \leq q/2 \\ h \neq 0}} \left| \Psi\left(\frac{h}{q}\right) \right|.$$

Applying the Poisson summation formula and the definition $\psi_M(x) = \psi((x - x_0)/M)$, we have

$$\Psi(y) = M \sum_{n \in \mathbb{Z}} \hat{\psi}(M(n + y))e(-(n + y)x_0),$$

where

$$\hat{\psi}(s) = \int_{\mathbb{R}} \psi(t)e(-st)\,dt.$$

By repeated integrations by parts, the assumption on the size of the derivatives of ψ gives the bounds

$$|\hat{\psi}(s)| \ll (\log M)^{O(1)}(1 + |s|)^{-A}$$

for any fixed $A \geq 0$, and therefore

$$|\Psi(y)| \ll M(\log M)^{O(1)}(1 + |y|M)^{-A} \tag{4-17}$$

for any fixed $A \geq 0$ and any $-\frac{1}{2} < y \leq \frac{1}{2}$. Taking, e.g., $A = 2$, we get

$$\sum_{\substack{-q/2 < h \leq q/2 \\ h \neq 0}} \left| \Psi\left(\frac{h}{q}\right) \right| \ll (\log M)^{O(1)} \sum_{1 \leq h \leq q/2} \frac{M}{(1 + |h|M/q)^2} \ll q(\log M)^{O(1)},$$

and therefore we obtain (4-12). From this, (4-13) follows immediately.

We now turn to (4-14). Fix $A > 0$ and $\varepsilon > 0$. Arguing as above, we have

$$\left| \sum_m \psi_M(m) f(m) - \frac{M'}{q} \sum_{m \in \mathbb{Z}/q\mathbb{Z}} f(m) \right|$$

$$\leq \frac{1}{q^{1/2}} \sum_{\substack{-q/2 < h \leq q/2 \\ h \neq 0}} \left| \Psi\left(\frac{h}{q}\right) \right| |\mathrm{FT}_q(f)(h)|$$

$$\ll (\log M)^{O(1)} \frac{M}{q^{1/2}} \sum_{0 < |h| \leq qM^{-1+\varepsilon}} |\mathrm{FT}_q(f)(h)|$$

$$+ (\log M)^{O(1)} \sum_{n \in \mathbb{Z}/q\mathbb{Z}} |f(n)| \sum_{|h| > qM^{-1+\varepsilon}} \frac{M}{q(1 + |h|M/q)^A}.$$

Changing A to a large value, we conclude that

$$\left| \sum_m \psi_M(m) f(m) - \frac{M'}{q} \sum_{m \in \mathbb{Z}/q\mathbb{Z}} f(m) \right|$$

$$\ll Mq^{-1/2}(\log M)^{O(1)} \sum_{0 < |h| \leq qM^{-1+\varepsilon}} |\mathrm{FT}_q(f)(h)| + M^{-A} \sum_{n \in \mathbb{Z}/q\mathbb{Z}} |f(n)|,$$

as claimed.

Finally, claim (ii) follows immediately from (4-14) by setting

$$f(m) := \sum_{\substack{i \in I \\ a_i = m \, (q)}} c_i, \quad \text{so that} \quad \mathrm{FT}_q(f)(h) = \frac{1}{\sqrt{q}} \sum_{i \in I} c_i e_q(a_i h). \qquad \square$$

Remark 4.11. In Section 7, we will use a slightly refined version, where the coefficients $\Psi(h/q)$ above are not estimated trivially.

By combining this lemma with Proposition 4.6, we can obtain nontrivial bounds for incomplete exponential sums of the form

$$\sum_n \psi_N(n) e_q(f(n))$$

for various moduli q, which are roughly of the shape

$$\sum_n \psi_N(n) e_q(f(n)) \ll q^{1/2+\varepsilon}$$

when $N \ll q$. A number of bounds of this type were used by Zhang [2014] to obtain his Type I and Type II estimates. However, it turns out that we can improve this bound for certain regimes of q, N when the modulus q is smooth, or at least densely divisible, by using the "q-van der Corput A-process" of [Heath-Brown 1978] and

[Graham and Ringrose 1990]. This method was introduced to handle incomplete *multiplicative* character sums, but it is also applicable to incomplete additive character sums. It turns out that these improved estimates lead to significant improvements in the Type I and Type II numerology over that obtained in [Zhang 2014].

Here is the basic estimate on incomplete one-dimensional exponential sums that we will need for the Type I and Type II estimates. Essentially the same bounds were obtained in [Heath-Brown 2001, Theorem 2].

Proposition 4.12 (incomplete additive character sums). *Let q be a squarefree integer, and let $f = P/Q \in \mathbb{Q}(X)$ with $P, Q \in \mathbb{Z}[X]$, such that the degree of Q (p) is equal to $\deg(Q)$ for all $p \mid q$. Assume that $\deg(P) < \deg(Q)$. Set $q_1 := q/(f, q)$. Let further $N \geq 1$ be given with $N \ll q^{O(1)}$ and let ψ_N be a function on $\mathbb{R}$ defined by*

$$\psi_N(x) = \psi\left(\frac{x - x_0}{N}\right),$$

where $x_0 \in \mathbb{R}$ and ψ is a smooth function with compact support satisfying

$$|\psi^{(j)}(x)| \ll \log^{O(1)} N$$

for all fixed $j \geq 0$, where the implied constant may depend on j.

(i) *(Polyá–Vinogradov + Ramanujan–Weil) We have the bound*

$$\sum_n \psi_N(n) e_q(f(n)) \ll q^\varepsilon \left(q_1^{1/2} + \frac{N}{q_1} \mathbf{1}_{N \geq q_1} \left| \sum_{n \in \mathbb{Z}/q_1\mathbb{Z}} e_{q_1}(f(n)/(f, q)) \right| \right) \quad (4\text{-}18)$$

for any $\varepsilon > 0$. In particular, lifting the $\mathbb{Z}/q_1\mathbb{Z}$ sum to a $\mathbb{Z}/q\mathbb{Z}$ sum, we have

$$\sum_n \psi_N(n) e_q(f(n)) \ll q^\varepsilon \left(q^{1/2} + \frac{N}{q} \left| \sum_{n \in \mathbb{Z}/q\mathbb{Z}} e_q(f(n)) \right| \right). \quad (4\text{-}19)$$

(ii) *(one van der Corput + Ramanujan–Weil) If $q = rs$, then we have the additional bound*

$$\sum_n \psi_N(n) e_q(f(n))$$

$$\ll q^\varepsilon \left((N^{1/2} r_1^{1/2} + N^{1/2} s_1^{1/4}) + \frac{N}{q_1} \mathbf{1}_{N \geq q_1} \left| \sum_{n \in \mathbb{Z}/q_1\mathbb{Z}} e_{q_1}(f(n)/(f, q)) \right| \right) \quad (4\text{-}20)$$

for any $\varepsilon > 0$, where $r_1 := (r, q_1)$ and $s_1 := (s, q_1)$. In particular, we have

$$\sum_n \psi_N(n) e_q(f(n)) \ll q^\varepsilon \left((N^{1/2} r^{1/2} + N^{1/2} s^{1/4}) + \frac{N}{q} \left| \sum_{n \in \mathbb{Z}/q\mathbb{Z}} e_q(f(n)) \right| \right). \quad (4\text{-}21)$$

In all cases, the implied constants depend on ε, $\deg(P)$, $\deg(Q)$ and the implied constants in the estimates for the derivatives of ψ.

Remark 4.13. The estimates obtained by completion of sums are usually inefficient in the regime $M = o(q)$, and they become trivial for $M \ll q^{1/2}$. For instance, when f is bounded in magnitude by 1, the trivial bound for the right-hand side of (4-13) is q, whereas the trivial bound for the left-hand side is of size about M, which means that one needs a cancellation at least by a factor q/M in the right-hand side to even recover the trivial bound. This becomes a prohibitive restriction if this factor is larger than $\sqrt{M}$. In this paper, this inefficiency is a major source of loss in our final exponents (the other main source being our frequent reliance on the Cauchy–Schwarz inequality, as each invocation of this inequality tends to halve all gains in exponents arising from application of the Riemann hypothesis over finite fields). It would thus be of considerable interest to find stronger estimates for incomplete exponential sums. But the only different (general) method we are aware of is the recent "sliding sum method" of Fouvry, Kowalski and Michel [Fouvry et al. 2013c], which however only improves on the completion technique when M is very close to $q^{1/2}$, and does not give stronger bounds than Lemma 4.9 and Proposition 4.12 in most ranges of interest. (Note however that uniformity of estimates is often even more crucial to obtaining good results, and for this purpose, the completion techniques are indeed quite efficient.)

Proof. We begin with some technical reductions. First of all, we may assume that q has no prime factor smaller than any fixed B depending on $\deg(P)$ and $\deg(Q)$, as the general case then follows by factoring out a bounded factor from q and splitting the summation over n into a bounded number of pieces.

Second, we also observe that, in all cases, we may replace f by $f/(f, q)$ and q by q_1 and (in the case when $q = rs$) r by r_1 and s by s_1, since if we write $q = q_1 q_2$ we have

$$e_q(f(n)) = e_{q_1}\left(\frac{P(n)}{q_2 Q(n)}\right).$$

Thus we can reduce to a situation where $(f, q) = 1$, so $q = q_1$, $r = r_1$ and $s = s_1$. In this case, the condition $\deg(P) < \deg(Q)$ implies also that $(f', q) = (f'', q) = 1$ by Lemma 4.5(ii), provided q has no prime factor less than some constant depending on $\deg(P)$ and $\deg(Q)$, which we may assume to be the case, as we have seen.

We now establish (4-18). We apply (4-14), and put the "main term" with $h = 0$ in the right-hand side, to get

$$\sum_n \psi_N(n) e_q(f(n)) \ll \frac{N^{1+\varepsilon}}{q} \sum_{|h| \leq q N^{-1+\varepsilon}} \left| \sum_{n \in \mathbb{Z}/q\mathbb{Z}} e_q(f(n) + hn) \right| + 1$$

for $\varepsilon > 0$ arbitrarily small (by selecting A large enough in (4-14) using the assumption $N \ll q^{O(1)}$).

If $N < q$, Proposition 4.6 applied for all h gives

$$\sum_n \psi_N(n)e_q(f(n)) \ll \frac{N^{1+\varepsilon}}{q^{1/2}} \sum_{0 \le |h| \le qN^{-1+\varepsilon}} (f' + h, q).$$

Since $(f'', q) = 1$, we also have $(f' + h, q) = 1$, and therefore

$$\sum_n \psi_N(n)e_q(f(n)) \ll q^{1/2}N^{2\varepsilon},$$

which implies (4-18). If $N \ge q$, on the other hand, we only apply Proposition 4.6 for $h \ne 0$, and we get in the same way

$$\sum_n \psi_N(n)e_q(f(n)) \ll \frac{N^{1+\varepsilon}}{q} \left| \sum_{n \in \mathbb{Z}/q\mathbb{Z}} e_q(f(n)) \right| + q^{1/2}N^{2\varepsilon},$$

which is again (4-18).

Consider now (4-20). We may assume that $N \le s$, since otherwise the claim follows simply from (4-18), and we may similarly assume that $r \le N$, since otherwise we can use the trivial bound

$$\sum_n \psi_N(n)e_q(f(n)) \ll N(\log N)^{O(1)} \ll r^{1/2}N^{1/2}(\log N)^{O(1)}.$$

Let $K := \lfloor N/r \rfloor$. Using translation invariance, we can write

$$\sum_n \psi_N(n)e_q(f(n)) = \frac{1}{K}\sum_n \sum_{k=1}^{K} \psi_N(n + kr)e_q(f(n + kr)).$$

Since $q = rs$, we have

$$e_q(f(n + kr)) = e_r(\bar{s}f(n))e_s(\bar{r}f(n + kr))$$

by Lemma 4.4 (and periodicity), and hence we obtain

$$\left| \sum_n \psi_N(n)e_q(f(n)) \right| \le \frac{1}{K}\sum_n \left| \sum_{k=1}^{K} \psi_N(n + kr)e_s(\bar{r}f(n + kr)) \right|$$

$$\ll \frac{N^{1/2}}{K}\left(\sum_n \left| \sum_{k=1}^{K} \psi_N(n + kr)e_s(\bar{r}f(n + kr)) \right|^2 \right)^{1/2},$$

where the factor $N^{1/2}$ arises because the summand is (as a function of n) supported on an interval of length $O(N)$. Expanding the square, we obtain

$$\left| \sum_n \psi_N(n)e_q(f(n)) \right|^2 \ll \frac{N}{K^2}\sum_{1 \le k,l \le K} A(k, l), \qquad (4\text{-}22)$$

where

$$A(k, l) = \sum_n \psi_N(n + kr)\overline{\psi_N(n + lr)}\, e_s\big(\overline{r}(f(n + kr) - f(n + lr))\big).$$

We have

$$A(k, k) = \sum_n |\psi_N(n + kr)|^2 \ll N(\log N)^{O(1)}.$$

and therefore

$$\sum_{1 \le k \le K} |A(k, k)| \ll KN(\log N)^{O(1)}. \tag{4-23}$$

It remains to handle the off-diagonal terms. For each $k \ne l$, we have

$$\frac{f(n + kr) - f(n + lr)}{r} = g(n),$$

where $g = P_1/Q_1 \in \mathbb{Q}(X)$ with integral polynomials

$$P_1(X) = P(X + kr)Q(X + lr) - Q(X + kr)P(X + lr),$$
$$Q_1(X) = rQ(X + kr)Q(X + lr).$$

Note that P_1 and Q_1 satisfy the assumptions of (4-18) with respect to the modulus s (although they might not be coprime).

We now claim that (provided all prime factors of q are large enough) we have

$$(s, g') \mid (s, k - l) \quad \text{and} \quad (s, g) \mid (s, k - l).$$

Indeed, since $\deg(P) < \deg(Q)$ and the degree of the reduction of Q modulo primes dividing q is constant, it is enough to show that $(s, g) \mid (s, k - l)$ by Lemma 4.5(ii). So suppose that a prime p divides (s, g). Then, by a change of variable, we have

$$p \mid \big(s, f(X + (k - l)r) - f(X)\big).$$

By induction, we thus have

$$p \mid \big(s, f(X + i(k - l)r) - f(X)\big)$$

for any integer i. If $p \nmid k - l$, then $(k - l)r$ generates $\mathbb{Z}/p\mathbb{Z}$ as an additive group, and we conclude that $p \mid (s, f(X + a) - f(X))$ for all $a \in \mathbb{Z}/p\mathbb{Z}$. This implies that $f(p)$ is constant where it is defined. But since $\deg(P) < \deg(Q)$ holds modulo p, for p large enough in terms of $\deg(Q)$, this would imply that $p \mid f$ (as in Lemma 4.5(ii)), contradicting the assumption $(s, f) = 1$. Thus we have $p \mid k - l$, and we conclude that $(s, g) \mid (s, k - l)$, and then $(s, g') \mid (s, k - l)$, as claimed.

By (4-18) and Proposition 4.6, we have

$$A(k, l) \ll q^\varepsilon \left(s^{1/2} + \frac{N}{s} \mathbf{1}_{N \geq s/(s,k-l)} \left| \sum_{n \in \mathbb{Z}/s\mathbb{Z}} e_s(g(n)) \right| \right)$$

$$\ll q^\varepsilon \left(s^{1/2} + \frac{N}{s^{1/2}} (s, k-l)^{1/2} \mathbf{1}_{N \geq s/(s,k-l)} \right).$$

Summing over k and l, we have

$$\sum_{1 \leq k \neq l \leq K} |A(k, l)| \ll q^\varepsilon K^2 s^{1/2} + q^\varepsilon N s^{-1/2} \sum_{1 \leq k \neq l \leq K} (s, k-l)^{1/2} \mathbf{1}_{N \geq s/(s,k-l)}. \quad (4\text{-}24)$$

We use the simple bound

$$\mathbf{1}_{N \geq s/(s,k-l)} \leq \sqrt{(s, k-l)} \sqrt{\frac{N}{s}}$$

to estimate the last sum as follows:

$$N s^{-1/2} \sum_{1 \leq k \neq l \leq K} (s, k-l)^{1/2} \mathbf{1}_{N \geq s/(s,k-l)} \leq \frac{N^{3/2}}{s} \sum_{1 \leq k \neq l \leq K} (s, k-l)$$

$$\ll N^{3/2} s^{-1} \times K^2 q^\varepsilon \ll K^2 s^{1/2} q^\varepsilon,$$

using Lemma 1.4 and the bound $N < s$. We combine this with (4-23) and (4-24) in the bound (4-22) to obtain

$$\left| \sum_n \psi_N(n) e_q(f(n)) \right|^2 \ll q^\varepsilon \frac{N}{K^2} (KN(\log N)^{O(1)} + K^2 s^{1/2}) \ll q^\varepsilon (Nr + N s^{1/2}),$$

from which (4-20) follows. □

Remark 4.14. (1) Assuming that $(f, q) = 1$, the first bound (4-18) is nontrivial (i.e., better than $O(N)$) as long as N is a bit larger than $q^{1/2}$. As for (4-20), we see that in the regime where the factorization $q = rs$ satisfies $r \approx q^{1/3} \approx s^{1/2}$, the bound is nontrivial in the significantly wider range where N is a bit larger than $q^{1/3}$.

(2) The procedure can also be generalized with similar results to more general q-periodic functions than $n \mapsto e_q(f(n))$, and this will be important for the most advanced Type I estimates (see Section 6J.1).

Remark 4.15. One can iterate the above argument and show that

$$\left| \sum_n \psi_N(n) e_q(f(n)) \right|$$

$$\ll q^\varepsilon \left(\sum_{i=1}^{l-1} N^{1-1/2^i} \tilde{r}_i^{1/2^i} + N^{1-1/2^{l-1}} \tilde{r}_l^{1/2^l} + \frac{N}{q_1} \mathbf{1}_{N \geq q_1} \left| \sum_{n \in \mathbb{Z}/q_1\mathbb{Z}} e_{q_1}(f(n)/(f,q)) \right| \right)$$

for any fixed $l \geq 1$ and any factorization $q = r_1 \cdots r_l$ with $\tilde{r}_i = (r_i, q_1)$; see [Graham and Ringrose 1990; Heath-Brown 2001]. However, we have found in practice that taking l to be 3 or higher (corresponding to two or more applications of the q-van der Corput A-process) ends up being counterproductive, mainly because the power of q that one can save over the trivial bound decays exponentially in l. However, it is possible that some other variation of the arguments (for instance, taking advantage of the Parseval identity, which would be a q-analogue of the van der Corput B-process) may give further improvements.

In our particular application, we only need a special case of Proposition 6.20. This is a strengthening of [Zhang 2014, Lemma 11], and it shows how an assumption of dense divisibility of a modulus may be exploited in estimates for exponential sums.

Corollary 4.16. *Let $N \geq 1$ and let ψ_N be a function on $\mathbb{R}$ defined by*

$$\psi_N(x) = \psi\left(\frac{x - x_0}{N}\right),$$

where $x_0 \in \mathbb{R}$ and ψ is a smooth function with compact support satisfying

$$|\psi^{(j)}(x)| \ll \log^{O(1)} N$$

for all fixed $j \geq 0$, where the implied constant may depend on j.

Let d_1, d_2 be squarefree integers, not necessarily coprime. Let c_1, c_2, l_1, l_2 be integers. Let $y \geq 1$ be a real number, and suppose that $[d_1, d_2]$ is y-densely divisible. Let d be a divisor of $[d_1, d_2]$ and let $a\ (d)$ be any residue class.

If $N \leq [d_1, d_2]^{O(1)}$, then we have

$$\left| \sum_{n=a\ (d)} \psi_N(n) e_{d_1}\left(\frac{c_1}{n+l_1}\right) e_{d_2}\left(\frac{c_2}{n+l_2}\right) \right|$$

$$\ll [d_1, d_2]^\varepsilon \left(d^{-1/2} N^{1/2} [d_1, d_2]^{1/6} y^{1/6} + d^{-1} \frac{(c_1, \delta'_1)}{\delta'_1} \frac{(c_2, \delta'_2)}{\delta'_2} N \right)$$

for any $\varepsilon > 0$, where $\delta_i := d_i / (d_1, d_2)$ and $\delta'_i := \delta_i / (d, \delta_i)$ for $i = 1, 2$. We also have the variant bound

$$\left| \sum_{n=a\ (d)} \psi_N(n) e_{d_1}\left(\frac{c_1}{n+l_1}\right) e_{d_2}\left(\frac{c_2}{n+l_2}\right) \right|$$

$$\ll [d_1, d_2]^\varepsilon \left(d^{-1/2} [d_1, d_2]^{1/2} + d^{-1} \frac{(c_1, \delta'_1)}{\delta'_1} \frac{(c_2, \delta'_2)}{\delta'_2} N \right).$$

In both cases the implied constant depends on ε.

Proof. Set $q = [d_1, d_2]$. We first consider the case $d = 1$, so that the congruence condition $n = a \ (d)$ is vacuous. Since $R = y^{1/3} q^{1/3} \leq yq$, the dense divisibility hypothesis implies that there exists a factorization $q = rs$ for some integers r, s such that

$$y^{-2/3} q^{1/3} \leq r \leq y^{1/3} q^{1/3}$$

and

$$y^{-1/3} q^{2/3} \leq s \leq y^{2/3} q^{2/3}.$$

Note now that, by the Chinese remainder theorem (as in Lemma 4.4), we can write

$$e_{d_1}\left(\frac{c_1}{n + l_1}\right) e_{d_2}\left(\frac{c_2}{n + l_2}\right) = e_q(f(n))$$

for a rational function $f = P/Q \in \mathbb{Q}(X)$ satisfying the assumptions of Proposition 4.12 (in particular $\deg(P) < \deg(Q)$). The first bound follows immediately from Proposition 4.12(ii), combined with the complete sum estimate

$$\left| \sum_{n \in \mathbb{Z}/[d_1, d_2]\mathbb{Z}} e_{d_1}\left(\frac{c_1}{n + l_1}\right) e_{d_2}\left(\frac{c_2}{n + l_2}\right) \right| \ll q^\varepsilon (c_1, \delta_1)(c_2, \delta_2)(d_1, d_2)$$

of Lemma 4.8. The second bound similarly follows from Proposition 4.12(i).

Now we consider the case when $d > 1$. Making the substitution $n = n'd + a$ and applying the previous argument (with N replaced by N/d, and with suitable modifications to x_0 and f), we reduce to showing that

$$\left| \sum_{\substack{n \in \mathbb{Z}/[d_1, d_2]\mathbb{Z} \\ n = a \ (d)}} e_{d_1}\left(\frac{c_1}{n + l_1}\right) e_{d_2}\left(\frac{c_2}{n + l_2}\right) \right| \ll q^\varepsilon (c_1, \delta_1')(c_2, \delta_2')(d_1', d_2'),$$

where $d_i' := d_i/(d, d_i)$ for $i = 1, 2$ (note that $d(d_1', d_2')/[d_1, d_2] = 1/(\delta_1' \delta_2')$). However, this again follows from Lemma 4.8 after making the change of variables $n = n'd + a$. $\square$

5. Type I and Type II estimates

Using the estimates of the previous section, we can now prove the Type I and Type II results of Theorem 2.8, with the exception of part (iii) of that theorem, for which we only make a preliminary reduction for now. The rest of the proof of that part, which depends on the concepts and results of Section 6, will be found in Section 8.

We recall the statements (see Definition 2.6):

Theorem 5.1 (new Type I and Type II estimates). *Let $\varpi, \delta, \sigma > 0$ be fixed quantities, let I be a bounded subset of $\mathbb{R}$, let $i \geq 1$ be fixed, let a (P_I) be a primitive congruence class, and let $M, N \gg 1$ be quantities with*

$$MN \asymp x \tag{5-1}$$

and

$$x^{1/2-\sigma} \ll N \ll x^{1/2}. \tag{5-2}$$

Let α, β be coefficient sequences located at scales M, N respectively, with β satisfying the Siegel–Walfisz property. Then we have the estimate

$$\sum_{\substack{d \in \mathscr{D}_I^{(i)}(x^\delta) \\ d \ll x^{1/2+2\varpi}}} |\Delta(\alpha \star \beta; a \, (d))| \ll x \log^{-A} x \tag{5-3}$$

for any fixed $A > 0$, provided that one of the following hypotheses holds:

   (i) $i = 1$, $54\varpi + 15\delta + 5\sigma < 1$, *and* $N \ll x^{1/2-2\varpi-c}$ *for some fixed $c > 0$.*

  (ii) $i = 2$, $56\varpi + 16\delta + 4\sigma < 1$, *and* $N \ll x^{1/2-2\varpi-c}$ *for some fixed $c > 0$.*

 (iii) $i = 4$, $\frac{160}{3}\varpi + 16\delta + \frac{34}{9}\sigma < 1$, $64\varpi + 18\delta + 2\sigma < 1$, *and* $N \ll x^{1/2-2\varpi-c}$ *for some fixed $c > 0$.*

 (iv) $i = 1$, $68\varpi + 14\delta < 1$, *and* $N \gg x^{1/2-2\varpi-c}$ *for some sufficiently small fixed $c > 0$.*

The proof of case (iii) *uses the general form of the Riemann hypothesis over finite fields* [Deligne 1980], *but the proofs of* (i), (ii), (iv) *only need the Riemann hypothesis for curves over finite fields.*

Before we begin the rigorous proof of Theorem 5.1, we give an informal sketch of our strategy of proof for these estimates, which is closely modeled on the arguments of [Zhang 2014]. The basic idea is to reduce the estimate (5-3) to a certain exponential sum estimate, of the type found in Corollary 4.16 (and, for the estimate (iii), in Corollary 6.24 of the next section). The main tools for these reductions are completion of sums (Lemma 4.9), the triangle inequality, and many techniques related to the Cauchy–Schwarz inequality (viewed in a broad sense), for instance, Vinogradov's bilinear form method, the q-van der Corput A-process, the method of Weyl differencing, and the dispersion method of Linnik.

5A. *Bilinear form estimates.* We begin with a short discussion of typical instances of applications of the Cauchy–Schwarz inequality (some examples already appeared in previous sections). We want to estimate a sum

$$\sum_{s \in S} c_s$$

of (typically) complex numbers c_s indexed by some finite set S of large size. Suppose we can parametrize S (possibly with repetition) by a *nontrivial* product set $A \times B$, i.e., by a product where neither factor is too small, or otherwise prove an inequality

$$\left| \sum_{s \in S} c_s \right| \leq \left| \sum_{a \in A} \sum_{b \in B} \alpha_a \beta_b k_{a,b} \right|$$

for certain coefficients α_a, β_b and $k_{a,b}$. The crucial insight is that one can often derive nontrivial estimates for an expression of this type with little knowledge of the coefficients α_a, β_b by exploiting the bilinear structure and studying the coefficients $k_{a,b}$.

Precisely, one can apply the Cauchy–Schwarz inequality to bound the right side by

$$\left(\sum_{a \in A} |\alpha_a|^2 \right)^{1/2} \left(\sum_{a \in A} \left| \sum_{b \in B} \beta_b k_{a,b} \right|^2 \right)^{1/2}.$$

The first factor in the above expression is usually easy to estimate, and the second factor can be expanded as

$$\left| \sum_{b,b' \in B} \beta_b \overline{\beta_{b'}} C(b,b') \right|^{1/2}, \qquad C(b,b') = \sum_{a \in A} k_{a,b} \overline{k_{a,b'}}.$$

One can then distinguish between the *diagonal contribution* defined by $b = b'$ and the *off-diagonal contribution* where $b \neq b'$. The contribution of the former is

$$\sum_{b \in B} \sum_{a \in A} |\beta_b|^2 |k_{a,b}|^2$$

which is (usually) not small, since there cannot be cancellation between these nonnegative terms. It may however be estimated satisfactorily, provided B is large enough for the diagonal $\{(b,b) : b \in B\}$ to be a "small" subset of the square $B \times B$. (In practice, there might be a larger subset of $B \times B$ than the diagonal where the coefficient $C(b,b')$ is not small, and that is then incorporated in the diagonal; in this paper, where b and b' are integers, it is the size of a greatest common divisor $(b - b', q)$ that will dictate which terms can be considered diagonal.)

On the other hand, the individual off-diagonal terms $C(b,b')$ can be expected to exhibit cancellation that makes them individually small. In order for the sum over $b \neq b'$ to remain of manageable size, one needs B to remain not too large. In order to balance the two contributions, it turns out to be extremely useful to have a flexible *family* of parametrizations $(a,b) \mapsto s$ of S by product sets $A \times B$, so that one can find a parametrization for which the set B is close to the optimum size arising from various estimates of the diagonal and nondiagonal parts. This idea of flexibility is a key idea at least since Iwaniec's discovery [1980] of the bilinear form of the error term in the linear sieve.

One of the key ideas in [Zhang 2014] is that if one is summing over smooth moduli, then such a flexible range of factorizations exists; to put it another way, the restriction to smooth moduli is essentially a "well-factorable" weight in the sense of Iwaniec. In this paper, we isolated the key property of smooth moduli needed for such arguments, namely, the property of *dense divisibility*. The general strategy is thus to keep exploiting the smoothness or dense divisibility of the moduli to split the sums over such moduli into a "well-factorable" form to which the Cauchy–Schwarz inequality may be profitably applied. (Such a strategy was already used to optimize the use of the q-van der Corput A-process in Corollary 4.16.)

5B. *Sketch of proofs.* We now give a more detailed, but still very informal, sketch of the proof of Theorem 5.1, omitting some steps and some terms for sake of exposition (e.g., smooth cutoffs are not mentioned). For simplicity we will pretend that the quantities ϖ, δ are negligible, although the quantity σ will still be of a significant size (note from Lemma 2.7 that we will eventually need to take σ to be at least $1/10$). The first step is to exploit the dense divisibility of the modulus d to factor it as $d = qr$, with q, r located at certain scales Q, R which we will specify later; with ϖ negligible, we expect QR to be approximately equal to $x^{1/2}$ but a bit larger. Our task is then to obtain a nontrivial bound on the quantity

$$\sum_{q \asymp Q} \sum_{r \asymp R} |\Delta(\alpha \star \beta; a\,(qr))|,$$

or equivalently to obtain a nontrivial bound on

$$\sum_{q \asymp Q} \sum_{r \asymp R} c_{q,r} \Delta(\alpha \star \beta; a\,(qr))$$

for an arbitrary bounded sequence $c_{q,r}$. We suppress here, and later, some additional information on the moduli q, r, e.g., that they are squarefree and coprime, to simplify this informal exposition. For similar reasons, we are being vague on what a "nontrivial bound" means, but roughly speaking, it should improve upon the "trivial bound" by a factor of $\log^{-A} x$, where A is very large (or arbitrarily large).

If we insert the definition (1-1), and denote generically by EMT the contribution of the second term in that definition (which is the "expected main term"), we see that we need a nontrivial bound on the quantity

$$\sum_{q \asymp Q} \sum_{r \asymp R} c_{q,r} \sum_{n=a\,(qr)} \alpha \star \beta(n) - \text{EMT}.$$

For simplicity, we will handle the r averaging trivially, and thus seek to control the sum

$$\sum_{q \asymp Q} c_{q,r} \sum_{n=a\,(qr)} \alpha \star \beta(n) - \text{EMT}$$

for a single $r \asymp R$. We rearrange this as

$$\sum_{m \asymp M} \alpha(m) \sum_{q \asymp Q} c_{q,r} \sum_{\substack{n \asymp N \\ nm = a \ (qr)}} \beta(n) - \text{EMT}.$$

Note that for fixed m coprime with q, the number of pairs (q, n) with $q \asymp Q$, $n \asymp N$, and $nm = a \ (qr)$ is expected to be about $(QN)/(QR) = N/R$. Thus, if we choose R to be a little bit less than N, e.g., $R = x^{-\varepsilon} N$, then the number of pairs (q, n) associated to a given value of m is expected to be nontrivial. This opens up the possibility of using the dispersion method of [Linnik 1963], as the diagonal contribution in that method is expected to be negligible. Accordingly, we apply Cauchy–Schwarz in the variable m, eliminating the rough coefficient sequence α, and end up with the task of controlling an expression of the shape

$$\sum_{m \asymp M} \left| \sum_{q \asymp Q} c_{q,r} \sum_{\substack{n \asymp N \\ nm = a \ (qr)}} \beta(n) - \text{EMT} \right|^2.$$

Opening the square as sketched above, this is equal to

$$\sum_{q_1, q_2 \asymp Q} c_{q_1,r} \overline{c_{q_2,r}} \sum_{n_1, n_2 \asymp N} \beta(n_1) \overline{\beta(n_2)} \left(\sum_{\substack{m \asymp M \\ n_1 m = a \ (q_1 r) \\ n_2 m = a \ (q_2 r)}} 1 - \text{EMT} \right).$$

Note that, since $a \ (qr)$ is a primitive residue class, the constraints $n_1 m = a \ (q_1 r)$ and $n_2 m = a \ (q_2 r)$ imply $n_1 = n_2 \ (r)$. Thus we can write $n_2 = n_1 + \ell r$ for some $\ell = O(N/R)$, which will be rather small (compare with the method of Weyl differencing).

For simplicity, we consider only[3] the case $\ell = 0$ here. We are thus led to the task of controlling sums such as

$$\sum_{q_1, q_2 \asymp Q} c_{q_1,r} \overline{c_{q_2,r}} \sum_{n \asymp N} \beta(n) \overline{\beta(n)} \left(\sum_{\substack{m \asymp M \\ nm = a \ (q_1 r) \\ nm = a \ (q_2 r)}} 1 - \text{EMT} \right). \qquad (5\text{-}4)$$

It turns out (using a technical trick of Zhang which we will describe below) that we can ensure that the moduli q_1, q_2 appearing here are usually coprime, in the sense that the contribution of the noncoprime pairs q_1, q_2 are negligible. Assuming this, we can use the Chinese remainder theorem to combine the two constraints $nm = a \ (q_1 r)$, $nm = a \ (q_2 r)$ into a single constraint $nm = a \ (q_1 q_2 r)$ on m. Now, we

---

[3]Actually, for technical reasons, in the rigorous argument we will dispose of the $\ell = 0$ contribution by a different method, so the discussion here should be viewed as an oversimplification.

note that if R is slightly less than N, then (since MN is close to x, and QR is close to $x^{1/2}$) the modulus $q_1 q_2 r$ is comparable to M. This means that the inner sum

$$\sum_{\substack{m \asymp M \\ nm = a \, (q_1 q_2 r)}} 1 - \mathrm{EMT}$$

is essentially a complete sum, and can therefore be very efficiently handled by Lemma 4.9. This transforms (5-4) into expressions such as

$$\sum_{0 < |h| \leq H} c_h \sum_{q_1, q_2 \asymp Q} c_{q_1, r} \overline{c_{q_2, r}} \sum_{n \asymp N} \beta(n) \overline{\beta(n)} e_{q_1 q_2 r} \left(\frac{ah}{n} \right),$$

where $H \approx Q^2 R / M$ is a fairly small quantity and the coefficients c_h are bounded. At this point, the contribution of the zero frequency $h = 0$ has canceled out with the expected main term EMT (up to negligible error).

This expression involves the essentially unknown (but bounded) coefficients $c_{q_1, r}$, $c_{q_2, r}$, $\beta(n)$, and, as before, we cannot do much more than eliminate them using the Cauchy–Schwarz inequality. This can be done in several ways here, depending on which variables are taken "outside" of the Cauchy–Schwarz inequality. For instance, if we take n to eliminate the $\beta(n)\overline{\beta(n)}$ term, we are led, after expanding the square and exchanging the sum in the second factor of the Cauchy–Schwarz inequality, to expressions such as

$$\sum_{0 < |h_1|, |h_2| \leq H} \sum \sum_{q_1, q_2, s_1, s_2 \asymp Q} \left| \sum_{n \asymp N} e_{q_1 q_2 r} \left(\frac{ah_1}{n} \right) e_{s_1 s_2 r} \left(-\frac{ah_2}{n} \right) \right|.$$

The sum over n has length N close to the modulus $[q_1 q_2 r, s_1 s_2 r] \approx Q^4 R$, and therefore can be estimated nontrivially using Corollary 4.16. As we will see, this arrangement of the Cauchy–Schwarz inequality is sufficient to establish the Type II estimate (iv).

The Type I estimates are obtained by a slightly different application of Cauchy–Schwarz. Indeed, note for instance that as the parameter σ (which occurs in the Type I condition, but not in Type II) gets larger, the length N in the sum may become smaller in comparison to the modulus $q_1 q_2 s_1 s_2 r$ in the exponential sum

$$\sum_{n \asymp N} e_{q_1 q_2 r} \left(\frac{ah_1}{n} \right) e_{s_1 s_2 r} \left(-\frac{ah_2}{n} \right),$$

and this necessitates more advanced exponential sum estimates to recover nontrivial cancellation. Here, the q-van der Corput A-method enlarges the range of parameters for which we can prove that such a cancellation occurs. This is one of the main reasons why our Type I estimates improve on those in [Zhang 2014]. (The other main reason is that we will adjust the Cauchy–Schwarz inequality to lower the modulus

in the exponential sum to be significantly smaller than $q_1 q_2 s_1 s_2 r \asymp Q^4 R$, while still keeping both the diagonal and off-diagonal components of the Cauchy–Schwarz estimate under control.)

5C. *Reduction to exponential sums.* We now turn to the details of the above strategy. We begin with preliminary manipulations (mostly following [Zhang 2014]) to reduce the estimate (5-3) to a certain exponential sum estimate. This reduction can be done simultaneously in the four cases (i), (ii), (iii), (iv), but the verification of the exponential sum estimate requires a different argument in each of the four cases.

In the remainder of this section $\varpi, \delta, \sigma, I, i, a, M, N, \alpha, \beta$ are as in Theorem 5.1. First of all, since β satisfies the Siegel–Walfisz property, the Bombieri–Vinogradov theorem (Theorem 2.9) implies

$$\sum_{d \leq x^{1/2} \log^{-B} x} |\Delta(\alpha \star \beta; a \ (d))| \ll x \log^{-A} x \qquad (5\text{-}5)$$

for any fixed $A > 0$ and some B depending on A. From this and dyadic decomposition, we conclude that to prove (5-3), it suffices to establish the estimate

$$\sum_{d \in \mathscr{D}_I^{(i)}(x^\delta) \cap [D, 2D]} |\Delta(\alpha \star \beta; a \ (d))| \ll x \log^{-A} x$$

for any fixed $A > 0$ and for all D such that

$$x^{1/2} \ll D \ll x^{1/2+2\varpi} \qquad (5\text{-}6)$$

(recall that this means $x^{1/2} \ll x^{o(1)} D$ and $D \ll x^{1/2+2\varpi+o(1)}$ for any $\varepsilon > 0$).

We now fix one such D. In the spirit of [Zhang 2014], we first restrict d to moduli which do not have too many small prime factors. Precisely, let

$$D_0 := \exp(\log^{1/3} x), \qquad (5\text{-}7)$$

and let $\mathscr{E}(D)$ be the set of $d \in [D, 2D]$ such that

$$\prod_{\substack{p \mid d \\ p \leq D_0}} p > \exp(\log^{2/3} x). \qquad (5\text{-}8)$$

We have (compare [Fouvry 1985, Lemme 4]):

Lemma 5.2. *For any fixed $A > 0$, and D obeying (5-6), we have*

$$|\mathscr{E}(D)| \ll D \log^{-A} x.$$

Proof. If $d \geq 1$ satisfies (5-8), then

$$\prod_{\substack{p \mid d \\ p \leq D_0}} p > \exp(\log^{2/3} x) = D_0^{\log^{1/3} x}.$$

In particular, d has at least $\log^{1/3} x$ prime factors, and therefore

$$\tau(d) \geq 2^{\log^{1/3} x}.$$

On the other hand, we have

$$\sum_{\substack{D \leq d \leq 2D \\ \tau(d) \geq \kappa}} 1 \leq \frac{1}{\kappa} \sum_{D \leq d \leq 2D} \tau(d) \ll \frac{D}{\kappa} \log x$$

for any $\kappa > 0$ by the standard bound

$$\sum_{D \leq d \leq 2D} \tau(d) \ll D \log x$$

(see (1-3)), and the result follows. $\qquad \square$

This allows us to dispose of these exceptional moduli:

Corollary 5.3. *We have*

$$\sum_{\substack{d \in \mathcal{D}_I^{(i)}(x^\delta) \\ d \in \mathcal{E}(D)}} |\Delta(\alpha \star \beta; a \, (d))| \ll x \log^{-A} x$$

for any fixed $A > 0$.

Proof. From (1-4) we derive the trivial bound

$$|\Delta(\alpha \star \beta; a \, (d))| \ll x D^{-1} \tau(d)^{O(1)} \log^{O(1)} x,$$

for every $d \asymp D$, and hence the Cauchy–Schwarz inequality gives

$$\sum_{\substack{d \in \mathcal{D}_I^{(i)}(x^\delta) \\ d \in \mathcal{E}(D)}} |\Delta(\alpha \star \beta; a \, (d))| \ll |\mathcal{E}(D)|^{1/2} x D^{-1} \log^{O(1)} x \left(\sum_{d \in \mathcal{E}(D)} \tau(d)^{O(1)} \right)^{1/2}$$

$$\ll x \log^{-A} x$$

by Lemma 5.2 and (1-3). $\qquad \square$

It therefore suffices to show that

$$\sum_{\substack{d \in \mathcal{D}_I^{(i)}(x^\delta) \\ d \in [D, 2D] \setminus \mathcal{E}(D)}} |\Delta(\alpha \star \beta; a \, (d))| \ll x \log^{-A} x \tag{5-9}$$

for any fixed $A > 0$.

Let $\varepsilon > 0$ be a small fixed quantity to be chosen later. From (5-2) and (5-6) we have

$$1 \leq x^{-3\varepsilon} N \leq D$$

for x large enough. Let $j \geq 0$ and $k \geq 0$ be fixed integers such that

$$i - 1 = j + k. \tag{5-10}$$

Then any integer $d \in \mathscr{D}_I^{(i)}(x^\delta)$ can by definition (see Definition 2.1) be factored as $d = qr$, where $q \in \mathscr{D}_I^{(j)}(x^\delta)$, $r \in \mathscr{D}_I^{(k)}(x^\delta)$, and

$$x^{-3\varepsilon-\delta} N \leq r \leq x^{-3\varepsilon} N.$$

Remark 5.4. The reason that r is taken to be slightly less than N is to ensure that a diagonal term is manageable when the time comes to apply the Cauchy–Schwarz inequality. The factor of 3 in the exponent is merely technical, and should be ignored on a first reading (ε will eventually be set to be very small, so the constants in front of ε will ultimately be irrelevant).

Let $d \in [D, 2D] \setminus \mathscr{E}(D)$, so that

$$s = \prod_{\substack{p \mid d \\ p \leq D_0}} p \ll 1.$$

Then replacing q by $q/(q, s)$ and r by $r(q, s)$, we obtain a factorization $d = qr$ where q has no prime factor $\leq D_0$ and

$$x^{-3\varepsilon-\delta} N \ll r \ll x^{-3\varepsilon} N. \tag{5-11}$$

By Lemma 2.10(0), (i), we have

$$q \in \mathscr{D}^{(j)}(sx^\delta) = \mathscr{D}^{(j)}(x^{\delta+o(1)}), \qquad r \in \mathscr{D}^{(k)}(sx^\delta) = \mathscr{D}^{(k)}(x^{\delta+o(1)}).$$

In particular, $q \in \mathscr{D}_J^{(j)}(x^{\delta+o(1)})$, where $J := I \cap (D_0, +\infty)$. As $i \geq 1$, we also have $qr = d \in \mathscr{D}_I(x^\delta) = \mathscr{D}_I^{(1)}(x^\delta)$.

Remark 5.5. The reason for removing all the small prime factors from q will become clearer later, when the Cauchy–Schwarz inequality is invoked to replace the single parameter q with two parameters q_1, q_2 in the same range. By excluding the small primes from q_1, q_2, this will ensure that q_1 and q_2 will almost always be coprime, which will make things much simpler.

The next step is to perform dyadic decompositions of the range of the q and r variables, which (in view of (5-1)) reduces the proof of (5-9) to the proof of the estimates

$$\sum_{\substack{q \in \mathscr{D}_J^{(j)}(x^{\delta+o(1)}) \cap [Q,2Q] \\ r \in \mathscr{D}_I^{(k)}(x^{\delta+o(1)}) \cap [R,2R] \\ qr \in \mathscr{D}_I(x^\delta)}} |\Delta(\alpha \star \beta; a\,(qr))| \ll MN \log^{-A} x$$

for any fixed $A > 0$ and any Q, R obeying the conditions

$$x^{-3\varepsilon-\delta} N \ll R \ll x^{-3\varepsilon} N, \tag{5-12}$$

$$x^{1/2} \ll QR \ll x^{1/2+2\varpi}. \tag{5-13}$$

We note that these inequalities also imply that

$$NQ \ll x^{1/2+2\varpi+\delta+3\varepsilon}. \tag{5-14}$$

For future reference we also claim the bound

$$RQ^2 \ll x. \tag{5-15}$$

In cases (i)–(iii) of Theorem 5.1, we have $\sigma + 4\varpi + \delta < \frac{1}{2}$ (with plenty of room to spare), and (5-15) then easily follows from (5-12), (5-13), and (5-2). For case (i), we have $6\varpi + \delta < \frac{1}{2}$, and we may argue as before, but with (5-2) replaced by the bound $N \gg x^{1/2-2\varpi-c}$.

Let Q, R be as above. We will abbreviate

$$\sum_q A_q = \sum_{q \in \mathscr{D}_J^{(j)}(x^{\delta+o(1)}) \cap [Q,2Q]} A_q \tag{5-16}$$

and

$$\sum_r A_r = \sum_{r \in \mathscr{D}_I^{(k)}(x^{\delta+o(1)}) \cap [R,2R]} A_r \tag{5-17}$$

for any summands A_q, A_r.

We now split the discrepancy by writing

$$\Delta(\alpha \star \beta; a\,(qr)) = \Delta_1(\alpha \star \beta; a\,(qr)) + \Delta_2(\alpha \star \beta; a\,(qr)),$$

where

$$\Delta_1(\alpha \star \beta; a\,(qr)) := \sum_{n=a\,(qr)} (\alpha \star \beta)(n) - \frac{1}{\varphi(q)} \sum_{\substack{(n,q)=1 \\ n=a\,(r)}} (\alpha \star \beta)(n),$$

$$\Delta_2(\alpha \star \beta; a\,(qr)) := \frac{1}{\varphi(q)} \sum_{\substack{(n,q)=1 \\ n=a\,(r)}} (\alpha \star \beta)(n) - \frac{1}{\varphi(qr)} \sum_{(n,qr)=1} (\alpha \star \beta)(n).$$

The second term can be dealt with immediately:

Lemma 5.6. *We have*

$$\sum_{q,r:qr\in\mathscr{D}_I(x^\delta)}\sum |\Delta_2(\alpha\star\beta;a\ (qr))| \ll NM\log^{-A}x$$

for any fixed $A > 0$.

Proof. Since $r \leq 2R \ll x^{1/2+o(1)-3\varepsilon}$, the Bombieri–Vinogradov theorem (Theorem 2.9), applied for each q to $\alpha_q \star \beta_q$, where $\alpha_q = \alpha\mathbf{1}_{(n,q)=1}$, $\beta_q = \mathbf{1}_{(n,q)=1}$, gives

$$\sum_{\substack{R\leq r\leq 2R \\ qr\in\mathscr{D}_I(x^\delta)}} \left| \sum_{\substack{(n,q)=1 \\ n=a\ (r)}} (\alpha\star\beta)(n) - \frac{1}{\varphi(r)} \sum_{(n,qr)=1} (\alpha\star\beta)(n) \right| \ll NM\log^{-A}x,$$

since β_q inherits the Siegel–Walfisz property from β. Dividing by $\varphi(q)$ and summing over $q \leq 2Q$, we get the result using the standard estimate

$$\sum_q \frac{1}{\varphi(q)} \ll \log x. \qquad \square$$

To deal with Δ_1, it is convenient to define

$$\Delta_0(\alpha\star\beta;a,b_1,b_2) = \sum_{\substack{n=a\ (r) \\ n=b_1\ (q)}} (\alpha\star\beta)(n) - \sum_{\substack{n=a\ (r) \\ n=b_2\ (q)}} (\alpha\star\beta)(n)$$

for all integers a, b_1, b_2 coprime to P_I. Indeed, we have

$$\sum_{\substack{q,r \\ qr\in\mathscr{D}_I(x^\delta)}}\sum |\Delta_1(\alpha\star\beta;a\ (qr))| \leq \frac{1}{\varphi(P_I)} \sum_{\substack{b\ (P_I) \\ (b,P_I)=1}} \sum_{\substack{q,r \\ qr\in\mathscr{D}_I(x^\delta)}}\sum |\Delta_0(\alpha\star\beta;a,a,b)|$$

by the triangle inequality and the Chinese remainder theorem. Hence it is enough to prove that

$$\sum_{\substack{q,r \\ qr\in\mathscr{D}_I(x^\delta)}}\sum |\Delta_0(\alpha\star\beta;a,b_1,b_2)| \ll NM\log^{-A}x \qquad (5\text{-}18)$$

for all a, b_1, b_2 coprime to P_I, and this will be our goal. The advantage of this step is that the two terms in Δ_0 behave symmetrically, in contrast to those in Δ_1 (or Δ), and this will simplify the presentation of the dispersion method: in the notation of [Bombieri et al. 1986; Linnik 1963; Zhang 2014], one only needs to control $\mathscr{S}_1$, and one avoids dealing explicitly with $\mathscr{S}_2$ or $\mathscr{S}_3$. This is mostly an expository simplification, however, since the estimation of $\mathscr{S}_1$ is always the most difficult part in applications of the dispersion method.

The fact that $r \leq R$ is slightly less than N ensures that the constraint $n = a\ (r)$ leaves room for nontrivial averaging of the variable n, and allows us to profitably use the dispersion method of Linnik. We begin by writing

$$\sum_{\substack{q,r \\ qr \in \mathcal{D}_I(x^\delta)}} |\Delta_0(\alpha \star \beta; a, b_1, b_2)| = \sum_{\substack{q,r \\ qr \in \mathcal{D}_I(x^\delta)}} c_{q,r} \left(\sum_{\substack{n=a\ (r) \\ n=b_1\ (q)}} (\alpha \star \beta)(n) - \sum_{\substack{n=a\ (r) \\ n=b_2\ (q)}} (\alpha \star \beta)(n) \right),$$

where $c_{q,r}$ are complex numbers of modulus 1. Expanding the Dirichlet convolution and exchanging the sums, we obtain

$$\sum_{\substack{q,r \\ qr \in \mathcal{D}_I(x^\delta)}} |\Delta_0(\alpha \star \beta; a, b_1, b_2)|$$

$$= \sum_r \sum_m \alpha(m) \left(\sum_{\substack{mn=a\ (r) \\ qr \in \mathcal{D}_I(x^\delta)}} c_{q,r} \beta(n) (\mathbf{1}_{mn=b_1\ (q)} - \mathbf{1}_{mn=b_2\ (q)}) \right).$$

By the Cauchy–Schwarz inequality applied to the r and m sums, (2-4), (2-6) and Lemma 1.3, we have

$$\sum_{\substack{q,r \\ qr \in \mathcal{D}_I(x^\delta)}} |\Delta_0(\alpha \star \beta; a, b_1, b_2)| \leq R^{1/2} M^{1/2} (\log x)^{O(1)}$$

$$\times \left(\sum_r \sum_m \psi_M(m) \left| \sum_{\substack{mn=a\ (r) \\ qr \in \mathcal{D}_I(x^\delta)}} c_{q,r} \beta(n) (\mathbf{1}_{mn=b_1\ (q)} - \mathbf{1}_{mn=b_2\ (q)}) \right|^2 \right)^{1/2}$$

for any smooth coefficient sequence ψ_M at scale M such that $\psi_M(m) \geq 1$ for m in the support of β. This means in particular that it is enough to prove the estimate

$$\sum_r \sum_m \psi_M(m) \left| \sum_{\substack{mn=a\ (r) \\ qr \in \mathcal{D}_I(x^\delta)}} c_{q,r} \beta(n) (\mathbf{1}_{mn=b_1\ (q)} - \mathbf{1}_{mn=b_2\ (q)}) \right|^2$$

$$\ll N^2 M R^{-1} \log^{-A} x \quad (5\text{-}19)$$

for any fixed $A > 0$, where ψ_M is a smooth coefficient sequence at scale M.

Let Σ denote the left-hand side of (5-19). Expanding the square, we find

$$\Sigma = \Sigma(b_1, b_1) - \Sigma(b_1, b_2) - \Sigma(b_2, b_1) + \Sigma(b_2, b_2), \quad (5\text{-}20)$$

where

$$\Sigma(b_1, b_2)$$

$$:= \sum_r \sum_m \psi_M(m) \sum_{\substack{q_1, q_2, n_1, n_2 \\ mn_1 = mn_2 = a\ (r) \\ q_1 r, q_2 r \in \mathcal{D}_I(x^\delta)}} \cdots \sum c_{q_1, r} \overline{c_{q_2, r}} \beta(n_1) \overline{\beta(n_2)} \mathbf{1}_{mn_1 = b_1\ (q_1)} \mathbf{1}_{mn_2 = b_2\ (q_2)}$$

for any integers b_1 and b_2 coprime to P_I (where the variables q_1 and q_2 are subject to the constraint (5-16)). We will prove that

$$\Sigma(b_1, b_2) = X + O(N^2 M R^{-1} \log^{-A} x) \qquad (5\text{-}21)$$

for all b_1 and b_2, where the main term X is independent of b_1 and b_2. From (5-20), the desired conclusion (5-19) then follows.

Since a is coprime to qr, so are the variables n_1 and n_2 in the sum. In particular, they satisfy the congruence $n_1 = n_2\ (r)$. We write $n_2 = n_1 + \ell r$ in the sum, rename n_1 as n, and therefore obtain

$$\Sigma(b_1, b_2) = \sum_r \sum_\ell \sum_{\substack{q_1, q_2 \\ q_1 r, q_2 r \in \mathcal{D}_I(x^\delta)}} \left(c_{q_1, r} \overline{c_{q_2, r}} \sum_n \beta(n) \overline{\beta(n + \ell r)} \right.$$
$$\left. \times \sum_m \psi_M(m) \mathbf{1}_{mn = b_1\ (q_1)} \mathbf{1}_{m(n + \ell r) = b_2\ (q_2)} \mathbf{1}_{mn = a\ (r)} \right)$$

after some rearranging (remembering that $(n, q_1 r) = (n + \ell r, q_2 r) = 1$). Note that the sum over ℓ is restricted to a range $0 \le |\ell| \ll L := N R^{-1}$.

We will now complete the sum in m (which is long since M is just a bit smaller than the modulus $[q_1, q_2] r \le Q^2 R$) using Lemma 4.9(ii), but first we handle separately the diagonal case $n_1 = n_2$, i.e., $\ell = 0$. This contribution, say $T(b_1, b_2)$, satisfies

$$|T(b_1, b_2)| \le \sum_r \sum_{\substack{q_1, q_2 \\ q_1 r, q_2 r \in \mathcal{D}_I(x^\delta)}} \sum_n |\beta(n)|^2 \sum_m \psi_M(m) \mathbf{1}_{mn = b_1\ (q_1)} \mathbf{1}_{mn = b_2\ (q_2)} \mathbf{1}_{mn = a\ (r)}$$

$$\ll \sum_{r \asymp R} \sum_{q_1, q_2 \asymp Q} \sum_{s \asymp x} \tau(s) \mathbf{1}_{s = b_1\ (q_1)} \mathbf{1}_{s = b_2\ (q_2)} \mathbf{1}_{s = a\ (r)}$$

$$\ll \sum_{r \asymp R} \sum_{q_1, q_2 \asymp Q} \frac{x}{r[q_1, q_2]} \ll x \ll N^2 M R^{-1} \log^{-A} x$$

(since $R Q^2 \ll x$ (from (5-15)) and $R \ll x^{-3\epsilon} N$).

Now we consider the contributions where $\ell \ne 0$. First, since n and $n + \ell r$ are coprime to $q_1 r$ and $q_2 r$ respectively, we have

$$\mathbf{1}_{mn = b_1\ (q_1)} \mathbf{1}_{m(n + \ell r) = b_2\ (q_2)} \mathbf{1}_{mn = a\ (r)} = \mathbf{1}_{m = \gamma\ ([q_1, q_2] r)} \qquad (5\text{-}22)$$

for some residue class $\gamma\ ([q_1, q_2] r)$ (which depends on b_1, b_2, ℓ, n and a). We will denote (q_1, q_2) by q_0, and observe that since q_1, q_2 have no prime factor less than D_0, we have either $q_0 = 1$ or $q_0 \ge D_0$. (The first case gives the principal contribution, and the reader may wish to assume that $q_0 = 1$ in a first reading.) The sum over n is further restricted by the congruence

$$\frac{b_1}{n} = \frac{b_2}{n + \ell r}\ (q_0), \qquad (5\text{-}23)$$

and we will use

$$C(n) := \mathbf{1}_{b_1/n \equiv (b_2)/n + \ell r \ (q_0)} \tag{5-24}$$

to denote the characteristic function of this condition (taking care of the fact that it depends on other parameters). Observe that, since q_0 is coprime to rb_1, this is the characteristic function of a union of at most $(b_1 - b_2, q_0, \ell r b_1) \le (q_0, \ell)$ congruence classes modulo q_0.

By applying Lemma 4.9(ii) to each choice of q_1, q_2, r, ℓ (where I is the range of the remaining parameter n) and summing, we derive

$$\Sigma(b_1, b_2) = \Sigma_0(b_1, b_2) + \Sigma_1(b_1, b_2) + O(MN^2 R^{-1} \log^{-A} x),$$

where

$$\begin{aligned}
&\Sigma_0(b_1, b_2) \\
&:= \left(\sum_m \psi_M(m) \right) \sum_r r^{-1} \sum_{\ell \ne 0} \sum_{\substack{q_1, q_2 \\ q_1 r, q_2 r \in \mathscr{D}_I(x^\delta)}} \frac{c_{q_1,r} \overline{c_{q_2,r}}}{[q_1, q_2]} \sum_n \beta(n) \overline{\beta(n + \ell r)} C(n)
\end{aligned}$$

and

$$\Sigma_1(b_1, b_2) \ll 1 + x^\varepsilon \widehat{\Sigma}_1(b_1, b_2)$$

with

$$\begin{aligned}
&\widehat{\Sigma}_1(b_1, b_2) \\
&:= \sum_r \sum_{\ell \ne 0} \sum_{\substack{q_1, q_2 \\ q_1 r, q_2 r \in \mathscr{D}_I(x^\delta)}} c_{q_1,r} \overline{c_{q_2,r}} \frac{1}{H} \sum_{1 \le |h| \le H} \left| \sum_n \beta(n) \overline{\beta(n + \ell r)} C(n) e_{[q_1,q_2]r}(\gamma h) \right|,
\end{aligned}$$

where

$$H := x^\varepsilon [q_1, q_2] r M^{-1} \ll x^\varepsilon Q^2 R M^{-1}. \tag{5-25}$$

We caution that H depends on q_1 and q_2, so one has to take some care if one is to interchange the h and q_1, q_2 summations.

Remark 5.7. Before going further, note that H is rather small, since M and R are close to $x^{1/2}$ and $\varepsilon > 0$ will be very small: precisely, we have

$$H \ll H_0 := x^\varepsilon \times (QR)^2 \times \frac{N}{R} \times \frac{1}{NM},$$

and using (5-12), (5-13) and (5-1), we see that

$$x^{4\varepsilon} \ll H_0 \ll x^{4\varpi + \varepsilon} (N/R) \ll x^{4\varpi + \delta + 4\varepsilon}. \tag{5-26}$$

As we will be using small values of $\varpi, \delta, \varepsilon$, one should thus think of H as being quite small compared to x.

We can deal immediately with $\Sigma_0(b_1, b_2)$. We distinguish between the contributions of q_1 and q_2 which are coprime, and the remainder. The first is independent of b_1 and b_2 (since these parameters are only involved in the factor $C(n) = \mathbf{1}_{b_1/n = b_2/(n+\ell r) \ (q_0)}$), which is then always 1) and it will be the main term X; thus

$$X := \left(\sum_m \psi_M(m)\right) \sum_r r^{-1} \sum_{\ell \neq 0} \sum_{\substack{q_1, q_2 \\ q_1 r, q_2 r \in \mathcal{D}_I(x^\delta) \\ (q_1, q_2) = 1}} \frac{c_{q_1, r} \overline{c_{q_2, r}}}{[q_1, q_2]} \sum_n \beta(n) \overline{\beta(n + \ell r)}.$$

The remaining contribution to $\Sigma_0(b_1, b_2)$, say $\Sigma_0'(b_1, b_2)$, is

$$\ll \frac{M(\log x)^{O(1)}}{R} \sum_{r \asymp R} \sum_{|\ell| \ll L} \sum_{\substack{1 \neq q_0 \ll Q \\ q_0 \in \mathcal{S}_J}} \frac{1}{q_0} \sum_{q_1, q_2 \asymp Q/q_0} \frac{1}{q_1 q_2} \sum_n (\tau(n) \tau(n + \ell r))^{O(1)} C(n).$$

We rearrange to sum over ℓ first (remember that $C(n)$ depends on ℓ also). Since rb_1 is coprime with q_0, the condition $b_1/n = b_2/(n + \ell r) \ (q_0)$ is a congruence condition modulo q_0 for ℓ, and therefore

$$\sum_{|\ell| \ll L} \tau(n + \ell r)^{O(1)} \mathbf{1}_{b_1/n = b_2/(n+\ell r) \ (q_0)} \ll \left(1 + \frac{L}{q_0}\right) \log^{O(1)} x = \left(1 + \frac{N}{q_0 R}\right) \log^{O(1)} x$$

by [Lemma 1.3](navigation). Since all $q_0 \neq 1$ in the sum satisfy $D_0 \leq q_0 \ll Q$, we get

$$\begin{aligned}
\Sigma_0'(b_1, b_2) &\ll \frac{MN(\log x)^{O(1)}}{R} \sum_{r \asymp R} \sum_{D_0 \leq q_0 \ll Q} \frac{1}{q_0}\left(1 + \frac{N}{q_0 R}\right) \sum_{q_1, q_2 \asymp Q/q_0} \frac{1}{q_1 q_2} \\
&\ll MN \log^{O(1)} x \sum_{D_0 \leq q_0 \ll Q} \frac{1}{q_0}\left(1 + \frac{N}{q_0 R}\right) \\
&\ll MN \log^{O(1)} x + \frac{1}{D_0} \frac{MN^2}{R} \log^{O(1)} x \\
&\ll MN^2 R^{-1} \log^{-A} x,
\end{aligned}$$

since $R \ll x^{-3\varepsilon} N$ and $D_0 \gg \log^A x$ for all $A > 0$.

Hence we have shown that

$$\Sigma(b_1, b_2) = X + O(x^\varepsilon |\widehat{\Sigma}_1(b_1, b_2)|) + O(MN^2 R^{-1} \log^{-A} x). \tag{5-27}$$

From the definition, and in particular the localization of r and the value of H, we have

$$\begin{aligned}
|\widehat{\Sigma}_1(b_1, b_2)| &\leq \sum_r \sum_{\ell \neq 0} \sum_{\substack{q_1, q_2 \\ q_1 r, q_2 r \in \mathcal{D}_I(x^\delta)}} \frac{1}{H} \sum_{0 < |h| \leq H} \left| \sum_n C(n) \beta(n) \overline{\beta(n + \ell r)} e_{[q_1, q_2] r}(\gamma h) \right| \\
&\ll x^{-\varepsilon} \frac{M}{RQ^2} \sum_{1 \leq |\ell| \ll L} \sum_{q_0 \ll Q} q_0 \sum_r \Upsilon_{\ell, r}(b_1, b_2; q_0), \tag{5-28}
\end{aligned}$$

where q_0 is again (q_1, q_2) and

$$\Upsilon_{\ell,r}(b_1, b_2; q_0) :=$$

$$\sum_{\substack{q_1, q_2 \asymp Q/q_0 \\ (q_1, q_2)=1}} \sum_{\substack{q_0 q_1, q_0 q_2 \in \mathscr{D}_I^{(j)}(x^{\delta+o(1)}) \\ q_0 q_1 r, q_0 q_2 r \in \mathscr{D}_I(x^\delta)}} \left(\mathbf{1}_{} \right.$$

$$\left. \times \sum_{1 \leq |h| \ll \frac{x^\varepsilon R Q^2}{q_0 M}} \left| \sum_n C(n) \beta(n) \overline{\beta(n+\ell r)} \Phi_\ell(h, n, r, q_0, q_1, q_2) \right| \right). \quad (5\text{-}29)$$

The latter expression involves the phase function Φ_ℓ, which we define for parameters $\boldsymbol{p} = (h, n, r, q_0, q_1, q_2)$ by

$$\Phi_\ell(\boldsymbol{p}) := e_r\left(\frac{ah}{nq_0q_1q_2} \right) e_{q_0q_1}\left(\frac{b_1 h}{nrq_2} \right) e_{q_2}\left(\frac{b_2 h}{(n+\ell r)rq_0q_1} \right). \quad (5\text{-}30)$$

Here we have spelled out and split, using (5-22) and the Chinese remainder theorem, the congruence class of γ modulo $[q_1, q_2]r$, and changed variables so that q_1 is q_0q_1, q_2 is q_0q_2 (hence $[q_1, q_2]r$ becomes $q_0q_1q_2r$). Moreover, the r summation must be interpreted using (5-17). It will be important for later purposes to remark that we also have

$$\widehat{\Sigma}_1(b_1, b_2) = 0$$

unless

$$\frac{x^\varepsilon Q^2 R}{q_0 M} \gg 1, \quad (5\text{-}31)$$

since otherwise the sum over h is empty.

Gathering these estimates, we obtain the following general reduction statement, where we pick a suitable value of (j, k) in each of the four cases of Theorem 5.1:

Theorem 5.8 (exponential sum estimates). *Let $\varpi, \delta, \sigma > 0$ be fixed quantities, let I be a bounded subset of $\mathbb{R}$, let $j, k \geq 0$ be fixed, let a (P_I), b_1 (P_I), b_2 (P_I) be primitive congruence classes, and let $M, N \gg 1$ be quantities satisfying the conditions (5-1) and (5-2). Let $\varepsilon > 0$ be a sufficiently small fixed quantity, and let Q, R be quantities obeying (5-12), (5-13). Let ℓ be an integer with $1 \leq |\ell| \ll N/R$, and let β be a coefficient sequence located at scale N.*

Further, let $\Phi_\ell(\boldsymbol{p})$ be the phase function defined by (5-30) for parameters $\boldsymbol{p} = (h, n, r, q_0, q_1, q_2)$, let $C(n)$ be the cutoff (5-24) and let $\Upsilon_{\ell,r}(b_1, b_2; q_0)$ be defined in terms of β, Φ, C by (5-29). Then we have

$$\sum_r \Upsilon_{\ell,r}(b_1, b_2; q_0) \lll x^{-\varepsilon} Q^2 R N(q_0, \ell) q_0^{-2} \quad (5\text{-}32)$$

for all $q_0 \in \mathscr{S}_I$, where the sum over r is over $r \in \mathscr{D}_I^{(k)}(x^{\delta+o(1)}) \cap [R, 2R]$, provided

that one of the following hypotheses is satisfied:

(i) $(j, k) = (0, 0)$, $54\varpi + 15\delta + 5\sigma < 1$, *and* $N \ll x^{1/2 - 2\varpi - c}$ *for some fixed* $c > 0$.

(ii) $(j, k) = (1, 0)$, $56\varpi + 16\delta + 4\sigma < 1$, *and* $N \ll x^{1/2 - 2\varpi - c}$ *for some fixed* $c > 0$.

(iii) $(j, k) = (1, 2)$, $\frac{160}{3}\varpi + 16\delta + \frac{34}{9}\sigma < 1$, $64\varpi + 18\delta + 2\sigma < 1$, *and* $N \ll x^{1/2 - 2\varpi - c}$ *for some fixed* $c > 0$.

(iv) $(j, k) = (0, 0)$, $68\varpi + 14\delta < 1$, *and* $N \gg x^{1/2 - 2\varpi - c}$ *for some sufficiently small fixed* $c > 0$.

The proof of the estimate (iii) *requires Deligne's form of the Riemann hypothesis for algebraic varieties over finite fields, but the proofs of* (i), (ii), (iv) *do not.*

Indeed, inserting this bound in (5-28) we obtain

$$x^\varepsilon |\widehat{\Sigma}(b_1, b_2)| \ll x^{-\varepsilon} MN \sum_{q_0 \ll Q} \frac{1}{q_0} \sum_{1 \le |\ell| \ll NR^{-1}} (q_0, \ell) \ll x^{-\varepsilon} MN^2 R^{-1}$$

(by Lemma 1.4, crucially using the fact that we have previously removed the $\ell = 0$ contribution), and hence using (5-27), we derive the goal (5-21).

Remark 5.9. As before, one should consider the $q_0 = 1$ case as the main case, so that the technical factors of q_0, (ℓ, q_0), and $C(n)$ should be ignored at a first reading; in practice, we will usually (though not always) end up discarding several powers of q_0 in the denominator in the final bounds for the $q_0 > 1$ case. The trivial bound for $\Upsilon_{\ell,r}(b_1, b_2; q_0)$ is about $(Q/q_0)^2 NH$, with $H = x^\varepsilon RQ^2 M^{-1} q_0^{-1}$. Thus one needs to gain about H over the trivial bound. As observed previously, H is quite small, and even a modestly nontrivial exponential sum estimate can suffice for this purpose (after using Cauchy–Schwarz to eliminate factors such as $\beta(n)\overline{\beta(n + \ell r)}$).

It remains to establish Theorem 5.8 in the four cases indicated. We will do this for (i), (ii), (iv) below, and defer the proof of (iii) to Section 8. In all four cases, one uses the Cauchy–Schwarz inequality to eliminate nonsmooth factors such as $\beta(n)$ and $\beta(n + \ell r)$, and reduces matters to incomplete exponential sum estimates. In the cases (i), (ii), (iv) treated below, the one-dimensional exponential sum estimates from Section 4D suffice; for the final case (iii), a multidimensional exponential sum estimate is involved, and we will prove it using Deligne's formalism of the Riemann hypothesis over finite fields, which we survey in Section 6.

5D. *Proof of Type II estimate.* We begin with the proof of Theorem 5.8(iv), which is the simplest of the four estimates to prove. We fix notation and hypotheses as in this statement.

To prove (5-32), we will not exploit any averaging in the variable r, and, more precisely, we will show that

$$\Upsilon_{\ell,r}(b_1, b_2; q_0) \ll x^{-\varepsilon} Q^2 N(q_0, \ell) q_0^{-2} \tag{5-33}$$

for each $q_0 \geq 1$, $r \asymp R$ and $\ell \ll N/R$. We abbreviate $\Upsilon = \Upsilon_{\ell,r}(b_1, b_2; q_0)$ in the remainder of this section, and set

$$H = x^{\varepsilon} R Q^2 M^{-1} q_0^{-1}.$$

By (5-29), we can then write

$$\Upsilon = \sum_{\substack{q_1, q_2 \asymp Q/q_0 \\ (q_1, q_2)=1}} \sum_{1 \leq |h| \leq H} c_{h,q_1,q_2} \sum_n C(n) \beta(n) \overline{\beta(n+\ell r)} \Phi_\ell(h, n, r, q_0, q_1, q_2) \tag{5-34}$$

for some coefficients c_{h,q_1,q_2} with modulus at most 1. We then exchange the order of summation to move the sum over n (and the terms $C(n)\beta(n)\overline{\beta(n+\ell r)}$) outside. Since $C(n)$ is the characteristic function of at most (q_0, ℓ) congruence classes modulo q_0 (as observed after (5-23)), we have

$$\sum_n C(n) |\beta(n)|^2 |\beta(n+\ell r)|^2 \ll N \frac{(q_0, \ell)}{q_0} \tag{5-35}$$

by Lemma 1.3 (and the Cauchy–Schwarz inequality), using the fact that $Q \leq N$.

By another application of the Cauchy–Schwarz inequality, and after inserting (by positivity) a suitable coefficient sequence $\psi_N(n)$, smooth at scale N and ≥ 1 for n in the support of $\beta(n)\overline{\beta(n+\ell r)}$, we obtain the bound

$$|\Upsilon|^2 \ll N \frac{(q_0, \ell)}{q_0} \sum_n \psi_N(n) C(n) \left| \sum_{\substack{q_1, q_2 \asymp Q/q_0 \\ (q_1, q_2)=1}} \sum_{1 \leq |h| \leq H} c_{h,q_1,q_2} \Phi_\ell(h, n, r, q_0, q_1, q_2) \right|^2$$

$$\ll N \frac{(q_0, \ell)}{q_0} \sum_{\substack{q_1, q_2, s_1, s_2 \asymp Q/q_0 \\ (q_1, q_2)=(s_1, s_2)=1}} \cdots \sum \sum_{1 \leq h_1, h_2 \leq |H|} |S_{\ell,r}(h_1, h_2, q_1, q_2, s_1, s_2)|,$$

where the exponential sum $S_{\ell,r} = S_{\ell,r}(h_1, h_2, q_1, q_2, s_1, s_2)$ is given by

$$S_{\ell,r} := \sum_n C(n) \psi_N(n) \Phi_\ell(h_1, n, r, q_0, q_1, q_2) \overline{\Phi_\ell(h_2, n, r, q_0, s_1, s_2)}. \tag{5-36}$$

We will prove the following estimate for this exponential sum (compare with [Zhang 2014, (12.5)]):

Proposition 5.10. *For any*

$$\boldsymbol{p} = (h_1, h_2, q_1, q_2, s_1, s_2)$$

with $(q_0 q_1 q_2 s_1 s_2, r) = 1$, *any* $\ell \neq 0$ *and* r *as above with*

$$q_0 q_i, \ q_0 s_i \ll Q, \quad r \ll R,$$

we have

$$|S_{\ell, r}(\boldsymbol{p})| \ll (q_0, \ell)\left(q_0^{-2} Q^2 R^{1/2} + \frac{N}{q_0 R}(h_1 s_1 s_2 - h_2 q_1 q_2, r)\right).$$

Assuming this, we obtain

$$|\Upsilon|^2 \ll N\left(\frac{(q_0, \ell)}{q_0}\right)^2 \sum \cdots \sum_{\substack{q_1, q_2, s_1, s_2 \asymp Q/q_0 \\ (q_1, q_2) = (s_1, s_2) = 1}} \sum\sum_{1 \leq h_1, h_2 \leq |H|} \left(\frac{1}{q_0}Q^2 R^{1/2} + \frac{N}{R}(h_1 s_1 s_2 - h_2 q_1 q_2, r)\right)$$

(since $S_{\ell, r} = 0$ unless $(q_0 q_1 q_2 s_1 s_2, r) = 1$, by the definition (5-30) and the definition of e_q in Section 4).

Making the change of variables $\Delta = h_1 s_1 s_2 - h_2 q_1 q_2$, and noting that each Δ has at most $\tau_3(\Delta) = |\{(a, b, c) : abc = \Delta\}|$ representations in terms of h_2, q_1, q_2 for each fixed h_1, s_1, s_2, we have

$$\sum \cdots \sum_{\substack{q_1, q_2, s_1, s_2 \asymp Q/q_0 \\ (q_1, q_2) = (s_1, s_2) = 1}} \sum\sum_{1 \leq h_1, h_2 \leq |H|} (h_1 s_1 s_2 - h_2 q_1 q_2, r)$$

$$\leq \sum_{|\Delta| \ll H(Q/q_0)^2} (\Delta, r) \sum \cdots \sum_{h_1, s_1, s_2} \tau_3(h_1 s_1 s_2 - \Delta)$$

$$\ll H\left(\frac{Q}{q_0}\right)^2 \sum_{0 \leq |\Delta| \ll H(Q/q_0)^2} (\Delta, r)$$

$$\ll H\left(\frac{Q}{q_0}\right)^2 \left(\frac{HQ^2}{q_0^2} + R\right)$$

by Lemma 1.3 (bounding $\tau_3 \leq \tau^2$) and Lemma 1.4. Therefore we obtain

$$|\Upsilon|^2 \ll N\frac{(q_0, \ell)^2}{q_0^2}\left\{\frac{H^2 Q^2 R^{1/2}}{q_0}\left(\frac{Q}{q_0}\right)^4 + \frac{H^2 N}{R}\left(\frac{Q}{q_0}\right)^4 + NH\left(\frac{Q}{q_0}\right)^2\right\}$$

$$\ll \frac{N^2 Q^4 (q_0, \ell)^2}{q_0^4}\left\{\frac{H^2 Q^2 R^{1/2}}{N} + \frac{H^2}{R} + \frac{H}{Q^2}\right\}$$

$$\ll \frac{N^2 Q^4 (q_0, \ell)^2}{q_0^4}\left\{x^{2\varepsilon}\frac{Q^6 R^{5/2}}{M^2 N} + x^{2\varepsilon}\frac{RQ^4}{M^2} + \frac{x^\varepsilon R}{M}\right\}, \tag{5-37}$$

where we have discarded some powers of $q_0 \geq 1$ in the denominator to reach the second and third lines. We now observe that

$$\frac{Q^6 R^{5/2}}{M^2 N} \asymp \frac{(NQ)(QR)^5}{x^2 R^{5/2}} \ll \frac{x^{1+12\varpi+\delta+3\varepsilon}}{R^{5/2}} \ll \frac{x^{1+12\varpi+7\delta/2+21\varepsilon/2}}{N^{5/2}},$$

$$\frac{Q^4 R}{M^2} \asymp \frac{N^2 R Q^4}{x^2} = \frac{(QR)(NQ)^3}{x^2 N} \ll \frac{x^{8\varpi+3\delta+9\varepsilon}}{N},$$

$$\frac{R}{M} \asymp \frac{NR}{x} \ll x^{-1-3\varepsilon} N^2 \ll x^{-3\varepsilon},$$

by (5-13) and (5-14) and the bound $N \ll M$. Under the Type II assumption that $N \gg x^{1/2-2\varpi-c}$ for a small enough $c > 0$ and that $\varepsilon > 0$ is small enough, we see that (5-37) implies (5-33) provided ϖ and δ satisfy

$$\begin{cases} 1 + 12\varpi + \frac{7\delta}{2} < \frac{5}{2}(\frac{1}{2} - 2\varpi), \\ 8\varpi + 3\delta < \frac{1}{2} - 2\varpi, \end{cases} \qquad \Longleftrightarrow \qquad \begin{cases} 68\varpi + 14\delta < 1, \\ 20\varpi + 6\delta < 1, \end{cases}$$

both of which are, indeed, consequences of the hypotheses of Theorem 5.8(iv) (the first implies the second because $\varpi > 0$ so $\delta < \frac{1}{14}$).

To finish this treatment of the Type II sums, it remains to prove the proposition.

Proof of Proposition 5.10. For fixed $(r, \ell, q_0, a, b_1, b_2)$ we can use (5-30) to express the phase Φ_ℓ in the form

$$\Phi_\ell(h, n, r, q_0, q_1, q_2) = e_r^{(1)}\left(\frac{h}{q_1 q_2 n}\right) e_{q_0 q_1}^{(2)}\left(\frac{h}{n q_2}\right) e_{q_2}^{(3)}\left(\frac{h}{(n+\tau)q_0 q_1}\right),$$

where $e_d^{(i)}$ denotes various nontrivial additive characters modulo d which may depend on $(r, \ell, q_0, a, b_1, b_2)$ and $\tau = \ell r$.

We set $\Phi_1(n) = \Phi_\ell(h_1, n, r, q_0, q_1, q_2)$ and $\Phi_2(n) = \Phi_\ell(h_2, n, r, q_0, s_1, s_2)$, and thus we have

$$\Phi_1(n)\overline{\Phi_2(n)} = e_r^{(1)}\left(\frac{h_1}{q_1 q_2 n} - \frac{h_2}{s_1 s_2 n}\right) e_{q_0 q_1}^{(2)}\left(\frac{h_1}{n q_2}\right) e_{q_0 s_1}^{(2)}\left(-\frac{h_2}{n s_2}\right)$$

$$\times e_{q_2}^{(3)}\left(\frac{h_1}{(n+\tau)q_0 q_1}\right) e_{s_2}^{(3)}\left(-\frac{h_2}{(n+\tau)q_0 s_1}\right), \quad (5\text{-}38)$$

and this can be written

$$\Phi_1(n)\overline{\Phi_2(n)} = e_{d_1}^{(4)}\left(\frac{c_1}{n}\right) e_{d_2}^{(5)}\left(\frac{c_2}{n+\tau}\right)$$

for some c_1 and c_2, where

$$d_1 := rq_0[q_1, s_1], \quad d_2 := [q_2, s_2].$$

Now, since $C(n)$ is the characteristic function of $\leq (q_0, \ell)$ residue classes modulo q_0, we deduce that

$$|S_{\ell,r}| = \left| \sum_n C(n) \psi_N(n) \Phi_1(n) \overline{\Phi_2(n)} \right| \leq (q_0, \ell) \max_{t \in \mathbb{Z}/q_0\mathbb{Z}} \left| \sum_{n \equiv t \, (q_0)} \psi_N(n) \Phi_1(n) \overline{\Phi_2(n)} \right|,$$

and by the second part of Corollary 4.16, we derive

$$|S_{\ell,r}| \ll (q_0, \ell) \left(\frac{[d_1, d_2]^{1/2}}{q_0^{1/2}} + \frac{N}{q_0} \frac{(c_1, \delta_1')}{\delta_1'} \frac{(c_2, \delta_2')}{\delta_2'} \right)$$

$$\ll (q_0, \ell) \left(R^{1/2} \left(\frac{Q}{q_0} \right)^2 + \frac{N}{q_0} \frac{(c_1, \delta_1')}{\delta_1'} \right),$$

where $\delta_i = d_i / (d_1, d_2)$ and $\delta_i' = \delta_i / (q_0, \delta_i)$, since

$$[d_1, d_2] \leq r q_0 q_1 q_2 s_1 s_2 \ll q_0 R \left(\frac{Q}{q_0} \right)^4, \qquad \frac{(c_2, \delta_2')}{\delta_2'} \leq 1.$$

Finally, we have

$$\frac{(c_1, \delta_1')}{\delta_1'} = \prod_{\substack{p|\delta_1 \\ p \nmid c_1, q_0}} p \leq \frac{(c_1, r)}{r}$$

(since $r \mid \delta_1$ and $(r, q_0) = 1$). But a prime p dividing r divides c_1 precisely when the r-component of (5-38) is constant, which happens exactly when $p \mid h_1 s_1 s_2 - h_2 q_1 q_2$, so that

$$S_{\ell,r} \ll (q_0, \ell) R^{1/2} \left(\frac{Q}{q_0} \right)^2 + \frac{(q_0, \ell)N}{q_0 R} (r, h_1 s_1 s_2 - h_2 q_1 q_2). \qquad \square$$

Remark 5.11. By replacing the lower bound $N \gg x^{1/2 - 2\varpi - c}$ with the lower bound $N \gg x^{1/2 - \sigma}$, the above argument also yields the estimate $\mathrm{Type}_I^{(1)}[\varpi, \delta, \sigma]$ whenever $48\varpi + 14\delta + 10\sigma < 1$. However, as this constraint does not allow σ to exceed $\frac{1}{10}$, one cannot use this estimate as a substitute for Theorem 2.8(ii) or Theorem 2.8(iii). If one uses the first estimate of Corollary 4.16 in place of the second, one can instead obtain $\mathrm{Type}_I^{(1)}[\varpi, \delta, \sigma]$ for the range $56\varpi + 16\delta + 6\sigma < 1$, which now does permit σ to exceed $\frac{1}{10}$, and thus gives some version of Zhang's theorem after combining with a Type III estimate. However, σ still does not exceed $\frac{1}{6}$, and so one cannot dispense with the Type III component of the argument entirely with this Type I estimate. By using a second application of q-van der Corput, though (i.e., using the $l = 3$ case of Proposition 4.12 rather than the $l = 2$ case), it is possible to raise σ above $\frac{1}{6}$, assuming sufficient amounts of dense divisibility; we leave the details to the interested reader. Thus it is in fact possible to obtain a nontrivial equidistribution estimate of the form $\mathrm{MPZ}[\varpi, \delta]$ using only the Type II

argument, if one is willing to use a sufficient number of applications of q-van der Corput, and using any nontrivial power savings on complete exponential sums as input. However, the Cauchy–Schwarz arguments used here are not as efficient in the Type I setting as the Cauchy–Schwarz arguments in the sections below, and so these estimates do not supersede their Type I counterparts.

5E. *Proof of first Type I estimate.* We will establish Theorem 5.8(i), which is the easiest of the Type I estimates to prove. The strategy follows closely that of the previous section. The changes, roughly speaking, are that the Cauchy–Schwarz argument is slightly modified (so that only the q_2 variable is duplicated, rather than both q_1 and q_2) and that we use an exponential sum estimate based on the first part of Corollary 4.16 instead of the second.

As before, we will establish the bound (5-33) for each individual r. We abbreviate again $\Upsilon = \Upsilon_{\ell,r}(b_1, b_2; q_0)$ and set

$$H = x^{\varepsilon} R Q^2 M^{-1} q_0^{-1}.$$

We begin with the formula (5-34) for Υ, move the q_1 and n sums outside, apply the Cauchy–Schwarz inequality (and insert a suitable smooth coefficient sequence $\psi_N(n)$ at scale N to the n sum), so that we get

$$|\Upsilon|^2 \le \Upsilon_1 \Upsilon_2$$

with

$$\Upsilon_1 := \sum_{q_1 \asymp Q/q_0} \sum_n C(n) |\beta(n)|^2 |\beta(n + \ell r)|^2 \ll \frac{N Q(q_0, \ell)}{q_0^2}$$

(as in (5-35)), and

$$\Upsilon_2 := \sum_n \psi_N(n) C(n) \sum_{\substack{q_1 \asymp Q/q_0}} \left| \sum_{\substack{q_2 \asymp Q/q_0 \\ (q_1, q_2) = 1}} \sum_{1 \le |h| \le H} c_{h, q_1, q_2} \Phi_{\ell}(h, n, r, q_0, q_1, q_2) \right|^2$$

$$= \sum_{\substack{q_1 \asymp Q/q_0}} \sum_{\substack{q_2, s_2 \asymp Q/q_0 \\ (q_1, q_2) = (q_1, s_2) = 1}} \sum_{1 \le h_1, h_2 \le |H|} c_{h_1, q_1, q_2} \overline{c_{h_2, q_1, s_2}} S_{\ell, r}(h_1, h_2, q_1, q_2, q_1, s_2),$$

where $S_{\ell,r}$ is the same sum (5-36) as before and the variables (q_1, q_2, s_2) are restricted by the condition $q_0 q_1 r, q_0 q_2 r, q_0 s_2 r \in \mathcal{D}_I(x^{\delta})$ (recall the definition (5-29)).

We will prove the following bound:

Proposition 5.12. *For any*

$$\boldsymbol{p} = (h_1, h_2, q_1, q_2, q_1, s_2)$$

with $(q_0 q_1 q_2 s_2, r) = 1$ and for any $\ell \neq 0$ and r as above with

$$q_0 q_i r, q_0 s_2 r \in \mathcal{D}_I(x^{\delta}) \quad and \quad q_0 q_i \ll Q, \ q_0 s_2 \ll Q, \ r \ll R,$$

we have

$$|S_{\ell,r}(\boldsymbol{p})| \ll q_0^{1/6} N^{1/2} x^{\delta/6} (Q^3 R)^{1/6} + R^{-1} N (h_1 s_2 - h_2 q_2, r).$$

We first conclude assuming this estimate: arguing as in the previous section to sum the greatest common divisors $(h_1 s_2 - h_2 q_2, r)$, we obtain

$$\Upsilon_2 \ll \left(\frac{Q}{q_0}\right)^3 H^2 \left\{ q_0^{1/6} N^{1/2} (Q^3 R)^{1/6} x^{\delta/6} + \frac{N}{R} \right\} + HN \left(\frac{Q}{q_0}\right)^2,$$

and therefore

$$|\Upsilon|^2 \ll \frac{NQ(q_0, \ell)}{q_0^2} \left\{ q_0^{1/6} \left(\frac{Q}{q_0}\right)^3 H^2 N^{1/2} (Q^3 R)^{1/6} x^{\delta/6} + \left(\frac{Q}{q_0}\right)^3 \frac{H^2 N}{R} + HN \left(\frac{Q}{q_0}\right)^2 \right\}$$

$$\ll \frac{N^2 Q^4 (q_0, \ell)^2}{q_0^4} \left\{ \frac{H^2 Q^{1/2} R^{1/6} x^{\delta/6}}{N^{1/2}} + \frac{H^2}{R} + \frac{H}{Q} \right\},$$

where we once again discard some powers of $q_0 \geq 1$ from the denominator. Using again (5-13) and (5-14) and $N \ll M$, we find that

$$\frac{H^2 Q^{1/2} R^{1/6} x^{\delta/6}}{N^{1/2}} \ll x^{\delta/6 + 2\varepsilon} \frac{R^{13/6} Q^{9/2}}{M^2 N^{1/2}} \ll x^{-2+\delta/6+2\varepsilon} \frac{N^{3/2} (QR)^{9/2}}{R^{7/3}}$$

$$\ll \frac{x^{1/4 + 9\varpi + 5\delta/2 + 9\varepsilon}}{N^{5/6}},$$

$$\frac{H^2}{R} \ll \frac{x^{8\varpi + 3\delta + 11\varepsilon}}{N},$$

$$\frac{H}{Q} \leq x^\varepsilon \frac{RQ}{M} \ll \frac{x^{1/2 + 2\varpi + \varepsilon}}{M} \ll x^{-c+\varepsilon},$$

and using the assumption $N \gg x^{1/2-\sigma}$ from (5-2), we will derive (5-33) if $c = 3\varepsilon$, $\varepsilon > 0$ is small enough, and

$$\begin{cases} \frac{1}{4} + 9\varpi + 5\frac{\delta}{2} < \frac{5}{6}(\frac{1}{2} - \sigma), \\ 8\varpi + 3\delta < \frac{1}{2} - \sigma, \end{cases} \qquad \Longleftrightarrow \qquad \begin{cases} 54\varpi + 15\delta + 5\sigma < 1, \\ 16\varpi + 6\delta + 2\sigma < 1. \end{cases}$$

For ϖ, δ, $\sigma > 0$, the first condition implies the second (as its coefficients are larger). Since the first condition is the assumption of Theorem 5.8(i), we are then done.

We now prove the exponential sum estimate.

Proof of Proposition 5.12. We set

$$\Phi_1(n) = \Phi_\ell(h_1, n, r, q_0, q_1, q_2), \qquad \Phi_2(n) = \Phi_\ell(h_2, n, r, q_0, q_1, s_2),$$

as in the proof of Proposition 5.10, and we write

$$\Phi_1(n) \overline{\Phi_2(n)} = e_{d_1}^{(4)} \left(\frac{c_1}{n}\right) e_{d_2}^{(5)} \left(\frac{c_2}{n+\tau}\right)$$

for some c_1 and c_2, where

$$d_1 := rq_0q_1, \qquad d_2 := [q_2, s_2].$$

Since rq_0q_1, rq_0q_2 and rq_0s_2 are x^δ-densely divisible, Lemma 2.10(ii) implies that the least common multiple $[d_1, d_2] = [rq_0q_1, rq_0q_2, rq_0s_2]$ is also x^δ-densely divisible.

Splitting again the factor $C(n)$ into residue classes modulo q_0, and applying the first part of Corollary 4.16 to each residue class, we obtain

$$|S_{\ell,r}| \ll (q_0, \ell) \left(\frac{N^{1/2}}{q_0^{1/2}} [d_1, d_2]^{1/6} x^{\delta/6} + \frac{N}{q_0} \frac{(c_1, \delta_1')}{\delta_1} \frac{(c_2, \delta_2')}{\delta_2'} \right),$$

where $\delta_i = d_i/(d_1, d_2)$ and $\delta_i' = \delta_i/(q_0, \delta_i)$. Again, as in the proof of Proposition 5.10, we conclude by observing that $[d_1, d_2] \le Q^3 R/q_0$ and $(c_2, \delta_2')/\delta_2' \le 1$, while

$$\frac{(c_1, \delta_1')}{\delta_1'} \le \frac{(c_1, r)}{r},$$

and inspection of the r-component of $\Phi_1(n)\overline{\Phi_2(n)}$ using (5-30) shows that a prime $p \mid r$ divides c_1 if and only if $p \mid h_1s_2 - h_2q_2$. $\qquad\square$

5F. *Proof of second Type I estimate.*

We finish this section with the proof of Theorem 5.8(ii). The idea is very similar to the previous Type I estimate, the main difference being that since q_1 (and q_2) is densely divisible in this case, we can split the sum over q_1 to obtain a better balance of the factors in the Cauchy–Schwarz inequality.

As before, we will prove the bound (5-33) for individual r, and we abbreviate $\Upsilon = \Upsilon_{\ell,r}(b_1, b_2; q_0)$ and set

$$H = x^\varepsilon R Q^2 M^{-1} q_0^{-1}.$$

We may assume that $H \ge 1$, since otherwise the bound is trivial. We note that q_0q_1 is, by assumption, $x^{\delta+o(1)}$-densely divisible, and therefore by Lemma 2.10(i) q_1 is y-densely divisible with $y = q_0 x^{\delta+o(1)}$. Furthermore we have

$$x^{-2\varepsilon} Q/H \gg x^{c-3\varepsilon}$$

by (5-13) and $M \gg x^{1/2+2\varpi+c}$, and

$$x^{-2\varepsilon} Q/H \ll q_1 y = q_1 q_0 x^{\delta+o(1)}$$

since $q_1q_0 \asymp Q$ and $H \ge 1$. Thus (assuming $c > 3\varepsilon$) we have the factorization

$$q_1 = u_1 v_1,$$

where u_1, v_1 are squarefree with

$$q_0^{-1} x^{-\delta-2\varepsilon} Q/H \ll u_1 \ll x^{-2\varepsilon} Q/H,$$
$$q_0^{-1} x^{2\varepsilon} H \ll v_1 \ll x^{\delta+2\varepsilon} H$$

(either from dense divisibility if $x^{-2\varepsilon} Q/H \ll q_1$, or taking $u_1 = q_1$, $v_1 = 1$ otherwise).

Define $\Upsilon_{U,V}$ to be

$$\sum_{1 \le |h| \le H} \sum_{u_1 \asymp U} \sum_{v_1 \asymp V} \sum_{\substack{q_2 \asymp Q/q_0 \\ (u_1 v_1, q_0 q_2)=1}} \left| \sum_n C(n) \beta(n) \overline{\beta(n+\ell r)} \Phi_\ell(h, n, r, q_0, u_1 v_1, q_2) \right|,$$

where u_1, v_1 are understood to be squarefree.

By dyadic decomposition of the sum over $q_1 = u_1 v_1$ in Υ, it is enough to prove that

$$\Upsilon_{U,V} \ll x^{-\varepsilon} (q_0, \ell) Q^2 N q_0^{-2} \tag{5-39}$$

whenever

$$q_0^{-1} x^{-\delta-2\varepsilon} Q/H \ll U \ll x^{-2\varepsilon} Q/H, \tag{5-40}$$
$$q_0^{-1} x^{2\varepsilon} H \ll V \ll x^{\delta+2\varepsilon} H, \tag{5-41}$$
$$UV \asymp Q/q_0. \tag{5-42}$$

We replace the modulus by complex numbers c_{h,u_1,v_1,q_2} of modulus at most 1, move the sum over n, u_1 and q_2 outside and apply the Cauchy–Schwarz inequality as in the previous sections to obtain

$$|\Upsilon_{U,V}|^2 \le \Upsilon_1 \Upsilon_2,$$

with

$$\Upsilon_1 := \sum_{\substack{u_1 \asymp U \\ q_2 \asymp Q/q_0}} \sum_n C(n) |\beta(n)|^2 |\beta(n+\ell r)|^2 \ll (q_0, \ell) \frac{NQU}{q_0^2}$$

as in (5-35) and

$$\Upsilon_2 := \sum_{\substack{u_1 \asymp U \\ q_2 \asymp Q/q_0}} \sum_n \psi_N(n) C(n) \left| \sum_{v_1 \asymp V; (u_1 v_1, q_0 q_2)=1} \sum_{1 \le |h| \le H} \left(c_{h,u_1,v_1,q_2} \right.\right.$$
$$\left.\left. \times \Phi_\ell(h, n, r, q_0, u_1 v_1, q_2) \right) \right|^2$$
$$= \sum_{\substack{u_1 \asymp U \\ q_2 \asymp Q/q_0}} \sum_{v_1, v_2 \asymp V; (u_1 v_1 v_2, q_0 q_2)=1} \sum_{1 \le |h_1|, |h_2| \le H} \left(c_{h_1,u_1,v_1,q_2} \overline{c_{h_2,u_1,v_2,q_2}} \right.$$
$$\left. \times T_{\ell,r}(h_1, h_2, u_1, v_1, v_2, q_2, q_0) \right),$$

where the exponential sum $T_{\ell,r}$ is a variant of $S_{\ell,r}$ given by

$$T_{\ell,r} := \sum_n C(n)\psi_N(n)\Phi_\ell(h_1, n, r, q_0, u_1 v_1, q_2)\overline{\Phi_\ell(h_2, n, r, q_0, u_1 v_2, q_2)}. \quad (5\text{-}43)$$

The analogue of Propositions 5.10 and 5.12 is:

Proposition 5.13. *For any*

$$\boldsymbol{p} = (h_1, h_2, u_1, v_1, v_2, q_2, q_0)$$

with $(u_1 v_1 v_2, q_0 q_2) = (q_0, q_2) = 1$, *any* $\ell \neq 0$ *and* r *as above, we have*

$$|T_{\ell,r}(\boldsymbol{p})| \ll (q_0, \ell)\left(q_0^{-1/2} N^{1/2} x^{\delta/3+\varepsilon/3}(RHQ^2)^{1/6} + \frac{N}{q_0 R}(h_1 v_2 - h_2 v_1, r)\right).$$

Assuming this, we derive as before that

$$\Upsilon_2 \ll (q_0, \ell) H^2 U V^2\left(\frac{Q}{q_0}\right)\left\{N^{1/2}(RHQ^2)^{1/6} x^{\delta/3+\varepsilon/3} + \frac{N}{R}\right\} + HNUV\left(\frac{Q}{q_0^2}\right),$$

and then

$$\begin{aligned}
|\Upsilon_{U,V}|^2 \\
\ll (q_0, \ell)^2 \frac{NQU}{q_0}&\left\{\frac{H^2 Q^3 N^{1/2}(HQ^2 R)^{1/6} x^{\delta/3+\varepsilon/3}}{U q_0^3} + \frac{H^2 N Q^3}{U R q_0^3} + HN\left(\frac{Q^2}{q_0^3}\right)\right\} \\
\ll (q_0, \ell)^2 \frac{N^2 Q^4}{q_0^4}&\left\{\frac{H^{13/6} Q^{1/3} R^{1/6} x^{\delta/3+\varepsilon/3}}{N^{1/2}} + \frac{H^2}{R} + \frac{H}{V q_0}\right\}
\end{aligned}$$

since $UV \asymp Q/q_0$, where we have again discarded a factor of q_0 in the first line. Using again (5-13), (5-14) and (5-41), we find that

$$\begin{aligned}
\frac{H^{13/6} Q^{1/3} R^{1/6} x^{\delta/3+\varepsilon/3}}{N^{1/2}} &\ll x^{\delta+5\varepsilon/2}\frac{R^{7/3} Q^{14/3}}{N^{1/2} M^{13/6}} \ll x^{1/6+28\varpi/3+\delta/3+5\varepsilon/2}\frac{N^{5/3}}{R^{7/3}} \\
&\ll \frac{x^{28\varpi/3+8\delta/3+1/6+19\varepsilon/2}}{N^{2/3}}, \\
\frac{H^2}{R} &\ll \frac{x^{8\varpi+3\delta+11\varepsilon}}{N}, \\
\frac{H}{V q_0} &\ll x^{-2\varepsilon},
\end{aligned}$$

and therefore (5-39) holds for sufficiently small ε provided

$$\begin{cases} \frac{28\varpi}{3} + \frac{8\delta}{3} + \frac{1}{6} < \frac{2}{3}\left(\frac{1}{2} - \sigma\right), \\ 8\varpi + 3\delta < \frac{1}{2} - \sigma, \end{cases} \quad \Longleftrightarrow \quad \begin{cases} 56\varpi + 16\delta + 4\sigma < 1, \\ 16\varpi + 6\delta + 2\sigma < 1. \end{cases}$$

Again the first condition implies the second, and the proof is completed. $\qquad\square$

Proof of Proposition 5.13. We proceed as in the previous cases. Setting

$$\Phi_1(n) := \Phi_\ell(h_1, n, r, q_0, u_1 v_1, q_2), \quad \Phi_2(n) := \Phi_\ell(h_2, n, r, q_0, u_1 v_2, q_2)$$

for brevity, we may write

$$\Phi_1(n)\overline{\Phi_2(n)} = e_{d_1}^{(4)}\left(\frac{c_1}{n}\right) e_{d_2}^{(5)}\left(\frac{c_2}{n+\tau}\right)$$

by (5-30) for some c_1 and c_2 and τ, where

$$d_1 := r q_0 u_1 [v_1, v_2], \quad d_2 := q_2.$$

Since $r q_0 u_1 v_1$, $r q_0 u_1 v_2$ and $r q_0 q_2$ are x^δ-densely divisible, Lemma 2.10(ii) implies that their gcd $[d_1, d_2]$ is also x^δ-densely divisible.

Splitting again the factor $C(n)$ into residue classes modulo q_0, and applying the first part of Corollary 4.16 to each residue class, we obtain

$$|T_{\ell,r}| \ll (q_0, \ell)\left(\frac{N^{1/2}}{q_0^{1/2}}[d_1, d_2]^{1/6} x^{\delta/6} + \frac{N}{q_0}\frac{(c_1, \delta_1')}{\delta_1'}\frac{(c_2, \delta_2')}{\delta_2'}\right),$$

where $\delta_i = d_i/(d_1, d_2)$ and $\delta_i' = \delta_i/(q_0, \delta_i)$. We conclude as before by observing that

$$[d_1, d_2] \ll QRUV^2 \ll x^{\delta+2\varepsilon}\frac{HQ^2R}{q_0},$$

by (5-41) and (5-42), that $(c_2, \delta_2)/\delta_2 \le 1$ and that $(c_1, \delta)/\delta_1 \le (c_1, r)/r$, where inspection of the r-component of $\Phi_1(n)\overline{\Phi_2(n)}$ using (5-30) shows that a prime $p \mid r$ divides c_1 if and only if $p \mid h_1 v_2 - h_2 v_1$. □

6. Trace functions and multidimensional exponential sum estimates

In this section (as in Section 4), we do not use the standard asymptotic convention (Definition 1.2), since we discuss general ideas that are of interest independently of the goal of bounding gaps between primes.

We will discuss some of the machinery and formalism of *ℓ-adic sheaves* $\mathscr{F}$ on curves[4] and their associated *Frobenius trace functions* $t_{\mathscr{F}}$. This will allow us to state and then apply the deep theorems of Deligne's general form of the Riemann hypothesis over finite fields for such sheaves. We will use these theorems to establish certain estimates for multivariable exponential sums which go beyond the one-dimensional estimates obtainable from Lemma 4.2 (specifically, the estimates we need are stated in Corollary 6.24 and Corollary 6.26).

---

[4]In our applications, the only curves U we deal with are obtained by removing a finite number of points from the projective line $\mathbb{P}^1$.

The point is that these Frobenius trace functions significantly generalize the rational phase functions $x \mapsto e_p(P(x)/Q(x))$ which appear in Lemma 4.2. They include more general functions, such as the hyper-Kloosterman sums

$$x \mapsto \frac{(-1)^{m-1}}{p^{\frac{m-1}{2}}} \sum_{\substack{y_1,\dots,y_m \in \mathbb{F}_p \\ y_1 \cdots y_m = x}} \cdots \sum e_p(y_1 + \cdots + y_m),$$

and satisfy a very flexible formalism. In particular, the class of Frobenius trace functions is (essentially) closed under basic operations such as pointwise addition and multiplication, complex conjugation, change of variable (pullback), and the normalized Fourier transform. Using these closure properties allows us to build a rich class of useful trace functions from just a small set of basic trace functions. In fact, the sheaves we actually use in this paper are ultimately obtained from only two sheaves: the Artin–Schreier sheaf and the third hyper-Kloosterman sheaf.[5] However, we have chosen to discuss more general sheaves in this section in order to present the sheaf-theoretic framework in a more natural fashion.

Because exponential sums depending on a parameter are often themselves trace functions, one can recast many multidimensional exponential sums (e.g.,

$$\sum_{x_1,\dots,x_n \in \mathbb{F}_p} e_p(f(x_1, \dots, x_n))$$

for some rational function $f \in \mathbb{F}_p(X_1, \dots, X_n))$ in terms of one-dimensional sums of Frobenius trace functions. As a very rough first approximation, [Deligne 1980] implies that the square root cancellation exhibited in Lemma 4.2 is also present for these more general sums of Frobenius trace functions, as long as certain degenerate cases are avoided. Therefore, at least in principle, this implies square root cancellation for many multidimensional exponential sums.

In practice, this is often not entirely straightforward, as we will explain. One particular issue is that the bounds provided by Deligne's theorems depend on a certain measure of complexity of the ℓ-adic sheaf defining the trace function, which is known as the *conductor* of a sheaf. In estimates for sums of trace functions, this conductor plays the same role that the degrees of the polynomials f, g play in Lemma 4.2. We will therefore have to expend some effort to control the conductors of various sheaves before we can extract usable estimates from Deligne's results.

This section is not self-contained, and assumes a certain amount of prior formal knowledge of the terminology of ℓ-adic cohomology on curves. For readers who are not familiar with this material, we would recommend as references such surveys

---

[5]One can even reduce the number of generating sheaves to one, because the sheaf-theoretic Fourier transform, combined with pullback via the inversion map $x \mapsto 1/x$, may be used to iteratively build the hyper-Kloosterman sheaves from the Artin–Schreier sheaf.

as [Iwaniec and Kowalski 2004, §11.11; Kowalski 2010; Fouvry et al. 2014c], and some of the books and papers of Katz, in particular [1980; 2001; 1988], as well as Deligne's own account [SGA 1977, Sommes trigonométriques]. We would like to stress that if the main results of the theory are assumed and the construction of some main objects (e.g., the Artin–Schreier and hyper-Kloosterman sheaves) is accepted, working with ℓ-adic sheaves essentially amounts to studying certain finite-dimensional continuous representations of the Galois group of the field $\mathbb{F}_p(X)$ of rational functions over $\mathbb{F}_p$.

Alternatively, for the purposes of establishing only the bounds on (incomplete) multivariable exponential sums used in the proofs of the main theorems of this paper (namely the bounds in Corollary 6.24 and Corollary 6.26), it is possible to ignore all references to sheaves, if one accepts the estimates on complete multidimensional exponential sums in Proposition 6.11 and Theorem 6.17 as "black boxes"; the estimates on incomplete exponential sums will be deduced from these results via completion of sums and the q-van der Corput A-process.

6A. *ℓ-adic sheaves on the projective line.* For p a prime, we fix an algebraic closure $\overline{\mathbb{F}}_p$ of $\mathbb{F}_p$ and denote by $k \subset \overline{\mathbb{F}}_p = \overline{k}$ a finite extension of $\mathbb{F}_p$. Its cardinality is usually denoted $|k| = p^{[k:\mathbb{F}_p]} = p^{\deg(k)} = q$. For us, the Frobenius element relative to k means systematically the *geometric Frobenius* Fr_k, which is the inverse in $\mathrm{Gal}(\overline{k}/k)$ of the *arithmetic Frobenius*, $x \mapsto x^q$ on $\overline{k}$.

We denote by $K = \mathbb{F}_p(t)$ the function field of the projective line $\mathbb{P}^1_{\mathbb{F}_p}$ and by $\overline{K} \supset \overline{\mathbb{F}}_p$ some separable closure; let $\overline{\eta} = \mathrm{Spec}(\overline{K})$ be the corresponding geometric generic point.

We fix another prime $\ell \neq p$, and we denote by $\iota : \overline{\mathbb{Q}}_\ell \hookrightarrow \mathbb{C}$ an algebraic closure of the field $\mathbb{Q}_\ell$ of ℓ-adic numbers, together with an embedding into the complex numbers. By an *ℓ-adic sheaf* $\mathscr{F}$ on a noetherian scheme X (in practice, a curve), we always mean a constructible sheaf of finite-dimensional $\overline{\mathbb{Q}}_\ell$-vector spaces with respect to the étale topology on X, and we recall that the category of ℓ-adic sheaves is abelian.

We will be especially interested in the case $X = \mathbb{P}^1_k$ (the projective line) and we will use the following notation for the translation, dilation, and fractional linear maps from $\mathbb{P}^1$ to itself:

$$[+l] : x \mapsto x + l,$$
$$[\times a] : x \mapsto ax,$$
$$\gamma : x \mapsto \gamma \cdot x = \frac{ax+b}{cx+d} \quad \text{for } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathbb{F}_p).$$

We will often transform a sheaf $\mathscr{F}$ on $\mathbb{P}^1_k$ by applying pullback by one of the above maps, and we denote these pullback sheaves by $[+l]^*\mathscr{F}$, $[\times a]^*\mathscr{F}$ and $\gamma^*\mathscr{F}$.

6A.1. *Galois representations.* The category of ℓ-adic sheaves on $\mathbb{P}_k^1$ admits a relatively concrete description in terms of representations of the Galois group $\mathrm{Gal}(\overline{K}/k.K)$. We recall some important features of it here, and we refer to [Katz 1980, 4.4] for a complete presentation.

For $j : U \hookrightarrow \mathbb{P}_k^1$ some nonempty open subset defined over k, we denote by $\pi_1(U)$ and $\pi_1^g(U)$ the *arithmetic* and *geometric fundamental groups* of U, which may be defined as the quotients of $\mathrm{Gal}(\overline{K}/k.K)$ and $\mathrm{Gal}(\overline{K}/\overline{k}.K)$, respectively, by the smallest closed normal subgroup containing all the inertia subgroups above the closed points of U. We have then a commutative diagram of short exact sequences of groups

$$
\begin{array}{ccccccccc}
1 & \longrightarrow & \mathrm{Gal}(\overline{K}/\overline{k}.K) & \longrightarrow & \mathrm{Gal}(\overline{K}/k.K) & \longrightarrow & \mathrm{Gal}(\overline{k}/k) & \longrightarrow & 1 \\
 & & \downarrow & & \downarrow & & \| & & \\
1 & \longrightarrow & \pi_1^g(U) & \longrightarrow & \pi_1(U) & \longrightarrow & \mathrm{Gal}(\overline{k}/k) & \longrightarrow & 1
\end{array}
\qquad (6\text{-}1)
$$

Given an ℓ-adic sheaf $\mathscr{F}$ on $\mathbb{P}_k^1$, there exists some nonempty (hence dense, in the Zariski topology) open set $j : U \hookrightarrow \mathbb{P}_k^1$ such that the pullback $j^*\mathscr{F}$ (the restriction of $\mathscr{F}$ to U) is *lisse*, or in other words, for which $j^*\mathscr{F}$ "is" a finite-dimensional continuous representation $\rho_\mathscr{F}$ of $\mathrm{Gal}(\overline{K}/k.K)$ factoring through $\pi_1(U)$

$$
\rho_\mathscr{F} : \mathrm{Gal}(\overline{K}/k.K) \twoheadrightarrow \pi_1(U) \to \mathrm{GL}(\mathscr{F}_{\overline{\eta}}),
$$

where the geometric generic stalk $\mathscr{F}_{\overline{\eta}}$ of $\mathscr{F}$ is a finite-dimensional $\overline{\mathbb{Q}}_\ell$-vector space. Its dimension is the (*generic*) *rank* of $\mathscr{F}$ and is denoted $\mathrm{rk}(\mathscr{F})$. There is a maximal (with respect to inclusion) open subset on which $\mathscr{F}$ is lisse, which will be denoted by $U_\mathscr{F}$.

We will freely apply the terminology of representations to ℓ-adic sheaves. The properties of $\rho_\mathscr{F}$ as a representation of the arithmetic Galois group $\mathrm{Gal}(\overline{K}/k.K)$ (or of the arithmetic fundamental group $\pi_1(U)$) will be qualified as "arithmetic", while the properties of its restriction $\rho_\mathscr{F}^g$ to the geometric Galois group $\mathrm{Gal}(\overline{K}/\overline{k}.K)$ (or the geometric fundamental group $\pi_1^g(U)$) will be qualified as "geometric". For instance, we will say that $\mathscr{F}$ is *arithmetically irreducible* (resp. *geometrically irreducible*) or *arithmetically isotypic* (resp. *geometrically isotypic*) if the corresponding arithmetic representation $\rho_\mathscr{F}$ (resp. the geometric representation $\rho_\mathscr{F}^g$) is.

We will be mostly interested in the geometric properties of a sheaf; therefore we will usually omit the adjective "geometric" in our statements, so that "isotypic" will mean "geometrically isotypic". We will always spell out explicitly when an arithmetic property is intended, so that no confusion can arise.

6A.2. *Middle-extension sheaves.* An ℓ-adic sheaf is a called a *middle-extension* sheaf if, for some (and in fact, for any) nonempty open subset $j : U \hookrightarrow \mathbb{P}^1_k$ such that $j^*\mathscr{F}$ is lisse, we have an arithmetic isomorphism

$$\mathscr{F} \simeq j_* j^* \mathscr{F},$$

or equivalently if, for every $\overline{x} \in \mathbb{P}^1(\overline{k})$, the specialization maps (see [Katz 1980, 4.4])

$$s_{\overline{x}} : \mathscr{F}_{\overline{x}} \to \mathscr{F}_{\overline{\eta}}^{I_{\overline{x}}}$$

are isomorphisms, where $I_{\overline{x}}$ is the inertia subgroup at $\overline{x}$. Given an ℓ-adic sheaf, its associated middle-extension is the sheaf

$$\mathscr{F}^{\mathrm{me}} = j_* j^* \mathscr{F}$$

for some nonempty open subset $j : U \hookrightarrow \mathbb{P}^1_k$ on which $\mathscr{F}$ is lisse. This sheaf is a middle-extension sheaf, and is (up to arithmetic isomorphism) the unique middle-extension sheaf whose restriction to U is arithmetically isomorphic to that of $\mathscr{F}$. In particular, $\mathscr{F}^{\mathrm{me}}$ does not depend on the choice of U.

6B. *The trace function of a sheaf.* Let $\mathscr{F}$ be an ℓ-adic sheaf on the projective line over $\mathbb{F}_p$. For each finite extension $k/\mathbb{F}_p$, $\mathscr{F}$ defines a complex valued function

$$x \mapsto t_{\mathscr{F}}(x; k)$$

on $k \cup \{\infty\} = \mathbb{P}^1(k)$, which is called the *Frobenius trace function*, or just *trace function*, associated with $\mathscr{F}$ and k. It is defined by

$$\mathbb{P}^1(k) \ni x \mapsto t_{\mathscr{F}}(x; k) := \iota(\mathrm{Tr}(\mathrm{Fr}_{x,k} | \mathscr{F}_{\overline{x}})).$$

Here $\overline{x} : \mathrm{Spec}(\overline{k}) \to \mathbb{P}^1_k$ denotes a geometric point above x, and $\mathscr{F}_{\overline{x}}$ is the stalk of $\mathscr{F}$ at that point, which is a finite-dimensional $\overline{\mathbb{Q}}_\ell$-vector space on which $\mathrm{Gal}(\overline{k}/k)$ acts linearly, and $\mathrm{Fr}_{x,k}$ denotes the geometric Frobenius of that Galois group. The trace of the action of this operator is independent of the choice of $\overline{x}$.

If $k = \mathbb{F}_p$, which is the case of importance for the applications in this paper, we will write $t_{\mathscr{F}}(x; p)$ or simply $t_{\mathscr{F}}(x)$ instead of $t_{\mathscr{F}}(x; \mathbb{F}_p)$.

If $x \in U_{\mathscr{F}}(k)$, the quantity $t_{\mathscr{F}}(x; k)$ is simply the trace of the geometric Frobenius conjugacy class of a place of $\overline{K}$ above x acting through the associated representation $\mathscr{F}_{\overline{\eta}}$, i.e., the value (under ι) of the character of the representation at this conjugacy class:

$$t_{\mathscr{F}}(x; k) = \iota(\mathrm{Tr}(\mathrm{Fr}_{x,k} | \mathscr{F}_{\overline{\eta}})).$$

If $\mathscr{F}$ is a middle-extension sheaf one has more generally

$$t_{\mathscr{F}}(x; k) = \iota(\mathrm{Tr}(\mathrm{Fr}_{x,k} | \mathscr{F}_{\overline{\eta}}^{I_{\overline{x}}})).$$

For any sheaf $\mathcal{F}$, the trace function of $\mathcal{F}$ restricted to $U_{\mathcal{F}}(k)$ coincides with the restriction of the trace function of $\mathcal{F}^{\mathrm{me}}$.

6B.1. *Purity and admissibility.* The following notion was introduced in [Deligne 1980].

Definition 6.1 (purity). For $i \in \mathbb{Z}$, an ℓ-adic sheaf on $\mathbb{P}^1_{\mathbb{F}_p}$ is *generically pure* (or *pure*, for short) of weight i if, for any $k/\mathbb{F}_p$ and any $x \in U_{\mathcal{F}}(k)$, the eigenvalues of $\mathrm{Fr}_{x,k}$ acting on $\mathcal{F}_{\overline{\eta}}$ are $\mathbb{Q}$-algebraic numbers whose Galois conjugates have complex absolute value equal to $q^{i/2} = |k|^{i/2}$.

Remark 6.2. Deligne proved (see [1980, (1.8.9)]) that if $\mathcal{F}$ is a generically pure middle-extension sheaf of weight i, then for any $k/\mathbb{F}_p$ and any $x \in \mathbb{P}^1(k)$, the eigenvalues of $\mathrm{Fr}_{x,k}$ acting on $\mathcal{F}_{\overline{\eta}}^{I_{\overline{x}}}$ are $\mathbb{Q}$-algebraic numbers whose Galois conjugates have complex absolute value $\leq q^{i/2}$.

In particular, if $\mathcal{F}$ is a middle-extension sheaf which is generically pure of weight i, then we get

$$|t_{\mathcal{F}}(x; k)| = \left| \iota\big(\mathrm{Tr}\big(\mathrm{Fr}_x \,|\, \mathcal{F}_{\overline{\eta}}^{I_{\overline{x}}}\big)\big)\right| \leq \mathrm{rk}(\mathcal{F})q^{i/2} \tag{6-2}$$

for any $x \in \mathbb{P}^1(k)$.

We can now describe the class of sheaves and trace functions that we will work with.

Definition 6.3 (admissible sheaves). Let k be a finite extension of $\mathbb{F}_p$. An *admissible sheaf* over k is a middle-extension sheaf on $\mathbb{P}^1_k$ which is pointwise pure of weight 0. An *admissible trace function* over k is a function $k \to \mathbb{C}$ which is equal to the trace function of some admissible sheaf restricted to $k \subset \mathbb{P}^1(k)$.

Remark 6.4. The weight-0 condition may be viewed as a normalization to ensure that admissible trace functions typically have magnitude comparable to 1. Sheaves which are pure of some other weight can be studied by reducing to the 0 case by the simple device of *Tate twists*. However, we will not need to do this, as we will be working exclusively with sheaves which are pure of weight 0.

6B.2. *Conductor.* Let $\mathcal{F}$ be a middle-extension sheaf on $\mathbb{P}^1_k$. The *conductor* of $\mathcal{F}$ is defined as

$$\mathrm{cond}(\mathcal{F}) := \mathrm{rk}(\mathcal{F}) + |(\mathbb{P}^1 - U_{\mathcal{F}})(\overline{k})| + \sum_{x \in (\mathbb{P}^1 - U_{\mathcal{F}})(\overline{k})} \mathrm{swan}_x(\mathcal{F}),$$

where $\mathrm{swan}_x(\mathcal{F})$ denotes the Swan conductor of the representation $\rho_{\mathcal{F}}$ at x, a nonnegative integer measuring the "wild ramification" of $\rho_{\mathcal{F}}$ at x (see, e.g., [Katz 1988, Definition 1.6] for the precise definition of the Swan conductor). If $\mathrm{swan}_x(\mathcal{F}) = 0$, one says that $\mathcal{F}$ is *tamely ramified* at x, and otherwise that it is *wildly ramified*.

The invariant $\mathrm{cond}(\mathcal{F})$ is a nonnegative integer (positive if $\mathcal{F} \neq 0$), and it measures the complexity of the sheaf $\mathcal{F}$ and of its trace function $t_{\mathcal{F}}$. For instance, if $\mathcal{F}$ is admissible, so that it is also pure of weight 0, then we deduce from (6-2) that

$$|t_{\mathcal{F}}(x; k)| \leq \mathrm{rk}(\mathcal{F}) \leq \mathrm{cond}(\mathcal{F}) \qquad (6\text{-}3)$$

for any $x \in k$.

6B.3. *Dual and tensor Product.* Given admissible sheaves $\mathcal{F}$ and $\mathcal{G}$ on $\mathbb{P}^1_k$, their tensor product, denoted by $\mathcal{F} \otimes \mathcal{G}$, is by definition the middle-extension sheaf associated to the tensor product representation $\rho_{\mathcal{F}} \otimes \rho_{\mathcal{G}}$ (computed over the intersection of $U_{\mathcal{F}}$ and $U_{\mathcal{G}}$, which is still a dense open set of $\mathbb{P}^1_k$). Note that this sheaf may be different from the tensor product of $\mathcal{F}$ and $\mathcal{G}$ as constructible sheaves (similarly to the fact that the product of two primitive Dirichlet characters is not necessarily primitive).

Similarly, the dual of $\mathcal{F}$, denoted $\check{\mathcal{F}}$, is defined as the middle extension sheaf associated to the contragredient representation $\widetilde{\rho_{\mathcal{F}}}$.

We have

$$U_{\mathcal{F}} \cap U_{\mathcal{G}} \subset U_{\mathcal{F} \otimes \mathcal{G}}, \qquad U_{\check{\mathcal{F}}} = U_{\mathcal{F}}.$$

It is not obvious, but true, that tensor products and duals of admissible sheaves are admissible. We then have

$$t_{\mathcal{F} \otimes \mathcal{G}}(x; k) = t_{\mathcal{F}}(x; k) t_{\mathcal{G}}(x; k), \qquad t_{\check{\mathcal{F}}}(x; k) = \overline{t_{\mathcal{F}}(x; k)} \qquad (6\text{-}4)$$

for $x \in U_{\mathcal{F}}(k) \cap U_{\mathcal{G}}(k)$ and $x \in \mathbb{P}^1(k)$, respectively. In particular, the product of two admissible trace functions $t_{\mathcal{F}}$ and $t_{\mathcal{G}}$ coincides with an admissible trace function outside a set of at most $\mathrm{cond}(\mathcal{F}) + \mathrm{cond}(\mathcal{G})$ elements, and the complex conjugate of an admissible trace function is again an admissible trace function.

We also have

$$\mathrm{cond}(\check{\mathcal{F}}) = \mathrm{cond}(\mathcal{F}) \qquad (6\text{-}5)$$

(which is easy to check from the definition of Swan conductors) and

$$\mathrm{cond}(\mathcal{F} \otimes \mathcal{G}) \ll \mathrm{rk}(\mathcal{F}) \, \mathrm{rk}(\mathcal{G}) \, \mathrm{cond}(\mathcal{F}) \, \mathrm{cond}(\mathcal{G}) \leq \mathrm{cond}(\mathcal{F})^2 \, \mathrm{cond}(\mathcal{G})^2, \qquad (6\text{-}6)$$

where the implied constant is absolute (which is also relatively elementary; see [Fouvry et al. 2014a, Proposition 8.2(2)] or [Fouvry et al. 2013b, Lemma 4.8]).

6C. *Irreducible components and isotypic decomposition.* Let k be a finite field, let $\mathcal{F}$ be an admissible sheaf over $\mathbb{P}^1_k$, and consider $U = U_{\mathcal{F}}$ and the corresponding open immersion $j : U \hookrightarrow \mathbb{P}^1_k$. A fundamental result of Deligne [1980, (3.4.1)] proves that $\rho_{\mathcal{F}}$ is then geometrically semisimple. Thus there exist lisse sheaves $\mathcal{G}$ on $U \times \bar{k}$, irreducible and pairwise nonisomorphic, and integers $n(\mathcal{G}) \geq 1$, such that

$$j^* \mathcal{F} \simeq \bigoplus_{\mathcal{G}} \mathcal{G}^{n(\mathcal{G})}$$

as an isomorphism of lisse sheaves on $U \times \bar{k}$ (the $\mathcal{G}$ might not be defined over k). Extending with j_* to $\mathbb{P}^1_{\bar{k}}$, we obtain a decomposition

$$\mathcal{F} \simeq \bigoplus_{\mathcal{G}} j_* \mathcal{G}^{n(\mathcal{G})},$$

where each $j_* \mathcal{G}$ is a middle-extension sheaf over $\bar{k}$. We call the sheaves $j_* \mathcal{G}$ the *geometrically irreducible components* of $\mathcal{F}$.

Over the open set $U_{\mathcal{F}}$, we can define the arithmetic semisimplification $\rho_{\mathcal{F}}^{ss}$ as the direct sum of the Jordan–Hölder arithmetically irreducible components of the representation $\rho_{\mathcal{F}}$. Each arithmetically irreducible component is either geometrically isotypic or induced from a proper finite index subgroup of $\pi_1(U_{\mathcal{F}})$. If an arithmetically irreducible component π is induced, it follows that the trace function of the middle-extension sheaf corresponding to π vanishes identically. Thus, if we denote by $\mathrm{Iso}(\mathcal{F})$ the set of middle-extensions associated to the geometrically isotypic components of $\rho_{\mathcal{F}}^{ss}$, we obtain the identity

$$t_{\mathcal{F}} = \sum_{\mathcal{G} \in \mathrm{Iso}(\mathcal{F})} t_{\mathcal{G}} \tag{6-7}$$

(indeed, these two functions coincide on $U_{\mathcal{F}}$ and are both trace functions of middle-extension sheaves), where each summand is admissible. For these facts, we refer to [Katz 1980, §4.4, §4.5] and [Fouvry et al. 2014a, Proposition 8.3].

6D. *Deligne's main theorem and quasiorthogonality.* The generalizations of complete exponential sums over finite fields that we consider are sums

$$S(\mathcal{F}; k) = \sum_{x \in k} t_{\mathcal{F}}(x; k)$$

for any admissible sheaf $\mathcal{F}$ over $\mathbb{P}^1_k$. By (6-3), we have the trivial bound

$$|S(\mathcal{F}; k)| \leq \mathrm{cond}(\mathcal{F})|k| = \mathrm{cond}(\mathcal{F})q.$$

Deligne's main theorem [1980, Théorème 1] provides strong nontrivial estimates for such sums, at least when p is large compared to $\mathrm{cond}(\mathcal{F})$.

Theorem 6.5 (sums of trace functions). *Let $\mathcal{F}$ be an admissible sheaf on $\mathbb{P}^1_k$, where $|k| = q$ and $U = U_{\mathcal{F}}$. We have*

$$S(\mathcal{F}; k) = q \, \mathrm{Tr}\big(\mathrm{Fr}_k \,|(\mathcal{F}_{\bar{\eta}})_{\pi_1^g(U)}\big) + O\big(\mathrm{cond}(\mathcal{F})^2 q^{1/2}\big),$$

where $(\mathcal{F}_{\bar{\eta}})_{\pi_1^g(U)}$ denotes the $\pi_1^g(U_{\mathcal{F}})$-coinvariant space[6] of $\rho_{\mathcal{F}}$, on which $\mathrm{Gal}(\bar{k}/k)$ acts canonically, and where the implied constant is effective and absolute.

---

[6]Recall that the coinvariant space of a representation of a group G is the largest quotient on which the group G acts trivially.

Proof. Using (6-3), we have

$$S(\mathscr{F}; k) = \sum_{x \in U(k)} t_{\mathscr{F}}(x; k) + O(\operatorname{cond}(\mathscr{F})^2),$$

where the implied constant is at most 1. The Grothendieck–Lefschetz trace formula (see, e.g., [Katz 1988, Chapter 3]) gives

$$S_{\mathscr{F}}(U, k) = \sum_{i=0}^{2} (-1)^i \operatorname{Tr}(\operatorname{Fr}_k \mid H_c^i(U \otimes_k \overline{k}, \mathscr{F})),$$

where $H_c^i(U \otimes_k \overline{k}, \mathscr{F})$ is the i-th compactly supported étale cohomology group of the base change of U to $\overline{k}$ with coefficients in $\mathscr{F}$, on which the global Frobenius automorphism Fr_k acts.

Since U is affine and $\mathscr{F}$ is lisse on U, it is known that $H_c^0(U \otimes_k \overline{k}, \mathscr{F}) = 0$. For $i = 1$, Deligne's main theorem shows that, because $\mathscr{F}$ is of weight 0, all eigenvalues of Fr_k acting on $H_c^1(U \times_k \overline{k}, \mathscr{F})$ are algebraic numbers with complex absolute value $\leq |k|^{1/2}$, so that

$$|\operatorname{Tr}(\operatorname{Fr}_k \mid H_c^1(U \otimes_k \overline{k}, \mathscr{F}))| \leq \dim(H_c^1(U \otimes_k \overline{k}, \mathscr{F})) q^{1/2}.$$

Using the Euler–Poincaré formula and the definition of the conductor, one easily obtains

$$\dim(H_c^1(U \otimes_k \overline{k}, \mathscr{F})) \ll \operatorname{cond}(\mathscr{F})^2$$

with an absolute implied constant (see, e.g., [Katz 1988, Chapter 2] or [Fouvry et al. 2013a, Theorem 2.4]).

Finally for $i = 2$, it follows from Poincaré duality that $H_c^2(U \otimes_k \overline{k}, \mathscr{F})$ is isomorphic to the Tate-twisted space of $\pi_1^g(U)$-coinvariants of $\mathscr{F}_{\overline{\eta}}$ (see, e.g., [Katz 1988, Chapter 2]), and hence the contribution of this term is the main term in the formula. $\qquad\square$

6D.1. *Correlation and quasiorthogonality of trace functions.* An important application of the above formula arises when estimating the *correlation* between the trace functions $t_{\mathscr{F}}$ and $t_{\mathscr{G}}$ associated to two admissible sheaves $\mathscr{F}, \mathscr{G}$, i.e., when computing the sum associated to the tensor product sheaf $\mathscr{F} \otimes \check{\mathscr{G}}$. We define the correlation sum

$$C(\mathscr{F}, \mathscr{G}; k) := \sum_{x \in k} t_{\mathscr{F}}(x; k) \overline{t_{\mathscr{G}}(x; k)}.$$

From (6-3) we have the trivial bound

$$|C_{\mathscr{F}, \mathscr{G}}(k)| \leq \operatorname{cond}(\mathscr{F}) \operatorname{cond}(\mathscr{G}) q.$$

The Riemann hypothesis allows us improve this bound when $\mathscr{F}, \mathscr{G}$ are "disjoint":

Corollary 6.6 (square root cancellation). *Let $\mathscr{F}, \mathscr{G}$ be two admissible sheaves on $\mathbb{P}^1_k$ for a finite field k. If $\mathscr{F}$ and $\mathscr{G}$ have no irreducible constituent in common, then we have*

$$|C(\mathscr{F}, \mathscr{G}; k)| \ll (\mathrm{cond}(\mathscr{F})\,\mathrm{cond}(\mathscr{G}))^4 q^{1/2},$$

where the implied constant is absolute. In particular, if in addition $\mathrm{cond}(\mathscr{F})$ and $\mathrm{cond}(\mathscr{G})$ are bounded by a fixed constant, then

$$|C(\mathscr{F}, \mathscr{G}; k)| \ll q^{1/2}.$$

Proof. We have

$$t_{\mathscr{F} \otimes \check{\mathscr{G}}}(x; k) = t_{\mathscr{F}}(x; k)\overline{t_{\mathscr{G}}(x; k)}$$

for $x \in U_{\mathscr{F}}(k) \cap U_{\mathscr{G}}(k)$ and

$$|t_{\mathscr{F} \otimes \check{\mathscr{G}}}(x; k)|, \quad |t_{\mathscr{F}}(x; k)\overline{t_{\mathscr{G}}(x; k)}| \leq \mathrm{cond}(\mathscr{F})\,\mathrm{cond}(\mathscr{G}).$$

Thus the previous proposition applied to the sheaf $\mathscr{F} \otimes \check{\mathscr{G}}$ gives

$$C(\mathscr{F}, \mathscr{G}; k) = S(\mathscr{F} \otimes \check{\mathscr{G}}; k) + O\big((\mathrm{cond}(\mathscr{F}) + \mathrm{cond}(\mathscr{G}))\,\mathrm{cond}(\mathscr{F})\,\mathrm{cond}(\mathscr{G})\big)$$

$$= q\,\mathrm{Tr}\big(\mathrm{Fr}_k\,|((\mathscr{F} \otimes \check{\mathscr{G}})_{\overline{\eta}})_{\pi_1^g(U)}\big) + O((\mathrm{cond}(\mathscr{F})\,\mathrm{cond}(\mathscr{G}))^4 q^{1/2})$$

using (6-5) and (6-6). We conclude by observing that, by Schur's Lemma and the geometric semisimplicity of admissible sheaves (proved by Deligne [1980, (3.4.1)]), our disjointness assumption on $\mathscr{F}$ and $\mathscr{G}$ implies that the coinvariant space vanishes. $\square$

6E. *The Artin–Schreier sheaf.* We will now start discussing specific important admissible sheaves. Let p be a prime and let $\psi : (\mathbb{F}_p, +) \to \mathbb{C}^\times$ be a nontrivial additive character. For any finite extension k of $\mathbb{F}_p$, we then have an additive character

$$\psi_k : \begin{cases} k \to \mathbb{C}^\times, \\ x \mapsto \psi(\mathrm{Tr}_{k/\mathbb{F}_p}(x)), \end{cases}$$

where $\mathrm{Tr}_{k/\mathbb{F}_p}$ is the trace map from k to $\mathbb{F}_p$.

One shows (see [Katz 1988, Chapter 4; SGA 1977, §1.4; Iwaniec and Kowalski 2004, pp. 302–303]) that there exists an admissible sheaf $\mathscr{L}_\psi$, called the *Artin–Schreier sheaf* associated to ψ, with the following properties:

- The sheaf $\mathscr{L}_\psi$ has rank 1, hence is automatically geometrically irreducible, and it is geometrically nontrivial.

- The sheaf $\mathscr{L}_\psi$ is lisse on $\mathbb{A}^1_{\mathbb{F}_p}$, and wildly ramified at ∞ with $\mathrm{swan}_\infty(\mathscr{L}_\psi) = 1$, so that in particular $\mathrm{cond}(\mathscr{L}_\psi) = 3$, independently of p and of the nontrivial additive character ψ.

- The trace function is given by the formula

$$t_{\mathcal{L}_\psi}(x; k) = \psi_k(x)$$

for every finite extension $k/\mathbb{F}_p$ and every $x \in \mathbb{A}^1(k) = k$, and

$$t_{\mathcal{L}_\psi}(\infty; k) = 0.$$

Let $f \in \mathbb{F}_p(X)$ be a rational function not of the shape $g^p - g + c$ for $g \in \mathbb{F}_p(X)$, $c \in \mathbb{F}_p$ (for instance whose zeros or poles have order prime to p). Then f defines a morphism $f : \mathbb{P}^1_{\mathbb{F}_p} \to \mathbb{P}^1_{\mathbb{F}_p}$, and we denote by $\mathcal{L}_{\psi(f)}$ the pull-back sheaf $f^*\mathcal{L}_\psi$, which we call the *Artin–Schreier sheaf associated to f and ψ*. Then $\mathcal{L}_{\psi(f)}$ has the following properties:

- It has rank 1, hence is geometrically irreducible, and it is geometrically non-trivial (because f is not of the form $g^p - g + c$ for some other function g, by assumption).

- It is lisse outside the poles of f, and wildly ramified at each pole with Swan conductor equal to the order of the pole, so that if the denominator of f has degree d (coprime to p) we have $\text{cond}(\mathcal{L}_{\psi(f)}) = 1 + e + d$, where e is the number of distinct poles of f.

- It has trace function given by the formula

$$t_{\mathcal{L}_{\psi(f)}}(x; k) = \psi\left(\text{tr}_{k/\mathbb{F}_p}(f(x))\right)$$

for any finite extension $k/\mathbb{F}_p$ and any $x \in \mathbb{P}^1(k)$ which is not a pole of f, and $t_{\mathcal{L}_{\psi(f)}}(x; k) = 0$ if x is a pole of f.

In particular, from Theorem 6.5, we thus obtain the estimate

$$\left| \sum_{x \in \mathbb{F}_p} \psi(f(x)) \right| \ll \deg(f)^2 p^{1/2}$$

for such f, which is a slightly weaker form of the Weil bound from Lemma 4.2. Note that this weakening, which is immaterial in our applications, is only due to the general formulation of Theorem 6.5, which did not attempt to obtain the best possible estimate for specific situations.

6F. *The ℓ-adic Fourier transform.* Let p be a prime, $k/\mathbb{F}_p$ a finite extension and ψ a nontrivial additive character of k. For a finite extension $k/\mathbb{F}_p$ and a function $x \mapsto t(x)$ defined on k, we define the *normalized Fourier transform* $\text{FT}_\psi t(x)$ by the formula

$$\text{FT}_\psi t(x) := -\frac{1}{q^{1/2}} \sum_{y \in k} t(y)\psi(xy)$$

(which is similar to (4-11) except for the sign). It is a very deep fact that, when applied to trace functions, this construction has a sheaf-theoretic incarnation. This was defined by Deligne and studied extensively by Laumon [1987] and Katz [1988]. However, a restriction on the admissible sheaves is necessary, in view of the following obstruction: if $t(x) = \psi(bx)$ for some $b \in k$, then its Fourier transform is a Dirac-type function

$$\mathrm{FT}_\psi(t)(x) = -q^{1/2}\delta_{-b}(x) = \begin{cases} -q^{1/2} & \text{if } x = -b, \\ 0 & \text{otherwise.} \end{cases}$$

But this cannot in general be an admissible trace function with bounded conductor as this would violate (6-2) at $x = -b$ if q is large enough. We make the following definition, as in [Katz 1988]:

Definition 6.7 (admissible Fourier sheaves). An admissible sheaf over $\mathbb{P}^1_k$ is a *Fourier sheaf* if its geometrically irreducible components are neither trivial nor Artin–Schreier sheaves $\mathcal{L}_\psi$ for some nontrivial additive character ψ.

Theorem 6.8 (sheaf-theoretic Fourier transform). *Let p be a prime and $k/\mathbb{F}_p$ a finite extension, and let ψ be a nontrivial additive character of k. Let $\mathcal{F}$ be an admissible ℓ-adic Fourier sheaf on $\mathbb{P}^1_k$. There exists an ℓ-adic sheaf*

$$\mathcal{G} = \mathrm{FT}_\psi(\mathcal{F}),$$

called the Fourier transform of $\mathcal{F}$, *which is also an admissible ℓ-adic Fourier sheaf, with the property that for any finite extension k'/k, we have*

$$t_\mathcal{G}(\,\cdot\,;k') = \mathrm{FT}_{\psi_{k'}}t_\mathcal{F}(\,\cdot\,;k);$$

in particular

$$t_\mathcal{G}(x;k) = -\frac{1}{\sqrt{|k|}}\sum_{y\in k} t_\mathcal{F}(y;k)\psi(xy).$$

Moreover, the following additional assertions hold:

- *The sheaf $\mathcal{G}$ is geometrically irreducible, or geometrically isotypic, if and only if $\mathcal{F}$ is.*

- *The Fourier transform is (almost) involutive, in the sense that we have a canonical arithmetic isomorphism*

$$\mathrm{FT}_\psi\mathcal{G} \simeq [\times(-1)]^*\mathcal{F}, \tag{6-8}$$

where $[\times(-1)]^*$ denotes the pull-back by the map $x \mapsto -x$.

- *We have*

$$\mathrm{cond}(\mathcal{G}) \leq 10\,\mathrm{cond}(\mathcal{F})^2. \tag{6-9}$$

Proof. These claims are established for instance in [Katz 1988, Chapter 8], with the exception of (6-9), which is proved in [Fouvry et al. 2014a, Proposition 8.2(1)]. □

6G. *Kloosterman sheaves.* Given a prime $p \geq 3$, a nontrivial additive character ψ of $\mathbb{F}_p$ and an integer $m \geq 1$, the m-th hyper-Kloosterman sums are defined by the formula

$$\mathrm{Kl}_m(x;k) := \frac{1}{q^{\frac{m-1}{2}}} \sum_{\substack{y_1,\dots,y_m \in k \\ y_1 \cdots y_m = x}} \psi_k(y_1 + \cdots + y_m) \tag{6-10}$$

for any finite extension $k/\mathbb{F}_p$ and any $x \in k$. Thus, we have for instance $\mathrm{Kl}_1(x;k) = \psi_k(x)$, while Kl_2 is essentially a classical Kloosterman sum.

The following deep result shows that, as functions of x, these sums are trace functions of admissible sheaves.

Proposition 6.9 (Deligne; Katz). *There exists an admissible Fourier sheaf $\mathcal{K}\ell_m$ such that, for any $k/\mathbb{F}_p$ and any $x \in k^\times$, we have*

$$t_{\mathcal{K}\ell_m}(x;k) = (-1)^{m-1}\,\mathrm{Kl}_m(x;k).$$

Furthermore:

- *$\mathcal{K}\ell_m$ is lisse on $\mathbb{G}_m = \mathbb{P}^1 - \{0,\infty\}$; if $m \geq 2$, it is tamely ramified at 0, and for $m = 1$ it is lisse at 0; for all $m \geq 1$, it is wildly ramified at ∞ with Swan conductor 1.*

- *$\mathcal{K}\ell_m$ is of rank m, and is geometrically irreducible.*

- *If p is odd, then the Zariski closure of the image $\rho_{\mathcal{K}\ell_m}(\pi_1^g(\mathbb{G}_m))$, which is called the geometric monodromy group of $\mathcal{K}\ell_m$, is isomorphic to SL_m if m is odd, and to Sp_m if m is even.*

It follows that $\mathrm{cond}(\mathcal{K}\ell_m) = m + 3$ for all $m \geq 2$ and all p, and that $\mathrm{cond}(\mathcal{K}\ell_1) = 3$.

Proof. All these results can be found in [Katz 1988]; more precisely, the first two points are part of Theorem 4.1.1 in [Katz 1988] and the last is part of Theorem 11.1 in the same reference. □

Remark 6.10. In particular, for $x \neq 0$, we get the estimate

$$|\mathrm{Kl}_m(x;k)| \leq m,$$

first proved by Deligne. Note that this exhibits square-root cancellation in the $(m-1)$-variable character sum defining $\mathrm{Kl}(x;k)$. For $x = 0$, it is elementary that

$$\mathrm{Kl}_m(0;k) = (-1)^{m-1}q^{-(m-1)/2}.$$

We have the following bounds for hyper-Kloosterman sums, where the case $m = 3$ is the important one for this paper:

Proposition 6.11 (estimates for hyper-Kloosterman sums). *Let $m \geq 2$ be an integer and ψ' an additive character of $\mathbb{F}_p$, which may be trivial. We have*

$$\left| \sum_{x \in \mathbb{F}_p^\times} \mathrm{Kl}_m(x; p) \psi'(x) \right| \ll p^{1/2}. \tag{6-11}$$

Further, let $a \in \mathbb{F}_p^\times$. If either $a \neq 1$ or ψ' is nontrivial, we have

$$\left| \sum_{x \in \mathbb{F}_p^\times} \mathrm{Kl}_m(x; p) \overline{\mathrm{Kl}_m(ax; p)} \psi'(x) \right| \ll p^{1/2}. \tag{6-12}$$

In these bounds, the implied constants depend only, and at most polynomially, on m.

Proof. The first bound (6-11) follows directly from Corollary 6.6 and (6-6) because $\mathcal{K}\ell_m$ is, for $m \geq 2$, geometrically irreducible of rank > 1, and therefore not geometrically isomorphic to the rank-1 Artin–Schreier sheaf $\mathcal{L}_{\psi'}$.

For the proof of (6-12), we use the identity[7]

$$\mathrm{Kl}_m(x) = \frac{1}{p^{1/2}} \sum_{y \in \mathbb{F}_p^\times} \mathrm{Kl}_{m-1}(y^{-1}) \psi(xy) = -\mathrm{FT}_\psi([y^{-1}]^* \, \mathrm{Kl}_{m-1})(x),$$

which is valid for all $x \in \mathbb{F}_p$ (including $x = 0$). If we let $b \in \mathbb{F}_p$ be such that $\psi'(x) = \psi(bx)$ for all x, then by the Plancherel formula, we deduce

$$\sum_{x \in \mathbb{F}_p} \mathrm{Kl}_m(x; p) \overline{\mathrm{Kl}_m(ax; p)} \psi'(x) = \sum_{y \in \mathbb{F}_p \setminus \{0, -b\}} \mathrm{Kl}_{m-1}(y^{-1}) \overline{\mathrm{Kl}_{m-1}(a(y+b)^{-1})}$$

$$= \sum_{\substack{y \in \mathbb{F}_p, \\ y \neq 0, -1/b}} \mathrm{Kl}_{m-1}(y; p) \overline{\mathrm{Kl}_{m-1}(\gamma \cdot y; p)},$$

where

$$\gamma := \begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix}.$$

We are in the situation of Corollary 6.6, with both sheaves $\mathcal{K}\ell_{m-1}$ and $\gamma^* \mathcal{K}\ell_{m-1}$ admissible and geometrically irreducible. If $m \geq 3$, $\mathcal{K}\ell_{m-1}$ is tamely ramified at 0 and wildly ramified at ∞, and $\gamma^* \mathcal{K}\ell_{m-1}$ is therefore tame at $\gamma^{-1}(0)$ and wild at $\gamma^{-1}(\infty)$, so that a geometric isomorphism $\mathcal{K}\ell_{m-1} \simeq \gamma^* \mathcal{K}\ell_{m-1}$ can only occur if $\gamma(0) = 0$ and $\gamma(\infty) = \infty$, or in other words if $b = 0$. If $b = 0$, we have $\gamma^* \mathcal{K}\ell_{m-1} = [\times a]^* \mathcal{K}\ell_{m-1}$ which is known to be geometrically isomorphic to $\mathcal{K}\ell_{m-1}$ if and only if $a = 1$, by [Katz 1988, Proposition 4.1.5]. Thus (6-12) follows from Corollary 6.6 for $m \geq 3$, using (6-6) and the formulas $\mathrm{cond}(\mathcal{K}\ell_{m-1}) = \mathrm{cond}(\gamma^* \mathcal{K}\ell_{m-1}) = m + 3$.

---

[7]One could use this identity to recursively build the hyper-Kloosterman sheaf from the Artin–Schreier sheaf, Theorem 6.8, and pullback via the map $x \mapsto 1/x$, if desired.

The case $m = 2$ is easy since the sum above is then simply

$$\sum_{\substack{y \in \mathbb{F}_p, \\ y \neq 0, -1/b}} \psi(y - ay/(by + 1)),$$

where the rational function $f(y) = y - ay/(by + 1)$ is constant if and only if $a = 1$, $b = 0$, so that we can use Lemma 4.2 in this case.   □

Remark 6.12. A similar result was proved by Michel [1998, Corollaire 2.9] using a different method. That method requires more information (the knowledge of the geometric monodromy group of $\mathcal{K}\ell_m$) but gives more general estimates. The case $m = 3$ is (somewhat implicitly) the result used in [Friedlander and Iwaniec 1985], which is proved by Birch and Bombieri in the Appendix to the same paper (with in fact two proofs, which are rather different and somewhat more ad hoc than the argument presented here). This same estimate is used by Zhang [2014] to control Type III sums.

6H. *The van der Corput method for trace functions.* Let $t = t_{\mathcal{F}}$ be the trace function associated to an admissible sheaf $\mathcal{F}$. In the spirit of Proposition 4.12, the q-van der Corput method, when applied to incomplete sums of t, followed by completion of sums, produces expressions of the form

$$\sum_{x \in \mathbb{F}_p} t(x) \overline{t(x + l)} \psi(hx)$$

for $(h, l) \in \mathbb{F}_p \times \mathbb{F}_p^{\times}$ and for some additive character ψ. We seek sufficient conditions that ensure square-root cancellation in the above sum, for any $l \neq 0$ and any h.

Observe that if

$$t(x) = \psi(ax^2 + bx),$$

then the sum is sometimes of size p. Precisely, this happens if and only if $h = 2al$. As we shall see, this phenomenon is essentially the only obstruction to square-root cancellation.

Definition 6.13 (no polynomial phase). For a finite field k and $d \geq 0$, we say that an admissible sheaf $\mathcal{F}$ over $\mathbb{P}_k^1$ has *no polynomial phase* of degree $\leq d$ if no geometrically irreducible component of $\mathcal{F}$ is geometrically isomorphic to a sheaf of the form $\mathcal{L}_{\psi(P(x))}$ where $P(X) \in \mathbb{F}_p[X]$ is a polynomial of degree $\leq d$.

Thus, for instance, an admissible sheaf is Fourier if and only if it has no polynomial phase of degree ≤ 1.

Remark 6.14. An obvious sufficient condition for $\mathcal{F}$ not to contain any polynomial phase (of any degree) is that each geometrically irreducible component of $\mathcal{F}$ be irreducible of rank ≥ 2, for instance if $\mathcal{F}$ itself is geometrically irreducible of rank ≥ 2.

The following inverse theorem is a variant of an argument of Fouvry, Kowalski and Michel [Fouvry et al. 2013a, Lemma 5.4].

Theorem 6.15. *Let $d \geq 1$ be an integer, and let p be a prime such that $p > d$. Let $\mathscr{F}$ be an isotypic admissible sheaf over $\mathbb{P}^1_{\mathbb{F}_p}$ with no polynomial phase of degree $\leq d$. Then either $\mathrm{cond}(\mathscr{F}) \geq p + 1$, or for any $l \in \mathbb{F}_p^\times$ the sheaf $\mathscr{F} \otimes [+l]^* \breve{\mathscr{F}}$ contains no polynomial phase of degree $\leq d - 1$.*

In all cases, for any $l \in \mathbb{F}_p^\times$ and any $P(X) \in \mathbb{F}_p[X]$ of degree $d - 1$, we have

$$\left| \sum_{x \in \mathbb{F}_p} t_{\mathscr{F}}(x + l) \overline{t_{\mathscr{F}}(x)} \psi(P(x)) \right| \ll p^{1/2}, \tag{6-13}$$

where the implied constant depends, at most polynomially, on $\mathrm{cond}(\mathscr{F})$ and on d. Furthermore, this estimate holds also if $l = 0$ and $P(x) = hx$ with $h \neq 0$.

Proof. First suppose that $l \neq 0$. Observe that if $\mathrm{cond}(\mathscr{F}) \geq p + 1$, the bound (6-13) follows from the trivial bound

$$|t_{\mathscr{F}}(x + l) \overline{t_{\mathscr{F}}(x)} \psi(P(x))| \leq \mathrm{rk}(\mathscr{F})^2 \leq \mathrm{cond}(\mathscr{F})^2,$$

and that if the sheaf $[+l]^* \mathscr{F} \otimes \breve{\mathscr{F}}$ contains no polynomial phase of degree $\leq d - 1$, then the bound is a consequence of Corollary 6.6.

We now prove that one of these two properties holds. We assume that $[+l]^* \mathscr{F} \otimes \breve{\mathscr{F}}$ contains a polynomial phase of degree $\leq d-1$, and will deduce that $\mathrm{cond}(\mathscr{F}) \geq p+1$.

Since $\mathscr{F}$ is isotypic, the assumption implies that there is a geometric isomorphism

$$[+l]^* \mathscr{F} \simeq \mathscr{F} \otimes \mathscr{L}_{\psi(P(x))}$$

for some polynomial $P(X) \in \mathbb{F}_p[X]$ of degree $\leq d - 1$. Then, considering the geometric irreducible component $\mathscr{G}$ of $\mathscr{F}$ (which is a sheaf on $\mathbb{P}^1_{\mathbb{F}_p}$) we also have

$$[+l]^* \mathscr{G} \simeq \mathscr{G} \otimes \mathscr{L}_{\psi(P(x))}. \tag{6-14}$$

If $\mathscr{G}$ is ramified at some point $x \in \mathbb{A}^1(\overline{k})$, then since $\mathscr{L}_{\psi(P(x))}$ is lisse on $\mathbb{A}^1(\overline{k})$, we conclude by iterating (6-14) that $\mathscr{G}$ is ramified at $x, x + l, x + 2l, \ldots, x + (p - 1)l$, which implies that $\mathrm{cond}(\mathscr{F}) \geq \mathrm{cond}(\mathscr{G}) \geq p + \mathrm{rk}(\mathscr{G})$. Thus there remains to handle the case when $\mathscr{G}$ is lisse outside ∞. It then follows from [Fouvry et al. 2013a, Lemma 5.4(2)] that either $\mathrm{cond}(\mathscr{G}) \geq \mathrm{rk}(\mathscr{G}) + p$, in which case $\mathrm{cond}(\mathscr{F}) \geq p + 1$ again, or that $\mathscr{G}$ is isomorphic (over $\overline{\mathbb{F}_p}$) to a sheaf of the form $\mathscr{L}_{\psi(Q(x))}$ for some polynomial of degree $\leq d$. Since $\mathscr{G}$ is a geometrically irreducible component of $\mathscr{F}$, this contradicts the assumption on $\mathscr{F}$.

Finally, consider the case where $l = 0$ and $P(x) = hx$ with $h \neq 0$. Using Corollary 6.6 and (6-6), the result holds for a given $h \in \mathbb{F}_p^\times$ unless the geometrically irreducible component $\mathcal{G}$ of $\mathcal{F}$ satisfies

$$\mathcal{G} \simeq \mathcal{G} \otimes \mathcal{L}_{\psi(hx)}.$$

Since $d \geq 1$, $\mathcal{F}$ is a Fourier sheaf, and hence so are $\mathcal{G}$ and $\mathcal{G} \otimes \mathcal{L}_{\psi(hx)}$. Taking the Fourier transform of both sides of this isomorphism, we easily obtain

$$[+h]^* \mathrm{FT}_\psi \mathcal{G} \simeq \mathrm{FT}_\psi \mathcal{G},$$

and it follows from [Fouvry et al. 2013a, Lemma 5.4(2)] again that $\mathrm{cond}(\mathrm{FT}_\psi \mathcal{G}) \geq p + 1$. Using the Fourier inversion formula (6-8) and (6-9), we derive

$$\mathrm{cond}(\mathcal{F}) \geq \mathrm{cond}(\mathcal{G}) \gg p^{1/2},$$

so that the bound (6-13) also holds trivially in this case. $\square$

Remark 6.16. For later use, we observe that the property of having *no polynomial phase of degree* ≤ 2 of an admissible sheaf $\mathcal{F}$ is invariant under the following transformations:

- Twists by an Artin–Schreier sheaf associated to a polynomial phase of degree ≤ 2, i.e., $\mathcal{F} \mapsto \mathcal{F} \otimes \mathcal{L}_{\psi(ax^2+bx)}$.

- Dilations and translations: $\mathcal{F} \mapsto [\times a]^* \mathcal{F}$ and $\mathcal{F} \mapsto [+b]^* \mathcal{F}$, where $a \in \mathbb{F}_p^\times$ and $b \in \mathbb{F}_p$.

- Fourier transforms, if $\mathcal{F}$ is Fourier: $\mathcal{F} \mapsto \mathrm{FT}_\psi \mathcal{F}$. Indeed, the Fourier transform of a sheaf $\mathcal{L}_{\psi(P(x))}$ with $\deg(P) = 2$ is geometrically isomorphic to $\mathcal{L}_{\psi(Q(x))}$ for some polynomial Q of degree 2.

6I. *Study of some specific exponential sums.* We now apply the theory above to some specific multidimensional exponential sums which appear in the refined treatment of the Type I sums in Section 8. For parameters $(a, b, c, d, e) \in \mathbb{F}_p$, with $a \neq c$, we consider the rational function

$$f(X, Y) := \frac{1}{(Y + aX + b)(Y + cX + d)} + eY \in \mathbb{F}_p(X, Y).$$

For a fixed nontrivial additive character ψ of $\mathbb{F}_p$ and for any $x \in \mathbb{F}_p$, we define the character sum

$$K_f(x; p) := -\frac{1}{p^{1/2}} \sum_{\substack{y \in \mathbb{F}_p \\ (y+ax+b)(y+cx+d) \neq 0}} \psi(f(x, y)). \tag{6-15}$$

For any $x \in \mathbb{F}_p$, the specialized rational function $f(x, Y) \in \mathbb{F}_p(Y)$ is nonconstant (it has poles in $\mathbb{A}^1_{\mathbb{F}_p}$), and therefore by Lemma 4.2 (or Theorem 6.5) we have

$$|K_f(x; p)| \leq 4. \tag{6-16}$$

We will prove the following additional properties of the sums $K_f(x; p)$:

Theorem 6.17. *For a prime p and parameters $(a, b, c, d, e) \in \mathbb{F}_p^5$ with $a \neq c$, the function $x \mapsto K_f(x; p)$ on $\mathbb{F}_p$ is the trace function of an admissible geometrically irreducible sheaf $\mathcal{F}$ whose conductor is bounded by a constant independent of p. Furthermore, $\mathcal{F}$ contains no polynomial phase of degree ≤ 2.*

In particular, we have

$$\left| \sum_{x \in \mathbb{F}_p} K_f(x; p) \psi(hx) \right| \ll p^{1/2} \tag{6-17}$$

for all $h \in \mathbb{F}_p$ and

$$\left| \sum_{x \in \mathbb{F}_p} K_f(x; p) \overline{K_f(x + l; p)} \psi(hx) \right| \ll p^{1/2} \tag{6-18}$$

for any $(h, l) \in \mathbb{F}_p^2 - \{(0, 0)\}$, where the implied constants are absolute.

Proof. Note that the estimates (6-17) and (6-18) follow from the first assertion (see Theorem 6.15).

We first normalize most of the parameters: we have

$$K_f(x; p) = -\frac{\psi(-eax - eb)}{p^{1/2}} \sum_{z \in \mathbb{F}_p} \psi\left(ez + \frac{1}{z(z + (c - a)x + d - b)} \right),$$

and by Remark 6.16, this means that we may assume that $c = d = 0$, $a \neq 0$. Furthermore, we have then

$$K_f(x; p) = K_{\tilde{f}}(ax + b; p),$$

where $\tilde{f}$ is the rational function f with parameters $(1, 0, 0, 0, e)$. Again by Remark 6.16, we are reduced to the special case $f = \tilde{f}$, i.e., to the sum

$$K_f(x; p) = -\frac{1}{p^{1/2}} \sum_{\substack{y \in \mathbb{F}_p \\ (y+x)y \neq 0}} \psi\left(\frac{1}{(y + x)y} + ey \right).$$

We will prove that the Fourier transform of K_f is the trace function of a geometrically irreducible Fourier sheaf with bounded conductor and no polynomial phase of degree ≤ 2. By the Fourier inversion formula (6-8) and (6-9), and the invariance of the property of not containing a polynomial phase of degree ≤ 2 under Fourier transform (Remark 6.16 again), this will imply the result for K_f.

For $z \in \mathbb{F}_p$, we have

$$\mathrm{FT}_\psi(K_f)(z) = \frac{1}{p} \sum_{y+x, y \neq 0} \sum \psi\left(\frac{1}{(y+x)y} + ey + zx\right).$$

If $z \neq 0$, the change of variables

$$y_1 := \frac{1}{(y+x)y}, \quad y_2 := z(y+x)$$

is a bijection

$$\{(x, y) \in \mathbb{F}_p \times \mathbb{F}_p : y(x+y) \neq 0\} \to \{(y_1, y_2) \in \mathbb{F}_p^\times \times \mathbb{F}_p^\times\}$$

(with inverse $y = z/(y_1 y_2)$ and $x = y_2/z - z/(y_1 y_2)$) which satisfies

$$\frac{1}{(y+x)y} + ey + zx = y_1 + \frac{ez}{y_1 y_2} + y_2 - \frac{z^2}{y_1 y_2} = y_1 + y_2 + \frac{z(e-z)}{y_1 y_2}$$

for $y(x+y) \neq 0$. Thus

$$\mathrm{FT}_\psi(K_f)(z) = \frac{1}{p} \sum_{y_1, y_2 \in \mathbb{F}_p^\times} \sum \psi\left(y_1 + y_2 + \frac{z(e-z)}{y_1 y_2}\right) = \mathrm{Kl}_3(z(e-z); p)$$

for $z(e-z) \neq 0$.

Similar calculations reveal that this identity also holds when $z = 0$ and $z = e$ (treating the doubly degenerate case $z = e = 0$ separately), i.e., both sides are equal to $1/p$ in these cases. This means that $\mathrm{FT}_\psi(K_f)$ is the trace function of the pullback sheaf

$$\mathcal{G}_f := \varphi^* \mathcal{K}\ell_3,$$

where φ is the quadratic map $\varphi : z \mapsto z(e-z)$.

The sheaf $\mathcal{G}_f$ has bounded conductor (it has rank 3 and is lisse on $U = \mathbb{P}^1_{\mathbb{F}_p} - \{0, e, \infty\}$, with wild ramification at ∞ only, where the Swan conductor can be estimated using [Katz 1988, 1.13.1], for $p \geq 3$). We also claim that $\mathcal{G}_f$ is geometrically irreducible. Indeed, it suffices to check that $\pi_1^g(U)$ acts irreducibly on the underlying vector space of $\rho_{\mathcal{K}\ell_3}$. But since $z \mapsto z(e-z)$ is a nonconstant morphism $\mathbb{P}^1_{\mathbb{F}_p} \to \mathbb{P}^1_{\mathbb{F}_p}$, $\pi_1^g(U)$ acts by a finite-index subgroup of the action of $\pi_1^g(\mathbb{G}_m)$ on $\mathcal{K}\ell_3$. Since the image of $\pi_1^g(\mathbb{G}_m)$ is Zariski-dense in SL_3 (as recalled in Proposition 6.9), which is a connected algebraic group, it follows that the image of $\pi_1^g(U)$ is also Zariski-dense in SL_3, proving the irreducibility.

Since $\mathcal{G}_f$ is geometrically irreducible of rank $3 > 1$, it does not contain any polynomial phase (see Remark 6.14), concluding the proof. $\square$

Remark 6.18. Another natural strategy for proving this theorem would be to start with the observation that the function $x \mapsto K_f(x; k)$ is the trace function of the constructible ℓ-adic sheaf

$$\mathcal{K}_f = R^1 \pi_{1,!} \mathcal{L}_{\psi(f)}(1/2), \qquad \mathcal{L}_{\psi(f)} = f^* \mathcal{L}_\psi,$$

where $\pi_1 : \mathbb{A}^2_{\mathbb{F}_p} \to \mathbb{A}^1_{\mathbb{F}_p}$ is the projection on the first coordinate and $R^1 \pi_{1,!}$ denotes the operation of higher-direct image with compact support associated to that map (and $\left(\frac{1}{2}\right)$ is a Tate twist). This is known to be mixed of weights ≤ 0 by [Deligne 1980], and it follows from the general results[8] of Fouvry, Kowalski and Michel in [Fouvry et al. 2013b] that the conductor of this sheaf is absolutely bounded as p varies. To fully implement this approach, it would still remain to prove that the weight-0 part of $\mathcal{K}_f$ is geometrically irreducible with no polynomial phase of degree ≤ 2. Although such arguments might be necessary in more advanced cases, the direct approach we have taken is simpler here.

Remark 6.19. In the remainder of this paper, we will only use the bounds (6-17) and (6-18) from Theorem 6.17. These bounds can also be expressed in terms of the Fourier transform $\mathrm{FT}_\psi(K_f)$ of K_f, since they are equivalent to

$$|\mathrm{FT}_\psi(K_f)(h)| \ll p^{1/2}$$

and

$$\left| \sum_{x \in \mathbb{F}_p} \mathrm{FT}_\psi(K_f)(x+h)\overline{\mathrm{FT}_\psi(K_f)(x)} \psi(-lx) \right| \ll p^{1/2},$$

respectively. As such, we see that it is in fact enough to show that $\mathrm{FT}_\psi(K_f)$, rather than K_f, is the trace function of a geometrically irreducible admissible sheaf with bounded conductor and no quadratic phase component. Thus, in principle, we could avoid any use of Theorem 6.8 in our arguments (provided that we took the existence of the Kloosterman sheaves for granted). However, from a conceptual point of view, the fact that K_f has a good trace function interpretation is more important than the corresponding fact for FT_ψ (for instance, the iterated van der Corput bounds in Remark 6.23 rely on the former fact rather than the latter).

6J. *Incomplete sums of trace functions.* In this section, we extend the discussion of Section 4 to general admissible trace functions. More precisely, given a squarefree integer q, we say that a q-periodic arithmetic function

$$t : \mathbb{Z} \to \mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$$

---

[8]Which were partly motivated by the current paper.

is an *admissible trace function* if we have

$$t(x) = \prod_{p|q} t(x; p) \tag{6-19}$$

for all x, where, for each prime $p \mid q$, $x \mapsto t(x; p)$ is the composition of reduction modulo p and the trace function associated to an admissible sheaf $\mathscr{F}_p$ on $\mathbb{P}^1_{\mathbb{F}_p}$.

An example is the case discussed in Section 4: for a rational function $f(X) = P(X)/Q(X) \in \mathbb{Q}(X)$ with $P, Q \in \mathbb{Z}[X]$ and a squarefree integer q such that $Q(q) \neq 0$, we can write

$$e_q(f(x)) = e_q\left(\frac{P(x)}{Q(x)}\right) = \prod_{p|q} e_p(\overline{q_p} f(x)), \quad \text{where } q_p = q/p$$

(by Lemma 4.4). In that case, we take

$$\mathscr{F}_p = \mathscr{L}_{\psi(f)}, \quad \text{where } \psi(x) = e_p(\overline{q_p}x).$$

Another example is given by the Kloosterman sums defined for q squarefree and $x \in \mathbb{Z}$ by

$$\mathrm{Kl}_m(x; q) = \frac{1}{q^{m-1/2}} \sum_{\substack{x_1,\ldots,x_m \in \mathbb{Z}/q\mathbb{Z} \\ x_1 \cdots x_m = x}} e_q(x_1 + \cdots + x_m), \tag{6-20}$$

for which we have

$$\mathrm{Kl}_m(x; q) = \prod_{p|q} \mathrm{Kl}_m(\overline{q_p}^m x; p) = \prod_{p|q} ([\times \overline{q_p}^m]^* \, \mathrm{Kl}_m(\,\cdot\,; p))(x),$$

and hence

$$\mathrm{Kl}_m(x; q) = (-1)^{(m-1)\Omega(q)} t(x),$$

where

$$t(x) = \prod_{p|q} (-1)^{m-1} t_{\mathscr{F}_p}(x; p) \quad \text{with } \mathscr{F}_p = [\times \overline{q_p}^m]^* \mathscr{K}\ell_m$$

is an admissible trace function modulo q.

Given a tuple of admissible sheaves $\mathscr{F} = (\mathscr{F}_p)_{p|q}$, we define the conductor $\mathrm{cond}(\mathscr{F})$ by

$$\mathrm{cond}(\mathscr{F}) = \prod_{p|q} \mathrm{cond}(\mathscr{F}_p).$$

Thus, for the examples above, the conductor is bounded by $C^{\Omega(q)}$ for some constant C depending only on f or m, respectively. This will be a general feature in applications.

6J.1. *A generalization of Proposition 4.12.* Thanks to the square root cancellation for complete sums of trace functions provided by Corollary 6.6, we may extend Proposition 4.12 to general admissible trace functions to squarefree moduli.

Proposition 6.20 (incomplete sum of trace function). *Let q be a squarefree natural number of polynomial size and let $t(\,\cdot\,; q) : \mathbb{Z} \to \mathbb{C}$ be an admissible trace function modulo q associated to admissible sheaves $\mathcal{F} = (\mathcal{F}_p)_{p|q}$.*
   Let further $N \geq 1$ be given with $N \ll q^{O(1)}$ and let ψ_N be a function on $\mathbb{R}$ defined by

$$\psi_N(x) = \psi\left(\frac{x - x_0}{N}\right),$$

where $x_0 \in \mathbb{R}$ and ψ is a smooth function with compact support satisfying

$$|\psi^{(j)}(x)| \ll \log^{O(1)} N$$

for all fixed $j \geq 0$, where the implied constant may depend on j.

  (i) *(Pólya–Vinogradov + Deligne) Assume that for every $p \mid q$ the sheaf $\mathcal{F}_p$ has no polynomial phase of degree ≤ 1. Then we have*

$$\left|\sum_n \psi_N(n) t(n; q)\right| \ll q^{1/2+\varepsilon}\left(1 + \frac{N}{q}\right) \tag{6-21}$$

  for any $\varepsilon > 0$.

 (ii) *(one van der Corput + Deligne) Assume that for every $p \mid q$ the sheaf $\mathcal{F}_p$ has no polynomial phase of degree ≤ 2. Then, for any factorization $q = rs$ and $N \leq q$, we have*

$$\left|\sum_n \psi_N(n) t(n; q)\right| \ll q^\varepsilon (N^{1/2} r^{1/2} + N^{1/2} s^{1/4}). \tag{6-22}$$

In all cases the implied constants depend on ε, $\mathrm{cond}(\mathcal{F})$ and the implied constants in the estimates for the derivatives of ψ.

Remark 6.21. In the context of Proposition 4.12, where $t(n; q) = e_q(P(n)/Q(n))$, the assumptions $\deg P < \deg Q$ and $\deg(Q(p)) = \deg(Q)$ (for all $p \mid q$) ensure that the sheaves $\mathcal{L}_{e_p(\overline{q_p}P(x)/Q(x))}$ do not contain any polynomial phase of any degree.

Remark 6.22. For future reference, we observe that in the proof of (6-22) below we will not use any of the properties of the functions $x \mapsto t(x; p)$ for $p \mid r$ for a given factorization $q = rs$, except for their boundedness.

Proof. For each $p \mid q$, the trace function $t_{\mathcal{F}_p}$ decomposes by (6-7) into a sum of at most $\mathrm{rk}(\mathcal{F}_p) \leq \mathrm{cond}(\mathcal{F}_p) \leq \mathrm{cond}(\mathcal{F})$ trace functions of isotypic admissible sheaves, and therefore $n \mapsto t(n; q)$ decomposes into a sum of at most $C^{\omega(q)}$ functions, each of which is an admissible trace function modulo q associated to isotypic admissible sheaves. Moreover, if no $\mathcal{F}_p$ contains a polynomial phase of degree $\leq d$, then all isotypic components share this property (in particular, since $d \geq 1$ for both statements, each component is also a Fourier sheaf). Thus we may assume without loss of generality that each $\mathcal{F}_p$ is isotypic.

We start with the proof of (6-21). By (4-12), we have

$$\left| \sum_n \psi_N(n) t(n; q) \right| \ll q^{1/2+\varepsilon} \left(1 + \frac{|N'|}{q} \right) \sup_{h \in \mathbb{Z}/q\mathbb{Z}} |\mathrm{FT}_q(t(h; q))|$$

$$\ll q^{1/2+\varepsilon} \left(1 + \frac{N}{q} \right) \sup_{h \in \mathbb{Z}/q\mathbb{Z}} |\mathrm{FT}_q(t(h; q))|$$

for any $\varepsilon > 0$, where $N' = \sum_n \psi_N(n)$. By Lemma 4.4, (6-19) and the definition of the Fourier transform, we have

$$\mathrm{FT}_q(t(\,\cdot\,; q))(h) = \prod_{p \mid q} \mathrm{FT}_p(t(\,\cdot\,; p))(\overline{q_p} h).$$

Since $t(\,\cdot\,; p) = t_{\mathcal{F}_p}$ is the trace function of a Fourier sheaf, we have

$$|\mathrm{FT}_p(t(\,\cdot\,; p))(\overline{q_p} h)| \leq 10 \, \mathrm{cond}(\mathcal{F}_p)^2 \leq 10 \, \mathrm{cond}(\mathcal{F})^2$$

for all h by (6-9) (or Corollary 6.6 applied to the sheaves $\mathcal{F}_p$ and $\mathcal{L}_{e_p(-\overline{q_p} x)}$). Combining these bounds, we obtain (6-21).

The proof of (6-22) follows closely that of (4-20). It is sufficient to prove this bound in the case $r \leq s$. We may also assume that $r \leq N \leq s$, since, otherwise, the result follows either from the trivial bound or (6-21). Then, denoting $K := \lfloor N/r \rfloor$, we write

$$\sum_n \psi_N(n) t(n; q) = \frac{1}{K} \sum_n \sum_{k=1}^K \psi_N(n + kr) t(n + kr; q).$$

Since $q = rs$, we have

$$t(n + kr; q) = t(n; r) t(n + kr; s),$$

where

$$t(n; r) = \prod_{p \mid r} t(n; p), \quad t(n; s) = \prod_{p \mid s} t(n; p)$$

are admissible trace functions modulo r and s, respectively. Hence

$$\left|\sum_n \psi_N(n)t(n;q)\right| \ll \frac{1}{K}\sum_n \left|\sum_{k=1}^K \psi_N(n+kr)t(n+kr;s)\right|$$

$$\ll \frac{N^{1/2}}{K}\left(\sum_n \left|\sum_{k=1}^K \psi_N(n+kr)t(n+kr;s)\right|^2\right)^{1/2}$$

$$\ll \frac{N^{1/2}}{K}\left(\sum_{1\le k,l\le K} A(k,l)\right)^{1/2},$$

where

$$A(k,l) = \sum_n \psi_N(n+kr)\overline{\psi_N(n+lr)}t(n+kr;s)\overline{t(n+lr;s)}.$$

The diagonal contribution satisfies

$$\sum_{1\le k\le K} A(k,k) \ll q^\varepsilon KN$$

for any $\varepsilon > 0$, where the implied constant depends on $\text{cond}(\mathscr{F})$.

Instead of applying (6-21) for the off-diagonal terms, it is slightly easier to just apply (4-12). For given $k \ne l$, since $kr, lr \ll N$, the sequence $\Psi_N(n) = \psi_N(n+kr)\overline{\psi_N(n+lr)}$ satisfies the assumptions of (4-12). Setting

$$w(n;s) = t(n+kr;s)\overline{t(n+lr;s)},$$

we obtain

$$|A(k,l)| = \left|\sum_n \Psi_N(n)w(n;s)\right| \ll q^\varepsilon s^{1/2}\sup_{h\in\mathbb{Z}/s\mathbb{Z}} |\text{FT}_s(w(\,\cdot\,;s))(h)|$$

by (4-12) (since $N \le s$). We have

$$\text{FT}_s(w(\,\cdot\,;s))(h) = \prod_{p|s} \text{FT}_p(w(\,\cdot\,;p))(\overline{s_p}h)$$

with $s_p = s/p$. For $p \mid k-l$, we use the trivial bound

$$|\text{FT}_p(w(\,\cdot\,;p))(\overline{s_p}h)| \ll p^{1/2},$$

and for $p \nmid k-l$, we have

$$\text{FT}_p(w(\,\cdot\,;p))(\overline{s_p}h) = \frac{1}{p^{1/2}}\sum_{x\in\mathbb{F}_p} t(x+kr;p)\overline{t(x+lr;p)}e_p(\overline{s_p}hx) \ll 1$$

by the change of variable $x \mapsto x + kq_1$ and (6-13), which holds for $\mathscr{F}_p$ by our assumptions. In all cases, the implied constant depends only on $\text{cond}(\mathscr{F}_p)$. Therefore

we have

$$A(k, l) \ll (k - l, s)^{1/2} q^{\varepsilon} s^{1/2},$$

and summing over $k \neq l$, we derive

$$\left| \sum_n \psi_N(n) e_q(f(n)) \right| \ll \frac{q^{\varepsilon} N^{1/2}}{K} \left(KN + s^{1/2} \sum_{1 \leq k \neq l \leq K} (k - l, s)^{1/2} \right)^{1/2}$$

$$\ll \frac{q^{\varepsilon} N^{1/2}}{K} (K^{1/2} N^{1/2} + s^{1/4} K),$$

which gives the desired conclusion (6-22).   □

Remark 6.23. Similarly to Remark 4.15, one can iterate the above argument and conclude that for any $l \geq 1$ and any factorization $q = q_1 \cdots q_l$

$$\left| \sum_n \psi_N(n) t(n; q) \right| \ll q^{\varepsilon} \left(\left(\sum_{i=1}^{l-1} N^{1 - 1/2^i} q_i^{1/2^i} \right) + N^{1 - 1/2^{l-1}} q_l^{1/2^l} \right),$$

assuming that $N < q$ and the $\mathcal{F}_p$ do not contain any polynomial phase of degree $\leq l$.

Specializing Proposition 6.20 to the functions in Theorem 6.17, we conclude:

Corollary 6.24. *Let $q \geq 1$ be a squarefree integer and let $K(\cdot; q)$ be given by*

$$K(x; q) := \frac{1}{q^{1/2}} \sum_{y \in \mathbb{Z}/q\mathbb{Z}} e_q(f(x, y)),$$

where

$$f(x, y) = \frac{1}{(y + ax + b)(y + cx + d)} + ey$$

and a, b, c, d, e are integers with $(a - c, q) = 1$. Let further $N \geq 1$ be given with $N \ll q^{O(1)}$ and let ψ_N be a function on $\mathbb{R}$ defined by

$$\psi_N(x) = \psi\left(\frac{x - x_0}{N} \right),$$

where $x_0 \in \mathbb{R}$ and ψ is a smooth function with compact support satisfying

$$|\psi^{(j)}(x)| \ll \log^{O(1)} N$$

for all fixed $j \geq 0$, where the implied constant may depend on j.

  Then we have

$$\left| \sum_n \psi_N(n) K(n; q) \right| \ll q^{1/2 + \varepsilon} \left(1 + \frac{N}{q} \right) \tag{6-23}$$

for any $\varepsilon > 0$.

Furthermore, for any factorization $q = rs$ and $N \leq q$, we have the additional bound

$$\left| \sum_n \psi_N(n) K(n; q) \right| \ll q^\varepsilon (N^{1/2} r^{1/2} + N^{1/2} s^{1/4}). \qquad (6\text{-}24)$$

Indeed, it follows from Theorem 6.17 and the assumption $(a - c, q) = 1$ that $K_f(\,\cdot\,; q)$ is an admissible trace function modulo q associated to sheaves which do not contain any polynomial phase of degree ≤ 2.

6J.2. *Correlations of hyper-Kloosterman sums of composite moduli.* Finally, we extend Proposition 6.11 to composite moduli:

Lemma 6.25 (correlation of hyper-Kloosterman sums). *Let s, r_1, r_2 be square-free integers with $(s, r_1) = (s, r_2) = 1$. Let $a_1 \in (\mathbb{Z}/r_1 s)^\times$, $a_2 \in (\mathbb{Z}/r_2 s)^\times$, and $n \in \mathbb{Z}/([r_1, r_2]s)\mathbb{Z}$. Then we have*

$$\sum_{h \in (\mathbb{Z}/s[r_1, r_2]\mathbb{Z})^\times} \mathrm{Kl}_3(a_1 h; r_1 s) \overline{\mathrm{Kl}_3(a_2 h; r_2 s)} e_{[r_1, r_2]s}(nh)$$

$$\ll (s[r_1, r_2])^\varepsilon s^{1/2} [r_1, r_2]^{1/2} (a_2 - a_1, n, r_1, r_2)^{1/2} (a_2 r_1^3 - a_1 r_2^3, n, s)^{1/2}$$

for any $\varepsilon > 0$, where the implied constant depends only on ε.

Proof. Let S be the sum to estimate. From Lemma 4.4, we get

$$\mathrm{Kl}_3(a_i h; r_i s) = \mathrm{Kl}_3(a_i \bar{s}^3 h; r_i) \, \mathrm{Kl}_3(a_i \bar{r}_i^3 h; s)$$

for $i = 1, 2$, as well as

$$e_{[r_1, r_2]s}(nh) = e_{[r_1, r_2]}(\bar{s}nh) e_s(\overline{[r_1, r_2]}nh),$$

and therefore $S = S_1 S_2$ with

$$S_1 = \sum_{h \in (\mathbb{Z}/[r_1, r_2]\mathbb{Z})^\times} \mathrm{Kl}_3(a_1 \bar{s}^3 h; r_1) \overline{\mathrm{Kl}_3(a_2 \bar{s}^3 h; r_2)} e_{[r_1, r_2]}(\bar{s}nh),$$

$$S_2 = \sum_{h \in (\mathbb{Z}/s\mathbb{Z})^\times} \mathrm{Kl}_3(a_1 \bar{r}_1^3 h; s) \overline{\mathrm{Kl}_3(a_2 \bar{r}_2^3 h; s)} e_s(\overline{[r_1, r_2]}nh).$$

Splitting further the summands as products over the primes dividing $[r_1, r_2]$ and s, respectively, we see that it is enough to prove the estimate

$$\left| \sum_{h \in (\mathbb{Z}/p\mathbb{Z})^\times} \mathrm{Kl}_3(b_1 h; d_1) \overline{\mathrm{Kl}_3(b_2 h; d_2)} e_p(mh) \right| \ll p^{1/2} (b_1 - b_2, m, d_1, d_2)^{1/2} \quad (6\text{-}25)$$

for p prime and integers $d_1, d_2 \geq 1$ such that $[d_1, d_2] = p$ is prime, and all $m \in \mathbb{Z}/p\mathbb{Z}$, and $b_1, b_2 \in (\mathbb{Z}/p\mathbb{Z})^\times$.

We now split into cases. First suppose that $d_2 = 1$, so that $d_1 = p$. Then we have $\mathrm{Kl}_3(b_2 h; d_2) = 1$, and the left-hand side of (6-25) simplifies to

$$\sum_{h \in (\mathbb{Z}/p\mathbb{Z})^\times} \mathrm{Kl}_3(b_1 h; p) e_p(mh) \ll p^{1/2}$$

by the first part of Proposition 6.11. Similarly, we obtain (6-25) if $d_1 = 1$.

If $d_1 = d_2 = p$ and $b_1 - b_2 = m = 0 \ (p)$, then the claim follows from the bound $|\mathrm{Kl}_3(h; p)| \ll 1$ (see Remark 6.10).

Finally, if $d_1 = d_2 = p$ and $b_1 - b_2 \neq 0 \ (p)$ or $m \neq 0 \ (p)$, then (6-25) is a consequence of the second part of Proposition 6.11. □

Finally, from this result, we obtain the following corollary:

Corollary 6.26 (correlation of hyper-Kloosterman sums, II). *Let s, r_1, r_2 be square-free integers with $(s, r_1) = (s, r_2) = 1$. Let $a_1 \in (\mathbb{Z}/r_1 s)^\times$, $a_2 \in (\mathbb{Z}/r_2 s)^\times$. Let further $H \geq 1$ be given with $H \ll (s[r_1, r_2])^{O(1)}$ and let ψ_H be a function on $\mathbb{R}$ defined by*

$$\psi_H(x) = \psi\left(\frac{x - x_0}{H}\right),$$

where $x_0 \in \mathbb{R}$ and ψ is a smooth function with compact support satisfying

$$|\psi^{(j)}(x)| \ll \log^{O(1)} H$$

for all fixed $j \geq 0$, where the implied constant may depend on j. Then we have

$$\left| \sum_{(h, s[r_1, r_2]) = 1} \Psi_H(h) \, \mathrm{Kl}_3(a_1 h; r_1 s) \overline{\mathrm{Kl}_3(a_2 h; r_2 s)} \right|$$

$$\ll (s[r_1, r_2])^\varepsilon \left(\frac{H}{[r_1, r_2]s} + 1\right) s^{1/2} [r_1, r_2]^{1/2} (a_2 - a_1, r_1, r_2)^{1/2} (a_2 r_1^3 - a_1 r_2^3, s)^{1/2}$$

for any $\varepsilon > 0$ and any integer n.

This exponential sum estimate will be the main estimate used for controlling Type III sums in Section 7.

Proof. This follows almost directly from Lemma 6.25 and the completion of sums in Lemma 4.9, except that we must incorporate the restriction $(h, s[r_1, r_2]) = 1$. We do this using Möbius inversion: the sum S to estimate is equal to

$$\sum_{\delta \mid s[r_1, r_2]} \mu(\delta) t_1(\delta) S_1(\delta),$$

where $t_1(\delta)$ satisfies $|t_1(\delta)| \leq \delta^{-2}$, because $\mathrm{Kl}_3(0; p) = p^{-1}$ for any prime p, and

$$S_1(\delta) = \sum_{\delta \mid h} \Psi_H(h) \, \mathrm{Kl}_3(\alpha_1 h; r_1 s/(\delta, r_1 s)) \overline{\mathrm{Kl}_3(\alpha_2 h; r_2 s/(\delta, r_2 s))}$$

$$= \sum_{h} \Psi_{H/\delta}(h) \, \mathrm{Kl}_3(\delta \alpha_1 h; r_1 s/(\delta, r_1 s)) \overline{\mathrm{Kl}_3(\delta \alpha_2 h; r_2 s/(\delta, r_2 s))}$$

for some $\alpha_i \in (\mathbb{Z}/r_i s/(\delta, r_i s)\mathbb{Z})^\times$. By Lemma 6.25 and Lemma 4.9, we have

$$\begin{aligned} &S_1(\delta) \\ &\ll (s[r_1, r_2])^\varepsilon \left(\frac{H}{\delta s[r_1, r_2]} + 1 \right) \left(\frac{s[r_1, r_2]}{\delta} \right)^{1/2} (a_2 - a_1, r_1, r_2)^{1/2} (a_2 r_1^3 - a_1 r_2^3, s)^{1/2} \end{aligned}$$

(the gcd factors for $S_1(\delta)$ are divisors of those for $\delta = 1$). Summing over $\delta \mid s[r_1, r_2]$ then gives the result. □

6J.3. *The Katz Sato–Tate law over short intervals.* In this section, which is independent of the rest of this paper, we give a sample application of the van der Corput method to Katz's equidistribution law for the angles of the Kloosterman sums $\mathrm{Kl}_2(n; q)$.

Given a squarefree integer $q \geq 1$ with $\omega(q) \geq 1$ prime factors, we define the *Kloosterman angle* $\theta(n; q) \in [0, \pi]$ by the formula

$$2^{\omega(q)} \cos(\theta(n; q)) = \mathrm{Kl}_2(n; q).$$

As a consequence of the determination of the geometric monodromy group of the Kloosterman sheaf $\mathcal{K}\ell_2$, Katz [1988] proved (among other things) a result which can be phrased as follows:

Theorem 6.27 (Katz's Sato–Tate equidistribution law). *As $p \to \infty$, the set of angles*

$$\{\theta(n; p) : 1 \leq n \leq p\} \subset [0, \pi]$$

becomes equidistributed on $[0, \pi]$ with respect to the Sato–Tate measure μ_{ST} with density

$$\frac{2}{\pi} \sin^2(\theta) \, d\theta,$$

i.e., for any continuous function $f : [0, \pi] \to \mathbb{C}$, we have

$$\int f(x) \, d\mu_{ST}(x) = \lim_{p \to +\infty} \frac{1}{p-1} \sum_{1 \leq n \leq p} f(\theta(n; p)).$$

By the Pólya–Vinogradov method one can reduce the length of the interval $[1, p]$:

Proposition 6.28. *For any $\varepsilon > 0$, the set of angles*

$$\{\theta(n; p) : 1 \le n \le p^{1/2+\varepsilon}\} \subset [0, \pi]$$

becomes equidistributed on $[0, \pi]$ with respect to the Sato–Tate measure μ_{ST} as $p \to +\infty$.

(In fact, using the "sliding sum method" [Fouvry et al. 2013c], one can reduce the range to $1 \le n \le p^{1/2} \Psi(p)$ for any increasing function Ψ with $\Psi(p) \to +\infty$).

As we show here, as a very special example of application of the van der Corput method, we can prove a version of Katz's Sato–Tate law for Kloosterman sums of composite moduli over shorter ranges:

Theorem 6.29. *Let q denote integers of the form $q = rs$ where r, s are two distinct primes satisfying*

$$s^{1/2} \le r \le 2s^{1/2}.$$

For any $\varepsilon > 0$, the set of pairs of angles

$$\{(\theta(n\bar{s}^2; r), \theta(n\bar{r}^2; s)) : 1 \le n \le q^{1/3+\varepsilon}\} \subset [0, \pi]^2$$

becomes equidistributed on $[0, \pi]^2$ with respect to the product measure $\mu_{ST} \times \mu_{ST}$ as $q \to +\infty$ among such integers.

Consequently the set

$$\{\theta(n; q) : 1 \le n \le q^{1/3+\varepsilon}\} \subset [0, \pi]$$

becomes equidistributed on $[0, \pi]$ with respect to the measure $\mu_{ST,2}$ obtained as the pushforward of the measure $\mu_{ST} \times \mu_{ST}$ by the map $(\theta, \theta') \mapsto \mathrm{acos}(\cos\theta \cos\theta')$.

Proof. The continuous functions

$$\mathrm{sym}_{k,k'}(\theta, \theta') := \mathrm{sym}_k(\theta)\, \mathrm{sym}_{k'}(\theta') = \frac{\sin((k+1)\theta)}{\sin\theta} \frac{\sin((k+1)\theta')}{\sin\theta'}$$

for $(k, k') \in \mathbb{N}_{\ge 0} - \{(0, 0)\}$ generate a dense subspace of the space of continuous functions on $[0, \pi]^2$ with mean 0 with respect to $\mu_{ST} \times \mu_{ST}$. Thus, by the classical Weyl criterion, it is enough to prove that

$$\sum_{1 \le n \le q^{1/3+\varepsilon}} \mathrm{sym}_k(\theta(\bar{s}^2 n; r))\, \mathrm{sym}_{k'}(\theta(\bar{r}^2 n; s)) = o(q^{1/3+\varepsilon}).$$

By a partition of unity, it is sufficient to prove that

$$\sum_n \Psi\left(\frac{n}{N}\right) \mathrm{sym}_k(\theta(\bar{s}^2 n; r))\, \mathrm{sym}_{k'}(\theta(\bar{r}^2 n; s)) \ll_{k,k'} q^{1/3+9\varepsilon/10} \qquad (6\text{-}26)$$

for any $N \leq q^{1/3+\varepsilon} \log q$ and any smooth function Ψ as above, where the subscript in $\ll_{k,k'}$ indicates that the implied constant is allowed to depend on k, k'. For any fixed (k, k'), the function

$$x \mapsto \mathrm{sym}_{k'}(\theta(\bar{r}^2 x; s))$$

is a trace function modulo s, namely, the trace function associated to the lisse sheaf obtained by composing the representation corresponding to the rank-2 pullback of the Kloosterman sheaf $[\times \bar{r}^2]^* \mathcal{K}\ell_2$ with the k-th symmetric power representation $\mathrm{sym}_{k'}: \mathrm{GL}_2 \to \mathrm{GL}_{k'+1}$. By [Katz 1988], this sheaf $\mathrm{sym}_{k'} \mathcal{K}\ell_2$ is nontrivial if $k' \geq 1$, and geometrically irreducible of rank $k' + 1 > 1$. Therefore, if $k' \geq 1$, the van der Corput method (6-22) (see also Remark 6.22) gives

$$\sum_n \Psi_N(n)\, \mathrm{sym}_k(\theta(\bar{s}^2 n; r))\, \mathrm{sym}_{k'}(\theta(\bar{r}^2 n; s)) \ll N^{1/2} q^{1/6} \ll_{k,k'} q^{1/3+9\varepsilon/10}.$$

Indeed, $\mathrm{sym}_{k'} \mathcal{K}\ell_2$, being geometrically irreducible of rank > 1, does not contain any quadratic phase.

If $k' = 0$ (so that the function modulo s is the constant function 1), then we have $k \geq 1$ and $\mathrm{sym}_k \mathcal{K}\ell_2$ is geometrically irreducible of rank > 1. Therefore it does not contain any linear phase, and by the Pólya–Vinogradov method (6-21), we obtain

$$\sum_n \Psi_N(n)\, \mathrm{sym}_k(\theta(\bar{s}^2 n; r))\, \mathrm{sym}_{k'}(\theta(\bar{r}^2 n; s)) \ll r^{1/2+\eta}(1 + N/r) \ll_\eta q^{1/6+\eta+\varepsilon}$$

for any $\eta > 0$. $\qquad\qquad\square$

7. The Type III estimate

In this section we establish Theorem 2.8(v). Let us recall the statement:

Theorem 7.1 (new Type III estimates). *Let $\varpi, \delta, \sigma > 0$ be fixed quantities, let I be a bounded subset of $\mathbb{R}$, let $i \geq 1$ be fixed, let $a\ (P_I)$ be a primitive congruence class, and let $M, N_1, N_2, N_3 \gg 1$ be quantities with*

$$MN_1 N_2 N_3 \asymp x, \tag{7-1}$$

$$N_1 N_2,\ N_1 N_3,\ N_2 N_3 \gg x^{1/2+\sigma}, \tag{7-2}$$

$$x^{2\sigma} \ll N_1,\ N_2,\ N_3 \ll x^{1/2-\sigma}. \tag{7-3}$$

Let $\alpha, \psi_1, \psi_2, \psi_3$ be smooth coefficient sequences located at scales M, N_1, N_2, N_3, respectively. Then we have the estimate

$$\sum_{\substack{d \in \mathcal{D}_I(x^\delta) \\ d \ll x^{1/2+2\varpi}}} |\Delta(\alpha \star \psi_1 \star \psi_2 \star \psi_3; a\ (d))| \ll x \log^{-A} x$$

for any fixed $A > 0$, provided that

$$\varpi < \tfrac{1}{12}, \quad \sigma > \tfrac{1}{18} + \tfrac{28}{9}\varpi + \tfrac{2}{9}\delta. \tag{7-4}$$

Our proof of this theorem is inspired in part by the recent work of Fouvry, Kowalski and Michel [Fouvry et al. 2014b], in which the value of the exponent of distribution of the ternary divisor function $\tau_3(n)$ in arithmetic progressions to large (prime) moduli is improved from the earlier results of [Fouvry and Iwaniec 1992] and [Heath-Brown 1986]. Our presentation is also more streamlined. The present argument moreover exploits the existence of an averaging over divisible moduli to derive further improvements to the exponent.

7A. *Sketch of proofs.* Before we give the rigorous argument, let us first sketch the solution of the model problem (in the spirit of Section 5B) of obtaining a nontrivial estimate for

$$\sum_{q \asymp Q} |\Delta(\psi_1 \star \psi_2 \star \psi_3, a\ (q))| \tag{7-5}$$

for Q slightly larger than $x^{1/2}$ in logarithmic scale (i.e., out of reach of the Bombieri–Vinogradov theorem). Here ψ_1, ψ_2, ψ_3 are smooth coefficient sequences at scales N_1, N_2, N_3, respectively, with $N_1 N_2 N_3 \asymp x$ and $N_1, N_2, N_3 \ll \sqrt{x}$, and q is implicitly restricted to suitably smooth or densely divisible moduli (we do not make this precise to simplify the exposition). The trivial bound for this sum is $\ll \log^{O(1)} x$, and we wish to improve it at least by a factor $\log^{-A} x$ for arbitrary fixed $A > 0$.

This problem is equivalent to estimating

$$\sum_{q \asymp Q} c_q \Delta(\psi_1 \star \psi_2 \star \psi_3, a\ (q))$$

when c_q is an arbitrary bounded sequence. As in Section 5B, we write EMT for unspecified main terms, and we wish to control the expression

$$\sum_{q \asymp Q} c_q \sum_{n=a\ (q)} \psi_1 \star \psi_2 \star \psi_3(n) - \mathrm{EMT}$$

to accuracy better than x. After expanding the convolution and completing the sums, this sum can be transformed to a sum roughly of the form

$$\frac{1}{H} \sum_{1 \le |h_i| \ll H_i} \sum \sum \sum_{q \asymp Q} c_q \sum_{\substack{n_1,n_2,n_3 \in \mathbb{Z}/q\mathbb{Z} \\ n_1 n_2 n_3 = a\ (q)}} \sum \sum e_q(h_1 n_1 + h_2 n_2 + h_3 n_3),$$

where $H_i := Q/N_i$ and $H := H_1 H_2 H_3 \asymp Q^3/x$, the main term having canceled out with the zero frequencies. As we are taking Q close to $x^{1/2}$, H is thus close to $x^{1/2}$ as well. Ignoring the degenerate cases when h_1, h_2, h_3 share a common factor with q, we see from (6-20) that

$$\sum_{\substack{n_1,n_2,n_3 \in \mathbb{Z}/q\mathbb{Z} \\ n_1 n_2 n_3 = a\,(q)}} e_q(h_1 n_1 + h_2 n_2 + h_3 n_3) = q\,\mathrm{Kl}_3(a h_1 h_2 h_3; q),$$

so we are now dealing essentially with the sum of hyper-Kloosterman sums

$$\frac{Q}{H} \sum_{1 \le |h_i| \ll H_i} \sum_{q \asymp Q} c_q\,\mathrm{Kl}_3(a h_1 h_2 h_3; q) = \frac{Q}{H} \sum_{1 \le |h| \ll H} \tilde{\tau}_3(h) \sum_{q \asymp Q} c_q\,\mathrm{Kl}_3(ah; q),$$

where

$$\tilde{\tau}_3(h) := \sum_{\substack{1 \le |h_i| \ll H_i \\ h_1 h_2 h_3 = h}} 1$$

is a variant of the divisor function τ_3.

A direct application of the deep Deligne bound

$$|\mathrm{Kl}_3(ah; q)| \ll 1 \tag{7-6}$$

for hyper-Kloosterman sums (see Remark 6.10) gives the trivial bound $\ll Q^2$, which just fails to give the desired result, so the issue is to find some extra cancellation in the phases of the hyper-Kloosterman sums.

One can apply immediately the Cauchy–Schwarz inequality to eliminate the weight $\tilde{\tau}_3(h)$, but it turns out to be more efficient to first use the assumption that q is restricted to densely divisible moduli and to factor q into rs where $r \asymp R$, $s \asymp S$, in which R and S are well-chosen in order to balance the diagonal and off-diagonal components resulting from the Cauchy–Schwarz inequality (it turns out that the optimal choices here will be $R, S \approx x^{1/4}$).

Applying this factorization, and arguing for each s separately, we are led to expressions of the form

$$\frac{Q}{H} \sum_{1 \le |h| \ll H} \tilde{\tau}_3(h) \sum_{r \asymp R} c_{rs}\,\mathrm{Kl}_3(ah; rs),$$

where we must improve on the bound $\ll QR$ coming from (7-6) for any given $s \asymp S$. If we then apply the Cauchy–Schwarz inequality to the sum over h, we get

$$\frac{Q}{H} \sum_{1 \le |h| \ll H} \tilde{\tau}_3(h) \sum_{r \asymp R} c_{rs} \, \mathrm{Kl}_3(ah; rs)$$

$$\ll \frac{Q}{H^{1/2}} \left(\sum_{1 \le |h| \ll H} \left| \sum_{r \asymp R} c_{rs} \, \mathrm{Kl}_3(ah; rs) \right|^2 \right)^{1/2}$$

$$\ll \frac{Q}{H^{1/2}} \left(\sum_{r_1, r_2 \asymp R} \sum_{1 \le |h| \ll H} \mathrm{Kl}_3(ah; r_1 s) \overline{\mathrm{Kl}_3(ah; r_2 s)} \right)^{1/2}.$$

The inner sum over h is now essentially of the type considered by Corollary 6.26, and this result gives an adequate bound. Indeed, the contribution of the diagonal terms $r_1 = r_2$ is $\ll RH$ (using (7-6)) and the contribution of each nondiagonal sum (assuming we are in the model case where r_1, r_2 are coprime, and the other greatest common divisors appearing in Corollary 6.26 are negligible) is

$$\sum_{1 \le |h| \ll H} \mathrm{Kl}_3(ah; r_1 s) \overline{\mathrm{Kl}_3(ah; r_1 s)} \ll (r_1 r_2 s)^{1/2} \ll R S^{1/2}$$

by Corollary 6.26, leading to a total estimate of size

$$\ll \frac{Q}{H^{1/2}} (R^{1/2} H^{1/2} + R^{3/2} S^{1/4}).$$

If $R = S \approx x^{1/4}$, this is very comfortably better than what we want, and this strongly suggests that we can take Q quite a bit larger than $x^{1/2}$.

Remark 7.2. It is instructive to run the same analysis for the fourth-order sum

$$\sum_{q \asymp Q} |\Delta(\psi_1 \star \psi_2 \star \psi_3 \star \psi_4, a \, (q))|,$$

where $\psi_1, \psi_2, \psi_3, \psi_4$ are smooth at scales N_1, N_2, N_3, N_4 with $N_1 \cdots N_4 \asymp x$ and $N_1, \ldots, N_4 \ll x^{1/2} \approx Q$. This is a model for the "Type IV" sums mentioned in Remark 3.2, and is clearly related to the exponent of distribution for the divisor function τ_4.

The quantity H is now of the form $H \approx Q^4/x \approx x$, and one now has to estimate the sum

$$\sum_{1 \le |h| \ll H} \tilde{\tau}_4(h) \sum_{q \asymp Q} c_q \, \mathrm{Kl}_4(ah; q)$$

to accuracy better than $Hx/Q^{3/2} \approx x^{5/4}$. If we apply the Cauchy–Schwarz inequality in the same manner after exploiting a factorization $q = rs$ with $r \asymp R$, $s \asymp S$ and $RS \asymp Q \approx x^{1/2}$, we end up having to control

$$\sum_{r_1, r_2 \asymp R} \left| \sum_{1 \le |h| \ll H} \mathrm{Kl}_4(ah; r_1 s) \overline{\mathrm{Kl}_4(ah; r_2 s)} \right|$$

with accuracy better than $(x^{5/4}/S)^2/H \approx x^{3/2}/S^2$. The diagonal contribution $r_1 = r_2$ is $\ll RH \approx x^{3/2}/S$, and the off-diagonal contribution is $\approx R^2(R^2S)^{1/2} \approx x^{3/2}/S^{5/2}$. However, even with the optimal splitting $S \approx 1$, $R \approx Q$, one cannot make both of these terms much smaller than the target accuracy of $x^{3/2}/S^2$. Thus the above argument does not improve upon the Bombieri–Vinogradov inequality for Type IV sums. (It is known, due to Linnik, that the exponent of distribution for τ_4 is at least $\frac{1}{2}$, in the stronger sense that the asymptotic formula holds for all moduli $\leq x^{1/2-\varepsilon}$ for $\varepsilon > 0$.) The situation is even worse, as the reader will check, for the Type V sums, in that one now cannot even recover Bombieri–Vinogradov with this method.

We will give the rigorous proof of Theorem 2.8(v) in the next two sections, by first performing the reduction to exponential sums, and then concluding the proof.

7B. *Reduction to exponential sums.* By Theorem 2.9 (the general version of the Bombieri–Vinogradov theorem) we have

$$\sum_{q \leq x^{1/2}\log^{-B(A)} x} |\Delta(\alpha \star \psi_1 \star \psi_2 \star \psi_3)| \ll x \log^{-A} x$$

for some $B(A) \geq 0$. We may therefore restrict our attention to moduli q in the range $x^{1/2}/\log^B x \leq q \ll x^{1/2+2\varpi}$.

We also write $N = N_1 N_2 N_3$. From (7-2) and (7-3), we deduce that

$$x^{3/4+3\sigma/2} \ll (N_1 N_2)^{1/2}(N_1 N_3)^{1/2}(N_2 N_3)^{1/2} = N \ll x^{3/2-3\sigma}. \tag{7-7}$$

It is convenient to restrict q to a finer-than-dyadic interval $\mathcal{I}(Q)$ in order to separate variables later using Taylor expansions. More precisely, for a small fixed $\varepsilon > 0$ and some fixed $c \geq 1$, we denote by $\mathcal{I} = \mathcal{I}(Q)$ a finer-than-dyadic interval of the type

$$\mathcal{I}(Q) := \{q : Q(1 - cx^{-\varepsilon}) \leq q \leq Q(1 + cx^{-\varepsilon})\},$$

(assuming, as always, that x is large, so that $cx^{-\varepsilon}$ is less than, say, $\frac{1}{2}$), and abbreviate

$$\sum_q A_q = \sum_{\substack{q \in \mathcal{D}_I(x^\delta) \\ q \in \mathcal{I}(Q)}} A_q$$

for given expression any A_q.

Theorem 7.1 will clearly follow if we prove that, for $\varepsilon > 0$ sufficiently small, we have

$$\sum_q |\Delta(\alpha \star \psi_1 \star \psi_2 \star \psi_3; a\,(q))| \ll x^{-2\varepsilon} MN \tag{7-8}$$

for all Q such that

$$x^{1/2} \ll Q \ll x^{1/2+2\varpi}. \tag{7-9}$$

We fix Q as above and denote by $\Sigma(Q; a)$ the left-hand side of (7-8). We have

$$\Sigma(Q; a) = \sum_q c_q \Delta(\alpha \star \psi_1 \star \psi_2 \star \psi_3; a\ (q))$$

for some sequence c_q with $|c_q| = 1$. We will prove that, for any $a\ (q)$, we have

$$\sum_q c_q \sum_{n=a\ (q)} (\alpha \star \psi_1 \star \psi_2 \star \psi_3)(n) = X + O(x^{-2\varepsilon+o(1)} MN) \qquad (7\text{-}10)$$

for some X that is independent of a (but that can depend on all other quantities, such as c_q, α, or ψ_1, ψ_2, ψ_3). Then (7-8) follows by averaging over all a coprime to P_I (as in the reduction to (5-18) in Section 5).

The left-hand side of (7-10), say $\Sigma_1(Q; a)$, is equal to

$$\Sigma_1(Q; a)$$
$$= \sum_q c_q \sum_{(m,q)=1} \alpha(m) \sum_{n_1,n_2,n_3} \psi_1(n_1)\psi_2(n_2)\psi_3(n_3) \mathbf{1}_{mn_1n_2n_3=a\ (q)}. \quad (7\text{-}11)$$

The next step is a variant of the completion of sums technique from Lemma 4.9. In that lemma, the Fourier coefficients of the cutoff functions were estimated individually using the fast decay of the Fourier transforms. In our current context, we want to keep track to some extent of their dependence on the variable q. Since we have restricted q to a rather short interval, we can separate the variables fairly easily using a Taylor expansion.

Note first that for $i = 1, 2, 3$, one has

$$N_i \ll x^{1/2-\sigma} \ll x^{-\sigma} Q,$$

so in particular ψ_i is supported in $(-q/2, q/2]$ if x is large enough. By discrete Fourier inversion, we have

$$\psi_i(x) = \frac{1}{q} \sum_{-q/2 < h \le q/2} \Psi_i\left(\frac{h}{q}\right) e\left(\frac{hx}{q}\right), \qquad (7\text{-}12)$$

where

$$\Psi_i(y) = \sum_n \psi_i(n) e(-ny)$$

is the analogue of the function Ψ in the proof of Lemma 4.9. As in that lemma, using the smoothness of ψ_i, Poisson summation, and integration by parts, we derive the bound

$$|\Psi_i(y)| \ll N_i (1 + N_i|y|)^{-C}$$

for any fixed $C \ge 0$ and any $-\frac{1}{2} \le y \le \frac{1}{2}$ (see (4-17)). More generally, we obtain

$$|\Psi_i^{(j)}(y)| \ll N_i^{1+j} (1 + N_i|y|)^{-C}$$

for any fixed $C \geq 0$, any $j \geq 0$ and any $-\frac{1}{2} \leq y \leq \frac{1}{2}$.

Denoting $H_i := Q/N_i \gg x^\sigma$, we thus have

$$\Psi_i^{(j)}\left(\frac{h}{q}\right) \ll x^{-100}$$

(say) for $x^{\varepsilon/2} H_i < |h| \leq q/2$ and all fixed j. On the other hand, for $|h| \leq x^{\varepsilon/2} H_i$ and $q \in \mathscr{I}$, a Taylor expansion using the definition of $\mathscr{I}$ and H_i gives

$$\frac{1}{q}\Psi_i\left(\frac{h}{q}\right) = \frac{1}{q}\sum_{j=0}^{J}\frac{1}{j!}\Psi_i^{(j)}(h/Q)\eta^j + O(N_i^{2+J}|\eta|^{J+1})$$

for any fixed J, where α is the q-dependent quantity

$$\eta := \frac{h}{q} - \frac{h}{Q} = \frac{h(Q-q)}{qQ} \ll x^{-\varepsilon}\frac{h}{Q} \ll x^{-\varepsilon/2}\frac{1}{N_i}.$$

Thus we obtain

$$\frac{1}{q}\Psi_i\left(\frac{h}{q}\right) = \frac{1}{q}\sum_{j=0}^{J}\frac{1}{j!}\Psi_i^{(j)}\left(\frac{h}{Q}\right)\left(\frac{h}{Q}\right)^j\left(\frac{q-Q}{q}\right)^j + O(x^{-(J+1)\varepsilon/2}N_i).$$

Taking J large enough, depending on $\varepsilon > 0$ but still fixed, this gives an expansion

$$\frac{1}{q}\Psi_i\left(\frac{h}{q}\right) = 1_{|h|<x^{\varepsilon/2}H_i}\frac{1}{H_i}\sum_{j=0}^{J}c_i(j,h)\frac{Q}{q}\left(\frac{q-Q}{q}\right)^j + O(x^{-100}), \qquad (7\text{-}13)$$

with coefficients that satisfy

$$c_i(j,h) = \frac{1}{j!}\Psi_i^{(j)}\left(\frac{h}{Q}\right)\left(\frac{h}{Q}\right)^j\frac{H_i}{Q} \ll 1,$$

as well as

$$\left(\frac{Q}{q}\right)\left(\frac{q-Q}{q}\right)^j \ll 1.$$

Let

$$H := H_1 H_2 H_3 = Q^3/N. \qquad (7\text{-}14)$$

Inserting (7-13) for $i = 1, 2, 3$ into (7-12) and the definition (7-11) of $\Sigma_1(Q; a)$, we see that $\Sigma_1(Q; a)$ can be expressed (up to errors of $O(x^{-100})$) as a sum of a bounded number (depending on ε) of expressions, each of the form

$$\Sigma_2(Q; a)$$
$$= \frac{1}{H}\sum_q \eta_q \sum_{(m,q)=1} \alpha(m) \sum_{\boldsymbol{h}} c(\boldsymbol{h}) \sum_{\boldsymbol{n}\in(\mathbb{Z}/q\mathbb{Z})^3} e_q(h_1 n_1 + h_2 n_2 + h_3 n_3)\boldsymbol{1}_{mn_1 n_2 n_3 = a\,(q)},$$

where η_q is a bounded sequence supported on $\mathscr{I} \cap \mathscr{D}_I(x^\delta)$, $\boldsymbol{h} := (h_1, h_2, h_3)$ and $c(\boldsymbol{h})$ are bounded coefficients supported on $|h_i| \leq x^{\varepsilon/2} H_i$, and $\boldsymbol{n}$ denotes (n_1, n_2, n_3). Our task is now to show that

$$\Sigma_2(Q; a) = X_2 + O(x^{-2\varepsilon + o(1)} MN)$$

for some quantity X_2 that can depend on quantities such as η_q, α, c, H, but which is independent of a.

We use $F(\boldsymbol{h}, a; q)$ to denote the hyper-Kloosterman type sum

$$F(\boldsymbol{h}, a; q) := \frac{1}{q} \sum_{\boldsymbol{n} \in ((\mathbb{Z}/q\mathbb{Z})^\times)^3} e_q(h_1 n_1 + h_2 n_2 + h_3 n_3) \mathbf{1}_{n_1 n_2 n_3 = a \, (q)} \qquad (7\text{-}15)$$

for $\boldsymbol{h} = (h_1, h_2, h_3) \in (\mathbb{Z}/q\mathbb{Z})^3$ and $a \in (\mathbb{Z}/q\mathbb{Z})^\times$ (note that the constraint $n_1 n_2 n_3 = a \, (q)$ forces n_1, n_2, n_3 to be coprime to q), so that

$$\Sigma_2(Q; a) = \frac{Q}{H} \sum_q \eta'_q \sum_{(m,q)=1} \alpha(m) \sum_{\boldsymbol{h}} c(\boldsymbol{h}) F(\boldsymbol{h}, a\overline{m}; q),$$

where $\eta'_q := (q/Q)\eta_q$ is a slight variant of η_q.

We next observe that $F(\boldsymbol{h}, a\overline{m}; q)$ is independent of a if $h_1 h_2 h_3 = 0$ (as can be seen by a change of variable). Thus the contribution X_2 to the sum from tuples $\boldsymbol{h}$ with $h_1 h_2 h_3 = 0$ is independent of a. The combination of these terms X_2 in the decomposition of $\Sigma_1(Q; a)$ in terms of instances of $\Sigma_2(Q; a)$ is the quantity X in (7-10). We denote by $\Sigma'_2(Q; a)$ the remaining contribution. Our task is now to show that

$$\Sigma'_2(Q, a) \ll x^{-2\varepsilon} MN. \qquad (7\text{-}16)$$

We must handle possible common factors of q and $h_1 h_2 h_3$ for $h_1 h_2 h_3 \neq 0$ (the reader may skip the necessary technical details and read on while assuming that q is always coprime to each of the h_i, so that all the b-factors appearing below become equal to 1).

For $i = 1, 2, 3$, we write

$$h_i = b_i l_i,$$

where $(l_i, q) = 1$ and $b_i \mid q^\infty$ (i.e., b_i is the product of all the primes in h_i, with multiplicity, that also divide q). We also write

$$b := \prod_{p \mid b_1 b_2 b_3} p = (h_1 h_2 h_3, q), \qquad (7\text{-}17)$$

so that we have a factorization $q = bd$, where $d \in \mathscr{D}_I(bx^\delta)$ by Lemma 2.10(i), since q is x^δ-densely divisible.

By Lemma 4.4, we have

$$F(\boldsymbol{h}, a\overline{m}; q) = F(\bar{d}\boldsymbol{h}, a\overline{m}; b)F(\bar{b}\boldsymbol{h}, a\overline{m}; d),$$

where $\bar{b}\boldsymbol{h} := (\bar{b}h_1, \bar{b}h_2, \bar{b}h_3)$. By an easy change of variable, the second factor satisfies

$$F(\bar{b}\boldsymbol{h}, a\overline{m}; d) = \mathrm{Kl}_3(ah_1h_2h_3\overline{mb^3}; d) = \mathrm{Kl}_3\left(\frac{ab_1b_2b_3}{b^3}\frac{l_1l_2l_3}{m}; d\right).$$

We observe that the residue class $ab_1b_2b_3\overline{mb^3}$ (d) is invertible.

Setting $\boldsymbol{b} := (b_1, b_2, b_3)$, $\boldsymbol{l} := (l_1, l_2, l_3)$, we can thus write

$$\Sigma_2'(Q; a) = \frac{Q}{H} \sum_{\boldsymbol{b}} \sum_{\boldsymbol{l}} c(\boldsymbol{b}, \boldsymbol{l}) \sum_{\substack{d \in \mathscr{D}_I(bx^\delta) \\ (d, bl_1l_2l_3)=1}} \eta_{bd}' \sum_{(m, bd)=1} \left(\alpha(m) F(\bar{d}\boldsymbol{h}, a\overline{m}; b) \right.$$
$$\left. \times \mathrm{Kl}_3\left(\frac{ab_1b_2b_3}{b^3}\frac{l_1l_2l_3}{m}; d\right) \right),$$

where b is defined as in (7-17), $c(\boldsymbol{b}, \boldsymbol{l}) := c(b_1l_1, b_2l_2, b_3l_3)$, and the sum over l_i is now over the range

$$0 < |l_i| \leq \frac{x^{\varepsilon/2} H_i}{b_i}. \tag{7-18}$$

To control the remaining factor of F, we have the following estimate, where we denote by $n^\flat$ the largest squarefree divisor of an integer $n \geq 1$ (the *squarefree radical of n*). Note that $b = (b_1 b_2 b_3)^\flat$.

Lemma 7.3. *Let the notation and hypotheses be as above.*

(1) *We have*

$$|F(\bar{d}\boldsymbol{h}, a\overline{m}; b)| \leq \frac{b_1^\flat b_2^\flat b_3^\flat}{b^2}.$$

(2) *The sum $F(\bar{d}\boldsymbol{h}, a\overline{m}; b)$ is independent of d and m.*

Proof. By further applications of Lemma 4.4 it suffices for (1) to show that

$$|F(\boldsymbol{c}, a; p)| \leq \frac{(c_1, p)(c_2, p)(c_3, p)}{p^2}$$

whenever p is prime, $\boldsymbol{c} = (c_1, c_2, c_3) \in (\mathbb{Z}/p\mathbb{Z})^3$, with $c_1c_2c_3 = 0$ (p), and $a \in (\mathbb{Z}/p\mathbb{Z})^\times$. Without loss of generality we may assume that $c_3 = 0$ (p), and then

$$F(\boldsymbol{c}, a; p) = \frac{1}{p} \sum_{n_1, n_2 \in (\mathbb{Z}/p\mathbb{Z})^\times} e_p(c_1n_1 + c_2n_2),$$

from which the result follows by direct computation of Ramanujan sums (see, e.g., [Iwaniec and Kowalski 2004, (3.5)]). Similarly, we see that the value of $F(\boldsymbol{c}, a; p)$ only depends on which c_i are divisible by p and which are not, and this gives (2). $\square$

This lemma leads to the estimate

$$|\Sigma_2'(Q; a)|$$

$$\ll \frac{Q}{H} \sum_{\mathbf{b}} \frac{b_1^\flat b_2^\flat b_3^\flat}{b^2} \sum_{\mathbf{l}} \left| \sum_{\substack{d \in \mathcal{D}_I(bx^\delta) \\ (bl_1 l_2 l_3, d) = 1}} \eta_{bd}' \sum_{(m, bd) = 1} \alpha(m) \, \mathrm{Kl}_3\left(\frac{ab_1 b_2 b_3 l_1 l_2 l_3}{b^3 m}; d\right) \right|$$

$$\ll \frac{Q}{H} \sum_{\mathbf{b}} \frac{b_1^\flat b_2^\flat b_3^\flat}{b^2} T(\mathbf{b}), \tag{7-19}$$

with

$$T(\mathbf{b}) := \sum_{0 < |\ell| \le x^{3\varepsilon/2} H / b_1 b_2 b_3} \tau_3(\ell) \left| \sum_{\substack{d \in b^{-1} \mathcal{D}_I(bx^\delta) \cap \mathcal{I} \\ (b\ell, d) = 1}} \eta_{bd}' \sum_{(m, bd) = 1} \alpha(m) \, \mathrm{Kl}_3\left(\frac{a\ell b_1 b_2 b_3}{b^3 m}; d\right) \right|;$$

following [Heath-Brown 1986] (particularly the arguments on p. 42), we have collected common values of $\ell = l_1 l_2 l_3$, and also replaced the bounded coefficients η_{bd}', supported on $\mathcal{I}$, with their absolute values. This is the desired reduction of Type III estimates to exponential sums.

7C. *End of the proof.* We now focus on estimating $T(\mathbf{b})$. First of all, we may assume that

$$\frac{Q}{b} \gg 1, \quad x^{3\varepsilon/2} \frac{H}{b_1 b_2 b_3} \gg 1, \tag{7-20}$$

since otherwise $T(\mathbf{b}) = 0$.

Let $y = bx^\delta$ and let S be a parameter such that

$$1 \le S \le y \frac{Q}{2b} = \frac{x^\delta Q}{2}. \tag{7-21}$$

The moduli d in the definition of $T(\mathbf{b})$ are y-densely divisible and we have $1 \le S \le dy$ (for x sufficiently large), so that there exists a factorization $d = rs$ with

$$y^{-1} S \le s \le S, \quad \frac{Q}{bS} \ll r \ll \frac{yQ}{bS},$$

and $(r, s) = 1$ (if $d < S \le dy$, we take $s = d$ and $r = 1$).

Thus we may write

$$T(\mathbf{b}) \ll \sum_{\substack{y^{-1} S \le s \le S \\ (b\ell, s) = 1}} \sum_{0 < |\ell| \le H_b} \tau_3(\ell) \left| \sum_{\substack{r \in \mathcal{S}_I \\ \frac{Q}{bS} \ll r \ll \frac{yQ}{bS} \\ (b\ell s, r) = 1}} \eta_{b, r, s}' \sum_{(m, brs) = 1} \alpha(m) \, \mathrm{Kl}_3\left(\frac{a\ell b_1 b_2 b_3}{b^3 m}; rs\right) \right|,$$

where $\eta'_{b,r,s}$ is some bounded sequence and

$$H_b := \frac{x^{3\varepsilon/2} H}{b_1 b_2 b_3}.$$

We apply the Cauchy–Schwarz inequality to the sum over s and l. As usual, we may insert a smooth coefficient sequence ψ_{H_b} at scale H_b, equal to 1 on $[-H_b, H_b]$, and derive

$$|T(b)|^2 \leq T_1 T_2,$$

where

$$T_1 := \sum_{y^{-1}S \leq s \leq S} \frac{1}{s} \sum_{0 < |\ell| \leq H_b} \tau_3(\ell)^2 \ll H_b$$

(by Equation (1-2)) and

$$T_2 := \sum_{y^{-1}S \leq s \leq S} \sum_{\ell} s \psi_{H_b}(\ell) \left| \sum_{\substack{r \in \mathscr{S}_I \\ \frac{Q}{bS} \ll r \ll \frac{yQ}{bS} \\ (b\ell, rs) = (r,s) = 1}} \eta'_{b,r,s} \sum_{(m,brs)=1} \alpha(m) \operatorname{Kl}_3 \left(\frac{a\ell b_1 b_2 b_3}{b^3 m} ; rs \right) \right|^2 .$$

We expand the square and find

$$|T_2| \leq \sum_{y^{-1}S \leq s \leq S} s \sum_{r_1, r_2} \sum_{m_1, m_2} |\alpha(m_1)| |\alpha(m_2)| |U(r_1, r_2, s, m_1, m_2)|,$$

where we have omitted the summation conditions

$$r_i \in \mathscr{S}_I; \quad \frac{Q}{bS} \ll r_i \ll \frac{yQ}{bS}; \quad (b\ell, r_i s) = (r_i, s) = (m_i, br_i s) = 1 \quad \text{for } i = 1, 2$$

on r_1, r_2 and m_1, m_2 for brevity, and where

$$U(r_1, r_2, s, m_1, m_2) := \sum_{\ell:(\ell, r_1 r_2 s)=1} \psi_{H_b}(\ell) \operatorname{Kl}_3 \left(\frac{a\ell b_1 b_2 b_3}{b^3 m_1} ; r_1 s \right) \overline{\operatorname{Kl}_3 \left(\frac{a\ell b_1 b_2 b_3}{b^3 m_2} ; r_2 s \right)}$$

is exactly the type of sum considered in Corollary 6.26 (recall that $ab_1 b_2 b_3$ is coprime to $r_1 r_2 s$).

We first consider the "diagonal terms", which here mean the cases where

$$\frac{ab_1 b_2 b_3}{b^3 m_1} r_2^3 - \frac{ab_1 b_2 b_3}{b^3 m_2} r_1^3 = \frac{ab_1 b_2 b_3}{b^3 m_1 m_2} (m_2 r_2^3 - m_1 r_1^3) = 0.$$

Using the Deligne bound $|\operatorname{Kl}_3(x; d)| \ll 1$ when $(d, x) = 1$ (Remark 6.10), this contribution T_2' satisfies the bound

$$T_2' \ll H_b \sum_{r_1, r_2} \sum \sum_{y^{-1}S \leq s \leq S} s \sum \sum_{\substack{m_1, m_2 \\ m_1 r_1^3 = m_2 r_2^3}} |\alpha(m_1)\alpha(m_2)|$$

$$\ll H_b M \sum_{Q/(bS) \ll r_1 \ll yQ/(bS)} \left(\frac{Q}{br_1}\right)^2$$

since each pair (r_1, m_1) determines $\ll 1$ pairs (r_2, m_2), and since s is, for each r_1, constrained to be $\asymp Q/(br_1)$ by the condition $r_1 s \asymp Q/b$. Summing, we obtain

$$T_2' \ll \frac{H_b M Q S}{b}. \tag{7-22}$$

We now turn to the off-diagonal case $m_1 r_1^3 - m_2 r_2^3 \neq 0$. By Corollary 6.26, we have

$$U(r_1, r_2, s, m_1, m_2)$$
$$\ll \left(\frac{H_b}{[r_1, r_2]s} + 1\right)(s[r_1, r_2])^{1/2}(r_1, r_2, m_2 - m_1)^{1/2}(m_1 r_1^3 - m_2 r_2^3, s)^{1/2}$$

in this case. We now sum these bounds to estimate the nondiagonal contribution T_2'' to T_2. This is a straightforward, if a bit lengthy, computation, and we state the result first:

Lemma 7.4. *We have*

$$T_2'' \ll \frac{M^2 Q^2}{b^2}\left(\frac{H_b b^{1/2}}{Q^{1/2}}\left(\frac{bS}{Q}\right)^{1/2} + \frac{Q^{1/2}}{b^{1/2}}\left(\frac{x^\delta Q}{S}\right)^{1/2}\right).$$

We first finish the proof of the Type III estimate using this. We first derive

$$T_2 = T_2' + T_2'' \ll \frac{MQH_b S}{b} + \frac{M^2 QS^{1/2}H_b}{b} + \frac{y^{1/2}M^2 Q^3}{b^3 S^{1/2}}.$$

We select the parameter S now, by optimizing it to minimize the sum of the first and last terms, subject to the constraint $S \leq (yQ)/(2b)$. Precisely, let

$$S = \min\left(\left(\frac{Q}{b}\right)^{4/3}\frac{y^{1/3}M^{2/3}}{H_b^{2/3}}, \frac{yQ}{2b}\right).$$

This satisfies (7-21) if x is large enough: we have $S \leq (yQ)/(2b)$ by construction, while $S \geq 1$ (for x large enough) follows either from $(yQ)/(2b) \gg y/2$ (see (7-20)), or from

$$\left(\frac{Q}{b}\right)^4 \frac{yM^2}{H_b^2} = \frac{(b_1 b_2 b_3)^2}{b^2}\frac{(MN)^2 x^{\delta - 3\varepsilon}}{bQ^2} \gg x^{2+\delta-3\varepsilon}Q^{-3} \gg x^{1/2+\delta-6\varpi-3\varepsilon} \gg x^\varepsilon$$

if $\varepsilon > 0$ is small enough (using $b \ll Q$ and $\varpi < \frac{1}{12}$).

This value of S leads to

$$|T(\boldsymbol{b})|^2 \ll H_{\boldsymbol{b}}\left(\frac{y^{1/3}H_{\boldsymbol{b}}^{1/3}M^{5/3}Q^{7/3}}{b^{7/3}} + \frac{y^{1/6}H_{\boldsymbol{b}}^{2/3}M^{7/3}Q^{5/3}}{b^{5/3}} + M^2\left(\frac{Q}{b}\right)^{5/2}\right)$$

(where the third term only arises if $S = (yQ)/(2b)$), which gives

$$T(\boldsymbol{b}) \ll \frac{x^{5\varepsilon/4}}{(b_1 b_2 b_3)^{1/2}b}(x^{\delta/6}H^{2/3}M^{5/6}Q^{7/6} + x^{\delta/12}H^{5/6}M^{7/6}Q^{5/6} + H^{1/2}MQ^{5/4})$$

using the definition of $H_{\boldsymbol{b}}$ and the bound $b_i \geq 1$ (to uniformize the three denominators involving b and $\boldsymbol{b}$).

We will shortly establish the following elementary fact:

Lemma 7.5. *The unsigned series*

$$\sum\sum\sum_{b_1,b_2,b_3 \geq 1} \frac{b_1^\flat b_2^\flat b_3^\flat}{(b_1 b_2 b_3)^{1/2}b^3}$$

converges to a finite value.

Now from (7-19) and this lemma, we get

$$\Sigma_2'(Q; a) \ll \frac{x^{5\varepsilon/4}Q}{H}(x^{\delta/6}H^{2/3}M^{5/6}Q^{7/6} + x^{\delta/12}H^{5/6}M^{7/6}Q^{5/6} + H^{1/2}MQ^{5/4}).$$

We now show that this implies (7-16) under suitable conditions on δ, ϖ and σ. Indeed, we have

$$\frac{x^{5\varepsilon/4}Q}{H}(x^{\delta/6}H^{2/3}M^{5/6}Q^{7/6} + x^{\delta/12}H^{5/6}M^{7/6}Q^{5/6} + H^{1/2}MQ^{5/4})$$
$$\ll MN(E_1 + E_2 + E_3),$$

where

$$E_1 := \frac{x^{5\varepsilon/4+\delta/6}Q^{13/6}}{H^{1/3}M^{1/6}N} = \frac{x^{5\varepsilon/4+\delta/6-1/6}Q^{7/6}}{N^{1/2}} \ll Q^{7/6}x^{5\varepsilon/4+\delta/6-3\sigma/4-13/24},$$

$$E_2 := \frac{x^{5\varepsilon/4+\delta/12}Q^{11/6}M^{7/6}}{H^{1/6}MN} = \frac{x^{5\varepsilon/4+\delta/12+1/6}Q^{4/3}}{N} \ll Q^{4/3}x^{5\varepsilon/4+\delta/12-3\sigma/2-7/12},$$

$$E_3 := \frac{x^{5\varepsilon/4}Q^{9/4}}{H^{1/2}N} = \frac{x^{5\varepsilon/4}Q^{3/4}}{N^{1/2}} \ll Q^{3/4}x^{5\varepsilon/4-3/8-3\sigma/4},$$

using the definition (7-14) of H and the lower bound (7-7) for N. Using $Q \ll x^{1/2+2\varpi}$, we see that we will have $E_1 + E_2 + E_3 \ll x^{-2\varepsilon}$ for some small positive $\varepsilon > 0$ provided

$$\begin{cases} \frac{7}{6}\left(\frac{1}{2}+2\varpi\right)+\frac{\delta}{6}-\frac{3\sigma}{4}-\frac{13}{24}<0, \\ \frac{4}{3}\left(\frac{1}{2}+2\varpi\right)+\frac{\delta}{12}-\frac{3\sigma}{2}-\frac{7}{12}<0, \\ \frac{3}{4}\left(\frac{1}{2}+2\varpi\right)-\frac{3\sigma}{4}-\frac{3}{8}<0, \end{cases} \iff \begin{cases} \sigma>\frac{28}{9}\varpi+\frac{2}{9}\delta+\frac{1}{18}, \\ \sigma>\frac{16}{9}\varpi+\frac{1}{18}\delta+\frac{1}{18}, \\ \sigma>2\varpi. \end{cases}$$

However, the first condition implies the second and third. Thus we deduce Theorem 7.1, provided that we prove the two lemmas above, which we will now do.

Proof of Lemma 7.4. We will relax somewhat the conditions on r_1, r_2 and s. We recall first that

$$\frac{Q}{bS}\ll r_1,r_2\ll\frac{yQ}{bS}=\frac{x^\delta Q}{S}.$$

Furthermore, the summation conditions imply $r_1 s\asymp Q/b\asymp r_2 s$, and in particular r_1 and r_2 also satisfy $r_1\asymp r_2$. In addition, as above, we have $s\asymp Q/(br_1)$ for a given r_1.

Using this last property to fix the size of s, we have

$$T_2''\ll\frac{Q}{b}\sum_{\frac{Q}{bS}\ll r_1\asymp r_2\ll\frac{yQ}{bS}}\sum\frac{1}{r_1}\left(\frac{H_b(br_1)^{1/2}}{(Q[r_1,r_2])^{1/2}}+\frac{(Q[r_1,r_2])^{1/2}}{(br_1)^{1/2}}\right)$$
$$\sum_{\substack{m_1,m_2\asymp M \\ r_1^3 m_1\neq r_2^3 m_2}}\sum(r_1,r_2,m_1-m_2)^{1/2}\sum_{s\asymp Q/(br_1)}(r_1^3 m_1-r_2^3 m_2,s)^{1/2}.$$

By Lemma 1.4, the inner sum is $\ll Q/(br_1)$ for all (r_1,r_2,m_1,m_2), and similarly, we get

$$\sum_{m_1,m_2\asymp M}\sum(r_1,r_2,m_1-m_2)^{1/2}\ll M^2+M(r_1,r_2)^{1/2},$$

so that

$$T_2''\ll\left(\frac{Q}{b}\right)^2\sum_{\frac{Q}{bS}\ll r_1\asymp r_2\ll\frac{yQ}{bS}}\sum\frac{1}{r_1^2}(M^2+M(r_1,r_2)^{1/2})\left(\frac{H_b(br_1)^{1/2}}{(Q[r_1,r_2])^{1/2}}+\frac{(Q[r_1,r_2])^{1/2}}{(br_1)^{1/2}}\right).$$

We set $r=(r_1,r_2)$ and write $r_i=rt_i$, and thus obtain

$$T_2''\ll\left(\frac{Q}{b}\right)^2\sum_{r\ll\frac{yQ}{bS}}\frac{M^2+r^{1/2}M}{r^2}\sum_{\frac{Q}{rbS}\ll t_1\asymp t_2\ll\frac{yQ}{rbS}}\sum\frac{1}{t_1^2}\left(\frac{H_b b^{1/2}}{(Qt_2)^{1/2}}+\frac{(Qt_2)^{1/2}}{b^{1/2}}\right)$$
$$\ll\left(\frac{Q}{b}\right)^2\sum_{r\ll\frac{yQ}{bS}}\frac{M^2+r^{1/2}M}{r^2}\sum_{\frac{Q}{rbS}\ll t_2\ll\frac{yQ}{rbS}}\left(\frac{H_b b^{1/2}}{Q^{1/2}t_2^{3/2}}+\frac{Q^{1/2}}{b^{1/2}t_2^{1/2}}\right)$$
$$\ll\left(\frac{MQ}{b}\right)^2\left(\frac{H_b b^{1/2}}{Q^{1/2}}\left(\frac{Q}{bS}\right)^{-1/2}+\frac{Q^{1/2}}{b^{1/2}}\left(\frac{yQ}{bS}\right)^{1/2}\right),$$

as claimed. (Note that it was important to keep track of the condition $r_1\asymp r_2$.) □

Proof of Lemma 7.5. If we write $t_i := b_i^{\flat}$, $b_i = t_i u_i$, then we have $t_i \mid b$, $u_i \mid t_i^{\infty}$ and

$$\frac{b_1^{\flat} b_2^{\flat} b_3^{\flat}}{(b_1 b_2 b_3)^{1/2} b^3} = \frac{1}{b^3} \prod_{i=1}^{3} \frac{t_i^{1/2}}{u_i^{1/2}},$$

and thus we can bound the required series by

$$\sum_{b \geq 1} \frac{1}{b^3} \left(\sum_{t \mid b} t^{1/2} \sum_{u \mid t^{\infty}} \frac{1}{u^{1/2}} \right)^3.$$

Using Euler products, we have

$$\sum_{u \mid t^{\infty}} \frac{1}{u^{1/2}} \leq \tau(t)^{O(1)}$$

and thus

$$\sum_{t \mid b} t^{1/2} \sum_{u \mid t^{\infty}} \frac{1}{u^{1/2}} \leq \tau(b)^{O(1)} b^{1/2},$$

and the claim now follows from another Euler product computation. $\qquad\square$

8. An improved Type I estimate

In this final section, we prove the remaining Type I estimate from Section 5, namely Theorem 5.1(iii). In Section 5C, we reduced this estimate to the exponential sum estimate of Theorem 5.8(iii).

8A. *First reduction.* The reader is invited to review the definition and notation of Theorem 5.8. We consider the sum

$$\Upsilon := \sum_{r} \Upsilon_{\ell,r}(b_1, b_2; q_0)$$

of (5-32) for each $1 \leq |\ell| \ll N/R$, where $\Upsilon_{\ell,r}$ was defined in (5-30) and the sum over r is restricted to $r \in \mathcal{D}_I^{(2)}(x^{\delta+o(1)}) \cap [R, 2R]$ (the property that r is doubly densely divisible being part of the assumptions of 5.8(iii)). Our task is to show the bound

$$\Upsilon \ll x^{-\varepsilon} Q^2 RN(q_0, \ell) q_0^{-2}$$

under the hypotheses of Theorem 5.8(iii).

In contrast to the Type I and II estimates of Section 5 (but similarly to the Type III estimate), we will exploit here the average over r, and hence the treatment will combine some features of all the methods used before.

As before, we set

$$H := x^{\varepsilon} R Q^2 M^{-1} q_0^{-1}. \tag{8-1}$$

We recall that, from (5-31), we have $H \gg 1$. We begin as in Section 5F by exploiting the x^δ-dense divisibility of $q_0 q_1$, which implies the $x^\delta q_0$-dense divisibility of q_1 by Lemma 2.10(i). Thus we reduce by dyadic decomposition to the proof of

$$\sum_r \Upsilon_{U,V} \ll x^{-\varepsilon}(q_0, \ell) R Q^2 N q_0^{-2} \tag{8-2}$$

(which corresponds to (5-39) with the average over r preserved), where

$$\Upsilon_{U,V} :=$$

$$\sum_{1 \leq |h| \leq H} \sum_{u_1 \asymp U} \sum_{v_1 \asymp V} \sum_{\substack{q_2 \asymp Q/q_0 \\ (u_1 v_1, q_0 q_2) = 1}} \left| \sum_n C(n) \beta(n) \overline{\beta(n + \ell r)} \Phi_\ell(h, n, r, q_0, u_1 v_1, q_2) \right|$$

as in Section 5F, whenever

$$q_0^{-1} x^{-\delta - 2\varepsilon} Q/H \ll U \ll x^{-2\varepsilon} Q/H, \tag{8-3}$$

$$q_0^{-1} x^{2\varepsilon} H \ll V \ll x^{\delta + 2\varepsilon} H, \tag{8-4}$$

$$UV \asymp Q/q_0 \tag{8-5}$$

(which are identical to the constraints (5-40), (5-41) and (5-42)), and whenever the parameters (ϖ, δ, σ) satisfy the conditions of Theorem 5.8(iii). As before, u_1, v_1 are understood to be squarefree.

We replace again the modulus by complex numbers c_{r,h,u_1,v_1,q_2} of modulus ≤ 1, which we may assume to be supported on parameters (r, h, u_1, v_1, q_2) with

$$(u_1 v_1, q_2) = 1$$

and with

$$q_0 u_1 v_1 r, \ q_0 q_2 r \ \text{ squarefree.}$$

(These numbers c_{r,h,u_1,v_1,q_2} are unrelated to the exponent c in Theorem 5.1.) We then move the sums over r, n, u_1 and q_2 outside and apply the Cauchy–Schwarz inequality as in the previous sections to obtain

$$\left| \sum_r \Upsilon_{U,V} \right|^2 \leq \Upsilon_1 \Upsilon_2$$

with

$$\Upsilon_1 := \sum_r \sum_{\substack{u_1 \asymp U \\ q_2 \asymp Q/q_0}} \sum_n C(n) |\beta(n)|^2 |\beta(n + \ell r)|^2 \ll (q_0, \ell) \frac{NQRU}{q_0^2}$$

(again as in (5-35)) and

$$\Upsilon_2 := \sum_{r} \sum_{\substack{u_1 \asymp U \\ q_2 \asymp Q/q_0}} \sum_{n} \psi_N(n) C(n) \left| \sum_{v_1 \asymp V} \sum_{1 \leq |h| \leq H} c_{h,r,u_1,v_1,q_2} \Phi_\ell(h,n,r,q_0,u_1 v_1, q_2) \right|^2$$

$$= \sum_{r} \sum_{\substack{u_1 \asymp U \\ q_2 \asymp Q/q_0}} \sum_{v_1,v_2 \asymp V} \sum_{1 \leq |h_1|,|h_2| \leq H} \left(c_{h_1,r,u_1,v_1,q_2} \overline{c_{h_2,r,u_1,v_2,q_2}} \right.$$
$$\left. \times T_{\ell,r}(h_1,h_2,u_1,v_1,v_2,q_2) \right),$$

where $T_{\ell,r}$ is defined by (5-43) and ψ_N is a smooth coefficient sequence at scale N.

The analysis of Υ_2 will now diverge from Section 5F. In our setting, the modulus r is doubly $x^{\delta+o(1)}$-densely divisible. As in the previous section, we will exploit this divisibility to split the average and apply the Cauchy–Schwarz inequality a second time.

Let D be a parameter such that

$$1 \ll D \ll x^\delta R, \tag{8-6}$$

which will be chosen and optimized later. By definition (see Definition 2.1) of doubly densely divisible integers, for each r, there exists a factorization $r = d r_1$ where

$$x^{-\delta} D \ll d \ll D$$

and where r_1 is $x^{\delta+o(1)}$-densely divisible (and $(d, r_1) = 1$, since r is squarefree). As before, in the case $D \geq R$ one can simply take $d = r$ and $r_1 = 1$.

We consider the sums

$$\Upsilon_3 := \sum_{\substack{d \asymp \Delta \\ (d,r_1)=1}} \sum_{1 \leq |h_1|,|h_2| \leq H} \sum_{\substack{v_1,v_2 \asymp V \\ (v_1 v_2, d r_1 q_0 u_1 q_2)=1}} |T_{\ell,d r_1}(h_1, h_2, u_1, v_1, v_2, q_2)|,$$

with d understood to be squarefree, for all Δ such that

$$\max(1, x^{-\delta} D) \ll \Delta \ll D \tag{8-7}$$

and all (r_1, u_1, q_2) such that

$$r_1 \asymp R/\Delta, \quad u_1 \asymp U, \quad q_2 \asymp Q/q_0, \tag{8-8}$$

and such that $r_1 q_0 u_1 q_2$ is squarefree and the integers r_1, $q_0 u_1 v_1$, $q_0 u_1 v_2$ and $q_0 q_2$ are $x^{\delta+o(1)}$-densely divisible.

For a suitable choice of D, we will establish the bound

$$\Upsilon_3 \ll (q_0, \ell) x^{-2\varepsilon} \Delta N V^2 q_0 \tag{8-9}$$

for all such sums. It then follows by dyadic subdivision of the variable d and by trivial summation over r_1, u_1 and q_2 that

$$\Upsilon_2 \ll (q_0, \ell) x^{-2\varepsilon} N V^2 q_0 \frac{RUQ}{q_0} = (q_0, \ell) x^{-2\varepsilon} N R U V^2 Q,$$

and hence that

$$\left| \sum_r \Upsilon_{U,V} \right|^2 \ll (q_0, \ell)^2 x^{-2\varepsilon} N^2 R^2 \left(\frac{Q}{q_0} \right)^4,$$

which gives the desired result.

We first write $\Upsilon_3 = \Upsilon_3' + \Upsilon_3''$, where Υ_3' is the diagonal contribution determined by $h_1 v_2 = h_2 v_1$. The number of quadruples (h_1, v_1, h_2, v_2) satisfying this condition is $\ll HV$ by the divisor bound, and therefore a trivial bound $\ll N$ for $T_{\ell,r}(h_1, h_2, u_1, v_1, v_2, q_2)$ gives

$$\Upsilon_3' \ll \Delta HNV \ll (q_0, \ell) x^{-2\varepsilon} \Delta N V^2 q_0$$

by (8-4). We now write

$$\Upsilon_3'' = \sum_{\substack{(h_1, v_1, h_2, v_2) \\ h_1 v_2 \neq h_2 v_1}} \sum \Upsilon_4(h_1, v_1, h_2, v_2),$$

where h_1, v_1, h_2, v_2 obey the same constraints as in the definition of Υ_3, and

$$\Upsilon_4(h_1, v_1, h_2, v_2) := \sum_{\substack{d \asymp \Delta \\ (d, r_1) = 1}} |T_{\ell, dr_1}(h_1, h_2, u_1, v_1, v_2, q_2)|.$$

We will shortly establish the following key estimate:

Proposition 8.1. *If $\varepsilon > 0$ is small enough, then we have*

$$\Upsilon_4(h_1, v_1, h_2, v_2) \ll (q_0, \ell) x^{-2\varepsilon} \Delta N H^{-2} q_0 (h_1 v_2 - h_2 v_1, q_0 q_2 r_1 u_1 [v_1, v_2]),$$

if we take

$$D := x^{-5\varepsilon} \frac{N}{H^4} \tag{8-10}$$

and if

$$\begin{cases} \frac{160}{3} \varpi + 16\delta + \frac{34}{9} \sigma < 1, \\ 64\varpi + 18\delta + 2\sigma < 1. \end{cases} \tag{8-11}$$

Assuming this proposition, we obtain

$$\Upsilon_3'' \ll (q_0, \ell) x^{-2\varepsilon} \Delta N V^2 q_0,$$

and hence (8-9), by the following lemma, which will be proved later:

Lemma 8.2. *We have*

$$\sum_{\substack{(h_1,v_1,h_2,v_2) \\ h_1v_2 \neq h_2v_1}} \sum (h_1v_2 - h_2v_1, q_0q_2r_1u_1[v_1, v_2]) \ll H^2V^2.$$

8B. *Reduction of Proposition 8.1 to exponential sums.* We now consider a specific choice of parameters r_1, u_1, q_2 and (h_1, v_1, h_2, v_2), so that $\Upsilon_4 = \Upsilon_4(h_1, v_1, h_2, v_2)$ is a sum with two variables which we write as

$$\Upsilon_4 = \sum_{d \asymp \Delta} \left| \sum_n \psi_N(n) C(n) \Psi(d, n) \right|,$$

where $C(n)$ restricts n to the congruence (5-23) and

$$\Psi(d, n) := \Phi_\ell(h_1, n, dr_1, q_0, u_1v_1, q_2) \overline{\Phi_\ell(h_2, n, dr_1, q_0, u_1v_2, q_2)}. \qquad (8\text{-}12)$$

We define D by (8-10), and we first check that this satisfies the constraints (8-6). Indeed, we first have

$$D = x^{-5\varepsilon} \frac{N}{H^4} = \frac{x^{-9\varepsilon} q_0^4 N M^4}{Q^8 R^4} \gg x^{-9\varepsilon - 16\varpi} \frac{R^4}{N^3} \gg x^{1/2 - \sigma - 16\varpi - 4\delta - 21\varepsilon}$$

by (5-2) and (5-12). Under the condition (8-11), this gives $D \gg 1$ if $\varepsilon > 0$ is taken small enough.

Moreover, since $H \gg 1$, we have

$$D = x^{-5\varepsilon} \frac{N}{H^4} \ll x^{-5\varepsilon} N \ll x^{-2\varepsilon + \delta} R \leq x^\delta R.$$

We apply the van der Corput technique with respect to the modulus d. Let

$$L := x^{-\varepsilon} \left\lfloor \frac{N}{\Delta} \right\rfloor. \qquad (8\text{-}13)$$

Note that from (8-6) and (5-12), it follows that $L \gg x^{-\varepsilon} N R^{-1} \geq 1$ for x sufficiently large.

For any l with $1 \leq l \leq L$, we have

$$\sum_n \psi_N(n) C(n) \Psi(d, n) = \sum_n \psi_N(n + dl) C(n + dl) \Psi(d, n + dl),$$

and therefore

$$|\Upsilon_4| \leq \frac{1}{L} \sum_{d \asymp \Delta} \sum_{n \ll N} \left| \sum_{l=1}^L \psi_N(n + dl) C(n + dl) \Psi(d, n + dl) \right|.$$

By the Cauchy–Schwarz inequality, for some smooth coefficient sequence ψ_Δ at scale Δ, we have

$$|\Upsilon_4|^2 \leq \frac{N\Delta}{L^2}|\Upsilon_5|, \tag{8-14}$$

where

$$\Upsilon_5 := \sum_{d \asymp \Delta} \psi_\Delta(d) \sum_n \left| \sum_{l=1}^{L} \psi_N(n+dl)C(n+dl)\Psi(d, n+dl) \right|^2.$$

Lemma 8.3. *Let*

$$m = q_0 r_1 u_1 [v_1, v_2] q_2.$$

There exist residue classes α (m) and β (m), independent of n and l, such that for all n and l we have

$$\Psi(d, n+dl) = \xi(n, d)e_m\left(\frac{\alpha}{d(n+(\beta+l)d)}\right),$$

where $|\xi(n, d)| \leq 1$. Moreover we have $(\alpha, m) = (h_1 v_2 - h_2 v_1, m)$.

Proof. From the definitions (8-12) and (5-30), if $\Psi(d, n)$ does not vanish identically, then we have

$$\Psi(d, n+dl)$$
$$= e_{dr_1}\left(\frac{a(h_1 - h_2)}{(n+dl)q_0 u_1 v_1 q_2}\right) e_{q_0 u_1 v_1}\left(\frac{b_1 h_1}{(n+dl)dr_1 q_2}\right) e_{q_0 u_1 v_2}\left(-\frac{b_1 h_2}{(n+dl)dr_1 q_2}\right)$$
$$\times e_{q_2}\left(\frac{b_2 h_1}{(n+dl+d\ell r_1)dr_1 q_0 u_1 v_1}\right) e_{q_2}\left(-\frac{b_2 h_2}{(n+dl+d\ell r_1)dr_1 q_0 u_1 v_2}\right).$$

By the Chinese remainder theorem, the first factor splits into a phase $e_d(\cdots)$ that is independent of l, and an expression involving e_{r_1}, which, when combined with the other four factors by another application of the Chinese remainder theorem, becomes an expression of the type

$$e_m\left(\frac{\alpha}{d(n+ld+\beta d)}\right)$$

for some residue classes α and β modulo m which are independent of l. Furthermore (α, m) is the product of primes p dividing m such that the product of these four factors is trivial, which (since $(q_2, q_0 u_1 [v_1, v_2]) = 1$) occurs exactly when $p \mid h_2 v_1 - h_1 v_2$ (recall that b_1 and b_2 are invertible residue classes). $\qquad\square$

Using this lemma, and the notation introduced there, it follows that

$$\left| \sum_{l=1}^{L} \psi_N(n+dl)C(n+dl)\Psi(d,n+dl) \right|^2$$

$$\leq \sum_{1\leq l_1,l_2\leq L} \psi_N(n+dl_1)\psi_N(n+dl_2)C(n+dl_1)C(n+dl_2)$$

$$e_m\left(\frac{\alpha}{d(n+\beta d+l_1 d)}\right)e_m\left(-\frac{\alpha}{d(n+\beta d+l_2 d)}\right)$$

$$= \sum_{1\leq l_1,l_2\leq L} \psi_N(n+dl_1)\psi_N(n+dl_2)e_m\left(\frac{\alpha(l_2-l_1)}{(n+\beta d+l_1 d)(n+\beta d+l_2 d)}\right),$$

and therefore, after shifting n by dl_1, writing $l := l_2 - l_1$, and splitting n, d into residue classes modulo q_0, that

$$\Upsilon_5 \leq \sum_{n_0,d_0\in\mathbb{Z}/q_0\mathbb{Z}} C(n_0)\Upsilon_5(n_0,d_0),$$

where

$$\Upsilon_5(n_0,d_0) := \sum_{\substack{|l|\leq L-1 \\ 1\leq l_1\leq L}} \sum \left| \sum_{d=d_0\,(q_0)} \psi_\Delta(d) \sum_{n=n_0\,(q_0)} \psi_N(n)\psi_N(n+dl) \right.$$

$$\left. \times e_m\left(\frac{\alpha l}{(n+\beta d)(n+(\beta+l)d)}\right) \right|. \quad (8\text{-}15)$$

Note that m is squarefree. Also, as m is the least common multiple of the $x^{\delta+o(1)}$-densely divisible quantities $r_1, q_0 u_1 v_1, q_0 u_1 v_2,$ and $q_0 q_2$, Lemma 2.10(ii) implies that m is also $x^{\delta+o(1)}$-densely divisible.

The contribution of $l = 0$ to $\Upsilon_5(n_0,d_0)$ is trivially

$$\ll \frac{NL\Delta}{q_0^2}, \quad (8\text{-}16)$$

and this gives a contribution of size

$$\ll \sqrt{(q_0,\ell)}\frac{N\Delta}{\sqrt{q_0 L}}$$

to Υ_4, as can be seen by summing over the $q_0(q_0,\ell)$ permitted residue classes $(n_0\,(q_0), d_0\,(q_0))$. Using (8-10), we have

$$\Delta \ll D = x^{-5\varepsilon}\frac{N}{H^4},$$

and we see from (8-13) that this contribution is certainly

$$\ll (q_0,\ell)x^{-2\varepsilon}\Delta NH^{-2}q_0,$$

and hence suitable for Proposition 8.1.

Let $\Upsilon_5'(n_0, d_0)$ and Υ_5' denote the remaining contributions to $\Upsilon_5(n_0, d_0)$ and Υ_5, respectively. It will now suffice to show that

$$\frac{N\Delta}{L^2}|\Upsilon_5'| \ll \left((q_0, \ell)x^{-2\varepsilon}\Delta N H^{-2}q_0\,(h_1 v_2 - h_2 v_1, q_0 q_2 r_1 u_1 [v_1, v_2])\right)^2. \quad (8\text{-}17)$$

We have

$$\Upsilon_5'(n_0, d_0) = \sum_{\substack{1 \le |l| \le L-1 \\ 1 \le l_1 \le L}} |\Upsilon_6(n_0, d_0)|, \quad (8\text{-}18)$$

where

$$\Upsilon_6(n_0, d_0)$$
$$:= \sum_{d=d_0\,(q_0)} \psi_\Delta(d) \sum_{n=n_0\,(q_0)} \psi_N(n)\psi_N(n+dl)e_m\left(\frac{\alpha l}{(n+\beta d)(n+(\beta+l)d)}\right). \quad (8\text{-}19)$$

For given $l \ne 0$ and l_1, the sum $\Upsilon_6(n_0, d_0)$ over n and d in (8-15) is essentially an incomplete sum in two variables of the type treated in Corollary 6.24. However, before we can apply this result, we must separate the variables n and d in $\psi_N(n+dl)$. As in the previous section, we can do this here using a Taylor expansion.

Let $J \ge 1$ be an integer. Performing a Taylor expansion to order J, we have

$$\psi_N(n+dl) = \psi\left(\frac{n+dl}{N}\right) = \sum_{j=0}^{J}\left(\frac{d}{\Delta}\right)^j \frac{1}{j!}\left(\frac{\Delta l}{N}\right)^j \psi^{(j)}\left(\frac{n}{N}\right) + O(x^{-\varepsilon J}),$$

since $dl \ll \Delta L \ll x^{-\varepsilon}N$ by (8-13). We can absorb the factor $(d/\Delta)^j$ into ψ_Δ, and after taking J large enough depending on ε, we see that we can express $\Upsilon_6(n_0, d_0)$ as a sum of finitely many sums

$$\Upsilon_6'(n_0, d_0) = \sum_{d=d_0(q_0)} \psi_\Delta(d) \sum_{n=n_1\,(q_0)} \psi_N'(n)e_m\left(\frac{\alpha l}{(n+\beta d)(n+(\beta+l)d)}\right)$$

for some residue classes n_1 (q_0), where ψ_Δ and ψ_N' are coefficient sequences smooth at scales Δ and N respectively, possibly different from the previous ones.

We will prove in Section 8D the following exponential sum estimate, using the machinery from Section 6:

Proposition 8.4. *Let m be a y-densely divisible squarefree integer of polynomial size for some $y \ge 1$, let Δ, $N > 0$ be of polynomial size, and let $\alpha, \beta, \gamma_1, \gamma_2, l \in \mathbb{Z}/m\mathbb{Z}$. Let ψ_Δ, ψ_N' be shifted smooth sequences at scale Δ and N respectively. Then for*

any divisor q_0 of m and for all residue classes d_0 (q_0) and n_0 (q_0), we have

$$\left| \sum_{d=d_0 \ (q_0)} \sum_{n=n_0 \ (q_0)} \psi_\Delta(d)\psi_N'(n)e_m\left(\frac{\alpha l}{(n+\beta d+\gamma_1)(n+(\beta+l)d+\gamma_2)}\right) \right|$$

$$\ll (\alpha l, m)\left(\frac{N}{q_0 m^{1/2}} + m^{1/2}\right)\left(1 + \left(\frac{\Delta}{q_0}\right)^{1/2}m^{1/6}y^{1/6} + \left(\frac{\Delta}{q_0}\right)m^{-1/2}\right). \quad (8\text{-}20)$$

We also have the bound

$$\left| \sum_{d=d_0 \ (q_0)} \sum_{n=n_0 \ (q_0)} \psi_\Delta(d)\psi_N'(n)e_m\left(\frac{\alpha l}{(n+\beta d+\gamma_1)(n+(\beta+l)d+\gamma_2)}\right) \right|$$

$$\ll (\alpha l, m)\left(\frac{N}{q_0 m^{1/2}} + m^{1/2}\right)\left(m^{1/2} + \left(\frac{\Delta}{q_0}\right)m^{-1/2}\right). \quad (8\text{-}21)$$

Remark 8.5. Suppose $q_0 = 1$ for simplicity. In practice, the dominant term on the right-hand side of (8-21) will be $(\alpha l, m)m^{1/2}\Delta^{1/2}m^{1/6}y^{1/6}$, which in certain regimes improves upon the bound of $((\alpha l, m)^{-1/2}m^{1/2})\Delta$ that is obtained by completing the sums in the variable n only without exploiting any additional cancellation in the variable d.

Note that if the phase

$$\frac{\alpha l}{(n+\beta d+\gamma_1)(n+(\beta+l)d+\gamma_2)}$$

was of the form $f(d) + g(n)$ for some nonconstant rational functions f and g, then the two-dimensional sum would factor into the product of two one-dimensional sums, and then the estimates we claim would basically follow from the one-dimensional bounds in Proposition 4.12. However, no such splitting is available, and so we are forced to use the genuinely multidimensional theory arising from Deligne's proof of the Riemann hypothesis over finite fields.

Applying Proposition 8.4, we have

$$\Upsilon_6'(n_0, d_0) \ll (\alpha l, m)\left(m^{1/2} + \frac{N/q_0}{m^{1/2}}\right)\left(1 + (\Delta/q_0)^{1/2}m^{1/6}x^{\delta/6} + \frac{\Delta/q_0}{m^{1/2}}\right),$$

as well as

$$\Upsilon_6'(n_0, d_0) \ll (\alpha l, m)\left(m^{1/2} + \frac{N/q_0}{m^{1/2}}\right)\left(m^{1/2} + \frac{\Delta/q_0}{m^{1/2}}\right).$$

Distinguishing the cases $N/q_0 \le m$ and $N/q_0 > m$, and summing over the finitely many cases of $\Upsilon_6'(n_0, d_0)$ that give $\Upsilon_6(n_0, d_0)$, we see that

$$\Upsilon_6(n_0, d_0)$$
$$\ll (\alpha l, m)\left\{m^{1/2}\left(1 + \left(\frac{\Delta}{q_0}\right)^{1/2}m^{1/6}x^{\delta/6} + \frac{\Delta/q_0}{m^{1/2}}\right) + \frac{N/q_0}{m^{1/2}}\left(m^{1/2} + \frac{\Delta/q_0}{m^{1/2}}\right)\right\}.$$

Note that $(\alpha l, m) \le (\alpha, m)(l, m)$ and hence, summing over l and l_1 in (8-18) (using Lemma 1.4), we get

$$\Upsilon_5'(n_0, d_0) \ll (\alpha, m)L^2\left\{m^{1/2} + \left(\frac{\Delta}{q_0}\right)^{1/2}m^{2/3}x^{\delta/6} + \frac{\Delta}{q_0} + \frac{N}{q_0} + \frac{N\Delta}{q_0^2 m}\right\}.$$

Next, summing over the $\le (q_0, \ell)q_0$ residue classes (n_0, d_0) allowed by the congruence restriction (5-23), we get

$$\Upsilon_5' \ll (q_0, \ell)(\alpha, m)L^2\left\{q_0 m^{1/2} + (q_0\Delta)^{1/2}m^{2/3}x^{\delta/6} + \Delta + N + \frac{N\Delta}{q_0 m}\right\},$$

and finally, by inserting some additional factors of q_0 and (q_0, ℓ), we derive

$$\frac{N\Delta}{L^2}|\Upsilon_5'| \ll (q_0, \ell)(\alpha, m)N\Delta\left\{q_0 m^{1/2} + (q_0\Delta)^{1/2}m^{2/3}x^{\delta/6} + \Delta + N + \frac{N\Delta}{q_0 m}\right\}$$

$$\ll (q_0, \ell)^2(\alpha, m)^2 q_0 N\Delta\left\{\Delta^{1/2}m^{2/3}x^{\delta/6} + \Delta + N + \frac{N\Delta}{m}\right\}.$$

In fact, since $\Delta \ll D \ll N$, we see that

$$\frac{N\Delta}{L^2}|\Upsilon_5'| \ll (q_0, \ell)^2(\alpha, m)^2 q_0 N\Delta\left\{\Delta^{1/2}m^{2/3}x^{\delta/6} + N + \frac{N\Delta}{m}\right\}.$$

We have $m = q_0 r_1 u_1 [v_1, v_2] q_2$ (see Lemma 8.3) and therefore (using (8-5) and (8-4)) we can bound m from above and below by

$$m \ll q_0 \times \frac{R}{\Delta} \times U \times V^2 \times \frac{Q}{q_0} \asymp \frac{Q^2 RV}{\Delta} \ll x^{\delta+2\varepsilon}\frac{Q^2 RH}{\Delta}$$

and

$$m \gg q_0 \times \frac{R}{\Delta} \times U \times V \times \frac{Q}{q_0} \asymp \frac{Q^2 R}{q_0\Delta},$$

which leads to

$$\frac{N\Delta}{L^2}|\Upsilon_5'| \ll (q_0, \ell)^2(\alpha, m)^2 q_0^2 N\Delta\left\{x^{5\delta/6+4\varepsilon/3}\frac{(Q^2 RH)^{2/3}}{\Delta^{1/6}} + N + \frac{N\Delta^2}{Q^2 R}\right\}$$

$$= (q_0, \ell)^2(\alpha, m)^2 q_0^2\frac{(N\Delta)^2}{H^4}\left\{x^{5\delta/6+2\varepsilon}\frac{H^4(Q^2 RH)^{2/3}}{N\Delta^{7/6}} + \frac{H^4}{\Delta} + \frac{H^4\Delta}{Q^2 R}\right\}$$

up to admissible errors. Since

$$\Delta^{-1} \ll \frac{x^\delta}{D} = x^{\delta+5\varepsilon}\frac{H^4}{N}, \quad \Delta \ll D = x^{-5\varepsilon}\frac{N}{H^4},$$

this leads to

$$\frac{N\Delta}{L^2}|\Upsilon_5'|$$
$$\ll (q_0, \ell)^2(\alpha, m)^2 q_0^2\frac{(N\Delta)^2}{H^4}\left\{x^{2\delta+8\varepsilon}\frac{H^{28/3}Q^{4/3}R^{2/3}}{N^{13/6}} + \frac{x^{\delta+5\varepsilon}H^8}{N} + \frac{x^{-5\varepsilon}N}{Q^2 R}\right\}$$

up to admissible errors. From the assumptions (5-2) and (5-13), we have

$$N \ll x^{1/2} \ll QR,$$

and thus

$$\frac{x^{-5\varepsilon} N}{Q^2 R} \ll x^{-5\varepsilon} Q^{-1} \ll x^{-5\varepsilon}.$$

On the other hand, from the value of H (see (8-1)) we get

$$x^{2\delta+8\varepsilon} \frac{H^{28/3} Q^{4/3} R^{2/3}}{N^{13/6}} \ll x^{2\delta+18\varepsilon} \frac{R^{10} Q^{20}}{M^{28/3} N^{13/6}} \ll x^{-28/3+2\delta+18\varepsilon} R^{10} Q^{20} N^{43/6},$$

$$\frac{x^{\delta+5\varepsilon} H^8}{N} \ll x^{\delta+13\varepsilon} \frac{R^8 Q^{16}}{N M^8} \ll x^{-8+\delta+13\varepsilon} N^7 Q^{16} R^8.$$

Using the other conditions $x^{1/2} \ll QR \ll x^{1/2+2\varpi}$ and

$$R \gg x^{-3\varepsilon-\delta} N, \quad N \gg x^{1/2-\sigma},$$

these quantities are in turn bounded respectively by

$$x^{2\delta+8\varepsilon} \frac{H^{28/3} Q^{4/3} R^{2/3}}{N^{13/6}} \leq x^{2/3+2\delta+40\varpi+18\varepsilon} \frac{N^{43/6}}{R^{10}} \ll x^{2/3+12\delta+40\varpi-17/6(1/2-\sigma)+48\varepsilon},$$

$$\frac{x^{\delta+5\varepsilon} H^8}{N} \leq x^{\delta+32\varpi+13\varepsilon} \frac{N^7}{R^8} \ll x^{9\delta+32\varpi+37\varepsilon-(1/2-\sigma)}.$$

Thus, by taking $\varepsilon > 0$ small enough, we obtain (8-17) (and hence Proposition 8.1) provided

$$\begin{cases} \frac{2}{3} + 12\delta + 40\varpi - \frac{17}{6}\left(\frac{1}{2}-\sigma\right) < 0, \\ 9\delta + 32\varpi - \left(\frac{1}{2}-\sigma\right) < 0, \end{cases} \iff \begin{cases} \frac{160}{3}\varpi + 16\delta + \frac{34}{9}\sigma < 1, \\ 64\varpi + 18\delta + 2\sigma < 1. \end{cases}$$

These are exactly the conditions claimed in Proposition 8.1.

8C. *Proof of Lemma 8.2.* This is a bit more complicated than the corresponding lemmas in Sections 5D–5F because the quantity $m = q_0 q_2 r_1 u_1 [v_1, v_2]$ depends also on v_1 and v_2.

We let $w := q_0 q_2 r_1 u_1$, so $m = w[v_1, v_2]$ and w is independent of (h_1, h_2, v_1, v_2) and coprime with $[v_1, v_2]$.

Since $(w, [v_1, v_2]) = 1$, we have

$$(h_1 v_2 - h_2 v_1, w[v_1, v_2]) = \sum_{\substack{d|h_1 v_2-h_2 v_1 \\ d|w[v_1,v_2]}} \varphi(d) \leq \sum_{d|w} d \sum_{\substack{e|[v_1,v_2] \\ de|h_1 v_2-h_2 v_1}} e,$$

and therefore

$$\sum_{\substack{(h_1,v_1,h_2,v_2) \\ h_1v_2 \neq h_2v_1}} \sum (h_1v_2 - h_2v_1, q_0q_2r_1u_1[v_1, v_2]) \leq \sum_{\substack{(h_1,v_1,h_2,v_2) \\ h_1v_2 \neq h_2v_1}} \sum_{d|w} \sum d \sum_{\substack{e|[v_1,v_2] \\ de|h_1v_2-h_2v_1}} e$$

$$\leq \sum_{d|w} d \sum_{\substack{(d,e)=1 \\ e \ll V^2 \\ e \text{ squarefree}}} e \sum_{\substack{([v_1,v_2],w)=1 \\ de|h_1v_2-h_2v_1 \\ e|[v_1,v_2] \\ h_1v_2 \neq h_2v_1}} 1.$$

The variable d is unrelated to the modulus d appearing previously in this section.

Let d, e be integers occurring in the outer sums, and (h_1, h_2, v_1, v_2) satisfying the other summation conditions. Then e is squarefree, and since $e \mid [v_1, v_2]$ and $e \mid h_1v_2 - h_2v_1$, any prime dividing e must divide one of (v_1, v_2), (h_1, v_1) or (h_2, v_2) (if it does not divide both v_1 and v_2, it is coprime to one of them, and $h_1v_2 - h_2v_1 = 0$ (p) gives one of the other divisibilities). Thus if we factor $e = e_1e_2e_3$, where

$$e_1 := \prod_{\substack{p|e \\ p|v_1 \\ p \nmid v_2}} p, \quad e_2 := \prod_{\substack{p|e \\ p \nmid v_1 \\ p|v_2}} p, \quad e_3 := \prod_{\substack{p|e \\ p|(v_1,v_2)}} p,$$

then these are coprime and we have

$$e_1 \mid h_1, \quad e_2 \mid h_2, \quad e_1e_3 \mid v_1, \quad e_2e_3 \mid v_2.$$

We write

$$h_1 = e_1\lambda_1, \quad h_2 = e_2\lambda_2, \quad v_1 = e_1e_3\nu_1, \quad v_2 = e_2e_3\nu_2.$$

Then we get

$$h_1v_2 - h_2v_1 = e(\lambda_1\nu_2 - \lambda_2\nu_1),$$

and since $de \mid h_1v_2 - h_2v_1$, it follows that $d \mid \lambda_1\nu_2 - \lambda_2\nu_1$.

Now fix some $e \ll V^2$. For each choice of factorization $e = e_1e_2e_3$, the number of pairs $(\lambda_1\nu_2, \lambda_2\nu_1)$ that can be associated to this factorization as above for some quadruple (h_1, h_2, v_1, v_2) is $\ll (HV/e)^2/d$, since each product $\lambda_1\nu_2, \lambda_2\nu_1$ is $\ll HV/e$, and d divides the difference. By the divisor bound, this gives $\ll (HV)^2/de^2$ for the number of quadruples (h_1, h_2, v_1, v_2). Summing over $d \mid w$ and e, we get a total bound

$$\ll (HV)^2\tau(w) \sum_{e \ll V^2} e^{-1} \ll H^2V^2,$$

as desired.

8D. *Proof of Proposition 8.4.* It remains to establish Proposition 8.4. We begin with the special case when $e = 1$ and $(\alpha l, m) = 1$. For simplicity, we set

$$f(n, d) = \frac{\alpha l}{(n + \beta d + \gamma_1)(n + (\beta + l)d + \gamma_2)}.$$

By completion of the sum over n (see Lemma 4.9(i)), we have

$$\sum_d \sum_n \psi_\Delta(d) \psi_N(n) e_m(f(n, d))$$

$$\ll \left(\frac{N}{m} + 1\right) \sup_{h \in \mathbb{Z}/m\mathbb{Z}} \left| \sum_d \psi_\Delta(d) \sum_{n \in \mathbb{Z}/m\mathbb{Z}} e_m(f(n, d) + hn) \right|$$

$$= \left(\frac{N}{\sqrt{m}} + \sqrt{m}\right) \sup_{h \in \mathbb{Z}/m\mathbb{Z}} \left| \sum_d \psi_\Delta(d) K_h(d; m) \right|,$$

where, for each $h \in \mathbb{Z}/m\mathbb{Z}$, we define

$$K_h(d; m) := \frac{1}{\sqrt{m}} \sum_{n \in \mathbb{Z}/m\mathbb{Z}} e_m(f(n, d) + hn).$$

By the first part of Corollary 6.24 (i.e., (6-23)), we get

$$\left| \sum_d \psi_\Delta(d) K_h(d; m) \right| \ll m^{1/2} + \Delta m^{-1/2}, \tag{8-22}$$

and this combined with (8-22) implies the second bound (8-21) (in the case $e = 1$, $(\alpha l, m) = 1$, that is). Furthermore, it also implies the first bound (8-20) for $\Delta > m^{2/3} y^{-1/3}$.

In addition, from the Chinese remainder theorem (Lemma 4.4) and (6-16), we deduce the pointwise bound

$$|K_h(d, m)| \ll 1 \tag{8-23}$$

which implies the trivial bound

$$\left| \sum_d \psi_\Delta(d) K_h(d; m) \right| \ll 1 + \Delta,$$

which gives (8-20) for $\Delta \le m^{1/3} y^{1/3}$. Thus we can assume that

$$m^{1/3} y^{1/3} \le \Delta \le m^{2/3} y^{-1/3} \le m.$$

We can then use the y-dense divisibility of m to factor m into $m_1 m_2$, where

$$y^{-2/3} m^{1/3} \le m_1 \le y^{1/3} m^{1/3},$$
$$y^{-1/3} m^{2/3} \le m_2 \le y^{2/3} m^{2/3}.$$

Now the second part of Corollary 6.24 (i.e., (6-24)) gives

$$\left| \sum_d \psi_\Delta(d) K_h(d; m) \right| \ll \Delta^{1/2} m_1^{1/2} + \Delta^{1/2} m_2^{1/4} \ll \Delta^{1/2} m^{1/6} y^{1/6},$$

which together with (8-22) gives (8-20).

This finishes the proof of Proposition 8.4 for the special case $e = 1$ and $(\alpha l, m) = 1$. The extension to a divisor $e \mid m$ is done exactly as in the proof of Corollary 4.16 in Section 4.

We now reduce to the case $(\alpha l, m) = 1$. Let

$$m' := m/(\alpha l, m),$$
$$y' := y(\alpha l, m),$$
$$\alpha' := \alpha/(\alpha l, m) = \frac{\alpha/(\alpha, m)}{(\alpha l, m)/(\alpha, m)},$$

where one computes the reciprocal of $(\alpha l, m)/(\alpha, m)$ inside $\mathbb{Z}/m'\mathbb{Z}$, so that α' is viewed as an element of $\mathbb{Z}/m'\mathbb{Z}$. The integer m' is y'-densely divisible by Lemma 2.10(ii), and it is also squarefree and of polynomial size. We have $(a'l, m') = 1$, and furthermore

$$\sum_d \sum_n \psi_\Delta(d) \psi_N(n) e_m(f(n, d))$$
$$= \sum_d \sum_n \psi_\Delta(d) \psi_N(n) e_{m'}(f'(n, d)) \prod_{p \mid (\alpha l, m)} (1 - \mathbf{1}_{p \mid (n+\beta d+\gamma_1)(n+(\beta+l)d+\gamma_2)}),$$

where

$$f'(n, d) = \frac{\alpha' l}{(n + \beta d + \gamma_1)(n + (\beta + l)d + \gamma_2)}$$

(here we use the convention explained at the end of Section 4A that leads to $e_p(\alpha x) = 1$ if p is prime, $\alpha = 0$ (p) and $x = +\infty \in \mathbb{P}^1(\mathbb{Z}/p\mathbb{Z})$).

Set

$$g(n, d) = (n + \beta d + \gamma_1)(n + (\beta + l)d + \gamma_2).$$

Then, expanding the product (as in inclusion-exclusion), we get

$$\sum_d \sum_n \psi_\Delta(d) \psi_N(n) e_m(f(n, d)) = \sum_{\delta \mid (\alpha l, m)} \mu(\delta) \sum_{\substack{d, n \\ \delta \mid g(n, d)}} \psi_\Delta(d) \psi_N(n) e_{m'}(f'(n, d))$$

(this usage of δ is unrelated to prior usages of δ in this section). Splitting the sum over n and d in residue classes modulo δ, this sum is then equal to

$$\sum_{\delta \mid (\alpha l, m)} \mu(\delta) \sum_{\substack{(d_0, n_0) \in (\mathbb{Z}/\delta\mathbb{Z})^2 \\ g(n_0, d_0) = 0 \ (\delta)}} \sum_{n = n_0 \ (n)} \sum_{d = d_0 \ (\delta)} \psi_\Delta(d) \psi_N(n) e_{m'}(f'(n, d)).$$

For each choice of (n_0, d_0), we can apply the previously proved case of Proposition 8.4 to deduce that

$$\sum_{n=n_0 \, (n)} \sum_{d=d_0 \, (\delta)} \psi_\Delta(d)\psi_N(n)e_{m'}(f'(n,d))$$

$$\ll \left(\sqrt{m'} + \frac{N}{\delta\sqrt{m'}}\right)\left(1 + \frac{\Delta^{1/2}}{\delta^{1/2}}(m'y')^{1/6} + \frac{\Delta}{\delta\sqrt{m'}}\right)$$

and

$$\sum_{n=n_0 \, (n)} \sum_{d=d_0 \, (\delta)} \psi_\Delta(d)\psi_N(n)e_{m'}(f'(n,d)) \ll \left(\sqrt{m'} + \frac{N}{\delta\sqrt{m'}}\right)\left(\sqrt{m'} + \frac{\Delta}{\delta\sqrt{m'}}\right).$$

Moreover, by the Chinese remainder theorem, there are $\ll \delta$ solutions $(n_0, d_0) \in (\mathbb{Z}/\delta\mathbb{Z})^2$ of $g(n_0, d_0) = 0 \, (\delta)$, and therefore we find

$$\sum_d \sum_n \psi_\Delta(d)\psi_N(n)e_m(f(n,d))$$

$$\ll \sum_{\delta | (\alpha l, m)} \delta\left(\sqrt{m'} + \frac{N}{\delta\sqrt{m'}}\right)\left(1 + \frac{\Delta^{1/2}}{\delta^{1/2}}(m'y')^{1/6} + \frac{\Delta}{\delta\sqrt{m'}}\right)$$

and

$$\sum_d \sum_n \psi_\Delta(d)\psi_N(n)e_m(f(n,d)) \ll \sum_{\delta | (\alpha l, m)} \delta\left(\sqrt{m'} + \frac{N}{\delta\sqrt{m'}}\right)\left(\sqrt{m'} + \frac{\Delta}{\delta\sqrt{m'}}\right).$$

It is now elementary to check that these give the bounds of Proposition 8.4 (note that $m'y' = my$).

About this project

This paper is part of the *Polymath project*, which was launched by Timothy Gowers in February 2009 as an experiment to see if research mathematics could be conducted by a massive online collaboration. The current project (which was administered by Terence Tao) is the eighth project in this series. Further information on the Polymath project can be found on the web site http://michaelnielsen.org/polymath1. Information about this specific project may be found at

http://michaelnielsen.org/polymath1/index.php?title=Bounded_gaps_between_primes

and a full list of participants and their grant acknowledgments may be found at

http://michaelnielsen.org/polymath1/index.php?title=Polymath8_grant_acknowledgments.

Acknowledgements

References

[Barban and Vehov 1969] M. B. Barban and P. P. Vehov, "Summation of multiplicative functions of polynomials", *Mat. Zametki* **5** (1969), 669–680. MR 40 #4221 Zbl 0192.39103

[Bombieri 1974] E. Bombieri, "Counting points on curves over finite fields (d'après S. A. Stepanov)", exposé no. 430, 234–241 in *Séminaire Bourbaki, 1972/1973*, Lecture Notes in Math. **383**, Springer, Berlin, 1974. MR 55 #2912 Zbl 0307.14011

[Bombieri et al. 1986] E. Bombieri, J. B. Friedlander, and H. Iwaniec, "Primes in arithmetic progressions to large moduli", *Acta Math.* **156**:3-4 (1986), 203–251. MR 88b:11058 Zbl 0588.10042

[Bombieri et al. 1987] E. Bombieri, J. B. Friedlander, and H. Iwaniec, "Primes in arithmetic progressions to large moduli, II", *Math. Ann.* **277**:3 (1987), 361–393. MR 88f:11085 Zbl 0625.10036

[Bombieri et al. 1989] E. Bombieri, J. B. Friedlander, and H. Iwaniec, "Primes in arithmetic progressions to large moduli, III", *J. Amer. Math. Soc.* **2**:2 (1989), 215–224. MR 89m:11087 Zbl 0674.10036

[Cochrane and Pinner 2006] T. Cochrane and C. Pinner, "Using Stepanov's method for exponential sums involving rational functions", *J. Number Theory* **116**:2 (2006), 270–292. MR 2006j:11113 Zbl 1093.11058

[Deligne 1974] P. Deligne, "La conjecture de Weil, I", *Inst. Hautes Études Sci. Publ. Math.* **43** (1974), 273–307. MR 49 #5013 Zbl 0287.14001

[Deligne 1980] P. Deligne, "La conjecture de Weil, II", *Inst. Hautes Études Sci. Publ. Math.* **52** (1980), 137–252. MR 83c:14017 Zbl 0456.14014

[Fouvry 1984] É. Fouvry, "Autour du théorème de Bombieri–Vinogradov", *Acta Math.* **152**:3-4 (1984), 219–244. MR 85m:11052 Zbl 0552.10024

[Fouvry 1985] É. Fouvry, "Sur le problème des diviseurs de Titchmarsh", *J. Reine Angew. Math.* **357** (1985), 51–76. MR 87b:11090 Zbl 0547.10039

[Fouvry and Iwaniec 1980] E. Fouvry and H. Iwaniec, "On a theorem of Bombieri–Vinogradov type", *Mathematika* **27**:2 (1980), 135–152. MR 82h:10057 Zbl 0469.10027

[Fouvry and Iwaniec 1983] E. Fouvry and H. Iwaniec, "Primes in arithmetic progressions", *Acta Arith.* **42**:2 (1983), 197–218. MR 84k:10035 Zbl 0517.10045

[Fouvry and Iwaniec 1992] É. Fouvry and H. Iwaniec, "The divisor function over arithmetic progressions", *Acta Arith.* **61**:3 (1992), 271–287. MR 93g:11089 Zbl 0764.11040

[Fouvry et al. 2013a] É. Fouvry, E. Kowalski, and P. Michel, "An inverse theorem for Gowers norms of trace functions over $\mathbf{F}_p$", *Math. Proc. Cambridge Philos. Soc.* **155**:2 (2013), 277–295. MR 3091520 Zbl 06203760

[Fouvry et al. 2013b] E. Fouvry, E. Kowalski, and P. Michel, "On the conductor of cohomological transforms", preprint, 2013. arXiv 1310.3603

[Fouvry et al. 2013c] E. Fouvry, E. Kowalski, and P. Michel, "The sliding-sum method for short exponential sums", preprint, 2013. arXiv 1307.0135

[Fouvry et al. 2014a] E. Fouvry, E. Kowalski, and P. Michel, "Algebraic twists of modular forms and Hecke orbits", preprint, 2014. arXiv 1207.0617

[Fouvry et al. 2014b] E. Fouvry, E. Kowalski, and P. Michel, "On the exponent of distribution of the ternary divisor function", *Mathematika* (online publication June 2014).

[Fouvry et al. 2014c] E. Fouvry, E. Kowalski, and P. Michel, "Trace functions over finite fields and their applications", preprint, 2014, http://www.math.ethz.ch/~kowalski/trace-functions-pisa.pdf. To appear in "Colloquium De Giorgi 2013 and 2014".

[Friedlander and Iwaniec 1985] J. B. Friedlander and H. Iwaniec, "Incomplete Kloosterman sums and a divisor problem", *Ann. of Math.* (2) **121**:2 (1985), 319–350. MR 86i:11050 Zbl 0572.10029

[Gallagher 1968] P. X. Gallagher, "Bombieri's mean value theorem", *Mathematika* **15** (1968), 1–6. MR 38 #5724 Zbl 0174.08103

[Goldston et al. 2009] D. A. Goldston, J. Pintz, and C. Y. Yıldırım, "Primes in tuples, I", *Ann. of Math.* (2) **170**:2 (2009), 819–862. MR 2011c:11146 Zbl 1207.11096

[Graham and Ringrose 1990] S. W. Graham and C. J. Ringrose, "Lower bounds for least quadratic nonresidues", pp. 269–309 in *Analytic number theory* (Allerton Park, IL, 1989), edited by B. C. Berndt et al., Progr. Math. **85**, Birkhäuser, Boston, 1990. MR 92d:11108 Zbl 0719.11006

[Heath-Brown 1978] D. R. Heath-Brown, "Hybrid bounds for Dirichlet L-functions", *Invent. Math.* **47**:2 (1978), 149–170. MR 58 #5549 Zbl 0362.10035

[Heath-Brown 1982] D. R. Heath-Brown, "Prime numbers in short intervals and a generalized Vaughan identity", *Canad. J. Math.* **34**:6 (1982), 1365–1377. MR 84g:10075 Zbl 0478.10024

[Heath-Brown 1986] D. R. Heath-Brown, "The divisor function $d_3(n)$ in arithmetic progressions", *Acta Arith.* **47**:1 (1986), 29–56. MR 88a:11088 Zbl 0549.10034

[Heath-Brown 2001] D. R. Heath-Brown, "The largest prime factor of $X^3 + 2$", *Proc. London Math. Soc.* (3) **82**:3 (2001), 554–596. MR 2001m:11158 Zbl 1023.11048

[Iwaniec 1980] H. Iwaniec, "A new form of the error term in the linear sieve", *Acta Arith.* **37** (1980), 307–320. MR 82d:10069 Zbl 0444.10038

[Iwaniec and Kowalski 2004] H. Iwaniec and E. Kowalski, *Analytic number theory*, American Mathematical Society Colloquium Publications **53**, American Mathematical Society, Providence, RI, 2004. MR 2005h:11005 Zbl 1059.11001

[Katz 1980] N. M. Katz, *Sommes exponentielles*, Astérisque **79**, Société Mathématique de France, Paris, 1980. MR 82m:10059 Zbl 0469.12007

[Katz 1988] N. M. Katz, *Gauss sums, Kloosterman sums, and monodromy groups*, Annals of Mathematics Studies **116**, Princeton University Press, 1988. MR 91a:11028 Zbl 0675.14004

[Katz 2001] N. M. Katz, "L-functions and monodromy: four lectures on Weil II", *Adv. Math.* **160**:1 (2001), 81–132. MR 2002c:11066 Zbl 1016.14011

[Kloosterman 1927] H. D. Kloosterman, "On the representation of numbers in the form $ax^2 + by^2 + cz^2 + dt^2$", *Acta Math.* **49**:3-4 (1927), 407–464. MR 1555249 Zbl 53.0155.01

[Kowalski 2010] E. Kowalski, "Some aspects and applications of the Riemann hypothesis over finite fields", *Milan J. Math.* **78**:1 (2010), 179–220. MR 2011g:11229 Zbl 1271.11113

[Laumon 1987] G. Laumon, "Transformation de Fourier, constantes d'équations fonctionnelles et conjecture de Weil", *Inst. Hautes Études Sci. Publ. Math.* **65** (1987), 131–210. MR 88g:14019 Zbl 0641.14009

[Linnik 1963] J. V. Linnik, *The dispersion method in binary additive problems*, American Mathematical Society, Providence, R.I., 1963. MR 29 #5804 Zbl 0112.27402

[Maynard 2013] J. Maynard, "Small gaps between primes", preprint, 2013. arXiv 1311.4600

[Michel 1998] P. Michel, "Minorations de sommes d'exponentielles", *Duke Math. J.* **95**:2 (1998), 227–240. MR 99i:11069 Zbl 0958.11056

[Montgomery and Vaughan 2007] H. L. Montgomery and R. C. Vaughan, *Multiplicative number theory, I: Classical theory*, Cambridge Studies in Advanced Mathematics **97**, Cambridge University Press, Cambridge, 2007. MR 2009b:11001 Zbl 1142.11001

[Mordell 1932] L. J. Mordell, "On a sum analogous to a Gauss's sum", *Q. J. Math., Oxf. Ser.* **3** (1932), 161–167. Zbl 0005.24603

[Motohashi 1976] Y. Motohashi, "An induction principle for the generalization of Bombieri's prime number theorem", *Proc. Japan Acad.* **52**:6 (1976), 273–275. MR 54 #10171 Zbl 0355.10035

[Motohashi and Pintz 2008] Y. Motohashi and J. Pintz, "A smoothed GPY sieve", *Bull. Lond. Math. Soc.* **40**:2 (2008), 298–310. MR 2009d:11132 Zbl 1278.11090

[Perel′muter 1969] G. I. Perel′muter, "Estimate of a sum along an algebraic curve", *Mat. Zametki* **5** (1969), 373–380. In Russian. MR 39 #2764 Zbl 0179.49903

[Pintz 2013] J. Pintz, "A note on bounded gaps between primes", preprint, 2013. arXiv 1306.1497

[Polymath 2014a] D. H. J. Polymath, "New equidistribution estimates of Zhang type, and bounded gaps between primes", preprint, 2014. arXiv 1402.0811v2

[Polymath 2014b] D. H. J. Polymath, "Variants of the Selberg sieve, and bounded intervals containing many primes", preprint, 2014. arXiv 1407.4897

[SGA 1977] P. Deligne (editor), *Cohomologie étale* (SGA $4\frac{1}{2}$), Lecture Notes in Math. **569**, Springer, Berlin, 1977. MR 57 #3132 Zbl 0345.00010

[Shiu 1980] P. Shiu, "A Brun–Titchmarsh theorem for multiplicative functions", *J. Reine Angew. Math.* **313** (1980), 161–170. MR 81h:10065 Zbl 0412.10030

[Siebert 1971] H. Siebert, "Einige Analoga zum Satz von Siegel–Walfisz", pp. 173–184 in *Zahlentheorie* (Tagung des Math. Forschungsinst., Oberwolfach, 1970), edited by M. Barner and W. Schwarz, Bibliographisches Inst., Mannheim, 1971. MR 51 #391 Zbl 0221.10041

[Vaughan 1977] R.-C. Vaughan, "Sommes trigonométriques sur les nombres premiers", *C. R. Acad. Sci. Paris Sér. A-B* **285**:16 (1977), A981–A983. MR 58 #16555 Zbl 0374.10025

[Weil 1948] A. Weil, *Sur les courbes algébriques et les variétés qui s'en déduisent*, Actualités Sci. Ind. **1041**, Hermann et Cie, Paris, 1948. MR 10,262c Zbl 0036.16001

[Zhang 2014] Y. Zhang, "Bounded gaps between primes", *Ann. of Math.* (2) **179**:3 (2014), 1121–1174. MR 3171761 Zbl 1290.11128

wouter.castryck@gmail.com   Departement Wiskunde, Katholieke Universiteit Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium

etienne.fouvry@math.u-psud.fr   Laboratoire de Mathématique, Campus d'Orsay, Université de Paris-Sud, Bâtiment 425 UMR 8628, 91405 Orsay Cedex, France

gharcos@renyi.hu   Alfréd Rényi Institute of Mathematics, 13–15 Reáltanoda utca, H-1053, Budapest, Hungary

kowalski@math.ethz.ch   Department of Mathematics, ETH Zurich, Rämistrasse 101, CH-8092 Zurich, Switzerland

philippe.michel@epfl.ch   Ecole Polytechnique Fédérale de Lausanne, SB-IMB-TAN, Station 8, CH-1015 Lausanne, Switzerland

paul.nelson@math.ethz.ch   Department of Mathematics, ETH Zurich, Rämistrasse 101, CH-8092 Zurich, Switzerland

epmath@tx.technion.ac.il   Technion Institute, 3200003 Haifa, Israel

pintz@renyi.hu   Alfréd Rényi Institute of Mathematics, 13–15 Reáltanoda utca, H-1053, Budapest, Hungary

drew@math.mit.edu   *Department of Mathematics, Massachusetts Institute*
                           of Technology, 77 Massachusetts Avenue,
                           Cambridge, MA 02139, United States

tao@math.ucla.edu   *Department of Mathematics,*
                           University of California Los Angeles, 405 Hilgard Avenue,
                           Los Angeles, CA 90095-1555, United States

xfxie@cs.cmu.edu   *The Robotics Institute, Carnegie Mellon University,*
                           Pittsburgh, PA 15213, United States

Relations between Dieudonné displays and crystalline Dieudonné theory

Eike Lau

We discuss the relation between crystalline Dieudonné theory and Dieudonné displays of p-divisible groups. The theory of Dieudonné displays is extended to the prime 2 without restriction, which implies that the classification of finite locally free group schemes by Breuil–Kisin modules holds for the prime 2 as well.

Introduction

Formal p-divisible groups G over a p-adically complete ring R are classified by Zink's nilpotent displays [Zink 2002; Lau 2008]. These are projective modules over the ring of Witt vectors $W(R)$ equipped with a filtration and with certain Frobenius-linear operators. A central point of the theory is a description of the Dieudonné crystal of G in terms of the nilpotent display associated to G.

Arbitrary p-divisible groups over R can be classified by displays only when R is a perfect ring. In certain cases, there is the following refinement.

Assume that R is a local Artin ring with perfect residue field k of characteristic p and with maximal ideal $\mathcal{N}_R$. Then $W(R)$ has a unique subring $\mathbb{W}(R)$, here called

the Zink ring of R, which is stable under the Frobenius and which sits in an exact sequence

$$0 \longrightarrow \widehat{W}(\mathcal{N}_R) \longrightarrow \mathbb{W}(R) \longrightarrow W(k) \longrightarrow 0,$$

where $\widehat{W}$ means Witt vectors with only finitely many nonzero components. Let us call R *odd* if $p > 2$ or if p annihilates R. The Verschiebung homomorphism v of $W(R)$, which appears in the definition of displays, stabilises the subring $\mathbb{W}(R)$ if and only if R is odd. In this case, Zink [2001a] defines Dieudonné displays over R as displays with $\mathbb{W}(R)$ in place of $W(R)$, and shows that they classify all p-divisible groups over R.

The restriction for $p = 2$ can be avoided with a small trick: The ring $\mathbb{W}(R)$ is always stable under the modified Verschiebung $\mathbb{v}(x) = v(u_0 x)$, where $u_0 \in W(R)$ is the unit defined by the relation $v(u_0) = p - [p]$. This allows to define Dieudonné displays without assuming that R is odd. It turns out that the Zink ring and Dieudonné displays can be defined for the following class of rings R, which we call *admissible*: the order of nilpotence of nilpotent elements of R is bounded, and R_{red} is a perfect ring of characteristic p.

Theorem A. *For each admissible ring R there is a functor*

$$\Phi_R : (p\text{-divisible groups over } R) \to (\text{Dieudonné displays over } R),$$

which is an equivalence of exact categories.

The equivalence easily extends to projective limits of admissible rings, which includes complete local rings with perfect residue field. If R is perfect, the theorem says that p-divisible groups over R are equivalent to Dieudonné modules. This is a result of Gabber, which is used in the proof. We repeat that for Artin rings (which is certainly the case of interest for most applications[1]), Theorem A is known when R is odd; in this case, Φ_R is the inverse of the functor BT of [Zink 2001a] and [Lau 2009]. But the present construction of the functor Φ_R based on the crystalline Dieudonné module is new, and also gives the following second result.

Let $\mathbb{D}(G)$ denote the covariant Dieudonné crystal of a p-divisible group G. Following [Zink 2001b], to a Dieudonné display $\mathscr{P}$ over an admissible ring R one can associate a crystal in locally free modules $\mathbb{D}(\mathscr{P})$.

Theorem B. *For a p-divisible group G over an admissible ring R with associated Dieudonné display $\mathscr{P} = \Phi_R(G)$, there is a natural isomorphism*

$$\mathbb{D}(G) \cong \mathbb{D}(\mathscr{P}).$$

---

[1] In subsequent work, Dieudonné displays over a larger class of base rings will be used to study the image of the crystalline Dieudonné functor over l.c.i. schemes.

This compatibility was not known before and can be useful in applications; see for example [Viehmann and Wedhorn 2013]. Our proofs of Theorems A and B are closely related. The main point is to construct the functor Φ_R and variants of it. Let $\mathbb{I}_R$ be the kernel of the natural homomorphism $\mathbb{W}(R) \to R$.

First, if R is an odd admissible ring, the ideal $\mathbb{I}_R$ carries natural divided powers. Thus the crystalline Dieudonné module of a p-divisible group over R can be evaluated at $\mathbb{W}(R)$, which gives a filtered F-V-module over $\mathbb{W}(R)$. We show that this construction can be extended to a functor Φ_R as in Theorem A. This is not evident because a filtered F-V-module does not in general determine a Dieudonné display. But the construction of Φ_R can be reduced to the case where R is a universal deformation ring; then the Dieudonné display is determined uniquely because p is not a zero divisor in $\mathbb{W}(R)$.

Next, for a divided power extension of admissible rings $S \to R$, one can define Dieudonné displays relative to $S \to R$, called triples in the work of Zink. They are modules over an extension $\mathbb{W}(S/R)$ of $\mathbb{W}(S)$. If R is odd and the divided powers are compatible with the canonical divided powers of p, then the evaluation of the crystalline Dieudonné module at the divided power extension $\mathbb{W}(S/R) \to R$ can be extended to a functor

$$\Phi_{S/R} : (p\text{-divisible groups over } R) \to (\text{Dieudonné displays for } S/R).$$

Again, this is not evident; the proof comes down to the fact that p is not a zero divisor in the Zink ring of the divided power envelope of the diagonal of the universal deformation space of a p-divisible group. Once the functors $\Phi_{S/R}$ are known to exist, Theorems A and B for odd admissible rings are straightforward consequences.

Now let R be an admissible ring which is not odd, so $p = 2$. In this case, the preceding constructions do not apply directly because the ideal $\mathbb{I}_R$ does not in general carry divided powers. This changes when $\mathbb{W}(R)$ is replaced by the slightly larger v-stabilised Zink ring $\mathbb{W}^+(R) = \mathbb{W}(R)[v(1)]$. With an obvious definition of v-stabilised Dieudonné displays, we get a functor

$$\Phi_R^+ : (2\text{-divisible groups over } R) \to (v\text{-stabilised Dieudonné displays over } R),$$

which is, however, not an equivalence. In order to construct a functor Φ_R as in Theorem A, we have to descend from $\mathbb{W}^+(R)$ to $\mathbb{W}(R)$. This can be reduced to the minimal case where $2\mathcal{N}_R = 0$. Then the ideal $\mathbb{I}_R$ carries exceptional divided powers, which allows us to evaluate the crystalline Dieudonné module at $\mathbb{W}(R)$. In order to get the functor Φ_R, we need some lift towards characteristic zero, which is provided by the fact that the exceptional divided powers exist on $\mathbb{I}_R/(v([4]))$ as soon as $4\mathcal{N}_R = 0$. Once Φ_R is known to exist in general, Theorems A and B follow again quite formally.

Breuil–Kisin modules. Now let R be a complete regular local ring with perfect residue field k of characteristic p. Theorem A implies that the classification of p-divisible groups over R by Breuil windows derived in [Vasiu and Zink 2010] and [Lau 2010] for odd p holds for $p = 2$ as well. Let us recall what this means: We write $R = \mathfrak{S}/E\mathfrak{S}$, where $\mathfrak{S}$ is a power series ring over $W(k)$ and where E has constant term p; we also have to choose an appropriate Frobenius lift σ on $\mathfrak{S}$. A Breuil window is a free $\mathfrak{S}$-module Q equipped with an $\mathfrak{S}$-linear map $\phi : Q \to Q^{(\sigma)}$ whose cokernel is annihilated by E; this is equivalent to the notion of a Breuil–Kisin module. As usual, one also gets a classification of finite locally free p-group schemes over R.

In the case of discrete valuation rings this completes the proof of a conjecture of Breuil [1998], which was proved in [Kisin 2006] if p is odd, and in [Kisin 2009] for connected p-divisible groups if $p = 2$. Shortly after the first version of this article was posted, independent proofs of Breuil's conjecture by W. Kim [2012] and T. Liu [2013] appeared online.

Assume that R has characteristic zero, and let S be the p-adic completion of the divided power envelope of the ideal $E\mathfrak{S} \subset \mathfrak{S}$. As a consequence of Theorem B, we show that for a p-divisible group over R the value of its crystalline Dieudonné module at S coincides with the base change of its Breuil window under $\sigma : \mathfrak{S} \to S$.

The functor BT. The original proof of Theorem A for odd local Artin rings in [Zink 2001a] depends on the construction of a functor BT from Dieudonné displays to p-divisible groups, which is a combination of the functor BT from nilpotent displays to formal p-divisible groups and a calculation of extensions. A modified construction of this functor is given in [Lau 2009]. Once the definition of Dieudonné displays for nonodd local Artin rings is available, all these arguments can be carried over almost literally to give an alternative proof of Theorem A in that case. In the present approach this construction serves only as an explicit description of the inverse of the functor Φ_R; this is used in [Lau 2012].

All rings are commutative with a unit unless the contrary is stated. For a p-divisible group G, we denote by $\mathbb{D}(G)$ the *covariant* Dieudonné crystal.

1. The Zink ring

In this section we study the Zink ring $\mathbb{W}(R)$, which was introduced in [Zink 2001a] under the notation $\hat{W}(R)$, and variants of $\mathbb{W}(R)$ in the presence of divided powers, following [Zink 2001b]. The definitions are stated in more generality, allowing arbitrary perfect rings instead of perfect fields. The modified Verschiebung v for $p = 2$ is new.

1A. *Preliminaries.* We fix a prime p. A commutative ring without unit N is called bounded nilpotent if there is a number n such that $x^n = 0$ for every $x \in N$. We will consider the following type of base rings.

Definition 1.1. A ring R is called *admissible* if its nilradical $\mathcal{N}_R$ is bounded nilpotent and if $R_{\mathrm{red}} = R/\mathcal{N}_R$ is a perfect ring of characteristic p.

Local Artin rings with perfect residue field are admissible. The ring $\mathcal{O}_{\mathbb{C}_p}/p$ is not admissible. We will also consider projective limits of admissible rings:

Definition 1.2. An *admissible topological ring* is a complete and separated topological ring R with linear topology such that the ideal $\mathcal{N}_R$ of topologically nilpotent elements is open, the ring $R_{\mathrm{red}} = R/\mathcal{N}_R$ is perfect of characteristic p, and for each open ideal N of R contained in $\mathcal{N}_R$, the quotient $\mathcal{N}_R/N$ is bounded nilpotent. Thus R is the projective limit of the admissible rings R/N.

Examples of admissible topological rings include complete local rings with perfect residue field. Admissible topological rings in which $\mathcal{N}_R$ is not topologically nilpotent arise from divided power envelopes; see Lemma 1.13.

Notation 1.3. For a commutative, not necessarily unitary ring A, let $W(A)$ be the ring of p-typical Witt vectors of A. We write f and v for the Frobenius and Verschiebung of $W(A)$. Let $I_A = v(W(A))$, let $w_i : W(A) \to A$ be given by the i-th Witt polynomial, and let $\hat{W}(A)$ be the group of all elements of $W(A)$ with nilpotent coefficients which are almost all zero.

Let us recall two well-known facts:

Lemma 1.4. *Let A be a perfect ring of characteristic p and let B be a ring with a bounded nilpotent ideal $J \subseteq B$. Every ring homomorphism $A \to B/J$ lifts to a unique ring homomorphism $W(A) \to B$.*

Proof. See [Grothendieck 1974, Chapitre IV, Proposition 4.3]; the ideal J there is assumed nilpotent, but the proof applies here as well. $\square$

Lemma 1.5 [Zink 2001b, Lemma 2.2]. *Let N be a nonunitary ring which is bounded nilpotent and annihilated by a power of p. Then $W(N)$ is bounded nilpotent and annihilated by a power of p.* $\square$

1B. *The Zink ring.* Let R be an admissible ring. By Lemma 1.4, the exact sequence

$$0 \longrightarrow W(\mathcal{N}_R) \longrightarrow W(R) \longrightarrow W(R_{\mathrm{red}}) \longrightarrow 0$$

has a unique ring homomorphism section $s : W(R_{\mathrm{red}}) \to W(R)$, which is f-equivariant by its uniqueness. Let

$$\mathbb{W}(R) = sW(R_{\mathrm{red}}) \oplus \hat{W}(\mathcal{N}_R).$$

Since $\hat{W}(\mathcal{N}_R)$ is an f-stable ideal of $W(R)$, the group $\mathbb{W}(R)$ is an f-stable subring of $W(R)$, which we call the Zink ring of R.

Lemma 1.6. *The ring $\mathbb{W}(R)$ is stable under the Verschiebung homomorphism*
$v : W(R) \to W(R)$ *if and only if $p \geq 3$ or $pR = 0$. In this case we have an
exact sequence*

$$0 \longrightarrow \mathbb{W}(R) \xrightarrow{v} \mathbb{W}(R) \xrightarrow{w_0} R \longrightarrow 0.$$

Proof. See [Zink 2001a, Lemma 2]. For some $r \geq 0$, the ring $R_0 = \mathbb{Z}/p^r\mathbb{Z}$ is a
subring of R, and we have $\mathbb{W}(R_0) = W(R_0) \cap \mathbb{W}(R)$. The calculation in [loc. cit.]
shows that the element $v(1) \in W(R_0)$ lies in $\mathbb{W}(R_0)$ if and only if $p \geq 3$ or $r = 1$.
For $a \in W(R_{\mathrm{red}})$ we have $v(s(f(a))) = v(f(s(a))) = v(1)s(a)$. Since $\widehat{W}(\mathcal{N}_R)$ is
stable under v and since f is surjective on $W(R_{\mathrm{red}})$, the first assertion of the lemma
follows. The sequence is an extension of

$$0 \longrightarrow W(R_{\mathrm{red}}) \xrightarrow{v} W(R_{\mathrm{red}}) \longrightarrow R_{\mathrm{red}} \longrightarrow 0$$

and

$$0 \longrightarrow \widehat{W}(\mathcal{N}_R) \xrightarrow{v} \widehat{W}(\mathcal{N}_R) \longrightarrow \mathcal{N}_R \longrightarrow 0,$$

which are both exact. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

With a slight modification the exception at the prime 2 can be removed. The ele-
ment $p - [p]$ of $W(\mathbb{Z}_p)$ lies in the image of v because it maps to zero in $\mathbb{Z}_p$. Moreover,
$v^{-1}(p - [p])$ maps to 1 in $W(\mathbb{F}_p)$, so this element is a unit in $W(\mathbb{Z}_p)$. We define

$$u_0 = \begin{cases} v^{-1}(2 - [2]) & \text{if } p = 2, \\ 1 & \text{if } p \geq 3. \end{cases}$$

The image of u_0 in $W(R)$ is also denoted by u_0. For $x \in W(R)$, let

$$\mathsf{v}(x) = v(u_0 x).$$

One could also take $u_0 = v^{-1}(p - [p])$ for all p, which would allow us to state
some results more uniformly, but for odd p this would be overcomplicated.

Lemma 1.7. *The ring $\mathbb{W}(R)$ is stable under $\mathsf{v} : W(R) \to W(R)$, and there is an
exact sequence*

$$0 \longrightarrow \mathbb{W}(R) \xrightarrow{\mathsf{v}} \mathbb{W}(R) \xrightarrow{w_0} R \longrightarrow 0.$$

Proof. By Lemma 1.6, we can assume that $p = 2$. For $a \in W(R_{\mathrm{red}})$, we have
$\mathsf{v}(s(f(a))) = v(u_0 f(s(a))) = v(u_0)s(a) = (2 - [2])s(a)$, which lies in $\mathbb{W}(R)$.
Since $\widehat{W}(\mathcal{N}_R)$ is stable under v and since f is surjective on $W(R_{\mathrm{red}})$, it follows
that $\mathbb{W}(R)$ is stable under v. The sequence is an extension of

$$0 \longrightarrow W(R_{\mathrm{red}}) \xrightarrow{\mathsf{v}} W(R_{\mathrm{red}}) \longrightarrow R_{\mathrm{red}} \longrightarrow 0$$

and

$$0 \longrightarrow \widehat{W}(\mathcal{N}_R) \xrightarrow{\mathsf{v}} \widehat{W}(\mathcal{N}_R) \longrightarrow \mathcal{N}_R \longrightarrow 0.$$

They are exact because in both cases $\mathsf{v} = v \circ u_0$, where u_0 acts bijectively. $\qquad\square$

1C. *The enlarged Zink ring.* Let us recall the logarithm of the Witt ring. For a divided power extension of rings $(B \to R, \delta)$ with kernel $\mathfrak{b} \subseteq B$, the δ-divided Witt polynomials define an isomorphism of $W(B)$-modules

$$\mathrm{Log} : W(\mathfrak{b}) \cong \mathfrak{b}^{\mathbb{N}},$$

where $x \in W(B)$ acts on $\mathfrak{b}^{\mathbb{N}}$ by $[b_0, b_1, \dots] \mapsto [w_0(x)b_0, w_1(x)b_1, \dots]$. The Frobenius and Verschiebung of $W(\mathfrak{b})$ act on $\mathfrak{b}^{\mathbb{N}}$ by

$$f([b_0, b_1, \dots]) = [pb_1, pb_2, \dots], \quad v([b_0, b_1, \dots]) = [0, b_0, b_1, \dots].$$

Moreover, Log induces an injective map $\widehat{W}(\mathfrak{b}) \to \mathfrak{b}^{(\mathbb{N})}$, which is bijective when the divided powers δ are nilpotent; see [Zink 2002, (149)] and the subsequent discussion. In general, let

$$\widetilde{W}(\mathfrak{b}) = \mathrm{Log}^{-1}(\mathfrak{b}^{(\mathbb{N})}).$$

This is an f-stable and v-stable ideal of $W(B)$ containing $\widehat{W}(\mathfrak{b})$.

Assume now that $(B \to R, \delta)$ is a divided power extension of admissible rings (it suffices to assume that R is admissible and that p is nilpotent in B, because then $\mathfrak{b}$ is bounded nilpotent due to the divided powers, so B is admissible as well). Let

$$\mathbb{W}(B, \delta) = \mathbb{W}(B) + \widetilde{W}(\mathfrak{b}).$$

This is an f-stable subring of $W(B)$, which we call the enlarged Zink ring of B with respect to the divided power ideal $(\mathfrak{b}, \delta)$. We also write $\mathbb{W}(B/R)$ for $\mathbb{W}(B, \delta)$. If the divided powers δ are nilpotent then $\mathbb{W}(B, \delta) = \mathbb{W}(B)$. We have the following analogues of Lemmas 1.7 and 1.6:

Lemma 1.8. *The ring $\mathbb{W}(B, \delta)$ is stable under $v : W(R) \to W(R)$, and there is an exact sequence*

$$0 \longrightarrow \mathbb{W}(B, \delta) \overset{\check{v}}{\longrightarrow} \mathbb{W}(B, \delta) \overset{w_0}{\longrightarrow} B \longrightarrow 0.$$

Proof. The ring $\mathbb{W}(B, \delta)$ is stable under v, because $\mathbb{W}(B)$ and $\widetilde{W}(\mathfrak{b})$ are; see Lemma 1.7. We have $\mathbb{W}(B, \delta)/\widetilde{W}(\mathfrak{b}) = \mathbb{W}(R)$. Thus, the exact sequence follows from the exactness of $0 \longrightarrow \widetilde{W}(\mathfrak{b}) \overset{\check{v}}{\longrightarrow} \widetilde{W}(\mathfrak{b}) \longrightarrow \mathfrak{b} \longrightarrow 0$ together with the exact sequence of Lemma 1.7. $\square$

Lemma 1.9. *The ring $\mathbb{W}(B, \delta)$ is stable under $v : W(R) \to W(R)$ if $p \geq 3$, or if $p \in \mathfrak{b}$ and the divided powers δ on $\mathfrak{b}$ induce the canonical divided powers on pB. In this case we have an exact sequence*

$$0 \longrightarrow \mathbb{W}(B, \delta) \overset{v}{\longrightarrow} \mathbb{W}(B, \delta) \overset{w_0}{\longrightarrow} B \longrightarrow 0.$$

Proof. If $p \geq 3$ then $\mathbb{W}(B, \delta)$ is stable under v because $\mathbb{W}(B)$ and $\widetilde{W}(\mathfrak{b})$ are stable under v; see Lemma 1.6. Assume that $p \in \mathfrak{b}$ and that δ induces the canonical divided powers on pB. Let $\xi = p - v(1) \in W(B)$. This element lies in $W(pB) \subseteq W(\mathfrak{b})$ and satisfies $\mathrm{Log}(\xi) = [p, 0, 0, \dots]$. Thus $\xi \in \widetilde{W}(\mathfrak{b})$, which implies that $v(1) \in \mathbb{W}(B, \delta)$. Using this, the proof of Lemma 1.6 shows that $\mathbb{W}(B, \delta)$ is stable under v. The exact sequence follows as usual. $\qquad\square$

1D. *The v-stabilised Zink ring.* Assume that $p = 2$. For an admissible ring R, let γ be the canonical divided powers on the ideal $2R$. We denote the associated enlarged Zink ring by

$$\mathbb{W}^+(R) = \mathbb{W}(R, \gamma) = \mathbb{W}(R) + \widetilde{W}(2R) \subseteq W(R).$$

The kernel of the projection $\mathbb{W}^+(R) \to W(R_{\mathrm{red}})$ will be denoted $\hat{W}^+(\mathcal{N}_R)$. In view of the following lemma, we call $\mathbb{W}^+(R)$ the v-stabilised Zink ring.

Lemma 1.10. *Let $p = 2$. We have*

$$\mathbb{W}^+(R) = \mathbb{W}(R) + \mathbb{W}(R)v(1).$$

The ring $\mathbb{W}^+(R)$ is equal to $\mathbb{W}(R)$ if and only if $2R = 0$. The $\mathbb{W}(R)$-module $\mathbb{W}^+(R)/\mathbb{W}(R)$ is an R_{red}-module generated by $v(1)$.

Proof. By Lemma 1.9, we have $v(1) \in \mathbb{W}^+(R)$. Clearly $2R = 0$ implies that $\mathbb{W}^+(R) = \mathbb{W}(R)$. In general, we consider the filtration

$$W(2\mathcal{N}_R) \subseteq W(2R) \subseteq W(R)$$

and the graded modules for the induced filtrations on $\mathbb{W}(R)$ and $\mathbb{W}^+(R)$. First, the restriction of the divided powers γ to the ideal $2\mathcal{N}_R$ is nilpotent, which implies that

$$\mathbb{W}^+(R) \cap W(2\mathcal{N}_R) = \widetilde{W}(2\mathcal{N}_R) = \hat{W}(2\mathcal{N}_R) = \mathbb{W}(R) \cap W(2\mathcal{N}_R).$$

Next we have $\mathbb{W}^+(R/2R) = \mathbb{W}(R/2R)$, or equivalently

$$\mathbb{W}^+(R)/\mathbb{W}^+(R) \cap W(2R) = \mathbb{W}(R)/\mathbb{W}(R) \cap W(2R).$$

Let $\mathfrak{c} = 2R/2\mathcal{N}_R$. By the preceding remarks, we have an isomorphism

$$\mathbb{W}^+(R)/\mathbb{W}(R) \cong \widetilde{W}(\mathfrak{c})/\hat{W}(\mathfrak{c}).$$

This is an $R/\mathcal{N}_R$-module. Assume that $2R \neq 0$, which implies that $\mathfrak{c} \neq 0$. For some ideal $\mathcal{N}_R \subseteq \mathfrak{b} \subseteq R$, multiplication by 2 induces an isomorphism $R/\mathfrak{b} \cong \mathfrak{c}$. Modulo 2, the divided Witt polynomials are $\tilde{w}_i(x) \equiv \gamma_2(x_{i-1}) + x_i$, so the isomorphism $\mathrm{Log}: W(\mathfrak{c}) \to \mathfrak{c}^{\mathbb{N}}$ takes the form

$$\mathrm{Log}(2a_0, 2a_1, \dots) = 2[a_0, a_0^2 + a_1, a_1^2 + a_2, a_2^2 + a_3, \dots],$$

with $a_i \in R/\mathfrak{b}$. It follows that $\widetilde{W}(\mathfrak{c})/\widehat{W}(\mathfrak{c})$ can be identified with the direct limit of the Frobenius homomorphism $R/\mathfrak{b} \to R/\mathfrak{b} \to \cdots$, which is isomorphic to $R/\sqrt{\mathfrak{b}}$. Under this identification, the element $\xi = 2 - v(1)$ of $\widetilde{W}(\mathfrak{c})$ maps to 1 in $R/\sqrt{\mathfrak{b}}$, because we have $\mathrm{Log}(\xi) = [2, 0, \ldots]$. Hence $\mathbb{W}^+(R)/\mathbb{W}(R)$ is generated by $v(1)$, with annihilator $\sqrt{\mathfrak{b}}$. $\square$

Assume again that $p = 2$. Let $(B \to R, \delta)$ be a divided power extension of admissible rings with kernel $\mathfrak{b} \subseteq B$ such that δ is compatible with the canonical divided powers γ on $2B$. Let δ^+ be the divided powers on $\mathfrak{b}^+ = \mathfrak{b} + 2B$ that extend δ and γ. In this case, we write

$$\mathbb{W}^+(B, \delta) = \mathbb{W}(B, \delta^+) = \mathbb{W}(B) + \widetilde{W}(\mathfrak{b}^+).$$

Clearly $\mathbb{W}(B, \delta) \subseteq \mathbb{W}^+(B, \delta) \supseteq \mathbb{W}^+(B)$. If the divided powers on $\mathfrak{b}^+/2B$ induced by δ are nilpotent, then $\mathbb{W}^+(B, \delta) = \mathbb{W}^+(B)$.

1E. *Passing to the limit.* The preceding considerations carry over to the topological case as follows. For an admissible topological ring R, let

$$\mathbb{W}(R) = \varprojlim_N \mathbb{W}(R/N),$$

the limit taken over all open ideals N of R with $N \subseteq \mathcal{N}_R$. Then Lemmas 1.6 and 1.7 hold for admissible topological rings. The enlarged Zink ring can be defined for topological divided power extensions in the following sense.

Definition 1.11. Let B and R be admissible topological rings. A topological divided power extension is a surjective ring homomorphism $B \to R$ whose kernel $\mathfrak{b}$ is equipped with divided powers δ such that $\mathfrak{b}$ is closed in B, the topology of R is the quotient topology of $B/\mathfrak{b}$, and the linear topology of B is induced by open ideals N for which $N \cap \mathfrak{b}$ is stable under δ. Let δ/N be the divided powers on $N/N \cap \mathfrak{b}$ induced by δ. We say that δ is topologically compatible with the canonical divided powers of p if the topology of B is induced by open ideals N such that δ/N is defined and compatible with the canonical divided powers of p.

Remark 1.12. The existence of divided powers on $\mathfrak{b}$ implies that $\mathfrak{b} \subseteq \mathcal{N}_B$. If B is a noetherian complete local ring, then every ideal $\mathfrak{b}$ of B is closed; moreover, if $\mathfrak{b}$ is given, for each n there is an open ideal $N \subseteq \mathfrak{m}_B^n$ such that $\mathfrak{b} \cap N$ is stable under arbitrary divided powers δ on $\mathfrak{b}$. Indeed, by Artin–Rees there is an l with $\mathfrak{m}_B^n \mathfrak{b} \supseteq \mathfrak{m}_B^l \cap \mathfrak{b}$; then take $N = \mathfrak{m}_B^n \mathfrak{b} + \mathfrak{m}_B^l$, which implies that $\mathfrak{b} \cap N = \mathfrak{m}_B^n \mathfrak{b}$.

Given a topological divided power extension of admissible topological rings $(B \to R, \delta)$ with kernel $\mathfrak{b} \subseteq B$, we define

$$\mathbb{W}(B, \delta) = \varprojlim_N \mathbb{W}(B/N, \delta/N),$$

where N runs through the open ideals of B contained in $\mathcal{N}_B$ such that $N \cap \mathfrak{b}$ is stable under δ. Lemmas 1.8 and 1.9 hold in the topological case.

Assume that $p = 2$. Then for an admissible topological ring, we put

$$\mathbb{W}^+(R) = \varprojlim_N \mathbb{W}^+(R/N),$$

the limit taken over all open ideals N of R contained in $\mathcal{N}_R$. If $(B \to R, \delta)$ is a topological divided power extension of admissible topological rings which is topologically compatible with the canonical divided powers of 2, we can define

$$\mathbb{W}^+(B, \delta) = \varprojlim_N \mathbb{W}^+(B/N, \delta/N),$$

where N runs through the open ideals of B contained in $\mathcal{N}_B$ such that δ/N is defined and compatible with the canonical divided powers of 2.

The following example of admissible topological rings is used in Section 3:

Lemma 1.13. *Let R be a ring which is I-adically complete for an ideal $I \subseteq R$ such that $K = R/I$ is a perfect ring of characteristic p. Assume that $I = J + pR$ for an ideal $J \subseteq R$ such that R/J^n has no p-torsion for each n. For a projective R-module t of finite type, we consider the complete symmetric algebra*

$$R[\![t]\!] = \prod_{n \geq 0} \mathrm{Sym}_R^n(t).$$

Let $(\mathfrak{a} \subseteq S, \delta)$ be the divided power envelope of the ideal $tR[\![t]\!] \subseteq R[\![t]\!]$, and let $\hat{S}$ be the I-adic completion of S. Then:

(i) *$\hat{S} \to R$ is naturally a topological divided power extension of admissible topological rings which is topologically compatible with the canonical divided powers of p.*

(ii) *$\hat{S}$ has no p-torsion.*

Proof. Let $\bar{R}_n = R/(p^n R + J^n)$ and $\bar{S}_n = S \otimes_R \bar{R}_n$. We have $S = R \oplus \mathfrak{a}$ and thus $\bar{S}_n = \bar{R}_n \oplus \bar{\mathfrak{a}}_n$ with $\bar{\mathfrak{a}}_n = \mathfrak{a} \otimes_R \bar{R}_n$; moreover, the ideal $\bar{\mathfrak{a}}_n$ carries divided powers δ_n induced by δ; see [Berthelot 1974, Chapitre I, Proposition 1.7.1]. In particular, $\bar{S}_n$ is admissible. Since $\hat{S} \to R$ is the projective limit over n of $\bar{S}_n \to \bar{R}_n$, to prove (i) it suffices to show that δ_n is compatible with the canonical divided powers of p. Now, $\mathrm{Spec}\, R \to \mathrm{Spec}\, R[\![t]\!]$ is a regular immersion by Lemma A.3, and thus S is flat over R by Proposition A.1. Since R has no p-torsion the same holds for S, so the divided powers on $\mathfrak{a}$ extend canonically to the ideal $\mathfrak{b} = \mathfrak{a} + pS$. We have $S/\mathfrak{b} = R/pR$. The assumptions imply that $\mathrm{Tor}_1^R(R/J^n, R/pR)$ is zero. Hence there is an exact sequence

$$0 \longrightarrow \mathfrak{b}/J^n \mathfrak{b} \longrightarrow S/J^n S \longrightarrow R/(pR + J^n) \longrightarrow 0,$$

which in turn gives an exact sequence

$$0 \longrightarrow (\mathfrak{b}/J^n\mathfrak{b})/p^n(S/J^nS) \longrightarrow S/(J^nS + p^nS) \longrightarrow R/(pR + J^n) \longrightarrow 0.$$

In both sequences the kernels carry divided powers which extend the canonical divided powers of p, since the ideals $J^n\mathfrak{b}$ of S and $p^n(S/J^nS)$ of S/J^nS are stable under the given divided powers. Thus the divided powers δ_n on $\bar{\mathfrak{a}}_n$ are compatible with the canonical divided powers of p, which proves (i).

Let $S_n = S/J^nS$, and let $\widehat{S}_n$ be its p-adic completion. Since S is flat over R and since R/J^n has no p-torsion, S_n and $\widehat{S}_n$ have no p-torsion. Using that $\widehat{S} = \varprojlim_n \widehat{S}_n$, it follows that $\widehat{S}$ has no p-torsion, which proves (ii).   □

1F. *Completeness.* For an admissible ring R, the Zink ring $\mathbb{W}(R)$ is p-adically complete. Indeed, $W(R_{\mathrm{red}})$ is p-adically complete, and $\widehat{W}(\mathcal{N}_R)$ is annihilated by a power of p because this holds for $W(\mathcal{N}_R)$ by Lemma 1.5. The following topological variant of this fact seems to be less obvious:

Proposition 1.14. *Let R be an I-adically complete ring such that the ideal I is finitely generated and $K = R/I$ is a perfect ring of characteristic p. Then the ring $\mathbb{W}(R)$ is p-adically complete. If $p = 2$, the ring $\mathbb{W}^+(R)$ is p-adically complete too.*

This is similar to [Zink 2002, Proposition 3], which says that $W(R)$ is p-adically complete if this holds for R.

Proof. The ring $W(R)$ is p-adically separated, because this holds for each $W(R/I^n)$. Thus $\mathbb{W}(R)$ is p-adically separated too. Let $S = W(K)[\![t_1, \ldots, t_r]\!]$, and let $S \to R$ be a homomorphism which maps $t_1, \ldots, t_r$ to a set of generators of I/I^2. Then $S \to R$ is surjective, and so is $\mathbb{W}(S) \to \mathbb{W}(R)$. Since $\mathbb{W}(R)$ is p-adically separated, in order to show that $\mathbb{W}(R)$ is p-adically complete we may assume that $R = S$. Consider the ideals $J_n = p^n W(R) + W(I^n)$ of $W(R)$ and $\mathbb{J}_n = \mathbb{W}(R) \cap J_n$ of $\mathbb{W}(R)$. Then

$$W(R)/J_n = W_n(K) \oplus W(I/I^n), \quad \mathbb{W}(R)/\mathbb{J}_n = W_n(K) \oplus \widehat{W}(I/I^n).$$

It follows that $W(R)$ and $\mathbb{W}(R)$ are complete and separated for the linear topologies generated by the ideals J_n and $\mathbb{J}_n$, respectively; moreover, $\mathbb{W}(R)$ is closed in $W(R)$. The ring $W(R)$ is also complete and separated for the linear topology generated by the ideals $J'_{n,m} = \mathrm{Ker}(W(R) \to W_m(R/I^n))$. The J-topology is finer than the J'-topology because $J_{2n} \subseteq J'_{n,n}$.

We claim that for each $r \geq 1$ the ideal $p^r W(R)$ of $W(R)$ is closed in the J'-topology. This is a variant of [Zink 2002, Lemma 6] with essentially the same proof. First, for $s \geq 1$, an element $x = (x_0, \ldots, x_m)$ of $W_{m+1}(R)$ satisfies $x_i \in I^s$ for all i if and only if $w_i(x) \in I^{i+s}$ for all i; see the proof of [Zink 2002, Lemma 4]. Then the proof of Lemma 5 in that work shows that an element $x \in W_m(R)$ is divisible

by p^r if and only if for each s the image $\bar{x} \in W_m(R/I^s)$ is divisible by p^r. Using this, the claim follows from the proof of [Zink 2002, Lemma 6].

Thus $p^r W(R)$ is closed in the finer J-topology as well. Assume that we have $p\mathbb{W}(R) = pW(R) \cap \mathbb{W}(R)$. Then $p^r\mathbb{W}(R)$ is closed in the $\mathbb{J}$-topology, which implies that $\mathbb{W}(R)$ is p-adically complete; see [Zink 2002, Lemma 7]. Thus for $p \geq 3$, the proof is completed by Lemma 1.15 below. For $p = 2$ the same reasoning shows that $\mathbb{W}^+(R)$ is p-adically complete. Now $\mathbb{W}^+(R)/\mathbb{W}(R)$ is isomorphic to K as abelian groups by the proof of Lemma 1.10. We get exact sequences

$$0 \longrightarrow K \longrightarrow \mathbb{W}(R)/2^n\mathbb{W}(R) \longrightarrow \mathbb{W}^+(R)/2^n\mathbb{W}^+(R) \longrightarrow K \longrightarrow 0,$$

where the transition maps from $n+1$ to n are zero on the left-hand K and the identity on the right-hand K. It follows that $\mathbb{W}(R)$ is p-adically complete as well. □

Lemma 1.15. *For a perfect ring K of characteristic p, let $R = W(K)[\![t_1, \ldots, t_r]\!]$, with the $(p, t_1, \ldots, t_r)$-adic topology. If $p \geq 3$ then*

$$pW(R) \cap \mathbb{W}(R) = p\mathbb{W}(R).$$

If $p = 2$ then

$$2W(R) \cap \mathbb{W}^+(R) = 2\mathbb{W}^+(R).$$

Proof. Assume $p = 2$. Let I be the kernel of $R \to K$ and let $\bar{I} = I/pR$. The filtration $0 \subseteq W(pR) \subseteq W(I) \subseteq W(R)$ induces a filtration of $\mathbb{W}(R)$ with successive quotients $\tilde{W}(pR) := \varprojlim_n \tilde{W}(pR/I^n pR)$ and $\hat{W}(\bar{I}) := \varprojlim_n \hat{W}(\bar{I}/\bar{I}^n)$ and $W(K)$. To prove the lemma it suffices to show that

$$pW(\bar{I}) \cap \hat{W}(\bar{I}) = p\hat{W}(\bar{I})$$

and

$$pW(pR) \cap \tilde{W}(pR) = p\tilde{W}(pR).$$

The first equality holds since multiplication by p on $W(\bar{I})$ is given by $(a_0, a_1, \ldots) \mapsto (0, a_0^p, a_1^p, \ldots)$, and for $a \in \bar{I}$ with $a^p \in \bar{I}^{pn}$ we have $a \in \bar{I}^n$. The second equality holds because the isomorphism $\mathrm{Log} : W(pR) \cong (pR)^{\mathbb{N}}$ induces an isomorphism between $\tilde{W}(pR)$ and the group of all sequences in $(pR)^{\mathbb{N}}$ that converge to zero I-adically. The proof for $p \geq 3$ is similar. □

1G. *Divided powers.* In Section 3, we will use that the augmentation ideals of the Zink ring and its variants carry natural divided powers, with some exception when $p = 2$; see also Section 4A.

Let us first recall the canonical divided powers on the Witt ring. If R is a $\mathbb{Z}_{(p)}$-algebra, then $W(R)$ is a $\mathbb{Z}_{(p)}$-algebra as well, and the ideal I_R carries divided powers γ which are determined by $(p-1)!\gamma_p(v(x)) = p^{p-2}v(x^p)$. Assume that $(B \to R, \delta)$ is a divided power extension of $\mathbb{Z}_{(p)}$-algebras with kernel $\mathfrak{b} \subseteq B$. Let $I_{B/R}$ be the kernel of $W(B) \to R$. If $i : \mathfrak{b} \to W(\mathfrak{b})$ is defined by $\mathrm{Log}(i(b)) =$

$[b, 0, 0, \ldots]$, we have $I_{B/R} = I_B \oplus i(\mathfrak{b})$, and the divided powers γ on I_B extend to divided powers $\gamma' = \gamma \oplus \delta$ on $I_{B/R}$ such that $\gamma'_n(i(b)) = i(\delta_n(b))$ for $b \in \mathfrak{b}$. If $p \in \mathfrak{b}$ and if δ extends the canonical divided powers of p, then $\gamma \oplus \delta$ extends the canonical divided powers of p, and f preserves $\gamma \oplus \delta$. This is clear when B has no p-torsion; the general case follows because $(B \to R)$ can be written as the quotient of a divided power extension $(B' \to R')$, where B' is the divided power algebra of a free module over a polynomial ring R'' over $\mathbb{Z}_{(p)}$, and $R' = R''/pR''$.

These facts extend to the Zink ring as follows:

Lemma 1.16. *Let* $\mathbb{I} \subseteq \mathbb{W}$ *be one of the following*:

(i) $\mathbb{I} = \mathbb{I}_R$ *and* $\mathbb{W} = \mathbb{W}(R)$ *for an admissible ring R with $p \geq 3$.*

(ii) $\mathbb{I} = \mathbb{I}_R^+$ *and* $\mathbb{W} = \mathbb{W}^+(R)$ *for an admissible ring R with $p = 2$.*

Then the divided powers γ on I_R induce divided powers on $\mathbb{I}$.

Proof. Since $\mathbb{W}$ is a $\mathbb{Z}_{(p)}$-algebra, it suffices to show that $\mathbb{I}$ is stable under the map $\gamma_p : I_R \to I_R$, which is true because $\mathbb{I} = v(\mathbb{W})$ by Lemmas 1.6 and 1.9. $\square$

Lemma 1.17. *Let* $(B \to R, \delta)$ *be a divided power extension of admissible rings with kernel $\mathfrak{b} \subseteq B$, and let $\mathbb{I}_{B/R}$ be the kernel of $\mathbb{W}(B, \delta) \to R$. Assume that $p \geq 3$; or that $p = 2$ and $p \in \mathfrak{b}$ and δ extends the canonical divided powers of p. Then the divided powers $\gamma \oplus \delta$ on $I_{B/R}$ induce divided powers on $\mathbb{I}_{B/R}$. If $p \in \mathfrak{b}$ and if δ extends the canonical divided powers of p, then the divided powers on $\mathbb{I}_{B/R}$ induced by $\gamma \oplus \delta$ extend the canonical divided powers of p and are preserved by f.*

Proof. Let $\mathbb{I}'_B$ be the kernel of $\mathbb{W}(B/R) \to B$. Then $\mathbb{I}_{B/R} = \mathbb{I}'_B \oplus i(\mathfrak{b})$, and we have $\mathbb{I}'_B = v(\mathbb{W}(B/R))$ by Lemma 1.9. Thus $\mathbb{I}'_B$ is stable under γ, and $\mathbb{I}_{B/R}$ is stable under $\gamma \oplus \delta$. The second assertion follows from the corresponding fact for the Witt ring. $\square$

2. Dieudonné displays

In this section, Dieudonné displays and a number of variants related to divided power extensions are defined. We use the formalism of frames and windows introduced in [Lau 2010]. First of all, let us recall a well-known fact:

Lemma 2.1. *Let A be a commutative, not necessarily unitary ring. For $x \in W(A)$ we have $f(x) \equiv x^p$ modulo $pW(A)$. Similarly, for $x \in \widehat{W}(A)$ we have $f(x) \equiv x^p$ modulo $p\widehat{W}(A)$.*

Proof. For $x \in W(R)$ write $x = [x_0] + v(y)$ with $x_0 \in R$ and $y \in W(R)$. Then $f(x) \equiv [x_0^p] \equiv x^p$ modulo $pW(R)$ because $fv = p$ and $v(y)^p = p^{p-1}v(y^p)$. The same calculation applies with $\widehat{W}$ in place of W. $\square$

2A. *Frames and windows.* We recall the notion of frames and windows from [Lau 2010], with some additions. A *preframe* is a quintuple

$$\mathcal{F} = (S, I, R, \sigma, \sigma_1)$$

where S and $R = S/I$ are rings, where $\sigma : S \to S$ is a ring endomorphism with $\sigma(a) \equiv a^p$ modulo pS, and where $\sigma_1 : I \to S$ is a σ-linear map of S-modules whose image generates S as an S-module. Then there is a unique element $\theta \in S$ with $\sigma(a) = \theta\sigma_1(a)$ for $a \in I$. The preframe $\mathcal{F}$ is called a *frame* if

$$I + pS \subseteq \mathrm{Rad}(S).$$

If, in addition, all projective R-modules of finite type can be lifted to projective S modules, then $\mathcal{F}$ is called a *lifting frame*.

A homomorphism of preframes or frames $\alpha : \mathcal{F} \to \mathcal{F}'$ is a ring homomorphism $\alpha : S \to S'$ with $\alpha(I) \subseteq I'$ such that $\sigma'\alpha = \alpha\sigma$ and $\sigma_1'\alpha = u \cdot \alpha\sigma_1$ for a unit $u \in S'$, which is then determined by α. It also follows that $\alpha(\theta) = u\theta'$. We say that α is a *u-homomorphism* of preframes or frames. If $u = 1$ then α is called *strict*.

Now let $\mathcal{F}$ be a frame. An *$\mathcal{F}$-window* is a quadruple

$$\mathscr{P} = (P, Q, F, F_1)$$

where P is a projective S-module of finite type with a submodule Q such that there exists a decomposition of S-modules $P = L \oplus T$ with $Q = L \oplus IT$, called a *normal decomposition*, and where $F : P \to P$ and $F_1 : Q \to P$ are σ-linear maps of S-modules with

$$F_1(ax) = \sigma_1(a)F(x)$$

for $a \in I$ and $x \in P$; we also assume that $F_1(Q)$ generates P as an S-module. Then $F(x) = \theta F_1(x)$ for $x \in Q$. If F is a lifting frame, every pair (P, Q) such that P is a projective S-module of finite type and P/Q is a projective R-module admits a normal decomposition. In general, for given (P, Q) together with a normal decomposition $P = L \oplus T$, giving σ-linear maps (F, F_1) which make an $\mathcal{F}$-window $\mathscr{P}$ is equivalent to giving a σ-linear isomorphism

$$\Psi : L \oplus T \to P$$

defined by F_1 on L and by F on T. The triple (L, T, Ψ) is called a *normal representation* of $\mathscr{P}$.

A frame homomorphism $\alpha : \mathcal{F} \to \mathcal{F}'$ induces a base change functor α_* from $\mathcal{F}$-windows to $\mathcal{F}'$-windows. In terms of normal representations, it is given by

$$(L, T, \Psi) \mapsto (S' \otimes_S L, S' \otimes_S T, \Psi')$$

with $\Psi'(s' \otimes l) = u\sigma'(s') \otimes \Psi(l)$ and $\Psi'(s' \otimes t) = \sigma'(s') \otimes \Psi(t)$.

A frame homomorphism $\alpha : \mathcal{F} \to \mathcal{F}'$ is called *crystalline* if the functor α_* is an equivalence of categories. For reference, we recall [Lau 2010, Theorem 3.2]:

Theorem 2.2. *Let $\alpha : \mathcal{F} \to \mathcal{F}'$ be a homomorphism of frames which induces an isomorphism $R \cong R'$ and a surjection $S \to S'$ with kernel $\mathfrak{a}$. We assume that there is a finite filtration of ideals $\mathfrak{a} = \mathfrak{a}_0 \supseteq \cdots \supseteq \mathfrak{a}_n = 0$ with $\sigma_1(\mathfrak{a}_i) \subseteq \mathfrak{a}_i$ and $\sigma(\mathfrak{a}_i) \subseteq \mathfrak{a}_{i+1}$, that σ_1 is elementwise nilpotent on each $\mathfrak{a}_i/\mathfrak{a}_{i+1}$, and that all projective S'-modules of finite type lift to projective S-modules of finite type. Then α is crystalline.* □

Let us recall the operator $V^{\sharp}$ of a window. For an S-module M we write $M^{(\sigma)} = S \otimes_{\sigma,S} M$. A filtered F-V-module over $\mathcal{F}$ is a quadruple

$$(P, Q, F^{\sharp}, V^{\sharp})$$

where P is a projective S-module of finite type, Q is a submodule of P such that P/Q is projective over R, and $F^{\sharp} : P^{(\sigma)} \to P$ and $V^{\sharp} : P \to P^{(\sigma)}$ are S-linear maps with $F^{\sharp}V^{\sharp} = \theta$ and $V^{\sharp}F^{\sharp} = \theta$.

Lemma 2.3. *There is a natural functor from $\mathcal{F}$-windows to filtered F-V-modules over $\mathcal{F}$, which is fully faithful if θ is not a zero divisor in S.*

Proof. The functor is $(P, Q, F, F_1) \mapsto (P, Q, F^{\sharp}, V^{\sharp})$, where $F^{\sharp}$ is the linearisation of F, and $V^{\sharp}$ is the unique S-linear map such that $V^{\sharp}(F_1(x)) = 1 \otimes x$ for $x \in Q$. Clearly this determines $V^{\sharp}$ if it exists. In terms of a normal representation (L, T, Ψ) of $\mathscr{P}$, thus $P = L \oplus T$, one can define $V^{\sharp} = (1 \oplus \theta)(\Psi^{\sharp})^{(-1)}$. The required relation $F^{\sharp}V^{\sharp} = \theta$ on P is equivalent to $F^{\sharp}V^{\sharp}F_1 = \theta F_1$ on Q, which is clear since $\theta F_1 = F$. The required relation $V^{\sharp}F^{\sharp} = \theta$ on $P^{(\sigma)}$ holds if and only if it holds after multiplication with $\sigma_1(a)$ for all $a \in I$. For $x \in P$ we calculate $\sigma_1(a)V^{\sharp}F^{\sharp}(1 \otimes x) = V^{\sharp}F_1(ax) = \sigma(a) \otimes x = \theta \sigma_1(a)(1 \otimes x)$.

Assume that θ is not a zero divisor in S. It suffices to show that the forgetful functors from windows to triples (P, Q, F) and from filtered F-V-modules to triples $(P, Q, F^{\sharp})$ are fully faithful. In the first case this holds because $\theta F_1 = F$. In the second case, for an endomorphism α of P with $\alpha F^{\sharp} = F^{\sharp}\alpha^{(\sigma)}$ we calculate $V^{\sharp}\alpha\theta = V^{\sharp}\alpha F^{\sharp}V^{\sharp} = V^{\sharp}F^{\sharp}\alpha^{(\sigma)}V^{\sharp} = \theta\alpha^{(\sigma)}V^{\sharp}$, which implies that $V^{\sharp}\alpha = \alpha^{(\sigma)}V^{\sharp}$. □

Finally, we recall the duality formalism. Let $\mathcal{F}$ denote the $\mathcal{F}$-window (S, I, σ, σ_1). A bilinear form between $\mathcal{F}$-windows

$$\beta : \mathscr{P} \times \mathscr{P}' \to \mathcal{F}$$

is an S-bilinear map $\beta : P \times P' \to S$ such that $\beta(Q \times Q') \subseteq I$ and $\beta(F_1 x, F_1' x') = \sigma_1(\beta(x, x'))$ for $x \in Q$ and $x' \in Q'$. For each $\mathscr{P}$, the functor $\mathscr{P}' \mapsto \mathrm{Bil}(\mathscr{P} \times \mathscr{P}', \mathcal{F})$ is represented by an $\mathcal{F}$-window $\mathscr{P}^t$, called the dual of $\mathscr{P}$. The tautological bilinear form $\mathscr{P} \times \mathscr{P}^t \to S$ is a perfect bilinear map $P \times P^t \to S$. There is a bijection between normal representations $P = L \oplus T$ and $P^t = L^t \oplus T^t$ determined by

$\langle L, L^t \rangle = 0 = \langle T, T^t \rangle$. The associated operators $\Psi : P \to P$ and $\Psi^t : P^t \to P^t$ are related by $\langle \Psi x, \Psi^t x' \rangle = \sigma \langle x, x' \rangle$.

There is also an obvious duality of filtered F-V-modules over $\mathcal{F}$: the dual of $\mathcal{M} = (P, Q, F^\sharp, V^\sharp)$ is $\mathcal{M}^t = (P^*, Q', V^{\sharp *}, F^{\sharp *})$, where $P^* = \mathrm{Hom}_S(P, S)$ and Q' is the submodule of all y in P^* with $y(Q) \subseteq I$. It is easy to see that the functor in Lemma 2.3 preserves duality.

2B. *Frames associated to the Witt ring.*
For an arbitrary ring R let $f_1 : I_R \to W(R)$ be the inverse of the Verschiebung v. Then

$$\mathscr{W}_R = (W(R), I_R, R, f, f_1)$$

is a preframe with $\theta = p$. If R is p-adically complete, $\mathscr{W}_R$ is a lifting frame because $W(R)$ is I_R-adically complete by [Zink 2002, Proposition 3], and windows over $\mathscr{W}_R$ are displays over R.

For a divided power extension of rings $(B \to R, \delta)$ with kernel $\mathfrak{b} \in B$, one can define a preframe

$$\mathscr{W}_{B/R} = (W(B), I_{B/R}, R, f, \tilde{f}_1)$$

with $I_{B/R} = I_B + W(\mathfrak{b})$ such that $\tilde{f}_1 : I_{B/R} \to W(B)$ is the unique extension of f_1 whose restriction to $W(\mathfrak{b})$ is given by $[a_0, a_1, a_2, \dots] \mapsto [a_1, a_2, \dots]$ in logarithmic coordinates; see Section 1C. The projection $\mathscr{W}_B \to \mathscr{W}_R$ factors into strict preframe homomorphisms $\mathscr{W}_B \to \mathscr{W}_{B/R} \to \mathscr{W}_R$.

As a special case, assume that R is a perfect ring of characteristic p. Then f is an automorphism of $W(R)$, and $I_R = pW(R)$. Let us define a Dieudonné module over R to be a triple (P, F, V) where P is a projective $W(R)$-module of finite type equipped with an f-linear endomorphism F and an f^{-1}-linear endomorphism V such that $FV = p$, or, equivalently, $VF = p$.

Lemma 2.4. *Displays over a perfect ring R are equivalent to Dieudonné modules over R.*

Proof. To a display (P, Q, F, F_1) we associate the Dieudonné module (P, F, V), where the linearisation of $V : P \to P$ is the operator $V^\sharp$ defined in Lemma 2.3. Then $VF_1 : Q \to P$ is the inclusion. Here F_1 is surjective since f is bijective. Thus $Q = V(P)$, and the functor is fully faithful; see Lemma 2.3. It remains to show that for every Dieudonné module (P, F, V) the R-module $M = P/V(P)$ is projective. For $\mathfrak{p} \in \mathrm{Spec}\, R$ let $\ell_M(\mathfrak{p})$ be the dimension of the fibre of M at $\mathfrak{p}$. Let $N = P/F(P)$. Then $\ell_M + \ell_N = \ell_{P/pP}$ as functions on $\mathrm{Spec}\, R$. Since M and N are of finite type and since P/pP is projective, the functions ℓ_M and ℓ_N are upper semicontinuous, and $\ell_{P/pP}$ is locally constant. It follows that ℓ_M is locally constant, which implies that M is projective because R is reduced. $\square$

2C. *Dieudonné frames.* For an admissible ring R in the sense of Definition 1.1, let $\mathbb{I}_R$ be the kernel of $w_0 : \mathbb{W}(R) \to R$, and let $\mathfrak{f}_1 : \mathbb{I}_R \to \mathbb{W}(R)$ be the inverse of v, which is well-defined by Lemma 1.7. If p is odd, then $\mathsf{v} = v$ and $\mathfrak{f}_1 = f_1$.

Lemma 2.5. *The quintuple*

$$\mathscr{D}_R = (\mathbb{W}_R, \mathbb{I}_R, R, f, \mathfrak{f}_1)$$

is a lifting frame.

We call $\mathscr{D}_R$ the Dieudonné frame associated to R.

Proof. In order that $\mathscr{D}_R$ is a preframe we need that $f(a) \equiv a^p$ modulo $p\mathbb{W}(R)$ for $a \in \mathbb{W}(R)$, which follows from Lemma 2.1 applied to $W(R_{\mathrm{red}})$ and to $\hat{W}(\mathcal{N}_R)$. Since $\hat{W}(\mathcal{N}_R)$ is a nilideal by Lemma 1.5 and since the quotient $\mathbb{W}(R)/\hat{W}(\mathcal{N}_R) = W(R_{\mathrm{red}})$ is p-adically complete with $pW(R_{\mathrm{red}}) = I_{R_{\mathrm{red}}}$, the kernel of $\mathbb{W}(R) \to R_{\mathrm{red}}$ lies in the radical of $\mathbb{W}(R)$, and projective R_{red}-modules of finite type lift to projective $\mathbb{W}(R)$-modules of finite type. It follows that $\mathscr{D}_R$ is a lifting frame. $\square$

The inclusion $\mathbb{W}(R) \to W(R)$ is a u_0-homomorphism of frames $\mathscr{D}_R \to \mathscr{W}_R$. Thus for $\mathscr{D}_R$ we have $\theta = p$ if p is odd and $\theta = 2u_0 = 2 - [4]$ if $p = 2$.

Definition 2.6. A Dieudonné display over R is a window over $\mathscr{D}_R$.

Thus a Dieudonné display is a quadruple $\mathscr{P} = (P, Q, F, F_1)$ where P is a projective $\mathbb{W}(R)$-module of finite type with a filtration $\mathbb{I}_R P \subseteq Q \subseteq P$ such that P/Q is a projective R-module, $F : P \to P$ and $F_1 : Q \to P$ are f-linear maps with $F_1(ax) = \mathfrak{f}_1(a)F(x)$ for $a \in \mathbb{I}_R$ and $x \in P$, and $F_1(Q)$ generates P. We write

$$\mathrm{Lie}(\mathscr{P}) = P/Q.$$

The *height* of $\mathscr{P}$ is the rank of the $\mathbb{W}(R)$-module P, and the *dimension* of $\mathscr{P}$ is the rank of the R-module $\mathrm{Lie}(\mathscr{P})$, both viewed as locally constant functions on $\mathrm{Spec}\, R$. As in the case of general frames, we also denote by $\mathscr{D}_R$ the Dieudonné display $(\mathbb{W}(R), \mathbb{I}_R, f, \mathfrak{f}_1)$ over R.

2D. *Relative Dieudonné frames.* Let $(B \to R, \delta)$ be a divided power extension of admissible rings with kernel $\mathfrak{b} \subseteq B$. Let $\mathbb{W}(B/R) = \mathbb{W}(B, \delta)$ as in Section 1C and let $\mathbb{I}_{B/R}$ be the kernel of the projection $\mathbb{W}(B/R) \to R$; thus

$$\mathbb{I}_{B/R} = \mathbb{I}_B + \tilde{W}(\mathfrak{b}).$$

Lemma 2.7. *There is a unique extension of* $\mathfrak{f}_1 : \mathbb{I}_B \to \mathbb{W}(B)$ *to an* f*-linear map* $\tilde{\mathfrak{f}}_1 : \mathbb{I}_{B/R} \to \mathbb{W}(B/R)$ *of* $\mathbb{W}(B/R)$*-modules such that the restriction of* $\tilde{\mathfrak{f}}_1$ *to* $\tilde{W}(\mathfrak{b})$ *is given by*

$$\tilde{\mathfrak{f}}_1([a_0, a_1, a_2, \dots]) = [w_0(u_0^{-1})a_1, w_1(u_0^{-1})a_2, \dots] \tag{2-1}$$

in logarithmic coordinates. The quintuple

$$\mathscr{D}_{B/R} = \mathscr{D}_{B/R,\delta} = (\mathbb{W}(B/R), \mathbb{I}_{B/R}, R, f, \tilde{\mathfrak{f}}_1)$$

is a lifting frame.

Proof. Clearly $\tilde{\mathfrak{f}}_1$ is determined by (2-1). Let $\mathbb{I}'_B$ be the kernel of $\mathbb{W}(B/R) \to B$. By Lemma 1.8, the inverse of $\mathtt{v}$ is an f-linear map $\mathfrak{f}'_1 : \mathbb{I}'_B \to \mathbb{W}(B/R)$ which extends $\mathfrak{f}_1$. In logarithmic coordinates, the restriction of $\mathtt{v}$ to $W(\mathfrak{b})$ is given by $[a_0, a_1, \dots] \mapsto [0, w_0(u_0)a_0, w_1(u_0)a_1, \dots]$. Thus $\mathfrak{f}'_1$ extends to the desired $\tilde{\mathfrak{f}}_1$. As in the proof of Lemma 2.5, the kernel of $\mathbb{W}(B/R) \to R_{\text{red}}$ lies in the radical of $\mathbb{W}(B/R)$, and projective R_{red}-modules of finite type lift to $\mathbb{W}(B/R)$. $\square$

We call $\mathscr{D}_{B/R}$ the relative Dieudonné frame associated to the divided power extension $(B/R, \delta)$, and $\mathscr{D}_{B/R}$-windows are called Dieudonné displays for B/R. There are natural strict frame homomorphisms

$$\mathscr{D}_B \longrightarrow \mathscr{D}_{B/R} \longrightarrow \mathscr{D}_R.$$

If the divided powers δ are nilpotent, then $\mathbb{W}(B) = \mathbb{W}(B/R)$.

Proposition 2.8. *The frame homomorphism $\mathscr{D}_{B/R} \to \mathscr{D}_R$ is crystalline.*

Proof. This follows from Theorem 2.2. Indeed, let $\mathfrak{a}$ denote the kernel of the surjective homomorphism $\mathbb{W}(B/R) \to \mathbb{W}(R)$; thus $\mathfrak{a} = \tilde{W}(\mathfrak{b}) \cong \mathfrak{b}^{(\mathbb{N})}$. The endomorphism $\tilde{\mathfrak{f}}_1$ of $\mathfrak{a}$ is elementwise nilpotent by (2-1). The required filtration of $\mathfrak{a}$ can be taken to be $\mathfrak{a}_i = p^i \mathfrak{a}$; this is a finite filtration by Lemma 1.5. We have $\tilde{\mathfrak{f}}_1(\mathfrak{a}_i) = \mathfrak{a}_i$ by (2-1), and $f(\mathfrak{a}_i) = \mathfrak{a}_{i+1}$ because the endomorphism f of $\mathfrak{a}$ is given by $[a_0, a_1, \dots] \mapsto [pa_1, pa_2, \dots]$ in logarithmic coordinates. $\square$

2E. *v*-stabilised Dieudonné frames. Assume that $p = 2$. The preceding constructions can be repeated with $\mathbb{W}^+$ and v in place of $\mathbb{W}$ and $\mathtt{v}$. More precisely, for an admissible ring R, let $\mathbb{I}_R^+$ be the kernel of $\mathbb{W}^+(R) \to R$ and let $f_1 : \mathbb{I}_R^+ \to \mathbb{W}^+(R)$ be the inverse of v, which is well-defined by Lemma 1.9. The v-stabilised Dieudonné frame associated to R is defined as

$$\mathscr{D}_R^+ = (\mathbb{W}^+(R), \mathbb{I}_R^+, R, f, f_1).$$

This is a lifting frame by the proof of Lemma 2.5. The inclusion $\mathbb{W}(R) \to \mathbb{W}^+(R)$ is a u_0-homomorphism of frames $\mathscr{D}_R \to \mathscr{D}_R^+$, which is invertible if and only if $2R = 0$. Windows over $\mathscr{D}_R^+$ are called v-stabilised Dieudonné displays over R.

Assume again that $p = 2$, and let $(B \to R, \delta)$ be a divided power extension of admissible rings with kernel $\mathfrak{b} \subseteq B$ which is compatible with the canonical divided powers of 2. Let $\mathbb{I}_{B/R}^+$ be the kernel of the natural map $\mathbb{W}_{B/R}^+ \to R$; thus

$$\mathbb{I}_{B/R}^+ = \mathbb{I}_B^+ + \tilde{W}(\mathfrak{b}).$$

There is a unique extension of $f_1 : \mathbb{I}_B^+ \to \mathbb{W}^+(B)$ to an f-linear map of $\mathbb{W}^+(B/R)$-modules $\tilde{f}_1 : \mathbb{I}_{B/R}^+ \to \mathbb{W}^+(B/R)$ such that its restriction to $\widetilde{W}(\mathfrak{b})$ is given by $[a_0, a_1, a_2, \dots] \mapsto [a_1, a_2, \dots]$ in logarithmic coordinates, and the quintuple

$$\mathscr{D}_{B/R}^+ = (\mathbb{W}^+(B/R), \mathbb{I}_{B/R}^+, R, f, \tilde{f}_1)$$

is a lifting frame. This follows from the proof of Lemma 2.7. We have a u_0-homomorphism of frames $\mathscr{D}_{B/R} \to \mathscr{D}_{B/R}^+$, which is invertible if and only if $2R = 0$, and strict frame homomorphisms

$$\mathscr{D}_B^+ \longrightarrow \mathscr{D}_{B/R}^+ \longrightarrow \mathscr{D}_R^+.$$

If the divided powers induced by δ on $(\mathfrak{b} + 2B)/2B$ are nilpotent, then $\mathbb{W}^+(B)$ is equal to $\mathbb{W}^+(B/R)$.

Corollary 2.9. *The frame homomorphism $\mathscr{D}_{B/R}^+ \to \mathscr{D}_R^+$ is crystalline.*

Proof. This follows from the proof of Proposition 2.8. $\qquad\qquad\square$

2F. *The crystals associated to Dieudonné displays.* Let R be an admissible ring. We denote the category of divided power extensions (Spec $A \to$ Spec B, δ), where A is an R-algebra which is an admissible ring, and where p is nilpotent in B, by $\mathrm{Cris}_{\mathrm{adm}}(R)$. Then the kernel of $B \to A$ is bounded nilpotent, so B is an admissible ring as well.

Let $\mathscr{P}$ be a Dieudonné display over R. For (Spec $A \to$ Spec B, δ) in $\mathrm{Cris}_{\mathrm{adm}}(R)$, we denote the base change of $\mathscr{P}$ to A by $\mathscr{P}_A$ and the unique Dieudonné display for B/A which lifts $\mathscr{P}_A$ by

$$\mathscr{P}_{B/A} = (P_{B/A}, Q_{B/A}, F, F_1);$$

see Proposition 2.8. A homomorphism of divided power extensions of admissible rings $\alpha : (B \to A, \delta) \to (B' \to A', \delta')$ induces a frame homomorphism $\mathscr{D}_\alpha : \mathscr{D}_{B/A} \to \mathscr{D}_{B'/A'}$, and we have a natural isomorphism

$$(\mathscr{D}_\alpha)_*(\mathscr{P}_{B/A}) \cong \mathscr{P}_{B'/A'}.$$

In more sophisticated terms, this can be expressed as follows: The frames $\mathscr{D}_{B/A}$ form a presheaf of frames $\mathscr{D}_{**}$ on $\mathrm{Cris}_{\mathrm{adm}}(R)$, and Proposition 2.8 implies that the category of Dieudonné displays over R is equivalent to the category of crystals in $\mathscr{D}_{**}$-windows on $\mathrm{Cris}_{\mathrm{adm}}(R)$. Then $\mathscr{P}_{B/A}$ is the value in (Spec $A \to$ Spec B, δ) of the crystal associated to $\mathscr{P}$.

For a Dieudonné display $\mathscr{P} = (P, Q, F, F_1)$ over R, we define the Witt crystal $\mathbb{K}(\mathscr{P})$ on $\mathrm{Cris}_{\mathrm{adm}}(R)$ by

$$\mathbb{K}(\mathscr{P})_{B/A} = P_{B/A}.$$

This is a projective $\mathbb{W}(B/A)$-module of finite type. The Dieudonné crystal $\mathbb{D}(\mathscr{P})$ on $\mathrm{Cris}_{\mathrm{adm}}(R)$ is defined by

$$\mathbb{D}(\mathscr{P})_{B/A} = P_{B/A} \otimes_{\mathbb{W}(B/A)} B.$$

This is a projective B-module of finite type. The Hodge filtration of $\mathscr{P}$ is the submodule

$$Q/\mathbb{I}_R P \subseteq P/\mathbb{I}_R P = \mathbb{D}(\mathscr{P})_{R/R}.$$

Corollary 2.10. *Let $(B \to R, \delta)$ be a nilpotent divided power extension of admissible rings. The category of Dieudonné displays over B is equivalent to the category of Dieudonné displays $\mathscr{P}$ over R together with a lift of the Hodge filtration of $\mathscr{P}$ to a direct summand of $\mathbb{D}(\mathscr{P})_{B/R}$.*

Proof. If the divided powers are nilpotent, then $\mathbb{W}(B/R) = \mathbb{W}(B)$, and lifts of windows under the frame homomorphism $\mathscr{D}_B \to \mathscr{D}_{B/R}$ are in bijection with lifts of the Hodge filtration. $\square$

The preceding definitions have a v-stabilised variant. Let $\mathrm{Cris}_{\mathrm{adm}}(R/\mathbb{Z}_p)$ be the full subcategory of $\mathrm{Cris}_{\mathrm{adm}}(R)$ where the divided powers are compatible with the canonical divided powers of p. Assume now that $p = 2$, and let $\mathscr{P}^+$ be a v-stabilised Dieudonné display over R, i.e., a window over $\mathscr{D}_R^+$. For $(\mathrm{Spec}\, A \to \mathrm{Spec}\, B, \delta)$ in $\mathrm{Cris}_{\mathrm{adm}}(R/\mathbb{Z}_2)$ we denote by $\mathscr{P}_A^+$ the base change of $\mathscr{P}^+$ to $\mathscr{D}_A^+$ and by

$$\mathscr{P}_{B/A}^+ = (P_{B/A}^+, Q_{B/A}^+, F, F_1)$$

the unique lift of $\mathscr{P}_A^+$ to a $\mathscr{D}_{B/A}^+$-window, which exists by Corollary 2.9. The v-stabilised Witt crystal $\mathbb{K}^+(\mathscr{P}^+)$ and the v-stabilised Dieudonné crystal $\mathbb{D}^+(\mathscr{P}^+)$ on $\mathrm{Cris}_{\mathrm{adm}}(R/\mathbb{Z}_2)$ are defined by $\mathbb{K}^+(\mathscr{P}^+)_{B/A} = P_{B/A}^+$ and

$$\mathbb{D}^+(\mathscr{P}^+)_{B/A} = P_{B/A}^+ \otimes_{\mathbb{W}^+(B/A)} B.$$

Corollary 2.11. *Assume that $p = 2$. Let $(B \to R, \delta)$ be a divided power extension of admissible rings which is compatible with the canonical divided powers of 2 such that the divided powers induced by δ on the kernel of $B/2B \to R/2R$ are nilpotent. Then the category of v-stabilised Dieudonné displays over B is equivalent to the category of v-stabilised Dieudonné displays $\mathscr{P}^+$ over R together with a lift of the Hodge filtration of $\mathscr{P}^+$ to a direct summand of $\mathbb{D}^+(\mathscr{P}^+)_{B/R}$.*

Proof. This is analogous to Corollary 2.10, using that $\mathbb{W}^+(B/R) = \mathbb{W}^+(B)$ under the given assumptions on δ; see the end of Section 1D. $\square$

Lemma 2.12. *Let $\mathscr{P}$ be a Dieudonné display over an admissible ring R with $p = 2$, and let $\mathscr{P}^+$ be its base change to $\mathscr{D}_R^+$. Then $\mathbb{D}(\mathscr{P}^+)$ is naturally isomorphic to the restriction of $\mathbb{D}(\mathscr{P})$ to $\mathrm{Cris}_{\mathrm{adm}}(R/\mathbb{Z}_2)$.*

Proof. For each $(\mathrm{Spec}\,A \to \mathrm{Spec}\,B, \delta)$ in $\mathrm{Cris}_{\mathrm{adm}}(R/\mathbb{Z}_2)$, the $\mathscr{D}_{B/A}^+$-window $\mathscr{P}_{B/A}^+$ is the base change of $\mathscr{P}_{B/A}$ by the frame homomorphism $\mathscr{D}_{B/A} \to \mathscr{D}_{B/A}^+$ by its uniqueness. The lemma follows easily. $\qquad\square$

Remark 2.13. Lemma 2.12 does not imply that the infinitesimal deformations of $\mathscr{P}$ and of $\mathscr{P}^+$ coincide: Let B be an admissible ring with $4B = 0$ and $2B \neq 0$ and let $R = B/2B$. The ideal $2B$ carries the canonical divided powers γ and the trivial divided powers δ. Corollary 2.10 applies to $(B \to R, \delta)$ but not to $(B \to R, \gamma)$, while Corollary 2.11 and Lemma 2.12 apply to $(B \to R, \gamma)$ but not to $(B \to R, \delta)$.

2G. *Passing to the limit.* The preceding considerations extend easily to the case of admissible topological rings with a countable base of topology. Let us begin with a standard lemma. For a ring A, let $\mathrm{V}(A)$ be the category of projective A-modules of finite type.

Lemma 2.14. *Let* $A = \varprojlim_{n \in \mathbb{N}} A_n$ *be an inverse limit of rings such that the transition maps* $\pi_n : A_n \to A_{n-1}$ *are surjective with* $\mathrm{Ker}(\pi_n) \subseteq \mathrm{Rad}(A_n)$. *Then the natural functor* $\rho : \mathrm{V}(A) \to \varprojlim_n \mathrm{V}(A_n)$ *is an equivalence.*

Proof. Since for $P \in \mathrm{V}(A)$ we have $P = \varprojlim_n (P \otimes_A A_n)$, the functor ρ is fully faithful. For a system of $P_n \in \mathrm{V}(A_n)$ with isomorphisms $P_n \otimes_A A_{n-1} \cong P_{n-1}$, we have to show that the A-module $P = \varprojlim_n P_n$ lies in $\mathrm{V}(A)$. Choose a surjective homomorphism $q_1 : A_1^r \to P_1$ and lift this to a compatible system of homomorphisms $q_n : A_n^r \to P_n$. All the q_n are surjective by Nakayama's Lemma. Let S_n be the set of linear sections of q_n. Since S_n carries a simply transitive action of $\mathrm{Hom}(P_n, \mathrm{Ker}(q_n))$, the reduction maps $S_n \to S_{n-1}$ are surjective. Thus the limit map $q : A^r \to P$ has a section, and we have $P \in \mathrm{V}(A)$. This proves that ρ is an equivalence. $\qquad\square$

For a ring A, let $\mathrm{BT}(A)$ be the category of p-divisible groups over A.

Lemma 2.15. *For an inverse limit* $A = \varprojlim_n A_n$ *as in* Lemma 2.14, *the natural functor* $\upsilon : \mathrm{BT}(A) \to \varprojlim_n \mathrm{BT}(A_n)$ *is an equivalence.*

Proof. See [Messing 1972, Chapter II, Lemma 4.16]. The functor ρ of Lemma 2.14 preserves tensor products, and a complex $P \to P' \to P'' \to 0$ in $\mathrm{V}(A)$ is exact if and only if its reduction to A_1 is exact. As in [Messing 1972, Chapter II, Lemma 4.16] it follows that υ is an equivalence. $\qquad\square$

For an admissible topological ring R, let $\mathscr{D}_R = \varprojlim_N \mathscr{D}_{R/N}$, where N runs through the open ideals of R contained in $\mathcal{N}_R$. As before, $\mathscr{D}_R$-windows are called Dieudonné displays over R.

Lemma 2.16. *If R is an admissible topological ring with a countable base of topology, then Dieudonné displays (or p-divisible groups) over R are equivalent to compatible systems of Dieudonné displays (or p-divisible groups) over R/N for each open ideal N contained in $\mathcal{N}_R$.*

Proof. One can write $R = \lim_{\underleftarrow{n \in \mathbb{N}}} R_n$ for a surjective system of admissible rings R_n with $R_{\mathrm{red}} = (R_n)_{\mathrm{red}}$ for each n. Then the case of p-divisible groups follows from Lemma 2.15, and the case of Dieudonné displays follows from Lemma 2.14 applied to R and to $\mathbb{W}(R) = \lim_{\underleftarrow{n}} \mathbb{W}(R_n)$; here the successive kernels are nilideals due to Lemma 1.5. See [Lau 2010, Lemma 2.1]. $\square$

3. From p-divisible groups to Dieudonné displays

In this section we define a functor from p-divisible groups over odd admissible rings to Dieudonné displays. In the nonodd case there is a v-stabilised version of this functor, which will serve as a first step towards the true functor in the next section. We begin with some preparation.

3A. *Finiteness over admissible rings.* We show that the categories of p-divisible groups or Dieudonné displays over an admissible ring R are the direct limit of the corresponding categories over the finitely generated $W(R_{\mathrm{red}})$-subalgebras of R, with fully faithful transition maps.

Proposition 3.1. *Every Dieudonné display over an admissible ring R is defined over a finitely generated $W(R_{\mathrm{red}})$-subalgebra of R. For an injective homomorphism of admissible rings $R \to S$ such that $R_{\mathrm{red}} \to S_{\mathrm{red}}$ is bijective, the base change of Dieudonné displays from R to S is fully faithful.*

Proof. For a ring A, let $V(A)$ be the category of projective A-modules of finite type. Since the ring $\mathbb{W}(R)$ is the filtered union of $\mathbb{W}(R')$, where R' runs through the finitely generated $W(R_{\mathrm{red}})$-subalgebras of R, the category $V(\mathbb{W}(R))$ is equivalent to the direct limit over R' of $V(\mathbb{W}(R'))$. Since a Dieudonné display over R can be given by $L, T \in V(\mathbb{W}(R))$ together with an f-linear automorphism Ψ of $L \oplus T$, the first assertion of the proposition follows. Similarly, every homomorphism of Dieudonné displays over R is defined over some finitely generated R'. Thus for the second assertion we may assume that $\mathcal{N}_S^r = 0$. Let $\bar{S} = S/\mathcal{N}_S^{r-1}$ and $\bar{R} = R/R \cap \mathcal{N}_S^{r-1}$. Let $R'' \subseteq S$ be the inverse image of $\bar{R} \subseteq \bar{S}$. By induction on r, the base change of Dieudonné displays from $\bar{R}$ to $\bar{S}$ is fully faithful. It follows that the base change from R'' to S is fully faithful as well. By Corollary 2.10, using trivial divided powers, Dieudonné displays over R or over R'' are equivalent to Dieudonné displays over $\bar{R}$ together with a lift of the Hodge filtration to R or to R'', respectively. Since $R \to R''$ is injective, it follows that the base change of Dieudonné displays from R to R'' is fully faithful. $\square$

For the case of p-divisible groups we first recall some standard facts.

Lemma 3.2. *Let $B \to A$ be a surjective ring homomorphism with kernel I such that $pI = 0$ and $x^p = 0$ for all $x \in I$. For an affine flat group scheme H over B, the kernel of $H(B) \to H(A)$ is annihilated by p.*

Proof. Let $B_0 = B/pB$ and $H_0 = H \otimes_B B_0$. The abelian group $B_0 \oplus I$ becomes a ring with multiplication $(a \oplus i)(a' \oplus i') = aa' \oplus (ai' + a'i + ii')$, and one can identify $B \times_A B$ with $B \times_{B_0} (B_0 \oplus I)$. Since the evaluation of affine schemes commutes with fibred products of rings, we obtain an isomorphism of abelian groups

$$\mathrm{Ker}(H(B) \to H(A)) \cong \mathrm{Ker}(H_0(B_0 \oplus I) \to H_0(B_0)).$$

The right-hand side lies in the kernel of the Frobenius F_{H_0} of H_0, which lies in $H_0[p]$ since $V_{H_0} \circ F_{H_0} = p$ by [SGA 1970, VII$_A$ 4.3]. This proves the lemma. $\square$

Lemma 3.3. *Let $B \to A$ be a surjective ring homomorphism with kernel I such that p is nilpotent in B and I is a nilideal. For a p-divisible group G over B, the homomorphism $G(B) \to G(A)$ is surjective.*

Proof. For a given $x \in G_n(A)$, since G_n is finitely presented there is a finitely generated ideal $I' \subseteq I$ such that x lifts to an element $x' \in G_n(B/I')$. Now we can use that G is formally smooth by [Messing 1972, Chapter II, Theorem 3.3.13]. $\square$

Lemma 3.4. *Let $B \to A$ be a surjective ring homomorphism whose kernel is bounded nilpotent and such that p is nilpotent in B. Then there is a number r such that for two p-divisible groups G and H over B, the reduction homomorphism $\mathrm{Hom}(G, H) \to \mathrm{Hom}(G_A, H_A)$ is injective with kernel annihilated by p^r.*

Proof. This is an easy consequence of Lemmas 3.2 and 3.3; see the proof of [Katz 1981, Lemma 1.1.3]. $\square$

Proposition 3.5. *Every p-divisible group over an admissible ring R is defined over a finitely generated $W(R_{\mathrm{red}})$-subalgebra of R. For an injective homomorphism of admissible rings $R \to S$ such that $R_{\mathrm{red}} \to S_{\mathrm{red}}$ is bijective, the base change of p-divisible groups from R to S is fully faithful.*

Proof. For a p-divisible group G over R, let $G_0 = G \otimes_R R_{\mathrm{red}}$. Using Lemma 3.4, we chose r such that for two p-divisible groups G and H over R, the cokernel of $\mathrm{Hom}(G, H) \to \mathrm{Hom}(G_0, H_0)$ is annihilated by p^r. Now let G be given, let G'' be a lift of G_0 to $W(R_{\mathrm{red}})$ and let $G' = G'' \otimes_{W(R_{\mathrm{red}})} R$. There are homomorphisms $\varphi : G' \to G$ and $\psi : G \to G'$ which each lift the multiplication $p^r : G_0 \to G_0$. Thus $\varphi\psi$ and $\psi\varphi$ are multiplication by p^{2r}. We obtain an isomorphism $G \cong G'/K_G$, where $K_G \subseteq G'$ is a finite locally free group scheme annihilated by p^{2r}; see Lemma 3.6 below. In particular K_G is finitely presented, and the first assertion of the proposition follows. To prove the second assertion, we consider two p-divisible groups G and H over R and a homomorphism $\varphi_0 : G_0 \to H_0$ over $R_{\mathrm{red}} = S_{\mathrm{red}}$. There is a unique lift of $p^r \varphi_0$ to a homomorphism $\psi : G \to H$, and there is a lift of φ_0 to R if ψ vanishes on $G[p^r]$. Since $R \to S$ is injective, this holds if and only if the scalar extension ψ_S vanishes on $G_S[p^r]$, which is equivalent to the existence of a lift of φ_0 to S. $\square$

Lemma 3.6. *Let $\varphi : G \to H$ and $\psi : H \to G$ be homomorphisms of p-divisible groups over a scheme S with $\varphi\psi = p^n$ and $\psi\varphi = p^n$. Then $\mathrm{Ker}(\varphi)$ and $\mathrm{Ker}(\psi)$ are finite locally free group schemes.*

Proof. Clearly $\mathrm{Ker}(\varphi)$ and $\mathrm{Ker}(\psi)$ are finite group schemes of finite presentation. Thus we may assume that $S = \mathrm{Spec}\, R$ for a local ring R with residue field k. Let $\mathrm{Ker}(\psi) = \mathrm{Spec}\, A$ and $G_n = \mathrm{Spec}\, B$. Choose elements $a_1, \ldots, a_{p^r} \in A$ which map to a k-basis of A_k, so they generate A as an R-module. We have a surjective homomorphism of fppf sheaves $\varphi : G_n \to \mathrm{Ker}(\psi)$. It follows that B_k is a locally free A_k-module of some rank p^s, thus a free A_k-module since A_k is finite. Choose $b_1, \ldots, b_{p^s} \in B$ which map to an A_k-basis of B_k. The elements $a_i b_j \in B$ map to a k-basis of B_k. Since B is a free R-module they form an R-basis of B. It follows that A is free over R with basis a_i. $\qquad\square$

3B. *Deformation rings.* Let $\Lambda \to K$ be a surjective ring homomorphism with finitely generated kernel $I \subseteq \Lambda$ such that Λ is I-adically complete. The ring K is not assumed to be a field. Let $\mathrm{Nil}_{\Lambda/K}$ be the category of Λ-algebras A together with a homomorphism of Λ-algebras $A \to K$ with nilpotent kernel. We consider covariant functors

$$F : \mathrm{Nil}_{\Lambda/K} \to (\mathrm{sets})$$

with the following properties (cf. [Schlessinger 1968]):

(3-1) The set $F(K)$ has precisely one element.

(3-2) For a surjective homomorphism $A_1 \to A$ in $\mathrm{Nil}_{\Lambda/K}$, the induced map $F(A_1) \to F(A)$ is surjective.

(3-3) For each pair of homomorphisms $A_1 \to A \leftarrow A_2$ in $\mathrm{Nil}_{\Lambda/K}$ such that one of them is surjective, the natural map $F(A_1 \times_A A_2) \to F(A_1) \times_{F(A)} F(A_2)$ is bijective. Then for each K-module N the set $F(K \oplus N)$ is naturally a K-module. In particular, $t_F = F(K[\varepsilon])$ is a K-module, which is called the tangent space of F.

(3-4) For each K-module N the natural homomorphism of K-modules $t_F \otimes_K N \to F(K \oplus N)$ is bijective.

(3-5) The K-module t_F is finitely presented.

The first three conditions imply that the functor $N \mapsto F(K \oplus N)$ preserves exact sequences of K-modules. Thus (3-4) is automatic if N is finitely presented. Moreover (3-1)–(3-4) imply that the K-module t_F is flat, so (3-5) implies that t_F is projective.

Proposition 3.7. *Assume that F satisfies* (3-1)–(3-5). *Then F is prorepresented by a complete Λ-algebra B. Let $\tilde{t}$ be a projective Λ-module of finite type which lifts t_F. Then B is isomorphic to the complete symmetric algebra $\Lambda[\![\tilde{t}^*]\!]$, where the $*$ means the dual. This is a power series ring over Λ if t_F is a free K-module.*

Proof. The K-module t_F is projective as explained above. Thus $\tilde{t}$ exists. Let $B = \Lambda[\![\tilde{t}^*]\!]$ and let $\bar{B} = K \oplus t_F^*$. We have an obvious projection $B \to \bar{B}$. Let $\bar{\xi} \in F(\bar{B}) = t_F \otimes t_F^* = \operatorname{End}(t_F)$ correspond to the identity of t_F and let $\xi \in F(B)$ be a lift of $\bar{\xi}$. We claim that the induced homomorphism of functors $\xi : B \to F$ is bijective. Note that the functor B satisfies (3-1)–(3-5). By induction it suffices to show that if $A \to \bar{A}$ is a surjection in $\operatorname{Nil}_{\Lambda/K}$ whose kernel N is a K-module of square zero and if $B(\bar{A}) \to F(\bar{A})$ is bijective, then $B(A) \to F(A)$ is bijective as well. We have a natural isomorphism $A \times_{\bar{A}} A \cong A \times_K (K \oplus N)$. It follows that the fibres of $B(A) \to B(\bar{A})$ and the fibres of $F(A) \to F(\bar{A})$ are principal homogeneous sets under the K-modules $B(K \oplus N)$ and $F(K \oplus N)$, respectively. The homomorphism $t_B \to t_F$ induced by ξ is bijective by construction, so $B(K \oplus N) \to F(K \oplus N)$ is bijective, and the proposition follows. $\square$

Corollary 3.8. *A homomorphism of functors which satisfy (3-1)–(3-5) is an isomorphism if and only if it induces an isomorphism on the tangent spaces.* $\square$

Remark 3.9. Let $\Lambda' \to K'$ be another pair as above and let $g : \Lambda' \to \Lambda$ be a ring homomorphism which induces a homomorphism $\bar{g} : K' \to K$. For given functors F on $\operatorname{Nil}_{\Lambda/K}$ and F' on $\operatorname{Nil}_{\Lambda'/K'}$, a homomorphism $h : F \to F'$ over g is a compatible system of maps

$$h(A) : F(A) \to F'(A \times_K K')$$

for A in $\operatorname{Nil}_{\Lambda/K}$; here $A \times_K K'$ is naturally an object of $\operatorname{Nil}_{\Lambda'/K'}$. If F and F' satisfy (3-1)–(3-5) and if B and B' are the complete algebras which prorepresent F and F', respectively, then h corresponds to a homomorphism $B' \to B$ compatible with g and $\bar{g}$. If $h(A)$ is bijective for all A, the induced homomorphism $B' \widehat{\otimes}_{\Lambda'} \Lambda \to B$ is an isomorphism.

Definition 3.10. Assume that p is nilpotent in $K = \Lambda/I$ as above. For a p-divisible group G over K let

$$\operatorname{Def}_G : \operatorname{Nil}_{\Lambda/K} \to (\text{sets})$$

be the deformation functor of G. This means that $\operatorname{Def}_G(A)$ is the set of isomorphism classes of p-divisible groups G' over A together with an isomorphism $G' \otimes_A K \cong G$. Let $t_G = \operatorname{Lie}(G^\vee) \otimes_K \operatorname{Lie}(G)$.

Proposition 3.11. *The functor Def_G is prorepresented by a complete Λ-algebra B. Explicitly, if $\tilde{t}$ is a projective Λ-module which lifts t_G, then B is isomorphic to the complete symmetric algebra $\Lambda[\![\tilde{t}^*]\!]$.*

We note that Lemma 2.15 gives a universal p-divisible group over B.

Proof. The functor Def_G satisfies (3-1)–(3-5) with tangent space t_G because for a surjective homomorphism $A' \to A$ in $\operatorname{Nil}_{\Lambda/K}$ whose kernel N is a K-module

of square zero and for $H \in \mathrm{Def}_G(A)$, the set of lifts of H to A' is a principally homogeneous set under the K-module $t_G \otimes_K N$ by [Messing 1972]. □

Remark 3.12. Let $g : \Lambda' \to \Lambda$ over $\bar{g} : K' \to K$ be as in Remark 3.9, such that p is nilpotent in K'. Let G over K be the base change of a p-divisible group G' over K'. For A in $\mathrm{Nil}_{\Lambda/K}$ we have a natural map

$$\mathrm{Def}_{G'}(A \times_K K') \to \mathrm{Def}_G(A).$$

This map is bijective, and its inverse is a homomorphism $\mathrm{Def}_G \to \mathrm{Def}_{G'}$ over g in the sense of Remark 3.9. If B and B' prorepresent Def_G and $\mathrm{Def}_{G'}$, respectively, we get an isomorphism $B' \widehat{\otimes}_{\Lambda'} \Lambda \cong B$.

Definition 3.13. Assume that $K = \Lambda/I$ is an admissible ring. For a Dieudonné display $\mathscr{P} = (P, Q, F, F_1)$ over K, we denote by

$$\mathrm{Def}_{\mathscr{P}} : \mathrm{Nil}_{\Lambda/K} \to \text{(sets)}$$

the deformation functor of $\mathscr{P}$. Let $t_{\mathscr{P}} = \mathrm{Hom}(Q/\mathbb{I}_K P, P/Q)$.

We are mainly interested in the case where K is perfect and $\Lambda = W(K)$. Then Dieudonné displays over K are displays because $\mathbb{W}(K) = W(K)$.

Proposition 3.14. *The functor $\mathrm{Def}_{\mathscr{P}}$ is prorepresented by a complete Λ-algebra B. Explicitly, if $\tilde{t}$ is a projective Λ-module which lifts $t_{\mathscr{P}}$, then B is isomorphic to the complete symmetric algebra $\Lambda[\![\tilde{t}^*]\!]$.*

We note that Lemma 2.16 gives a universal Dieudonné display over B.

Proof. The functor $\mathrm{Def}_{\mathscr{P}}$ satisfies (3-1)–(3-5) with tangent space $t_{\mathscr{P}}$ because for a surjective homomorphism $A' \to A$ in $\mathrm{Nil}_{\Lambda/K}$ whose kernel N is a K-module of square zero and for $\mathscr{P}' \in \mathrm{Def}_{\mathscr{P}}(A)$, the set of lifts of $\mathscr{P}'$ to A' is a principally homogeneous set under the K-module $t_{\mathscr{P}} \otimes_K N$ by Corollary 2.10. □

Remark 3.15. Let $g : \Lambda' \to \Lambda$ over $\bar{g} : K' \to K$ be as in Remark 3.9, such that K and K' are admissible rings. Assume that $\mathscr{P}$ is the base change of a Dieudonné display $\mathscr{P}'$ over K'. If B and B' represent $\mathrm{Def}_{\mathscr{P}}$ and $\mathrm{Def}_{\mathscr{P}'}$, respectively, then $B' \widehat{\otimes}_{\Lambda'} \Lambda \cong B$. This is analogous to Remark 3.12.

3C. *Crystals and frames.* Let $\mathcal{F} = (S, I, R, \sigma, \sigma_1)$ be a frame as in Section 2A, such that S and R are p-adically complete, S has no p-torsion, I carries divided powers, and $\sigma = p\sigma_1$ on I. Thus (S, σ) is a frame for each $R/p^n R$ in the sense of [Zink 2001b]. By a well-known construction, the crystalline Dieudonné functor allows us to associate to a p-divisible group over R an $\mathcal{F}$-window; this is explained in the proof of [Zink 2001b, Theorem 1.6] for the Dieudonné crystal of a nilpotent display, and in [Kisin 2006; 2009] for p-divisible groups.

The construction goes as follows. First, one can define a filtered F-V-module; here it is not necessary to assume that S has no p-torsion.

Construction 3.16. Let $\mathcal{F} = (S, I, R, \sigma, \sigma_1)$ be a frame such that S and R are p-adically complete, I is equipped with divided powers δ which are compatible with the canonical divided powers of p, and $\sigma = p\sigma_1$ on I. Let δ' be the divided powers on $I' = I + pS$ which extend δ and the canonical divided powers of p. We assume that σ preserves δ', which is automatic if S has no p-torsion. Then one can define a functor

$$\Phi^o : (p\text{-divisible groups over } R) \to (\text{filtered } F\text{-}V\text{-modules over } \mathcal{F}),$$

$$G \mapsto (P, Q, F^\sharp, V^\sharp)$$

as follows. Let $R_0 = R/pR$ and let σ_0 be its Frobenius endomorphism. For a given p-divisible group G over R put $P = \mathbb{D}(G)_{S/R} = \mathbb{D}(G_0)_{S/R_0}$, where $\mathbb{D}(G)$ is the *covariant*[2] Dieudonné crystal of G, and let Q be the kernel of the natural map $P \to \mathrm{Lie}(G)$. Since σ preserves δ', there is a natural isomorphism

$$P^{(\sigma)} \cong \mathbb{D}(\sigma_0^* G_0)_{S/R_0}.$$

Thus we can define $V^\sharp : P \to P^{(\sigma)}$ to be induced by the Frobenius $F : G_0 \to \sigma_0^* G_0$ and $F^\sharp : P^{(\sigma)} \to P$ to be induced by the Verschiebung $V : \sigma_0^* G_0 \to G_0$.

In the second step one associates F_1.

Proposition 3.17. *Let $\mathcal{F}$ be a frame as in the beginning of* Section 3C. *For a p-divisible group G over R let $\Phi^o(G) = (P, Q, F^\sharp, V^\sharp)$ be the filtered F-V-module over $\mathcal{F}$ given by* Construction 3.16. *There is a unique $F_1 : Q \to P$ such that (P, Q, F, F_1) is an $\mathcal{F}$-window, and it gives back $V^\sharp$ by the functor of* Lemma 2.3.

Proof. We have functors $(P, Q, F, F_1) \mapsto (P, Q, F^\sharp, V^\sharp) \mapsto (P, Q, F^\sharp)$, which are fully faithful; see Lemma 2.3. Thus we have to show that $F(Q)$ lies in pP so that $F_1 = p^{-1}F$ is well-defined, that $F_1(Q)$ generates P, and that the pair (P, Q) admits a normal decomposition. Since R and S are p-adically complete and since the kernel of $S/pS \to R/pR$ is a nilideal due to its divided powers, all projective R-modules of finite type lift to S. Thus a normal decomposition exists. The existence of F_1 and the surjectivity of its linearisation are proved in [Kisin 2006, Lemma A.2] if S is local with perfect residue field, but the proof can be easily adapted to the general case. To prove surjectivity, for each maximal ideal of S, which necessarily comes from a maximal ideal $\mathfrak{m}$ of R, we choose an embedding of $R/\mathfrak{m}$ into a perfect field k. There is a ring homomorphism $\alpha : S \to W(k)$ which lifts $R \to k$ such that $f\alpha = \alpha\sigma$; it can be constructed as $S \to W(S) \to W(k)$. Then α is a homomorphism of frames $\mathcal{F} \to \mathscr{W}_k$, and the assertion is reduced to the case of $\mathscr{W}_k$, which is classical. $\square$

---

[2] This differs from the notation of [Berthelot et al. 1982], where $\mathbb{D}(G)$ is contravariant. One can switch between the covariant and contravariant crystals by passing to the dual of G or of $\mathbb{D}(G)$, which amount to the same thing by the crystalline duality theorem [Berthelot et al. 1982, 5.3].

Remark 3.18. The surjectivity of F_1 in the proof of Proposition 3.17 can also be deduced from the crystalline duality theorem. Let $P = L \oplus T$ be a normal decomposition and let $\Psi : P \to P$ be given by F_1 on L and by F on T. We have to show that the linearisation $\Psi^{\sharp} : P^{(\sigma)} \to P$ is an isomorphism. Let (P', Q', F', F_1') be the quadruple associated to the Cartier dual $G^{\vee}$. The duality theorem gives a perfect pairing $P \times P' \to S$ such that $\langle F(x), F'(x') \rangle = p\sigma\langle x, x' \rangle$. It follows that $\langle F(x), F_1'(x') \rangle = \sigma\langle x, x' \rangle$ and $\langle F_1(x), F'(x') \rangle = \sigma\langle x, x' \rangle$ whenever this makes sense. The unique decomposition $P' = L' \oplus T'$ with $\langle L, L' \rangle = 0 = \langle T, T' \rangle$ is a normal decomposition of P', and the dual of the associated $\Psi'^{\sharp}$ is an inverse of $\Psi^{\sharp}$.

3D. *The Dieudonné display associated to a p-divisible group.* For an admissible ring R with $p \geq 3$, we consider the Dieudonné frame $\mathscr{D}_R$ defined in Lemma 2.5. The ring $\mathbb{W}(R)$ is p-adically complete by the remark preceding Proposition 1.14. By Lemma 1.17 the ideal $\mathbb{I}_R$ carries natural divided powers compatible with the canonical divided powers of p, and the induced divided powers on the kernel of $\mathbb{W}(R) \to R/pR$ are preserved by the Frobenius. Thus Construction 3.16 gives a functor

$$\Phi_R^o : (p\text{-divisible groups over } R) \to (\text{filtered } F\text{-}V\text{-modules over } \mathscr{D}_R)$$

which is compatible with base change in R.

Theorem 3.19. *For each admissible ring R with $p \geq 3$, there is a unique functor*

$$\Phi_R : (p\text{-divisible groups over } R) \to (\textit{Dieudonné displays over } R)$$

which is compatible with base change in R such that the filtered F-V-module over $\mathscr{D}_R$ associated to $\Phi_R(G)$ is equal to $\Phi_R^o(G)$. In particular there is a natural isomorphism $\mathrm{Lie}(G) \cong \mathrm{Lie}(\Phi_R(G))$.

Proof. Clearly $\Phi_R^o(G) = (P, Q, F^{\sharp}, V^{\sharp})$ is functorial in R and G. We have to show that there is a unique operator $F_1 : Q \to P$ which is functorial in R and G such that $\Phi_R(G) = (P, Q, F, F_1)$ is a Dieudonné display over R.

Let $K = R_{\mathrm{red}}$ and $\Lambda = W(K)$. Let $\overline{G} = G \otimes_R K$ and let B be the complete Λ-algebra which prorepresents the functor $\mathrm{Def}_{\overline{G}}$ on $\mathrm{Nil}_{\Lambda/K}$; see Proposition 3.11. Let $\mathscr{G}$ be the universal deformation of G over B. If I denotes the kernel of $B \to K$, we can define

$$\Phi_B^o(\mathscr{G}) = \varprojlim_n \Phi_{B/I^n}^o(\mathscr{G} \otimes_B B/I^n).$$

On the other hand, the ring $\mathbb{W}(B)$ is p-adically complete by Proposition 1.14. Therefore we can also define $\Phi_B^o(\mathscr{G})$ be a direct application of Construction 3.16, and this agrees with the limit definition. The ring $\mathbb{W}(B)$ has no p-torsion because

B has no p-torsion. Thus by Proposition 3.17 there is a unique operator F_1 which makes $\Phi^o_B(\mathscr{G})$ into a Dieudonné display $\Phi_B(\mathscr{G})$ over B.

By Proposition 3.5 there is a unique homomorphism $B \to R$ of augmented algebras such that $G = \mathscr{G} \otimes_B R$ as deformations of $\overline{G}$. Necessarily we define $\Phi_R(G)$ as the base change of $\Phi_B(\mathscr{G})$ under $B \to R$. It remains to show that $\Phi_R(G)$ is functorial in R and G.

Assume that G is the base change of a p-divisible group G' over R' under a homomorphism of admissible rings $R' \to R$. Let K', Λ', $\overline{G}'$, B', $\mathscr{G}'$ have the obvious meaning. We have a natural homomorphism of $W(K')$-algebras $B' \to B$ together with an isomorphism $\mathscr{G}' \otimes_{B'} B \cong \mathscr{G}$; see Remark 3.12. By the uniqueness of F_1 over B, we see that $\Phi_B(\mathscr{G})$ coincides with the base change of $\Phi_{B'}(\mathscr{G}')$. It follows that $\Phi_R(G)$ is the base change of $\Phi_{R'}(G')$.

Assume that $u : G \to G_1$ is a homomorphism of p-divisible groups over R. Let $\overline{G}_1$, B_1, $\mathscr{G}_1$ have the obvious meaning. We have to show that $\Phi^o_R(u)$ commutes with F_1. We may assume that u is an isomorphism because otherwise one can pass to the automorphism $\begin{pmatrix} 1 & 0 \\ u & 1 \end{pmatrix}$ of $G \oplus G_1$. This reasoning uses that the natural isomorphism $\Phi^o_R(G \oplus G_1) = \Phi^o_R(G) \oplus \Phi^o_R(G_1)$ preserves the operators F_1 defined on the three modules, which follows from the uniqueness of F_1 over the ring which prorepresents $\mathrm{Def}_{\overline{G}} \times \mathrm{Def}_{\overline{G}_1}$. An isomorphism $u : G \to G_1$ induces an isomorphism $\bar{u} : \overline{G} \cong \overline{G}_1$, which gives an isomorphism $B \cong B_1$ together with an isomorphism $\tilde{u} : \mathscr{G} \otimes_B B_1 \cong \mathscr{G}_1$ that lifts $\bar{u}$. By the uniqueness of F_1 over B_1 it follows that $\Phi^o_{B_1}(\tilde{u})$ preserves F_1. Since u is the base change of $\tilde{u}$ by the homomorphism $B_1 \to R$ defined by G_1, it follows that $\Phi^o_R(u)$ preserves F_1 as well. $\square$

In order to analyse the action of the functors Φ_R on infinitesimal deformations, we need the following extension of Theorem 3.19. Let $(R' \to R, \delta)$ be a divided power extension of admissible rings with $p \geq 3$ which is compatible with the canonical divided powers of p. Again, the ring $\mathbb{W}(R'/R)$ is p-adically complete, and $\mathbb{I}_{R'/R}$ carries natural divided powers compatible with the canonical divided powers of p such that f preserves their extension to the kernel of $\mathbb{W}(R'/R) \to R/pR$. Thus Construction 3.16 gives a functor

$$\Phi^o_{R'/R} : (p\text{-divisible groups over } R) \to (\text{filtered } F\text{-}V\text{-modules over } \mathscr{D}_{R'/R})$$

which is compatible with base change in the triple $(R' \to R, \delta)$.

Theorem 3.20. *Assume that $p \geq 3$. For each divided power extension of admissible rings $(R' \to R, \delta)$ compatible with the canonical divided powers of p, there is a unique functor*

$$\Phi_{R'/R} : (p\text{-divisible groups over } R) \to (\text{Dieudonné displays for } R'/R)$$

*which is compatible with base change in the triple $(R' \to R, \delta)$ such that the filtered
F-V-module over $\mathscr{D}_{R'/R}$ associated to $\Phi_{R'/R}(G)$ is equal to $\Phi^o_{R'/R}(G)$.*

Proof. For a given p-divisible group G over R we choose a lift to a p-divisible
group G' over R', which exists by [Illusie 1985, Théorème 4.4]. The Dieudonné
display $\Phi_{R'}(G')$ is well-defined by Theorem 3.19, and necessarily $\Phi_{R'/R}(G)$ is
defined as the base change of $\Phi_{R'}(G')$ by the frame homomorphism $\mathscr{D}_{R'} \to \mathscr{D}_{R'/R}$.
We have to show that the operator F_1 on $\Phi^o_{R'/R}(G)$ defined in this way does not
depend on the choice of G'. If this is proved it follows easily that $\Phi_{R'/R}(G)$
is functorial in G and in $(R' \to R, \delta)$; here, instead of arbitrary homomorphisms of
p-divisible groups, it suffices to treat isomorphisms.

Let $K, \Lambda, \overline{G}, B, \mathscr{G}$ be as in the proof of Theorem 3.19. We have an isomorphism
$B \cong \Lambda[\![t]\!]$ for a finitely generated projective Λ-module t. Let $C = B \widehat{\otimes}_\Lambda B$. The
automorphism $\tau = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ of $t \oplus t$ defines an isomorphism

$$C = \Lambda[\![t \oplus t]\!] \xrightarrow{\tau} \Lambda[\![t \oplus t]\!] = B[\![t_B]\!]$$

under which the multiplication homomorphism $\mu : C \to B$ corresponds to the
augmentation $B[\![t_B]\!] \to B$ defined by $t_B \mapsto 0$. Let I be the kernel of $B \to K$, let
S be the divided power envelope of the ideal $t_B B[\![t_B]\!] \subseteq B[\![t_B]\!]$, and let C' be the
I-adic completion of S. By Lemma 1.13, μ extends to a divided power extension
of admissible topological rings $\mu' : C' \to B$ which is topologically compatible with
the canonical divided powers of p.[3]

Assume that G_1 and G_2 are two lifts of G to p-divisible groups over R'. Let $\mathscr{G}_1$
and $\mathscr{G}_2$ be the p-divisible groups over C which are the base change of $\mathscr{G}$ by the
two natural homomorphisms $B \to C$. By Proposition 3.5 there are well-defined
homomorphisms $\bar{\alpha} : B \to R$ and $\alpha : C \to R'$ such that $G = \mathscr{G} \otimes_{B,\bar{\alpha}} R$ and
$G_i = \mathscr{G}_i \otimes_{C,\alpha} R'$ as deformations of $\overline{G}$. We have the commutative diagram of rings
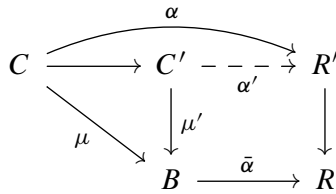
where α' is constructed as follows. There is a unique homomorphism $\alpha'' : S \to R'$
which extends α and which commutes with the divided powers on the kernel of
$S \to B$ and of $R' \to R$. Each of the two homomorphisms $B \to C \to R'$ factors
over B/I^n for some n. Thus α'' induces a homomorphism $S/I^n S \to R'$, which

---

[3]The construction of C' seems to depend on choosing one of the two natural maps $B \to C$, but
actually it is independent of the choice as the I-adic topologies defined on S by these two maps
coincide.

gives the required α'. We obtain the following commutative diagram of frames, where ι is given by $C \to C'$, and ι' is given by the identity of R':

$$
\begin{array}{ccc}
\mathscr{D}_C & \xrightarrow{\ \iota\ } & \mathscr{D}_{C'/B} \\
\alpha \downarrow & & \downarrow \alpha' \\
\mathscr{D}_{R'} & \xrightarrow{\ \iota'\ } & \mathscr{D}_{R'/R}
\end{array}
$$

We have to show that the isomorphism of filtered F-V-modules over $\mathscr{D}_{R'/R}$

$$
\iota'_*(\Phi^o_{R'}(G_1)) \cong \Phi^o_{R'/R}(G) \cong \iota'_*(\Phi^o_{R'}(G_2)) \tag{3-6}
$$

commutes with the operator F_1 defined on the outer terms by the functor $\Phi_{R'}$. The construction of Φ^o can be extended to topological divided power extensions of admissible topological rings by passing to the projective limit. Then (3-6) arises by α'_* from the natural isomorphism of filtered F-V-modules over $\mathscr{D}_{C'/B}$

$$
\iota_*(\Phi^o_C(\mathscr{G}_1)) \cong \Phi^o_{C'/B}(\mathscr{G}) \cong \iota_*(\Phi^o_C(\mathscr{G}_2)). \tag{3-7}
$$

Since α_* preserves F_1 it suffices to show that (3-7) commutes with the operators F_1 defined on the outer terms by the functor Φ_C. This follows from the relation $pF_1 = F$ because $\mathbb{W}(C'/B)$ has no p-torsion by Lemma 1.13. $\square$

Corollary 3.21. *Assume that $p \geq 3$. For a p-divisible group G over an admissible ring R with associated Dieudonné display $\mathscr{P} = \Phi_R(G)$, there is a natural isomorphism of crystals on* $\mathrm{Cris}_{\mathrm{adm}}(R/\mathbb{Z}_p)$

$$
\mathbb{D}(G) \cong \mathbb{D}(\mathscr{P})
$$

which is compatible with the natural isomorphism $\mathrm{Lie}(G) \cong \mathrm{Lie}(\mathscr{P})$.

The category $\mathrm{Cris}_{\mathrm{adm}}$ and the crystal $\mathbb{D}(\mathscr{P})$ were defined in Section 2F.

Proof. Let $(R' \to R, \gamma)$ be a divided power extension of admissible rings with $p \geq 3$ compatible with the canonical divided powers of p. The Dieudonné display $\Phi_{R'/R}(G)$ given by Theorem 3.20 is the unique lift of $\mathscr{P}$ under the crystalline frame homomorphism $\mathscr{D}_{R'/R} \to \mathscr{D}_R$. By the construction of the underlying filtered F-V-module $\Phi^o_{R'/R}(G)$ and by the definition of the crystal $\mathbb{K}(\mathscr{P})$ in Section 2F we obtain a natural isomorphism of $\mathbb{W}(R'/R)$-modules

$$
\mathbb{D}(G)_{\mathbb{W}(R')/R} \cong \mathbb{K}(\mathscr{P})_{R'/R}.
$$

The tensor product with the projection $\mathbb{W}(R'/R) \to R'$, which is a homomorphism of divided power extensions of R, gives a natural isomorphism of R'-modules $\mathbb{D}(G)_{R'/R} \cong \mathbb{D}(\mathscr{P})_{R'/R}$ which is compatible with the natural isomorphism $\mathrm{Lie}(G) \cong \mathrm{Lie}(\mathscr{P})$. $\square$

Now Theorem B for odd primes can be deduced quite formally:

Corollary 3.22. *Assume that $p \geq 3$. For a p-divisible group G over an admissible ring R with associated Dieudonné display $\mathscr{P} = \Phi_R(G)$, there is a natural isomorphism of crystals on $\mathrm{Cris}_{\mathrm{adm}}(R)$*

$$\mathbb{D}(G) \cong \mathbb{D}(\mathscr{P})$$

which is compatible with the natural isomorphism $\mathrm{Lie}(G) \cong \mathrm{Lie}(\mathscr{P})$.

Here the covariant Dieudonné crystal $\mathbb{D}(G)$ can be defined for divided power extensions that are not necessarily compatible with the canonical divided powers of p by [Mazur and Messing 1974, Chapter II §9]; see also [Berthelot et al. 1982, §1.4].

Proof. Let $\mathbb{D}'(G) = \mathbb{D}(\Phi_R(G))$. Consider a divided power extension $R' \to R$ of admissible rings which need not be compatible with the canonical divided powers of p. We claim that for two lifts G_1 and G_2 of G to R' the following diagram of natural isomorphisms commutes:

$$\begin{array}{ccccc}
\mathbb{D}(G_2)_{R'/R'} & \xleftarrow{\sim} & \mathbb{D}(G)_{R'/R} & \xrightarrow{\sim} & \mathbb{D}(G_1)_{R'/R'} \qquad (3\text{-}8) \\
\downarrow{\scriptstyle\sim} & & & & \downarrow{\scriptstyle\sim} \\
\mathbb{D}'(G_2)_{R'/R'} & \xrightarrow{\sim} & \mathbb{D}'(G)_{R'/R} & \xleftarrow{\sim} & \mathbb{D}'(G_1)_{R'/R'}
\end{array}$$

This gives a well-defined isomorphism $\alpha(G) : \mathbb{D}(G)_{R'/R} \cong \mathbb{D}'(G)_{R'/R}$. It is easy to see that $\alpha(G)$ is compatible with the natural isomorphism $\mathrm{Lie}(G) \cong \mathrm{Lie}(\mathscr{P})$, that $\alpha(G)$ is functorial in the divided power extension $R' \to R$ and that $\alpha(G \oplus H) = \alpha(G) \oplus \alpha(H)$. In order to show that α is functorial in G it suffices to consider isomorphisms. So let $u : G \to H$ be an isomorphism of p-divisible groups over R. We can choose lifts G_1 of G and H_1 of H to R' such that u extends to $\tilde{u} : G_1 \to H_1$. Then the following diagram shows that α commutes with u:

$$\begin{array}{ccccccc}
\mathbb{D}(G)_{R'/R} & \xrightarrow{\sim} & \mathbb{D}(G_1)_{R'/R'} & \xrightarrow{\sim} & \mathbb{D}'(G_1)_{R'/R'} & \xrightarrow{\sim} & \mathbb{D}'(G)_{R'/R} \\
{\scriptstyle\mathbb{D}(u)}\downarrow & & {\scriptstyle\mathbb{D}(\tilde{u})}\downarrow & & \downarrow{\scriptstyle\mathbb{D}'(\tilde{u})} & & \downarrow{\scriptstyle\mathbb{D}'(u)} \\
\mathbb{D}(H)_{R'/R} & \xrightarrow{\sim} & \mathbb{D}(H_1)_{R'/R'} & \xrightarrow{\sim} & \mathbb{D}'(H_1)_{R'/R'} & \xrightarrow{\sim} & \mathbb{D}'(H)_{R'/R}
\end{array}$$

It remains to show that (3-8) commutes. Let $K, \Lambda, \overline{G}, B$ be as in the proof of Theorem 3.19. Let $C = B \,\widehat{\otimes}_\Lambda\, B$ and C' be as in the proof of Theorem 3.20 so that the multiplication homomorphism $\mu : C \to B$ extends to a topological divided power extension $\mu' : C' \to B$ of admissible topological rings without p-torsion which is topologically compatible with the canonical divided powers of p. We have homomorphisms $B \to R$ defined by G and $C \to R'$ defined by (G_1, G_2),

which extend to a homomorphism of divided power extensions from $(C' \to B)$ to $(R' \to R)$. Thus (3-8) is the base change of a similar diagram for $(C' \to B)$, which commutes by Corollary 3.21. □

3E. A v-stabilised variant. Let R be an admissible ring with $p = 2$. The v-stabilised Zink ring $\mathbb{W}^+(R)$ considered in Section 1D and in Section 2E is 2-adically complete, and its ideal $\mathbb{I}_R^+$ carries natural divided powers which are compatible with the canonical divided powers of 2. The proof of Theorem 3.19 with $\mathbb{W}^+$ in place of $\mathbb{W}$ shows the following:

Proposition 3.23. *For each admissible ring R with $p = 2$ there is a unique functor*

$$\Phi_R^+ : (2\text{-divisible groups over } R) \to (\mathscr{D}_R^+\text{-windows})$$

which is compatible with base change in R such that the filtered F-V-module over $\mathscr{D}_R^+$ associated to $\Phi_R^+(G)$ is given by Construction 3.16. □

Corollary 3.24. *For each admissible ring R with $p = 2$ and $2R = 0$ there is a unique functor*

$$\Phi_R : (2\text{-divisible groups over } R) \to (\text{Dieudonné displays over } R)$$

which is compatible with base change in R such that the filtered F-V-module over $\mathscr{D}_R$ associated to $\Phi_R(G)$ is given by Construction 3.16.

Proof. Proposition 3.23 gives the functors Φ_R since $\mathscr{D}_R^+ = \mathscr{D}_R$ when $2R = 0$. The uniqueness follows as in the proof of Theorem 3.19, using $B/2B$ instead of B. □

Let $(R' \to R, \delta)$ be a divided power extension of admissible rings with $p = 2$ which is compatible with the canonical divided powers of 2. The ring $\mathbb{W}^+(R'/R)$ is 2-adically complete, and its ideal $\mathbb{I}_{R'/R}^+$ carries natural divided powers compatible with the canonical divided powers of 2. The proof of Theorem 3.20 with $\mathbb{W}^+$ in place of $\mathbb{W}$ gives the following:

Proposition 3.25. *For each divided power extension of admissible rings $(R' \to R, \delta)$ with $p = 2$ such that δ is compatible with the canonical divided powers of 2 there is a unique functor*

$$\Phi_{R'/R}^+ : (2\text{-divisible groups over } R) \to (\mathscr{D}_{R'/R}^+\text{-windows})$$

which is functorial in the triple $(R' \to R, \delta)$ such that the filtered F-V-module over $\mathscr{D}_{R'/R}^+$ associated to $\Phi_{R'/R}^+(G)$ is given by Construction 3.16. □

The proof of Corollary 3.21 then shows the following:

Corollary 3.26. *Assume that $p = 2$. For a 2-divisible group G over an admissible ring R with associated v-stabilised Dieudonné display $\mathscr{P}^+ = \Phi_R^+(G)$ there is a natural isomorphism of crystals on* $\mathrm{Cris}_{\mathrm{adm}}(R/\mathbb{Z}_2)$

$$\mathbb{D}(G) \cong \mathbb{D}^+(\mathscr{P}^+)$$

which is compatible with the natural isomorphism $\mathrm{Lie}(G) \cong \mathrm{Lie}(\mathscr{P}^+)$. $\qquad\square$

There is no analogue of Corollary 3.22 for Φ_R^+ because $\mathbb{D}^+(\mathscr{P}^+)$ is only a crystal on $\mathrm{Cris}_{\mathrm{adm}}(R/\mathbb{Z}_2)$ and not on $\mathrm{Cris}_{\mathrm{adm}}(R)$, but see Corollary 4.10.

4. From 2-divisible groups to Dieudonné displays

In this section we construct a functor Φ_R from p-divisible groups over an admissible ring R with $p = 2$ to Dieudonné displays. When $2R = 0$, this has been done in the previous section, and the extension to all R is unique, as will be shown in the end of this section. The construction relies on the following definition of divided powers on the ideal $\mathbb{I}_R \subseteq \mathbb{W}(R)$ when $4R = 0$.

4A. *Divided powers on Zink rings.* We note that for a $\mathbb{Z}_{(2)}$-algebra B and an ideal $\mathfrak{b} \subseteq B$, divided powers on $\mathfrak{b}$ are equivalent to a map $\gamma : \mathfrak{b} \to \mathfrak{b}$ such that

$$\gamma(xy) = x^2\gamma(y) \text{ for } x \in B \text{ and } y \in \mathfrak{b}, \tag{4-1}$$
$$\gamma(x + y) = \gamma(x) + xy + \gamma(y) \text{ for } x, y \in \mathfrak{b}. \tag{4-2}$$

Here (4-1) and (4-2) also give $2\gamma(x) = x^2$ for $x \in \mathfrak{b}$, since we can calculate

$$4\gamma(x) = \gamma(2x) = \gamma(x + x) = 2\gamma(x) + x^2.$$

For an admissible ring R with $p = 2$, the canonical divided powers on the ideal $I_R \subseteq W(R)$ defined by $\gamma(v(a)) = v(a^2)$ induce divided powers on $\mathbb{I}_R \subseteq \mathbb{W}(R)$ only if $2R = 0$; see Section 1G. Using $\mathbbitv$ instead of v we get a little further.

Proposition 4.1. *For an admissible ring R with $p = 2$ we consider the map*

$$\gamma : \mathbb{I}_R \to \mathbb{I}_R, \qquad \gamma(\mathbbitv(a)) = \mathbbitv(a^2).$$

If $4R = 0$, then γ defines divided powers on $\mathbb{I}_R$ which are compatible with the canonical divided powers of 2, and the corresponding extension of γ to $\mathbb{I}_R + 2\mathbb{W}(R)$ is stable under the Frobenius f of $\mathbb{W}(R)$.

If $8R = 0$, let $U \subseteq \mathbb{W}(R)$ be the set of all Witt vectors of the form $v([x]) = (0, x, 0, \ldots)$ with $x \in 4R$. This is an ideal. Let $\widetilde{S} = \mathbb{W}(R)/U$. Then γ induces divided powers on the ideal $\mathbb{I}_R/U$ of $\widetilde{S}$, which can naturally be extended to divided powers on $\mathbb{I}_R/U + 2\widetilde{S}$ that commute with the endomorphism σ on $\widetilde{S}$ induced by f, and the extended divided powers stabilise the ideal $2\widetilde{S}$.

Proof. We will only consider the case $8R = 0$ and show that the extended divided powers satisfy $\gamma(2) = 2 - [4]$. Then the case $4R = 0$ follows.

Since $4R$ is an ideal of square zero, we have $\hat{W}(4R) = (4R)^{(\mathbb{N})}$ as $W(R)$-modules, where $W(R)$ acts on the i-th component of the right-hand side by the i-th Witt polynomial w_i, and f annihilates $\hat{W}(4R)$. Thus U is an ideal of $\mathbb{W}(R)$, and f induces $\sigma : \tilde{S} \to \tilde{S}$. Let us show that γ factors over a map $\mathbb{I}_R/U \to \mathbb{I}_R$. Indeed, for $a \in \mathbb{W}(R)$ and $x \in 4R$ we have

$$\gamma(\mathsf{v}(a) + v([x])) = \gamma(\mathsf{v}(a) + \mathsf{v}([x])) = \mathsf{v}((a + [x])^2) = \mathsf{v}(a^2) = \gamma(\mathsf{v}(a));$$

here $v([x]) = \mathsf{v}([x])$ because u_0 maps to 1 in $W(\mathbb{F}_2)$ and thus $u_0[x] = [x]$. Let us verify axiom (4-1) for the map $\gamma : \mathbb{I}_R \to \mathbb{I}_R$. For $a, b \in \mathbb{W}(R)$ we have

$$\gamma(a\mathsf{v}(b)) = \gamma(\mathsf{v}(f(a)b)) = \mathsf{v}(f(a^2)b^2) = a^2\mathsf{v}(b^2) = a^2\gamma(\mathsf{v}(b)).$$

Consider now axiom (4-2). For $a, b \in \mathbb{W}(R)$ we calculate

$$\gamma(\mathsf{v}(a) + \mathsf{v}(b)) = \mathsf{v}((a + b)^2) = \gamma(\mathsf{v}(a)) + \mathsf{v}(2ab) + \gamma(\mathsf{v}(b))$$

so that $\mathsf{v}(2ab)$ has to be related with $\mathsf{v}(a)\mathsf{v}(b)$, which is

$$\mathsf{v}(a)\mathsf{v}(b) = v(u_0 a)v(u_0 b) = v(2u_0^2 ab) = \mathsf{v}(2u_0 ab).$$

Since $2u_0 = 2 - [4]$, we get

$$\mathsf{v}(2ab) - \mathsf{v}(a)\mathsf{v}(b) = \mathsf{v}([4]ab) = v([4a_0 b_0]) \in U.$$

Thus (4-2) holds for $\gamma : \mathbb{I}_R/U \to \mathbb{I}_R/U$, and γ defines divided powers on this ideal. We want to extend these to divided powers on the ideal

$$\tilde{I} = \mathbb{I}_R/U + 2\tilde{S} = \mathbb{I}_R/U + \hat{W}(2R)/U.$$

Let

$$\mathfrak{b} = \{(2a_0, 4a_1, 0, \dots) \mid a_i \in R\} \subseteq \hat{W}(2R).$$

This is an ideal of $\mathbb{W}(R)$ with $\mathbb{I}_R \cap \mathfrak{b} = U$, and we have

$$\tilde{I} = \mathbb{I}_R/U \oplus \mathfrak{b}/U.$$

Thus the extension of γ to $\tilde{I}$ corresponds to giving arbitrary divided powers on $\mathfrak{b}/U \cong 2R$. We take $\gamma([2a]) = [-2a^2]$ for $a \in R$. Using $\mathsf{v}(1) = 2 - [2]$ we obtain

$$\gamma(2) = \gamma([2] + \mathsf{v}(1)) = [-2] + \mathsf{v}(1) = [-2] + 2 - [2] = 2 - [4]$$

in $\tilde{S}$, as announced. Let us show that $\gamma\sigma = \sigma\gamma$ on $\tilde{I}$: for $a \in \mathbb{W}(R)$, we have

$$\gamma\sigma(\mathsf{v}(a)) = \gamma((2 - [4])a) = \gamma(2 - [4])a^2 = \gamma(2)a^2$$
$$= (2 - [4])a^2 = \sigma(\mathsf{v}(a^2)) = \sigma\gamma(\mathsf{v}(a)).$$

Finally we have $[4] = 2[2]$ in $\mathfrak{b}/U$, which implies that $\gamma(2) \in 2\tilde{S}$. This finishes the proof of Proposition 4.1.   □

Remark 4.2. The proof shows that the extension of γ is uniquely determined by the condition that it commutes with σ. By choosing $\gamma([2a]) = [2a^2]$, we get an extension with $\gamma(2) = 2$ but which does not commute with σ.

Let R be an admissible ring with $4R = 0$. Proposition 4.1 implies that its Dieudonné frame $\mathcal{D}_R$ satisfies the hypotheses of Construction 3.16 so that we obtain a functor

$$\Phi_R^o : (\text{2-divisible groups over } R) \to (\text{filtered } F\text{-}V\text{-modules over } \mathcal{D}_R).$$

However, we cannot argue as in Theorem 3.19 in order to get a $\mathcal{D}_R$-window, because the divided powers on $\mathbb{I}_R$ do not exist for universal deformation rings, and thus Proposition 3.17 cannot be applied directly. The following modification will be sufficient for our purpose.

4B. *A frame lift.* Assume we are given a strict frame homomorphism

$$\mathcal{F}' = (S', I', R', \sigma', \sigma_1') \xrightarrow{\pi} \mathcal{F} = (S, I, R, \sigma, \sigma_1)$$

such that both $\pi : S' \to S$ and $I' \to I$ are surjective, and an ideal $U \subseteq \text{Ker}(\pi)$ which is stable under σ'. Let

$$\tilde{S} = S'/U, \quad J = \text{Ker}(\pi)/U, \quad \tilde{I} = \text{Ker}(\tilde{S} \to R);$$

thus $S = \tilde{S}/J$ and $R = \tilde{S}/\tilde{I}$. Let $\tilde{\sigma} : \tilde{S} \to \tilde{S}$ be the homomorphism induced by σ', let $\theta' \in S'$ be the element defined by the relation $\sigma' = \theta'\sigma_1'$ on I', and let $\tilde{\theta} \in \tilde{S}$ and $\theta \in S$ be its images. We assume that $\mathcal{F}$ satisfies the conditions of Construction 3.16; i.e., S and R are p-adically complete, I carries divided powers compatible with the canonical divided powers of p and with σ, and $\theta = p$. Then Construction 3.16 gives a functor

$$\Phi^o : (p\text{-divisible groups over } R) \to (\text{filtered } F\text{-}V\text{-modules over } \mathcal{F}).$$

We also assume that the following conditions are satisfied.

(4-3) We have $\tilde{\sigma}(J) = 0$ and $J = \{x \in \tilde{S} \mid px = 0\}$.

(4-4) We have $\tilde{\theta} = p\tilde{u}$ for a unit $\tilde{u} \in \tilde{S}$.

(4-5) The ideal $\tilde{I} + p\tilde{S}$ is equipped with divided powers which lift the given divided powers on $I + pS$, which commute with $\tilde{\sigma}$, and which stabilise the ideal $p\tilde{S}$.

(4-6) There is an ideal $\mathfrak{a} \subseteq S$ with $\sigma(\mathfrak{a}) \subseteq \mathfrak{a} \subseteq \text{Rad } S$ such that the ring $S/\mathfrak{a}$ has no p-torsion.

If S has no p-torsion one can take $\mathcal{F}' = \mathcal{F}$ and all axioms are clear. The following extends Proposition 3.17. Note that the prime p is arbitrary here.

Proposition 4.3. *In this situation there is a well-defined functor*

$$\Phi : (p\text{-divisible groups over } R) \to (\mathcal{F}\text{-windows})$$

such that for $\Phi^o(G) = (P, Q, F^\sharp, V^\sharp)$ *the filtered F-V-module associated to $\Phi(G)$ is equal to* $(P, Q, F^\sharp, uV^\sharp)$, *where* $u \in S$ *is the image of* $\tilde{u} \in \widetilde{S}$.

Proof. Conditions (4-3) and (4-4) imply that multiplication by $\tilde{\theta}$ on $\widetilde{S}$ induces an injective map $\tilde{\theta} : S \to \widetilde{S}$ with image $\tilde{\theta}\widetilde{S} = p\widetilde{S}$. Moreover $\tilde{\sigma}$ induces a homomorphism $\tilde{\sigma} : S \to \widetilde{S}$ that lifts σ. The relation $\theta'\sigma_1' = \sigma'$ on I' gives $\tilde{\theta} \circ \sigma_1 = \tilde{\sigma}$ as maps $I \to \widetilde{S}$.

Let G be a p-divisible group over R and let $\Phi^o(G)$ be as above, i.e.,

$$P = \mathbb{D}(G)_{S/R} = \mathbb{D}(G_{R_0})_{S/R_0}$$

with $R_0 = R/pR$, the submodule $Q \subseteq P$ is the kernel of $P \to \mathrm{Lie}\, G$, and $F^\sharp$ and $V^\sharp$ are induced by the Verschiebung and Frobenius of G_{R_0}. The proof of [Kisin 2006, Lemma A.2] shows that $F(Q) \subseteq pP$. Let us recall the argument: for $S_0 = S/pS$, the kernel of $S_0 \to R_0$ is a nilideal because it carries divided powers. By [Illusie 1985, Théorème 4.4] there is a lift G_{S_0} of G_{R_0} to S_0, and we have $P = \mathbb{D}(G_{S_0})_{S/S_0}$. Let $Q_1 = \mathrm{Ker}(P \to \mathrm{Lie}\, G_{S_0})$. Then $Q \subseteq Q_1 + IP$, and the image of F applied to both summands lies in pP.

Since $pJ = 0$ and S is p-adically complete, so is $\widetilde{S}$. By (4-5) we can define

$$\widetilde{P} = \mathbb{D}(G)_{\widetilde{S}/R} = \mathbb{D}(G_{R_0})_{\widetilde{S}/R_0}.$$

Here we use the (dual of the) Dieudonné crystal of [Mazur and Messing 1974, Chapter II §9], which is defined for divided power extensions that are not necessarily compatible with the canonical divided powers of p; see also [Berthelot et al. 1982, §1.4]. Let $\widetilde{Q} \subseteq \widetilde{P}$ be the kernel of $\widetilde{P} \to \mathrm{Lie}\, G$; this is the inverse image of Q under the projection $\widetilde{P} \to \widetilde{P}/J\widetilde{P} = P$. Again, the Verschiebung and Frobenius of G_{R_0} induce $\widetilde{S}$-linear maps $\widetilde{F}^\sharp : \widetilde{P}^{(\tilde{\sigma})} \to \widetilde{P}$ and $\widetilde{V}^\sharp : \widetilde{P} \to \widetilde{P}^{(\tilde{\sigma})}$. Since the divided powers stabilise the ideal $p\widetilde{S}$, the argument of [Kisin 2006, Lemma A.2] again shows that $\widetilde{F}(\widetilde{Q}) \subseteq p\widetilde{P} = \tilde{\theta}\widetilde{P}$, where $\widetilde{F} : \widetilde{P} \to \widetilde{P}$ is the $\tilde{\sigma}$-linear map corresponding to $\widetilde{F}^\sharp$. Since $\tilde{\sigma}$ annihilates J, the map $\widetilde{F}$ induces a map $\widetilde{F} : P \to \widetilde{P}$ which lifts F. Let $F_1 : Q \to P$ be the composition

$$F_1 : Q \xrightarrow{\widetilde{F}} \tilde{\theta}\widetilde{P} \underset{\sim}{\xleftarrow{\tilde{\theta}}} P,$$

i.e., $F_1 = \tilde{\theta}^{-1} \circ \widetilde{F}$. We define $\Phi(G) = (P, Q, F, F_1)$. In order that this is an $\mathcal{F}$-window we have to verify that

(i) for $x \in P$ and $a \in I$ we have $F_1(ax) = \sigma_1(a)F(x)$;

(ii) the image $F_1(Q)$ generates P.

Moreover, $uV^\sharp$ is the operator associated to $\Phi(G)$ as we have claimed if and only if

(iii) for $x \in Q$ we have $uV^\sharp(F_1(x)) = 1 \otimes x$ in $P^{(\sigma)}$.

The equation in (i) is equivalent to $\widetilde{F}(ax) = \tilde{\theta}(\sigma_1(a)F(x))$. Since $\widetilde{F}(ax) = \tilde{\sigma}(a)\widetilde{F}(x)$ and $\tilde{\sigma} = \tilde{\theta} \circ \sigma_1$, this is clear. To prove (ii), it suffices to show that for each maximal ideal $\mathfrak{m}$ of R and perfect field extension $R/\mathfrak{m} \subseteq k$ the vector space $\widetilde{P} \otimes_{\tilde{S}} k$ is generated by $F_1(Q)$. Using (4-6) we get a sequence of σ-equivariant maps $S \to \widetilde{S}/\mathfrak{a} \to W(\widetilde{S}/\mathfrak{a}) \to W(k)$; the second arrow exists uniquely since $\widetilde{S}/\mathfrak{a}$ has no p-torsion and carries a Frobenius lift induced by $\tilde{\sigma}$; see [Bourbaki 1983, IX, §1.2, Proposition 2] and the explanation following [Zink 2001b, Theorem 4]. By functoriality we are reduced to the case where $\mathcal{F}' = \mathcal{F} = \mathscr{W}_k$, which is classical. Assertion (iii) is equivalent to $\tilde{u}\widetilde{V}^\sharp(\widetilde{F}(x)) = \tilde{\theta}(1 \otimes x)$ for $x \in Q$, which holds since $\widetilde{V}^\sharp(\widetilde{F}(x)) = p(1 \otimes x)$ in $\widetilde{P}^{(\tilde{\sigma})}$ for all $x \in \widetilde{P}$. □

Now we construct an example for Proposition 4.3 with $p = 2$. Let A_{red} be a perfect ring of characteristic 2 and let $A = W(A_{\mathrm{red}})[\![t]\!]$, where t is a finitely generated projective $W(A_{\mathrm{red}})$-module. Let $\mathfrak{m} = (2, t)$ be the kernel of $A \to A_{\mathrm{red}}$. We write $A_n = A/2^n A$ and $A_{n+} = A/2^n \mathfrak{m}$. Only the rings

$$A_{2+} \to A_{1+} \to A_1$$

will play a role. We consider the frames $\mathcal{F}' = \mathscr{D}_{A_{2+}} \to \mathcal{F} = \mathscr{D}_{A_{1+}}$, i.e.,

$$S = \mathbb{W}(A_{1+}), \quad I = \mathbb{I}_{A_{1+}}, \quad R = A_{1+},$$
$$S' = \mathbb{W}(A_{2+}), \quad I' = \mathbb{I}_{A_{2+}}, \quad R' = A_{2+}.$$

Then $\theta' = 2 - [4]$ in S' and thus $\theta = 2$ in S. Let $U \subseteq S'$ be the ideal of all Witt vectors $v([x])$ with $x \in 4A_{2+}$, and let $\widetilde{S} = S'/U$. As above, we write

$$J = \mathrm{Ker}(\widetilde{S} \to S) = \widehat{W}(2\mathfrak{m}/4\mathfrak{m})/U$$

and

$$\widetilde{I} = \mathrm{Ker}(\widetilde{S} \to R) = (I' + \widehat{W}(2\mathfrak{m}/4\mathfrak{m}))/U.$$

Proposition 4.4. *These data satisfy the axioms* (4-3)–(4-6).

Proof. The divided powers required in (4-5) are given by Proposition 4.1. Since $2\mathfrak{m}/4\mathfrak{m} \subseteq A_{2+}$ is an ideal of square zero, we have

$$J' := \widehat{W}(2\mathfrak{m}/4\mathfrak{m}) = (2\mathfrak{m}/4\mathfrak{m})^{(\mathbb{N})}$$

as $W(A_{2+})$-modules, where $W(A_{2+})$ acts on the i-th component of the right-hand side via the i-th Witt polynomial. We have $\sigma'(J') = 2J' = 0$ and $J = J'/U$. Thus $\tilde{\sigma} : \widetilde{S} \to \widetilde{S}$ is defined and vanishes on J, and $2J = 0$.

Lemma 4.5. *Multiplication by 2 induces an isomorphism of groups*

$$\hat{W}(2A_{1+}) \xrightarrow{\sim} \hat{W}(4A_{2+})/U.$$

Proof. The divided Witt polynomials for the canonical divided powers of 2 give an isomorphism $\mathrm{Log} : W(2A) \cong 2A^{\mathbb{N}}$. The composition

$$W(2A) \xrightarrow{\mathrm{Log}} 2A^{\mathbb{N}} \longrightarrow (2A/4A)^{\mathbb{N}}$$

is given by $(2a_0, 2a_1, \dots) \mapsto 2[a_0, a_0^2 + a_1, a_1^2 + a_2, \dots]$, while the composition

$$W(4A) \xrightarrow{\mathrm{Log}} 4A^{\mathbb{N}} \longrightarrow (4A/8A)^{\mathbb{N}}$$

is simply $(4a_0, 4a_1, \dots) \mapsto 4[a_0, a_1, \dots]$. It follows that the homomorphism of the lemma is isomorphic to the homomorphism

$$A_{\mathrm{red}}^{(\mathbb{N})} \to A_{\mathrm{red}}^{(\mathbb{N})}, \quad (a_0, a_1, \dots) \mapsto (a_0, a_1^2 + a_2, a_2^2 + a_3, \dots).$$

Since A_{red} is perfect this map is bijective. $\qquad\square$

Let us continue with the proof of Proposition 4.4. To verify (4-3), let $x \in \tilde{S}$ with $2x = 0$. Since $\mathbb{W}(A_1)$ has no 2-torsion we have $x \in \hat{W}(2A_{2+})/U$. Lemma 4.5 implies that $x \in J$, and (4-3) is proved. Let $u = 1 - [2]$ in $\mathbb{W}(A_{2+})$, which is a unit. By the proof of Lemma 4.5 we have $2u = 2 - (4, 4, 0, \dots) \equiv 2 - [4] = \theta'$ modulo U, which proves (4-4). In (4-6) we can take $\mathfrak{a} = \hat{W}(2A_{1+})$. $\qquad\square$

4C. *The Dieudonné display associated to a 2-divisible group.* Let $p = 2$ and let $u = 1 - [2]$ in $\mathbb{W}(\mathbb{Z}_2)$. We begin to construct the functor Φ_R in an initial case. Recall that Φ_R^o was defined in the end of Section 4A when $4R = 0$.

Proposition 4.6. *For each admissible ring R with $p = 2$ and $2\mathcal{N}_R = 0$ there is a functor*

$$\Phi_R : (\text{2-divisible groups over } R) \to (\mathscr{D}_R\text{-windows})$$

compatible with base change in R such that for $\Phi_R^o(G) = (P, Q, F^\sharp, V^\sharp)$ the filtered F-V-module associated to $\Phi_R(G)$ is equal to $(P, Q, F^\sharp, uV^\sharp)$.

Proof. This is similar to the proof of Theorem 3.19. Propositions 4.3 and 4.4 give the desired system of functors Φ_R for topological admissible rings R of the type $R = A_{1+}$ as above. For a p-divisible group G over an admissible ring R as in the proposition, let $\Lambda = W(R_{\mathrm{red}})$ and $\bar{G} = G \otimes_R R_{\mathrm{red}}$. Let A be the Λ-algebra that prorepresents the functor $\mathrm{Def}_{\bar{G}}$ on $\mathrm{Nil}_{\Lambda/K}$ (this A was denoted by B in Section 3), let $\mathscr{G}$ over A be the universal deformation, and let $\mathscr{G}_{1+} = \mathscr{G} \otimes_A A_{1+}$. The unique homomorphism of Λ-algebras $A \to R$ with $G = \mathscr{G} \otimes_A R$ as deformations of $\bar{G}$ factors over a homomorphism $A_{1+} \to R$, and we define $\Phi_R(G)$ as the base change of $\Phi_{A_{1+}}(\mathscr{G}_{1+})$ under this map. We have to show that the operator F_1 attached to

$\Phi_R^o(G)$ in this way is functorial in G and in R. This is analogous to the proof of Theorem 3.19, using that F_1 is functorial with respect to homomorphisms of rings of the type A_{1+}. □

For an admissible ring R, let $i : \mathscr{D}_R \to \mathscr{D}_R^+$ be the natural homomorphism.

Proposition 4.7. *Let R be an admissible ring with $p = 2$ and $2\mathcal{N}_R = 0$. For each 2-divisible group G over R there is a natural isomorphism of $\mathscr{D}_R^+$-windows*

$$\Phi_R^+(G) \cong i_* \Phi_R(G).$$

The functor Φ_R^+ was defined in Proposition 3.23.

Proof. Let us write $\Psi_R(G) = i_* \Phi_R(G)$ so that we have two functors

$$\Phi_R^+, \Psi_R : (\text{2-divisible groups over } R) \to (\mathscr{D}_R^+\text{-windows}).$$

When $2R = 0$, thus $\mathscr{D}_R = \mathscr{D}_R^+$, these functors coincide by the uniqueness assertion of Corollary 3.24. The rest is quite formal. Let $R_1 = R/2R$. For a p-divisible group G over R, let $G_1 = G \otimes_R R_1$ and $\overline{G} = G \otimes_R R_{\mathrm{red}}$. The canonical divided powers of 2 make $R \to R_1$ into a divided power extension. By Corollaries 2.11 and 3.26, the $\mathscr{D}_R^+$-windows $\Phi_R^+(G)$ and $\Psi_R(G)$ correspond to two lifts of the Hodge filtration of G to $\mathbb{D}(G_1)_{R/R_1}$. Their difference is measured by a homomorphism of R_1-modules

$$h_G' : V(G_1) \to \mathrm{Lie}(G_1) \otimes_R 2R,$$

where $V(G)$ is the kernel of $\mathbb{D}(G)_R \to \mathrm{Lie}(G)$. We have to show that h_G' is zero for all G. Since h_G' is functorial in R we may assume that $R = A_{1+}/\mathfrak{m}^n$ for some $n \geq 2$, where A is the universal deformation ring of $\overline{G}$. Then $2R$ is a free R_{red}-module of rank one, so h_G' corresponds to an element

$$h_G \in \mathrm{Hom}(V(\overline{G}), \mathrm{Lie}(\overline{G})).$$

Now an injective homomorphism $R_{\mathrm{red}} \to R_{\mathrm{red}}'$ gives an injective homomorphism of the associated rings $R \to R'$, while a product decomposition $R_{\mathrm{red}} = \prod R_{i,\mathrm{red}}$ gives $R = \prod R_i$. Since R_{red} embeds into the product of its localisations at minimal prime ideals, we may assume that $k := R_{\mathrm{red}}$ is a field. There is a deformation $\overline{G}'$ of $\overline{G}$ over $R_{\mathrm{red}}' := k[[x]]^{\mathrm{per}}$ with ordinary generic fibre. Let A' be its universal deformation ring and let G' over $R' = A_{1+}'/\mathfrak{m}^n$ be given by the universal deformation. By functoriality it suffices to show that $h_{G'} = 0$. Again we can pass to the field of fractions $k((x))^{\mathrm{per}}$. Thus we are left to show that $h_G = 0$ if G is ordinary over $R = A_{1+}/\mathfrak{m}^n$, where $k = R_{\mathrm{red}}$ is a perfect field. There is a deformation G'' of G_k over $R'' := W_2(k)$ which decomposes into the direct sum of its étale and multiplicative part. Let $R \to R''$ be the unique homomorphism such that $G'' = G \otimes_R R''$ as deformations of G_k. Since this does not change h_G we may

replace G by G''. Since Ψ_R and Φ_R^+ both preserve direct sums we may assume that G is étale or of multiplicative type. Then h_G vanishes since $\mathrm{Hom}(V(\bar{G}), \mathrm{Lie}(\bar{G}))$ is zero. $\qquad\square$

Lemma 4.8. *Let R be an admissible ring with $p = 2$ and let $R_{1+} = R/2\mathcal{N}_R$. The commutative diagram of frames*

$$
\begin{array}{ccc}
\mathscr{D}_R & \xrightarrow{\ i\ } & \mathscr{D}_R^+ \\
\downarrow & & \downarrow \\
\mathscr{D}_{R_{1+}} & \xrightarrow{\ i\ } & \mathscr{D}_{R_{1+}}^+
\end{array}
$$

is Cartesian on each component of the frames, and the associated diagram of window categories is 2-Cartesian.

Proof. The vertical arrows are surjective, and the horizontal arrows are injective with equal cokernel by Lemma 1.10 and its proof. Thus the diagram of frames is Cartesian on each component. For a ring A, let $V(A)$ be the category of projective A-modules of finite type. The functor

$$
V(\mathbb{W}(R)) \to V(\mathbb{W}^+(R)) \times_{V(\mathbb{W}^+(R_{1+}))} V(\mathbb{W}(R_{1+}))
$$

is fully faithful since the diagram is Cartesian, and it is essentially surjective since $V(\mathbb{W}(R)) \to V(\mathbb{W}(R_{1+}))$ and $V(\mathbb{W}^+(R)) \to V(\mathbb{W}^+(R_{1+}))$ are bijective on isomorphism classes and surjective on automorphism groups. It follows easily that the diagram of window categories is 2-Cartesian. $\qquad\square$

Theorem 4.9. *For each admissible ring R with $p = 2$ there is a functor*

$$
\Phi_R : (\text{2-divisible groups over } R) \to (\mathscr{D}_R\text{-windows})
$$

compatible with base change in R such that Φ_R is given by Proposition 4.6 when $2\mathcal{N}_R = 0$, and such that there is a natural isomorphism of $\mathscr{D}_R^+$-windows

$$
\Phi_R^+(G) \cong i_* \Phi_R(G).
$$

Proof. This is clear from Propositions 3.23 and 4.7 and Lemma 4.8. $\qquad\square$

Corollary 4.10. *Let $p = 2$. For each 2-divisible group G over an admissible ring R with associated Dieudonné display $\mathscr{P} = \Phi_R(G)$, there is a natural isomorphism of crystals on $\mathrm{Cris}_{\mathrm{adm}}(R)$*

$$
\mathbb{D}(G) \cong \mathbb{D}(\mathscr{P})
$$

which is compatible with the natural isomorphism $\mathrm{Lie}(G) \cong \mathrm{Lie}(\mathscr{P})$.

Proof. We have a natural isomorphism of crystals on $\text{Cris}_{\text{adm}}(R/\mathbb{Z}_2)$

$$\mathbb{D}(G) \cong \mathbb{D}(\Phi_R^+(G)) \cong \mathbb{D}(\Phi_R(G))$$

by Corollary 3.26, Theorem 4.9, and Lemma 2.12. The isomorphism of crystals on $\text{Cris}_{\text{adm}}(R)$ follows as in the proof of Corollary 3.22. $\square$

4D. *Uniqueness of the functor Φ_R.*

Proposition 4.11. *Assume that for each admissible ring R with $p = 2$ we have a functor*

$$\Phi_R' : (2\text{-divisible groups over } R) \to (\mathscr{D}_R\text{-windows})$$

compatible with base change in R such that $\Phi_R' = \Phi_R$ when $2R = 0$. Then there is a natural isomorphism $\Phi_R' \cong \Phi_R$ which is functorial in R and equal to the identity when $2R = 0$.

Proof. We first show that $\Phi_R' \cong \Phi_R$ when $4R = 0$. Let $R_1 = R/2R$. For a p-divisible group G over R, let $G_1 = G \otimes_R R_1$ and let

$$\mathscr{P}_1 = \Phi_{R_1}(G_1) = (P, Q, F, F_1)$$

be its Dieudonné display. If we take the trivial divided powers on the ideal $2R$, Corollary 2.10 implies that the difference between $\Phi_R(G)$ and $\Phi_R'(G)$ as lifts of $\mathscr{P}_1$ is measured by a homomorphism

$$h_G' : Q/\mathbb{I}_{R_1}P \to P/Q \otimes_{R_1} 2R.$$

Let $V(G) = \text{Lie}(G^\vee)^\vee$. By Corollary 3.24 and by the construction of $\Phi_R^o(G)$ we can view h_G' as a homomorphism

$$h_G : V(G_1) \to \text{Lie}(G_1) \otimes_{R_1} 2R.$$

We want to show that $h_G = 0$. We may assume that $R = A/(\mathfrak{m}^n + 4A)$, where A is the universal deformation ring of $G \otimes_R R_{\text{red}}$ and $\mathfrak{m}$ is the kernel of $A \to R_{\text{red}}$. As in the proof of Proposition 4.7, one reduces to the case where $k = R_{\text{red}}$ is a field and G is ordinary. Assume that G is an extension $0 \longrightarrow \mu_{p^\infty} \longrightarrow G \longrightarrow \mathbb{Q}_p/\mathbb{Z}_p \longrightarrow 0$. Then $V(G) = V(\mathbb{Q}_p/\mathbb{Z}_p) = R$ and $\text{Lie}(G) = \text{Lie}(\mu_{p^\infty}) = R$, so that $h_G \in 2R$. Thus $G \mapsto h_G$ defines a map $g : \underline{\text{Ext}}^1(\mathbb{Q}_p/\mathbb{Z}_p, \mu_{p^\infty}) \to \mathbb{G}_a$ of functors on the category of local Artin rings with residue field k and annihilated by 4. It is easy to see that g is additive. Here $\underline{\text{Ext}}^1(\ ) = \mu_{p^\infty}$, and it follows that $g = 0$. This implies easily that $h_G' = 0$ when G is ordinary. Thus $\Phi_R \cong \Phi_R'$ when $4R = 0$. If for some $n \geq 1$ we know that $\Phi_R \cong \Phi_R'$ when $2^n R = 0$, the same reasoning shows that $\Phi_R \cong \Phi_R'$ when $2^{n+1}R = 0$, and the proposition follows. $\square$

5. Equivalence of categories

Let R be an admissible ring. Dieudonné displays over R_{red} are displays, and they are equivalent to Dieudonné modules over R_{red} by Lemma 2.4. Under this equivalence, the functor $\Phi_{R_{\mathrm{red}}}$ corresponds to $\Phi^o_{R_{\mathrm{red}}}$.

Proposition 5.1. *For an admissible ring R the following diagram of categories is 2-Cartesian:*

$$
\begin{array}{ccc}
(\textit{p-divisible groups over } R) & \xrightarrow{\ \Phi_R\ } & (\textit{Dieudonné displays over } R) \\
\big\downarrow & & \big\downarrow \\
(\textit{p-divisible groups over } R_{\mathrm{red}}) & \xrightarrow{\ \Phi_{R_{\mathrm{red}}}\ } & (\textit{Dieudonné modules over } R_{\mathrm{red}})
\end{array}
$$

Proof. The categories of p-divisible groups and Dieudonné displays over R are the direct limit of the corresponding categories over all finitely generated $W(R_{\mathrm{red}})$-algebras contained in R; see Section 3A. Thus we may assume that the ideal $\mathcal{N}_R$ is nilpotent. If $\mathfrak{a} \subseteq R$ is an ideal equipped with nilpotent divided powers and if the proposition holds for $R/\mathfrak{a}$, then it holds for R. This follows from the comparison of crystals in Corollaries 3.21 and 4.10, since lifts from $R/\mathfrak{a}$ to R of p-divisible groups and of Dieudonné displays are both classified by lifts of the Hodge filtration by [Messing 1972] and by Corollary 2.10. When $\mathfrak{a}^2 = 0$, we can take the trivial divided powers on $\mathfrak{a}$. The result follows by induction on the order of nilpotence of $\mathcal{N}_R$. □

Remark 5.2. Since p-divisible groups and Dieudonné displays over a perfect ring K have universal deformation rings which are twisted power series rings over $\Lambda = W(K)$, in order to prove Proposition 5.1 the case $R = K[\varepsilon]$ is sufficient. In particular, for $p = 2$ this means that as soon as the functors Φ_R defined in Corollary 3.24 when $2R = 0$ are known to exist for all R, Proposition 5.1 is automatic. This reasoning does not apply to the functors Φ_R^+ (which also extend the functors Φ_R for $2R = 0$ to all R but which are not an equivalence in general) because the deformation functors of v-stabilised Dieudonné displays are not prorepresentable.

We have the following result of Gabber, which is classical when R_{red} is a field. It is also proved in [Lau 2013, Corollary 6.5].

Theorem 5.3. *The functor $\Phi_{R_{\mathrm{red}}}$ is an equivalence of categories.* □

Corollary 5.4. *For every admissible ring R the functor Φ_R is an equivalence of exact categories.*

Proof. By Theorem 5.3 and Proposition 5.1, the functor Φ_R is an equivalence of categories. A short sequence $0 \longrightarrow A \longrightarrow B \longrightarrow C \longrightarrow 0$ of p-divisible groups or of Dieudonné displays over R is exact if and only if all its scalar extensions to

perfect fields are exact. Thus Φ_R and its inverse preserve exact sequences, since this holds over perfect fields. □

This proves Theorem A. Using Lemmas 2.15 and 2.16, we also get:

Corollary 5.5. *For every admissible topological ring R with a countable base of topology, p-divisible groups over R are equivalent to Dieudonné displays over R.* □

Finally we note the following consequence of the crystalline duality theorem. The duality of windows is recalled in the end of Section 2A.

Corollary 5.6. *Let G be a p-divisible group over an admissible ring R and let $G^\vee$ be its Cartier dual. There is a natural isomorphism*

$$\Phi_R(G^\vee) \cong \Phi_R(G)^t.$$

Proof. Assume first that p is odd. The crystalline duality theorem [Berthelot et al. 1982, 5.3] gives an isomorphism of filtered F-V-modules $\Phi_R^o(G^\vee)^t \cong \Phi_R^o(G)$. Since the functor from windows to filtered F-V-modules preserves duality, the uniqueness part of Theorem 3.19 implies that this isomorphism preserves F_1, i.e., it is an isomorphism of Dieudonné displays $\Phi_R(G^\vee)^t \cong \Phi_R(G)$. For $p = 2$, using the uniqueness part of Corollary 3.24 we similarly get an isomorphism of Dieudonné displays $\Phi_R(G^\vee)^t \cong \Phi_R(G)$ when $2R = 0$. Then Proposition 4.11 gives such an isomorphism for all R. □

6. Breuil–Kisin modules

We recall the main construction of [Lau 2010] without restriction on p. Let R be a complete regular local ring with maximal ideal $\mathfrak{m}_R$ and with perfect residue field k of characteristic p. Choose a representation $R = \mathfrak{S}/E\mathfrak{S}$ with

$$\mathfrak{S} = W(k)[\![x_1, \ldots, x_r]\!]$$

such that E is a power series with constant term p. Let $J \subset \mathfrak{S}$ be the ideal generated by $x_1, \ldots, x_r$. Choose a ring endomorphism $\sigma : \mathfrak{S} \to \mathfrak{S}$ which lifts the Frobenius of $\mathfrak{S}/p\mathfrak{S}$ such that $\sigma(J) \subseteq J$. Let $\sigma_1 : E\mathfrak{S} \to \mathfrak{S}$ be defined by $\sigma_1(Ex) = \sigma(x)$ for $x \in \mathfrak{S}$. These data define a frame

$$\mathscr{B} = (\mathfrak{S}, E\mathfrak{S}, R, \sigma, \sigma_1).$$

For each positive integer a, let $R_a = R/\mathfrak{m}_R^a$ and $\mathfrak{S}_a = \mathfrak{S}/J^a$. We have frames

$$\mathscr{B}_a = (\mathfrak{S}_a, E\mathfrak{S}_a, R_a, \sigma, \sigma_1),$$

where σ and σ_1 are induced by the corresponding operators of $\mathscr{B}$.

The frames $\mathscr{B}$ and $\mathscr{B}_a$ are related with the Witt and Dieudonné frames of R and of R_a as follows. Let $\delta : \mathfrak{S} \to W(\mathfrak{S})$ be the unique lift of the identity of $\mathfrak{S}$

such that $f\delta = \delta\sigma$, or equivalently $w_n\delta = \sigma^n$ for $n \geq 0$; see [Bourbaki 1983, IX, §1.2, Proposition 2] and the explanation following [Zink 2001b, Theorem 4]. The composition of δ with the projection $W(\mathfrak{S}) \to W(R)$ is a ring homomorphism

$$\varkappa : \mathfrak{S} \to W(R)$$

which lifts the projection $\mathfrak{S} \to R$ such that $f\varkappa = \varkappa\sigma$. The same construction gives compatible homomorphisms

$$\varkappa_a : \mathfrak{S}_a \to W(R_a)$$

for $a \geq 1$, which induce $\varkappa$ in the projective limit. Since the element $\varkappa(E)$ maps to zero in R it lies in the image of $v : W(R) \to W(R)$. Let

$$u = v^{-1}(\varkappa(E)) = f_1(\varkappa(E)).$$

We will denote the image of u in $W(R_a)$ also by u.

Lemma 6.1. *The element $u \in W(R)$ is a unit. The homomorphisms $\varkappa$ and $\varkappa_a$ are u-homomorphisms of frames $\varkappa : \mathscr{B} \to \mathscr{W}_R$ and $\varkappa_a : \mathscr{B}_a \to \mathscr{W}_{R_a}$.*

Proof. See [Lau 2010, Proposition 6.1]. Since $W(R) \to W(k)$ is a local homomorphism, in order to show that u is a unit we can work with $\varkappa_1$, i.e., consider the case where $R = k$ and $\mathfrak{S} = W(k)$. Then $E = p$ and $u = 1$. In order that $\varkappa$ and $\varkappa_a$ are u-homomorphisms of frames we need that $f_1\varkappa = u \cdot \varkappa\sigma_1$. For $x \in \mathfrak{S}$ we calculate $f_1(\varkappa(Ex)) = f_1(\varkappa(E)\varkappa(x)) = f_1(\varkappa(E)) \cdot f(\varkappa(x)) = u \cdot \varkappa(\sigma(x)) = u \cdot \varkappa(\sigma_1(Ex))$, as required. $\square$

Let $\bar{\sigma}$ be the semilinear endomorphism of the free $W(k)$-module J/J^2 induced by σ. Since σ induces the Frobenius modulo p, $\bar{\sigma}$ is divisible by p.

Proposition 6.2. *The following conditions are equivalent*:

(i) *The image of $\varkappa : \mathfrak{S} \to W(R)$ lies in $\mathbb{W}(R)$.*

(ii) *The image of $\delta : \mathfrak{S} \to W(\mathfrak{S})$ lies in $\mathbb{W}(\mathfrak{S})$.*

(iii) *The endomorphism $p^{-1}\bar{\sigma}$ of J/J^2 is nilpotent modulo p.*

Remark 6.3. In the special case $\sigma(x_i) = x_i^p$ the conditions of Proposition 6.2 hold. This is easy to see directly: we have $\delta(x_i) = [x_i]$, which gives (i) and (ii), moreover (iii) holds since $\bar{\sigma}$ is zero.

Proof of Proposition 6.2. For odd p the equivalence between (i) and (iii) is [Lau 2010, Proposition 9.1]; its proof shows that (i) $\implies$ (iii) $\implies$ (ii) $\implies$ (i). The proof also applies for $p = 2$ if [Lau 2010, Lemma 9.2] is replaced by Lemma 6.4. $\square$

Lemma 6.4. *For $x \in \mathfrak{S}$ let $\tau(x) = (\sigma(x) - x^p)/p$. Let $\mathfrak{m}$ be the maximal ideal of $\mathfrak{S}$. For $n \geq 0$, the map τ preserves $\mathfrak{m}^n J$ and induces a σ-linear endomorphism $\mathrm{gr}_n(\tau)$ of the k-module $\mathrm{gr}_n(J) = \mathfrak{m}^n J / \mathfrak{m}^{n+1} J$. The endomorphism $\mathrm{gr}_0(\tau)$ is equal to $p^{-1}\bar{\sigma}$ modulo p. For $n \geq 1$, there is a surjective k-linear map*

$$\pi_n : \mathrm{gr}_n(J) \to \mathrm{gr}_0(J)$$

such that $\mathrm{gr}_0(\tau)\pi_n = \pi_n \mathrm{gr}_n(\tau)$ and such that $\mathrm{gr}_n(\tau)$ vanishes on $\mathrm{Ker}(\pi_n)$. In particular, $p^{-1}\bar{\sigma}$ is nilpotent modulo p if and only if $\mathrm{gr}_0(\tau)$ is nilpotent, which implies that $\mathrm{gr}_n(\tau)$ is nilpotent for each n.

Proof. We have $\sigma(J) \subseteq J^p + pJ \subseteq \mathfrak{m}J$ and thus $\sigma(\mathfrak{m}^n J) \subseteq \mathfrak{m}^{n+1} J$. It follows that $p\tau(\mathfrak{m}^n J) \subseteq p\mathfrak{S} \cap \mathfrak{m}^{n+1} J = p\mathfrak{m}^n J$ and $\tau(\mathfrak{m}^n J) \subseteq \mathfrak{m}^n J$. For $x, y \in \mathfrak{m}^n J$ the element $\tau(x+y) - \tau(x) - \tau(y)$ is a multiple of xy and thus lies in $\mathfrak{m}^{2n+1} J$. Hence τ induces an additive endomorphism $\mathrm{gr}_n(\tau)$ of $\mathrm{gr}_n(J)$. It is σ-linear because for $a \in \mathfrak{S}$ and $x \in \mathfrak{m}^n J$ the element $\tau(ax) - \sigma(a)\tau(x) = \tau(a)x^p$ lies in $\mathfrak{m}^{pn} J^p \subseteq \mathfrak{m}^{n+1} J$. Let us write $\sigma(x_i) = x_i^p + py_i$ with $y_i \in J$. We have $\tau(x_i) = y_i$ and $p^{-1}\bar{\sigma}(x_i) \equiv y_i$ modulo J^2. Thus $\mathrm{gr}_0(\tau)$ coincides with $p^{-1}\bar{\sigma}$ modulo p.

For each $n \geq 0$, a basis of $\mathrm{gr}_n(J)$ is given by all elements $p^b \underline{x}^{\underline{c}}$ with $\underline{c} \in \mathbb{N}^r$ and $1 \leq |\underline{c}| \leq n+1$ and $b + |\underline{c}| = n+1$. Let $n \geq 1$, and define π_n to be the k-linear map with $\pi_n(p^n x_i) = x_i$ and $\pi_n(p^b \underline{x}^{\underline{c}}) = 0$ if $|\underline{c}| > 1$. Then $\mathrm{gr}_n(\tau)$ vanishes on $\mathrm{Ker}(\pi_n)$ because $\sigma(J) \subseteq \mathfrak{m}J$, thus $\sigma(J^2) \subseteq \mathfrak{m}^2 J^2$, and because for $x \in \mathfrak{m}^n J$ we have $x^p \in \mathfrak{m}^{n+2} J$. The relation $\mathrm{gr}_0(\tau)\pi_n = \pi_n \mathrm{gr}_n(\tau)$ holds since $\tau(p^n x_i) \equiv p^{n-1} x_i^p + p^n y_i$ modulo $\mathfrak{m}^{n+1}(J)$. The last assertion of the lemma is immediate. $\square$

Lemma 6.5. *If the equivalent conditions of* Proposition 6.2 *hold, then $\varkappa$ and $\varkappa_a$ are u-homomorphisms of frames*

$$\varkappa : \mathscr{B} \to \mathscr{D}_R, \quad \varkappa_a : \mathscr{B}_a \to \mathscr{D}_{R_a},$$

where the unit $\mathsf{u} \in \mathbb{W}(R)$ is given by

$$\mathsf{u} = \mathsf{v}^{-1}(\varkappa(E)) = \mathbb{f}_1(\varkappa(E)).$$

In $W(R)$ we have $\mathsf{u} = u$ if p is odd and $\mathsf{u} = (v^{-1}(2 - [2]))^{-1}u$ if $p = 2$.

Proof. The proof of Lemma 6.1 with f_1 replaced by $\mathbb{f}_1$ shows that u is a unit of $\mathbb{W}(R)$ and that $\varkappa$ and $\varkappa_a$ are u-homomorphisms of frames as indicated. The relation between u and u follows from the fact that $\mathscr{D}_R \to \mathscr{W}_R$ is a u_0-homomorphism, where $u_0 = 1$ if p is odd and $v(u_0) = 2 - [2]$ if $p = 2$. $\square$

Theorem 6.6. *If the equivalent conditions of* Proposition 6.2 *hold, the frame homomorphisms $\varkappa : \mathscr{B} \to \mathscr{D}_R$ and $\varkappa_a : \mathscr{B}_a \to \mathscr{D}_{R_a}$ are crystalline.*

Proof. The proof for odd p in [Lau 2010, Theorem 9.3] works almost literally for $p = 2$ as well. Let us recall the essential parts of the argument. Fix an integer

$a \geq 1$. One can define a factorisation of the projection $\mathscr{B}_{a+1} \to \mathscr{B}_a$ into strict frame homomorphisms

$$\mathscr{B}_{a+1} \xrightarrow{\iota} \tilde{\mathscr{B}}_{a+1} \xrightarrow{\pi} \mathscr{B}_a \qquad (6\text{-}1)$$

such that $\mathscr{B}_{a+1} = (\mathfrak{S}_{a+1}, \tilde{I}, R_a, \sigma, \tilde{\sigma}_1)$. This determines $\tilde{I}$ and $\tilde{\sigma}_1$ uniquely as follows. Let $\bar{J}^a = J^a / J^{a+1}$. We have $\tilde{I} = E\mathfrak{S}_{a+1} + \bar{J}^a$ and $E\mathfrak{S}_{a+1} \cap \bar{J}^a = p\bar{J}^a$. The endomorphism $\bar{\sigma}$ of $\bar{J}^a$ induced by σ is divisible by p^a, and the operator $\tilde{\sigma}_1 : \tilde{I} \to \mathfrak{S}_{a+1}$ is the unique extension of σ_1 such that $\tilde{\sigma}_1(x) = p^{-1}\bar{\sigma}(x)$ for $x \in \bar{J}^a$. On the other hand, we consider the factorisation

$$\mathscr{D}_{R_{a+1}} \xrightarrow{\iota'} \mathscr{D}_{R_{a+1}}/R_a \xrightarrow{\pi'} \mathscr{D}_{R_a} \qquad (6\text{-}2)$$

with respect to the trivial divided powers on the kernel $\mathfrak{m}_R^a / \mathfrak{m}_R^{a+1}$. Then $\varkappa_{a+1}$ is a $\mathfrak{u}$-homomorphism of frames $\tilde{\mathscr{B}}_{a+1} \to \mathscr{D}_{R_{a+1}}/R_a$. Indeed, the only condition to be verified is that for $x \in \bar{J}^a$ we have

$$\tilde{\mathbb{f}}_1(\varkappa_{a+1}(x)) = \mathfrak{u} \cdot \varkappa_{a+1}(\tilde{\sigma}_1(x)) \qquad (6\text{-}3)$$

in the k-vector space $\hat{W}(\mathfrak{m}_R^a / \mathfrak{m}_R^{a+1})$. On this space $\mathfrak{u}$ acts as the identity. Let $y = (y_0, y_1, \dots)$ in $W(\bar{J}^a)$ be defined by $y_n = \tilde{\sigma}_1^n(x)$. Then $\delta(x) = y$ because the Witt polynomials give $w_n(y) = p^n \tilde{\sigma}_1^n(x) = \sigma^n(x) = w_n(\delta(x))$ as required. Thus $\varkappa_{a+1}(x)$ is the reduction of y. Since $\tilde{\mathbb{f}}_1$ acts on $\hat{W}(\mathfrak{m}_R^a / \mathfrak{m}_R^{a+1})$ by a shift to the left, the relation (6-3) follows. We obtain compatible $\mathfrak{u}$-homomorphisms of frames $\varkappa_* : (6\text{-}1) \to (6\text{-}2)$. The homomorphisms π and π' are crystalline; see the proof of [Lau 2010, Theorem 9.3]. Lifts of windows under ι and under ι' are both classified by lifts of the Hodge filtration from R_a to R_{a+1} in a compatible way. Thus if $\varkappa_a$ is crystalline then so is $\varkappa_{a+1}$, and Theorem 6.6 follows by induction, using that $\varkappa_1$ is an isomorphism. $\qquad\square$

Following the terminology of [Vasiu and Zink 2010], a Breuil window relative to $\mathfrak{S} \to R$ is a pair (Q, ϕ) where Q is a free $\mathfrak{S}$-module of finite rank and where $\phi : Q \to Q^{(\sigma)}$ is an $\mathfrak{S}$-linear map with cokernel annihilated by E. For such (Q, ϕ) there is a unique linear map $\psi : Q^{(\sigma)} \to Q$ with $\psi\phi = E$; the pairs (Q, ψ) are usually called Breuil–Kisin modules or Kisin modules. The category of $\mathscr{B}$-windows is equivalent to the category of Breuil windows relative to $\mathfrak{S} \to R$ by the assignment $(P, Q, F, F_1) \mapsto (Q, \phi)$, where ϕ is the composition of the inclusion $Q \to P$ with the inverse of $F_1^\sharp : Q^{(\sigma)} \cong P$; see [Lau 2010, Lemma 8.2].

Corollary 6.7. *If the equivalent conditions of* Proposition 6.2 *hold, there is an equivalence of exact categories between p-divisible groups over R and Breuil windows relative to $\mathfrak{S} \to R$.*

Proof. This is analogous to [Lau 2010, Corollary 8.3], using Corollary 5.4. $\qquad\square$

Following [Vasiu and Zink 2010] again, a Breuil module relative to $\mathfrak{S} \to R$ is a triple (M, ϕ, ψ) where M is a finitely generated $\mathfrak{S}$-module annihilated by a power of p and of projective dimension at most one and where $\phi : M \to M^{(\sigma)}$ and $\psi : M^{(\sigma)} \to M$ are $\mathfrak{S}$-linear maps with $\phi\psi = E$ and $\psi\phi = E$. If R has characteristic zero, such triples are equivalent to pairs (M, ϕ) or (M, ψ); see [Lau 2010, Lemma 8.6]. Again, the pairs (M, ψ) are usually called Breuil–Kisin modules or Kisin modules.

Corollary 6.8. *If the equivalent conditions of* Proposition 6.2 *hold, there is an equivalence of exact categories between commutative finite locally free group schemes of p-power order over R and Breuil modules relative to $\mathfrak{S} \to R$.*

Proof. This is analogous to [Lau 2010, Theorem 8.5].   □

Example 6.9. Let $R = W(k)$ and $\mathfrak{S} = W(k)[\![t]\!]$ with $\sigma(t) = t^p$. Define $\mathfrak{S} \to R$ by $t \mapsto p$; thus $E = p - t$. We have $\varkappa(E) = p - [p]$ and thus $u = v^{-1}(p - [p])$. Assume that $p = 2$. Then $u = u_0$, and $\mathcal{B} \to \mathcal{D}_R$ is a strict frame homomorphism. This example has motivated the definition of Dieudonné displays for $p = 2$.

7. Breuil–Kisin modules and crystals

We keep the notation of Section 6 and assume that the equivalent conditions of Proposition 6.2 hold. Assume that R has characteristic zero. Let S be the p-adic completion of the divided power envelope of the ideal $E\mathfrak{S} \subset \mathfrak{S}$, and let I be the kernel of $S \to R$. Since $\sigma : \mathfrak{S} \to \mathfrak{S}$ preserves the ideal (E, p), it extends to $\sigma : S \to S$. It is easy to see that $\sigma(I) \subseteq pS$, thus $\sigma : S \to S$ is a Frobenius lift again.

Proposition 7.1. *Let (Q, ϕ) be a Breuil window relative to $\mathfrak{S} \to R$ and let G be the associated p-divisible group over R; see* Corollary 6.7. *There is a natural isomorphism*

$$\mathbb{D}(G)_{S/R} \cong S \otimes_{\mathfrak{S}} Q^{(\sigma)}$$

such that the Hodge filtration of $\mathbb{D}(G)_{S/R}$ corresponds to the submodule generated by $\phi(Q) + IQ^{(\sigma)}$, and the Frobenius of $\mathbb{D}(G)_{S/R}$ corresponds to the σ-linear endomorphism of $Q^{(\sigma)}$ defined by $x \mapsto 1 \otimes \phi^{-1}(Ex)$.

In Kisin's theory (when R is one-dimensional) the analogous result is immediate from the construction. To prove Proposition 7.1, we consider the frame

$$\mathcal{S} = (S, I, R, \sigma, \sigma_1)$$

with $\sigma_1(x) = \sigma(x)/p$ for $x \in I$. The inclusion $\mathfrak{S} \to S$ is a u-homomorphism of frames $\iota : \mathcal{B} \to \mathcal{S}$ with $u = \sigma(E)/p \in S$. This element is a unit as required, since the arrow $\mathfrak{S} \to R$ is mapped surjectively onto $W(k) \to k$, which gives a local homomorphism $S \to W(k)$ that maps u to 1. Recall that we have frames

$\mathscr{D}_R \to \mathscr{D}_R^+$ when $p = 2$ and let us write $\mathscr{D}_R^+ = \mathscr{D}_R$ when $p \geq 3$. Then we have the commutative diagram of frames

$$
\begin{array}{ccc}
\mathscr{B} & \xrightarrow{\ \iota\ } & \mathscr{S} \\
\varkappa \downarrow & & \downarrow \varkappa_S \\
\mathscr{D}_R & \longrightarrow & \mathscr{D}_R^+
\end{array}
$$

Indeed, since $\mathbb{W}^+(R) \to R$ is a divided power extension of p-adically complete rings, the ring homomorphism $\varkappa : \mathfrak{S} \to \mathbb{W}^+(R)$ extends to $\varkappa_S : S \to \mathbb{W}^+(R)$, which is a strict frame homomorphism $\mathscr{S} \to \mathscr{D}_R^+$. Here $\varkappa$ is crystalline by Theorem 6.6. The proof of Proposition 7.1 will use the following fact:

Theorem 7.2. *The frame homomorphism $\varkappa_S$ is crystalline.*

This is a variant of the main result of [Zink 2001b]. It is easy to see that S is an admissible topological ring in the sense of Definition 1.2 if and only if $r = 1$, i.e., if R is a discrete valuation ring. In that case, the methods of [Zink 2001b] apply directly, but additional effort is needed to prove Theorem 7.2 in general. The proof is postponed to the next section.

Proof of Proposition 7.1. Let $\mathscr{P}_0 = (P, Q, F, F_1)$ be the $\mathscr{B}$-window associated to (Q, ϕ); thus $P = Q^{(\sigma)}$, the inclusion map $Q \to P$ is ϕ, and $F : P \to P$ is the σ-linear endomorphism of $Q^{(\sigma)}$ defined by $x \mapsto 1 \otimes \phi^{-1}(Ex)$. By definition we have $\Phi_R(G) = \varkappa_*(\mathscr{P}_0)$, which implies that $\Phi_R^+(G) = \varkappa_{S*}\iota_*(\mathscr{P}_0)$; here we use Theorem 4.9 when $p = 2$. On the other hand, the frames $\mathscr{S}$ and $\mathscr{D}_R^+$ both satisfy the hypotheses of the beginning of Section 3C. Thus Construction 3.16 and Proposition 3.17 applied to G give an $\mathscr{S}$-window $\mathscr{P}_1$ with an isomorphism $\varkappa_{S*}(\mathscr{P}_1) \cong \Phi_R^+(G)$, using the characterisation of Φ_R^+ in Theorem 3.19 and Proposition 3.23. Since the base change functor $\varkappa_{S*}$ is fully faithful by Theorem 7.2, the isomorphism $\varkappa_{S*}(\mathscr{P}_1) \cong \Phi_R^+(G) \cong \varkappa_{S*}\iota_*(\mathscr{P}_0)$ descends to an isomorphism $\mathscr{P}_1 \cong \iota_*(\mathscr{P}_0)$, which proves the proposition. □

7A. *Proof of Theorem 7.2.* Let us begin with a closer look on the p-adically complete ring S. For $m \geq 0$ let $S_{\langle m \rangle} \subseteq S$ be the closure of the $\mathfrak{S}$-algebra generated by $E^i / i!$ for $i \leq p^m$.

Proposition 7.3. *For $m \geq 1$, there is a surjective homomorphism of $\mathfrak{S}$-algebras $\mathfrak{S}[\![t_1, \ldots, t_m]\!] \to S_{\langle m \rangle}$ defined by $t_i \mapsto E^{p^i} / p^i!$.*

In particular, $S_{\langle m \rangle}$ is a noetherian complete local ring.

Lemma 7.4. *Let A be a noetherian complete local ring with a descending sequence of ideals $A \supseteq \mathfrak{a}_0 \supseteq \mathfrak{a}_1 \supseteq \cdots$. Then $A \to \varprojlim_i A/\mathfrak{a}_i$ is surjective.*

Proof. Let $\mathfrak{m}$ be the maximal ideal of A. For each r, the images of $\mathfrak{a}_i \to A/\mathfrak{m}^r$ stabilise for $i \to \infty$ to an ideal $\bar{\mathfrak{a}}_r \subseteq A/\mathfrak{m}^r$. We have

$$\varprojlim_i A/\mathfrak{a}_i = \varprojlim_{i,r} A/(\mathfrak{a}_i + \mathfrak{m}^r) = \varprojlim_r (A/\mathfrak{m}^r)/\bar{\mathfrak{a}}_r.$$

Since the ideals $\bar{\mathfrak{a}}_r$ form a surjective system, taking the limit over r of the exact sequences $0 \longrightarrow \bar{\mathfrak{a}}_r \longrightarrow A/\mathfrak{m}^r \longrightarrow (A/\mathfrak{m}^r)/\bar{\mathfrak{a}}_r \longrightarrow 0$ proves the lemma. $\qquad\square$

Proof of Proposition 7.3. Since the image of $E^{p^i}/p^i!$ in $S/p^n S$ is nilpotent, there is a well-defined homomorphism $\pi_{m,n} : \mathfrak{S}[\![t_1, \ldots, t_m]\!] \to S/p^n S$ with $t_i \mapsto E^{p^i}/p^i!$. By definition, $S_{\langle m \rangle}$ is the projective limit over n of the image of $\pi_{n,m}$. The proposition follows by Lemma 7.4. $\qquad\square$

Let $K = W(k) \otimes \mathbb{Q}$ and $\mathfrak{S}_{\mathbb{Q}} = K[\![x_1, \ldots, x_r]\!]$. Since $\sigma : \mathfrak{S} \to \mathfrak{S}$ preserves the ideal $J = (x_1, \ldots, x_r)$, it extends to a homomorphism $\sigma : \mathfrak{S}_{\mathbb{Q}} \to \mathfrak{S}_{\mathbb{Q}}$. For $r = 1$ it is easy to describe S and $S_{\langle m \rangle}$ as explicit subrings of $\mathfrak{S}_{\mathbb{Q}}$, since instead of the divided powers of E one can take the divided powers of x_1^e, where e is defined by $pR = \mathfrak{m}_R^e$. For $r \geq 2$ the situation is more complicated.

Proposition 7.5. *The natural embedding $\mathfrak{S} \to \mathfrak{S}_{\mathbb{Q}}$ extends to an injective homomorphism $S \to \mathfrak{S}_{\mathbb{Q}}$ that commutes with σ.*

Thus $S_{\langle m \rangle}$ is the image of $\mathfrak{S}[\![t_1, \ldots, t_m]\!] \to \mathfrak{S}_{\mathbb{Q}}$ as in Proposition 7.3.

Proof of Proposition 7.5. Recall that $J = (x_1, \ldots, x_r)$ as an ideal of $\mathfrak{S}$. Choose $E' \in J^e$ with $E - E' \in p\mathfrak{S}$ such that e is maximal; thus $p \in \mathfrak{m}_R^e \setminus \mathfrak{m}_R^{e+1}$. Let us write $\mathrm{gr}^i_{E'}(\mathfrak{S}) = E'^i \mathfrak{S}/E'^{i+1}\mathfrak{S}$, etc.

Lemma 7.6. *The map of graded rings $\mathrm{gr}_{E'}(\mathfrak{S}) \to \mathrm{gr}_{E'}(\mathfrak{S}_{\mathbb{Q}})$ is injective.*

Proof. It suffices to show that $\mathfrak{S}/E'\mathfrak{S} \to \mathfrak{S}_{\mathbb{Q}}/E'\mathfrak{S}_{\mathbb{Q}}$ is injective. The choice of E' implies that the image of E' in the regular local rings $\mathfrak{S}/p\mathfrak{S}$ and $\mathfrak{S}_{\mathbb{Q}}$ lies in the same power of the maximal ideals. Therefore the k-dimension of $\mathfrak{S}/(p\mathfrak{S} + E'\mathfrak{S} + J^n)$ is equal to the K-dimension of $\mathfrak{S}_{\mathbb{Q}}/(E'\mathfrak{S}_{\mathbb{Q}} + J^n\mathfrak{S}_{\mathbb{Q}})$. Since the last module is isomorphic to $\mathfrak{S}/(E'\mathfrak{S} + J^n) \otimes \mathbb{Q}$, it follows that $\mathfrak{S}/(E'\mathfrak{S} + J^n)$ is a free $W(k)$-module and injects into $\mathfrak{S}_{\mathbb{Q}}/(E'\mathfrak{S}_{\mathbb{Q}} + J^n\mathfrak{S}_{\mathbb{Q}})$. Since $\mathfrak{S}/E'\mathfrak{S}$ and $\mathfrak{S}_{\mathbb{Q}}/E'\mathfrak{S}_{\mathbb{Q}}$ are J-adically complete the lemma follows. $\qquad\square$

Let $S_0 \subseteq \mathfrak{S}_{\mathbb{Q}}$ be the $\mathfrak{S}$-algebra generated by $E'^i/i!$ for $i \geq 1$, or equivalently by $E^i/i!$ for $i \geq 1$, so S is the p-adic completion of S_0. Let $S_{0,n}$ be the image of $S_0 \to \mathfrak{S}_{\mathbb{Q}}/E'^n\mathfrak{S}_{\mathbb{Q}}$ and let $\widetilde{S} = \varprojlim_n S_{0,n}$. Each $S_{0,n}$ is a noetherian complete local ring with residue field k and thus a p-adically complete ring. Since $S_{0,n}$ has no p-torsion it follows that $\widetilde{S}$ is p-adically complete. We obtain a homomorphism $S \to \widetilde{S} \subseteq \mathfrak{S}_{\mathbb{Q}}$ which extends $S_0 \subseteq \widetilde{S} \subseteq \mathfrak{S}_{\mathbb{Q}}$.

Lemma 7.7. *We have $S_0 \cap p\widetilde{S} = pS_0$ inside $\widetilde{S}$.*

Proof. Let $x \in S_0 \cap p\tilde{S}$ be given. We have to show that x lies in pS_0. Assume that $x \neq 0$ and choose an expression $(\star)$ $x = \sum_{i=0}^{s} a_i E'^{m_i}/n_i!$ with $a_i \in \mathfrak{S}$ such that $n_0 < \cdots < n_s$. We use induction on $n_s - n_0$.

Suppose E' divides a_0 in $\mathfrak{S}$. Then $a_0 E'^{m_0}/n_0! = a_0' E'^{m_0'}/(n_0')!$ with $n_0' = n_0 + 1$ and $a_0' = n_0' a_0/E'$. If $s > 0$ this allows us to find a new expression of x of the type $(\star)$ with a smaller value of $n_s - n_0$, and we are done by induction. If $s = 0$ we replace the expression $(\star)$ by $x = a_0' E'^{m_0'}/n_0'!$; call this a modification of the first type.

Suppose E' does not divide a_0 in $\mathfrak{S}$. Lemma 7.6 implies that the image of x in $\mathrm{gr}_{E'}^{n_0}(\mathfrak{S}_{\mathbb{Q}})$ is nonzero. In S_{0,n_0+1} we have $x = py$. Choose an expression $y = \sum_{i=\ell}^{n_0} c_i E'^i/i!$ with $c_i \in \mathfrak{S}$ such that ℓ is maximal. Then E' does not divide c_ℓ in $\mathfrak{S}$, and Lemma 7.6 implies that y has nonzero image in $\mathrm{gr}_{E'}^\ell(\mathfrak{S}_{\mathbb{Q}})$. Thus $\ell = n_0$. Using Lemma 7.6 again, it follows that the image of a_0 in $\mathfrak{S}/E'\mathfrak{S}$ is divisible by p. Let $a_0 = pb_0 + b_1 E'$ with $b_i \in \mathfrak{S}$ and let $x' = x - pb_0 E'^{m_0}/n_0!$. Then $x - x' \in pS_0$; thus $x' \in S_0 \cap p\tilde{S}$, and we have to show that $x' \in pS_0$. If $s > 0$ we get an expression of x' of the type $(\star)$ with a smaller value of $n_s - n_0$, and we are done by induction. If $s = 0$ we replace x by x' and take for $(\star)$ the expression $x' = a_0' E'^{m_0'}/n_0'!$ with $n_0' = n_0 + 1$ and $a_0' = n_0' b_1$; call this a modification of the second type.

If $s > 0$ the inductive step is already finished. So we may assume that $s = 0$. We successively apply modifications of the first or second type depending on whether E' divides a_0. After at most p steps, the new value of a_0 becomes divisible by p, and thus x lies in pS_0. $\qquad\square$

Lemma 7.7 implies that $S_0/p^n S_0 \to \tilde{S}/p^n \tilde{S}$ is injective, so the projective limit $S \to \tilde{S}$ is injective, and thus $S \to \mathfrak{S}_{\mathbb{Q}}$ is injective. In order that this map commutes with σ it suffices to show that $S \to \mathfrak{S}_{\mathbb{Q}}/J^n\mathfrak{S}_{\mathbb{Q}}$ commutes with σ for each n; this is true since $S_0 \to \mathfrak{S}_{\mathbb{Q}}/J^n\mathfrak{S}_{\mathbb{Q}}$ commutes with σ, and the image of this map is p-adically complete. Thus Proposition 7.5 is proved. $\qquad\square$

We turn to the frames associated to the rings S and $S_{\langle m \rangle}$.

Lemma 7.8. *For $m \geq 1$ we have a subframe of $\mathscr{S} = (S, I, R, \sigma, \sigma_1)$,*

$$\mathscr{S}_{\langle m \rangle} = (S_{\langle m \rangle}, I_{\langle m \rangle}, R, \sigma, \sigma_1).$$

Proof. Necessarily $I_{\langle m \rangle} = I \cap S_{\langle m \rangle}$. We have to show that $\sigma : S \to S$ stabilises $S_{\langle m \rangle}$ and that $\sigma_1 = p^{-1}\sigma : I \to S$ maps $I_{\langle m \rangle}$ into $S_{\langle m \rangle}$. We will show that $\sigma(S)$ and $\sigma_1(I)$ are even contained in $S_{\langle 1 \rangle}$. Namely, we have $\sigma(E) = px$ with $x \in \mathfrak{S}[E^p/p]$. Thus $\sigma_1(E^i/i!) = (p \cdot i!)^{-1}(px)^i$ lies in $\mathfrak{S}[E^p/p]$, using that $1 + v_p(i!) \leq i$ for $i \geq 1$. Since $I/p^n I$ is the kernel of $S/p^n S \to R/p^n R$, this ideal is generated as an $\mathfrak{S}$-module by the elements $E^i/i!$ for $i \geq 1$. Thus the image of the map $I/p^{n+1}I \to S/p^n S$ induced by σ_1 lies in the image of $S_{\langle 1 \rangle}$, and it follows that $\sigma_1(I) \subseteq S_{\langle 1 \rangle}$. Since $S = \mathfrak{S} + I$, we get $\sigma(S) \subseteq S_{\langle 1 \rangle}$. $\qquad\square$

Proposition 7.9. *For $m \geq 1$ the inclusion $\mathscr{S}_{\langle m \rangle} \to \mathscr{S}$ is crystalline.*

Proof. This is a formal consequence of the relations $\sigma(S) \subseteq S_{\langle m \rangle}$ and $\sigma_1(I) \subseteq S_{\langle m \rangle}$ verified in the proof of Lemma 7.8.

Indeed, let $\mathscr{P} = (P, Q, F, F_1)$ be an $\mathscr{S}$-window. Choose a normal decomposition $P = L \oplus T$, and let $\Psi : L \oplus T \to P$ be the σ-linear isomorphism defined by F_1 on L and by F on T. Then $P_{\langle m \rangle} := S_{\langle m \rangle} \Psi(L \oplus T)$ is a free $S_{\langle m \rangle}$-module with $S \otimes_{S_{\langle m \rangle}} P_{\langle m \rangle} = P$. Moreover, $F_1(Q) \subseteq P_{\langle m \rangle}$ and $F(P) \subseteq P_{\langle m \rangle}$. Let $Q_{\langle m \rangle} = Q \cap P_{\langle m \rangle}$. Then $P_{\langle m \rangle}/Q_{\langle m \rangle} = P/Q$ is a projective R-module. Let $P_{\langle m \rangle} = L_{\langle m \rangle} \oplus T_{\langle m \rangle}$ be a normal decomposition and let $\Psi_{\langle m \rangle} : L_{\langle m \rangle} \oplus T_{\langle m \rangle} \to P_{\langle m \rangle}$ be the σ-linear map defined by F_1 on $L_{\langle m \rangle}$ and by F on $T_{\langle m \rangle}$. In order that the quadruple $\mathscr{P}_{\langle m \rangle} = (P_{\langle m \rangle}, Q_{\langle m \rangle}, F, F_1)$ is an $\mathscr{S}_{\langle m \rangle}$-window with base change $\mathscr{P}$ we need that the determinant of $\Psi_{\langle m \rangle}$ is invertible. But the determinant of $\Psi_{\langle m \rangle}$ becomes invertible in S because $\mathscr{P}$ is a window, and $S_{\langle m \rangle} \to S$ is a local homomorphism. Thus the base change functor from $\mathscr{S}_{\langle m \rangle}$-windows to $\mathscr{S}$-windows is essentially surjective.

In order that the functor is fully faithful it suffices to show that it induces a bijection $\mathrm{End}(\mathscr{P}_{\langle m \rangle}) \to \mathrm{End}(\mathscr{P})$. Clearly the map is injective. We have to show that every $h \in \mathrm{End}(\mathscr{P})$ stabilises $P_{\langle m \rangle}$. But $h(F_1(Q)) = F_1(h(Q)) \subseteq F_1(Q) \subseteq P_{\langle m \rangle}$, and $F_1(Q)$ generates $P_{\langle m \rangle}$ as an $S_{\langle m \rangle}$-module. This proves the proposition. $\square$

Proposition 7.10. *For $m \geq 1$, the composition $\mathscr{S}_{\langle m \rangle} \longrightarrow \mathscr{S} \xrightarrow{\varkappa_S} \mathscr{D}_R^+$ is crystalline.*

This is the main step in the proof of Theorem 7.2. The proof of Proposition 7.10 is a variant of the proof of Theorem 6.6.

Proof. We choose e such that $p \in \mathfrak{m}_R^e \setminus \mathfrak{m}_R^{e+1}$, and consider the index set $N = \{1, 2, \ldots\} \cup \{e+\}$, ordered by the natural order of $\mathbb{Z}$ and $e < e+ < e+1$. For $n \in N$ let $n+ \in N$ be its successor. Let $\mathfrak{m}_R^{e+} = \mathfrak{m}_R^{e+1} + pR$. For $n \in N$ let $R_n = R/\mathfrak{m}_R^n$. We equip the ideal $\mathfrak{m}_R^n/\mathfrak{m}_R^{n+}$ of R_{n+} with the trivial divided powers if $n \neq e+$ and with the canonical divided powers of p if $n = e+$; these are again trivial if p is odd. In all cases the divided powers are compatible with the canonical divided powers of p, and we obtain frames

$$\mathscr{D}_{R_{n+}/R_n}^+ = (\mathbb{W}^+(R_{n+}), \mathbb{I}_{R_{n+}/R_n}^+, R_n, f, \tilde{f}_1).$$

Let T_n be the image of $S_{\langle m \rangle} \xrightarrow{\varkappa_S} \mathbb{W}^+(R) \longrightarrow \mathbb{W}^+(R_n)$. Since $\varkappa_S \sigma = f \varkappa_S$, the ring T_n is stable under f. Let K_n be the kernel of $T_n \to R_n$ and let $\tilde{K}_n$ be the kernel of $T_{n+} \to R_n$.

We claim that $\tilde{f}_1(\tilde{K}_n) \subseteq T_{n+}$.

To prove this, let M_n be the kernel of $S_{\langle m \rangle} \to R_n$, so $\tilde{K}_n$ is the image of $M_n \to \mathbb{W}^+(R_{n+})$. Since $\varkappa_S \sigma = f \varkappa_S$ and since f_1 is f-linear it suffices to show that a set of generators x_i of the ideal M_n with images $\varkappa_S(x_i) = \bar{x}_i \in \tilde{K}_n$ satisfies $f_1(\bar{x}_i) \in T_{n+}$. Since $\mathfrak{m}_R = JR$, for $n \neq e+$ the ideal M_n is generated by $I_{\langle m \rangle}$ and J^n, while M_{e+} is generated by $I_{\langle m \rangle}$ and J^{e+1} and p. We check these generators case by case.

First, for $x \in I_{\langle m \rangle}$ we have $f_1(\bar{x}) \in T_{n+}$ because $\mathscr{S}_{\langle m \rangle} \to \mathscr{D}_{R n+}^+$ is a frame homomorphism.

Assume that $n \neq e+$. The homomorphism $\delta : J^n / J^{n+1} \to W(J^n / J^{n+1})$ is given by $\delta(x) = (x, \sigma_1(x), (\sigma_1)^2(x), \dots)$. Indeed, applying the Witt polynomial w_n to this equation gives $\sigma^n(x) = p^n (\sigma_1)^n(x)$, which is true. Since the divided powers on $\mathfrak{m}_R^n / \mathfrak{m}_R^{n+}$ are trivial, the endomorphism $\tilde{f}_1$ of $W(\mathfrak{m}_R^n / \mathfrak{m}_R^{n+})$ is given by a shift to the left. Thus the map $\varkappa_S : J^n / J^{n+1} \to W(\mathfrak{m}_R^n / \mathfrak{m}_R^{n+})$ satisfies $\varkappa_S \sigma_1 = \tilde{f}_1 \varkappa_S$, and we see that $f_1(\bar{x}) \in T_{n+}$ for $x \in J^n$.

Assume now that $n = e+$. Since J^{e+1} maps to zero in $W(R_{e+1})$, it remains to show that $\tilde{f}_1(p) \in T_{n+}$. Now $\mathrm{Log}(p - v(1)) = [p, 0, 0, \dots]$; see the proof of Lemma 1.9. Thus $\tilde{f}_1(p) = f_1(v(1)) = 1$, and the claim is proved.

We obtain frames

$$\mathscr{T}_n = (T_n, K_n, R_n, f, f_1), \quad \mathscr{T}_{n+/n} = (T_{n+}, \tilde{K}_n, R_n, f, \tilde{f}_1),$$

and a commutative diagram of frames with strict homomorphisms

$$
\begin{array}{ccccc}
\mathscr{T}_{n+} & \xrightarrow{\psi'} & \mathscr{T}_{n+/n} & \xrightarrow{\pi'} & \mathscr{T}_n \\
\downarrow{\iota_{n+}} & & \downarrow & & \downarrow{\iota_n} \\
\mathscr{D}_{R n+}^+ & \xrightarrow{\psi} & \mathscr{D}_{R n+/R n}^+ & \xrightarrow{\pi} & \mathscr{D}_{R n}^+
\end{array}
$$

Here π is crystalline because the hypotheses of Theorem 2.2 are satisfied; see the proof of Corollary 2.9. Since the vertical arrows are injective, it follows that π' satisfies the hypotheses of Theorem 2.2 as well, and thus π' is crystalline. Moreover, lifts of windows under ψ and under ψ' correspond to lifts of the Hodge filtration from R_n to R_{n+} in the same way. Since ι_1 is bijective, it follows that ι_n is crystalline for each n. Consider the limit frame

$$\mathscr{T} = \varprojlim_n \mathscr{T}_n = (T, K, R, f, f_1).$$

The inclusion $\iota : \mathscr{T} \to \mathscr{D}_R^+$ is the projective limit over n of ι_n and thus crystalline. Since $S_{\langle m \rangle}$ is noetherian by Proposition 7.3, Lemma 7.4 implies that $T = \varprojlim_n T_n$ is the image of $\varkappa_S : S_{\langle m \rangle} \to \mathbb{W}^+(R)$. If $\varkappa_S$ is injective, we get $\mathscr{S}_{\langle m \rangle} = \mathscr{T}$, so $\mathscr{S}_{\langle m \rangle} \to \mathscr{D}_R^+$ is crystalline as required.

Since we have not proved that $\varkappa_S$ is injective we need an extra argument. Let $\mathfrak{a}$ be the kernel of $\varkappa_S : S_{\langle m \rangle} \to \mathbb{W}^+(R)$ and let $\mathfrak{a}_n = \mathfrak{a} \cap J^n \mathfrak{S}_{\mathbb{Q}}$ for $n \geq 1$; here we use that S is a subring of $\mathfrak{S}_{\mathbb{Q}}$ by Proposition 7.5. We have $\mathfrak{a} = \mathfrak{a}_1$. The ideals $\mathfrak{a}_n$ of $S_{\langle m \rangle}$ are stable under σ, and they are also stable under σ_1 since $S_{\langle m \rangle} / \mathfrak{a}$ and $\mathfrak{a}_n / \mathfrak{a}_{n+1}$ have no p-torsion. Thus we can define frames $\mathscr{S}_{\langle m \rangle, n} = (S_{\langle m \rangle} / \mathfrak{a}_n, I_{\langle m \rangle} / \mathfrak{a}_n, R, \sigma, \sigma_1)$. We have $\mathscr{S}_{\langle m \rangle, 1} = \mathscr{T}$, and the projective limit over n of $\mathscr{S}_{\langle m \rangle, n}$ is isomorphic to $\mathscr{S}_{\langle m \rangle}$

by [Lemma 7.4](#) and [Proposition 7.5](#). The ideal $\mathfrak{a}_n/\mathfrak{a}_{n+1}$ is a finitely generated $W(k)$-submodule of $(J^n/J^{n+1}) \otimes \mathbb{Q}$. Since the conditions of [Proposition 6.2](#) are satisfied, the endomorphism σ_1 of $\mathfrak{a}_n/\mathfrak{a}_{n+1}$ is p-adically nilpotent. Thus $\mathscr{S}_{\langle m \rangle, n+1} \to \mathscr{S}_{\langle m \rangle, n}$ is crystalline; see the proof of [Lau 2010, Theorem 9.3]. It follows that $\mathscr{S}_{\langle m \rangle} \to \mathscr{T}$ is crystalline, so $\mathscr{S}_{\langle m \rangle} \to \mathscr{D}_R^+$ is crystalline too. $\qquad\square$

[Theorem 7.2](#) follows from Propositions [7.9](#) and [7.10](#). $\qquad\square$

Remark 7.11. Assume that $r = 1$; i.e., R is a discrete valuation ring. If $pR = \mathfrak{m}_R^e$, the ring S is the p-adic completion of $W(k)[\![t]\!][\{t^{em}/m!\}_{m \geq 1}]$. It is easy to see that each quotient $S/p^n S$ is admissible, so the p-adically complete ring S is an admissible topological ring. In particular, $\mathbb{W}^+(S)$ is defined. Since we assumed that the image of $\delta : \mathfrak{S} \to W(\mathfrak{S})$ lies in $\mathbb{W}(\mathfrak{S})$, it follows that the image of $\delta : S \to W(S)$ lies in $\mathbb{W}^+(S)$, using that $\mathbb{W}^+(S) \to R$ is the projective limit of the divided power extensions $\mathbb{W}^+(S/p^n S) \to R/p^n R$ and that each $\mathbb{W}^+(S/p^n S)$ is p-adically complete. If p is odd this means that $\mathscr{S}$ is a Dieudonné frame in the sense of [Zink 2001b, Definition 3.1], and [Theorem 7.2](#) becomes a special case of [Zink 2001b, Theorem 3.2]. For $p = 2$ the proof of [loc. cit.] works as well. The starting point is the construction of an inverse functor of $\varkappa_{S*}$; it maps a $\mathscr{D}_R^+$-window $\mathscr{P}$ to the value of its crystal $\mathbb{D}^+(\mathscr{P})_{S/R}$, equipped with an appropriate $\mathscr{S}$-window structure.

If $r \geq 2$, the ring S is not admissible and thus the crystal of a $\mathscr{D}_R^+$-window can not be evaluated at S/R. However, one can define by hand a subframe $\mathscr{D}_{S/R}^+$ of $\mathscr{W}_{S/R}$ such that $\mathscr{D}_{S/R}^+ \to \mathscr{D}_R^+$ is crystalline. This allows us to evaluate the crystal at S/R and to define an inverse functor of $\varkappa_{S*}$ as before. The underlying ring of $\mathscr{D}_{S/R}^+$ is defined as follows. Let $S_{m,n}$ be the image of $S_{\langle m \rangle} \to S/p^n S$ and let $I_{m,n}$ be the kernel of $S_{m,n} \to R/p^n R$. The divided Witt polynomials define an isomorphism $\mathrm{Log} : W(I/p^n I) \cong (I/p^n I)^{\mathbb{N}}$, and our ring is $\varprojlim_n \varinjlim_m$ of the rings $\mathbb{W}^+(S_{0,n}) + \mathrm{Log}^{-1}((I_{m,n})^{\langle \mathbb{N} \rangle})$. The $\varprojlim_n$ of these rings for fixed $m \geq 1$ gives a frame $\mathscr{D}_{S_{\langle m \rangle}/R}^+$ with a crystalline homomorphism to $\mathscr{D}_R^+$. This allows us to construct the inverse functor from $\mathscr{D}_R^+$-windows to $\mathscr{S}_{\langle m \rangle}$-windows. We leave out the details.

8. Rigidity of p-divisible groups

In this section, we record a rigidity property of the category of p-divisible groups that will be used in [Section 9](#). As preparation, for a local ring R we consider the additive category $\mathcal{F}_R$ of commutative finite locally free p-group schemes over R. It is known that $\mathcal{F}_R$ is equivalent to the full subcategory of the bounded derived category of the exact category of p-divisible groups over R formed by the complexes of length one which are isogenies; see [Kisin 2006, (2.3.5)]. In elementary terms this equivalence can be expressed as follows:

Proposition 8.1. *For a local ring R, let $\mathcal{I}_R$ be the category with isogenies of p-divisible groups over R as objects and homomorphisms of complexes modulo homotopy as homomorphisms. The set S of quasi-isomorphisms in $\mathcal{I}_R$ allows a calculus of right fractions. In particular, the localised category $S^{-1}\mathcal{I}_R$ is additive. It is equivalent to the additive category $\mathcal{F}_R$.*

For completeness let us prove this directly.

Proof. Let $\widetilde{\mathcal{I}}_R$ be the category with isogenies of p-divisible groups over R as objects and homomorphisms of complexes as homomorphisms. We denote isogenies by $X = [X^0 \to X^1]$. Let $h^0(X)$ be the kernel of $X^0 \to X^1$. A homomorphism $f : X \to Y$ in $\widetilde{\mathcal{I}}_R$ is homotopic to zero if and only if $h^0(f)$ is the zero map; the homotopy is unique if it exists. We claim:

($\star$) For each homomorphism $f : X \to Y$ in $\widetilde{\mathcal{I}}_R$, one can find a quasi-isomorphism $t : Z \to X$ in $\widetilde{\mathcal{I}}_R$ and a homomorphism $g : Z \to Y$ in $\widetilde{\mathcal{I}}_R$ which is an epimorphism in both components such that ft is homotopic to g. Namely, embed $h^0(X)$ into $Z^0 = X^0 \oplus Y^0$ by $(1, f)$ and put $Z^1 = Z^0/h^0(X)$. Define t and g by the projections $Z^0 \to X^0$ and $Z^0 \to Y^0$. There is a homotopy between ft and g because $ft = g$ on $h^0(Z)$, and ($\star$) is proved.

Next, for given homomorphisms $X \xrightarrow{f} Y \xleftarrow{s} Y'$ in $\widetilde{\mathcal{I}}_R$, where s is a quasi-isomorphism, one can find an isogeny X' with a homomorphism $g : X' \to Y'$ and a quasi-isomorphism $t : X' \to X$ such that ft is homotopic to sg. Indeed, by ($\star$) we can assume that the components of f are epimorphisms. Then take $X' = X \times_Y Y'$ componentwise. It follows easily that S allows a calculus of right fractions. We have an additive functor $h^0 : S^{-1}\mathcal{I} \to \mathcal{F}$. It is surjective on isomorphism classes by a theorem of Raynaud [Berthelot et al. 1982, Théorème 3.1.1]. Let X and Y be isogenies. The functor h^0 is full, because for a given homomorphism $f_0 : h^0(X) \to h^0(Y)$, the construction in ($\star$) allows us to represent f_0 as $h^0(gt^{-1})$. The functor is faithful because if a right fraction $gt^{-1} : X \to Y$ induces zero on h^0 then g induces zero on h^0, and thus g is homotopic to zero. $\square$

Let (Art) be the category of local Artin schemes with perfect residue field of characteristic p, and let $(p\text{-div}) \to (\text{Art})$ be the fibred category of p-divisible groups over schemes in (Art).

Lemma 8.2. *Assume that u is an exact automorphism of the fibred category $(p\text{-div})$ over (Art) such that for the group $E = \mathbb{Q}_p/\mathbb{Z}_p$ over $\operatorname{Spec} \mathbb{F}_p$ there is an isomorphism $u(E) \cong E$. Then u is isomorphic to the identity functor.*

Proof. For each U in (Art) we are given a functor $G \mapsto G^u$ from the category of p-divisible groups over U to itself, which preserves short exact sequences, compatible with base change in U, such that $\operatorname{Hom}(G, H) \cong \operatorname{Hom}(G^u, H^u)$. We have to show that there is an natural isomorphism $G^u \cong G$ for all G, compatible with base change

in U. Let $(p\text{-fin}) \to (\text{Art})$ be the fibred category of commutative finite locally free p-group schemes over schemes in (Art). By [Proposition 8.1](), u induces an automorphism of $(p\text{-fin})$ over (Art). Let $H \in (p\text{-fin})$ over $U \in (\text{Art})$ be given. Assume that p^n annihilates H and H^u. For each $T \to U$ in (Art) there is a natural isomorphism

$$H(T) \cong \text{Hom}_T(\mathbb{Z}/p^n\mathbb{Z}, H_T) \cong \text{Hom}_T(\mathbb{Z}/p^n\mathbb{Z}, H_T^u) \cong H^u(T),$$

using that $(\mathbb{Z}/p^n\mathbb{Z})^u \cong \mathbb{Z}/p^n\mathbb{Z}$. Since commutative finite locally free group schemes over U form a full subcategory of the category of abelian presheaves on $(\text{Art})/U$, we get a natural isomorphism $H \cong H^u$, which induces a natural isomorphism $G \cong G^u$ for all p-divisible groups G over U. $\qquad\square$

9. The reverse functor

We fix an admissible ring R which is local of dimension zero; thus $k = R_{\text{red}}$ is a perfect field of characteristic p. In this case, one can write down an inverse of the functor Φ_R as follows. The construction appears in [Lau 2009]() when $p \geq 3$ or $pR = 0$ and extends to the general case with appropriate changes.

Definition 9.1. Let $\mathcal{J}_R$ be the category of R-algebras A such that $\mathcal{N}_A$ is bounded nilpotent and A_{red} is the union of finite dimensional k-algebras.

We call a ring homomorphism $A \to B$ ind-étale if B is the filtered direct limit of étale A-algebras.

Lemma 9.2. *Every $A \in \mathcal{J}_R$ is admissible. The category $\mathcal{J}_R$ is stable under tensor products. If $A \to B$ is an ind-étale or a quasi-finite ring homomorphism with $A \in \mathcal{J}_R$ then $B \in \mathcal{J}_R$.*

Proof. Since a reduced finite k-algebra is étale and thus perfect, every A in $\mathcal{J}_R$ is admissible. Let $A \to B$ a ring homomorphism with $A \in \mathcal{J}_R$. Then $\mathcal{N}_A B$ is bounded nilpotent, so B is lies in $\mathcal{J}_R$ if and only if $B/\mathcal{N}_A B$ lies in $\mathcal{J}_R$. For given homomorphisms $B \leftarrow A \to C$ in $\mathcal{J}_R$ we have to show that $B \otimes_A C$ lies in $\mathcal{J}_R$. By the preceding remark, we may assume that A, B, C are reduced. Then $B \otimes_A C$ is the direct limit of étale k-algebras and thus lies in $\mathcal{J}_R$. Let $g : A \to B$ be an ind-étale or quasi-finite ring homomorphism with $A \in \mathcal{J}_R$. In order to show that $B \in \mathcal{J}_R$ we may assume that A is reduced. Then every finitely generated k-subalgebra of A is étale. Thus each étale A-algebra is defined over an étale k-subalgebra of A. If g is ind-étale it follows that B lies in $\mathcal{J}_R$. Assume that g is quasi-finite. Then B is defined over an étale k-subalgebra of A. Since all finite k-algebras lie in $\mathcal{J}_R$ and since $\mathcal{J}_R$ is stable under tensor products, it follows that $B \in \mathcal{J}_R$. $\qquad\square$

Let $S = \text{Spec } R$ and let $\mathcal{J}_S$ be the category of affine S-schemes $\text{Spec } A$ with $A \in \mathcal{J}_R$. If τ is either ind-étale or fpqc, we consider the τ-topology on $\mathcal{J}_S$ in which coverings are finite families of morphisms $(\text{Spec } B_i \to \text{Spec } A)$ such that

the associated homomorphism $A \to \prod_i B_i$ is faithfully flat, and ind-étale if τ is ind-étale. Let $\widetilde{S}_{\mathcal{J},\tau}$ be the category of τ-sheaves on $\mathcal{J}_S$.

Lemma 9.3. *The presheaf of rings $\mathbb{W}$ on $\mathcal{J}_S$ is an fpqc sheaf.*

Proof. See [Lau 2009, Lemma 1.5]. Since the presheaf W is an fpqc sheaf, it suffices to show that for an injective ring homomorphism $A \to B$ in $\mathcal{J}_R$ we have $\mathbb{W}(A) = \mathbb{W}(B) \cap W(A)$ in $W(B)$. This is easily verified using that $A_{\mathrm{red}} \to B_{\mathrm{red}}$ is injective and $\widehat{W}(\mathcal{N}_A) = \widehat{W}(\mathcal{N}_B) \cap W(A)$ in $W(\mathcal{N}_B)$. $\qquad\square$

Let $\mathscr{P}$ be a Dieudonné display over R. For $\operatorname{Spec} A \in \mathcal{J}_S$ let $\mathscr{P}_A = (P_A, Q_A, F, F_1)$ be the base change of $\mathscr{P}$ to A. We define two complexes $Z(\mathscr{P})$ and $Z'(\mathscr{P})$ of presheaves of abelian groups on $\mathcal{J}_S$ by

$$Z(\mathscr{P})(\operatorname{Spec} A) = [Q_A \xrightarrow{F_1-1} P_A], \tag{9-1}$$

$$Z'(\mathscr{P})(\operatorname{Spec} A) = [Q_A \xrightarrow{F_1-1} P_A] \otimes [\mathbb{Z} \to \mathbb{Z}[1/p]], \tag{9-2}$$

such that $Z(\mathscr{P})$ lies in degrees $0, 1$ and $Z'(\mathscr{P})$ lies in degrees $-1, 0, 1$, so the second tensor factor lies in degrees $-1, 0$.

Proposition 9.4. *The components of $Z'(\mathscr{P})$ are fpqc sheaves on $\mathcal{J}_S$. The ind-étale (and thus the fpqc) cohomology sheaves of $Z'(\mathscr{P})$ vanish outside degree zero, and the cohomology sheaf in degree zero is represented by a well-defined p-divisible group $\mathrm{BT}_R(\mathscr{P})$ over R. This defines an additive and exact functor*

$$\mathrm{BT}_R : (\textit{Dieudonné displays over } R) \to (\textit{p-divisible groups over } R).$$

One can also express the definition of the functor BT_R by the formula

$$\mathrm{BT}_R(\mathscr{P}) = [Q \xrightarrow{F_1-1} P] \otimes^L \mathbb{Q}_p/\mathbb{Z}_p$$

in the derived category of either ind-étale or fpqc abelian sheaves on $\mathcal{J}_S$.

Proof. This is essentially proved in [Lau 2009], but we recall the arguments for completeness and because there is a small modification when $p = 2$. To begin with, p-divisible groups over R form a full subcategory of the abelian presheaves on $\mathcal{J}_S$ because finite group schemes over R lie in $\mathcal{J}_S$; see Lemma 9.2. Hence BT_R is a well-defined additive functor if the assertions on the cohomology of $Z'(\mathscr{P})$ hold. Since an exact sequence of Dieudonné displays over R induces an exact sequence of the associated complexes of presheaves Z', the functor BT_R is exact if it is defined.

The components of $Z(\mathscr{P})$ and $Z'(\mathscr{P})$ are fpqc sheaves on $\mathcal{J}_S$ by Lemma 9.3. These complexes carry two filtrations. First, a Dieudonné display is called étale if $Q = P$, and nilpotent if $V^\sharp$ is topologically nilpotent. Every Dieudonné display over R is naturally an extension of an étale by a nilpotent Dieudonné display, which induces exact sequences of the associated complexes $Z(\dots)$ and $Z'(\dots)$. Thus we

may assume that $\mathscr{P}$ is étale or nilpotent. Second, for every $\mathscr{P}$ we have an exact sequence of complexes of presheaves

$$0 \longrightarrow \hat{Z}(\mathscr{P}) \longrightarrow Z(\mathscr{P}) \longrightarrow Z_{\mathrm{red}}(\mathscr{P}) \longrightarrow 0,$$

defined by $Z_{\mathrm{red}}(\mathscr{P})(\mathrm{Spec}\, A) = Z(\mathscr{P})(\mathrm{Spec}\, A_{\mathrm{red}})$. The same holds for Z' instead of Z. We write $\hat{Z}(\mathscr{P}) = [\hat{Q} \to \hat{P}]$.

Assume that $\mathscr{P}$ is étale. Then $F_1 : P \to P$ is an f-linear isomorphism. Thus $F_1 : \hat{P} \to \hat{P}$ is elementwise nilpotent, and the complex $\hat{Z}(\mathscr{P})$ is acyclic. It follows that $Z(\mathscr{P})$ is quasi-isomorphic to the complex $Z_{\mathrm{red}}(\mathscr{P}) = Z_{\mathrm{red}}$, which is the projective limit of the complexes $Z_{\mathrm{red},n} = Z_{\mathrm{red}}/p^n Z_{\mathrm{red}}$. In the étale topology, each $Z_{\mathrm{red},n}$ is a surjective homomorphism of sheaves whose kernel is a locally constant sheaf G_n of free $\mathbb{Z}/p^n\mathbb{Z}$-modules of rank equal to the rank of P. The system $(G_n)_n$ defines an étale p-divisible group G over R, and Z_{red} is quasi-isomorphic to $T_p G = \varprojlim G_n$ as ind-étale sheaves. It follows that $Z'(\mathscr{P}) \simeq Z'_{\mathrm{red}}(\mathscr{P})$ is quasi-isomorphic to the complex $[T_p G \to T_p G \otimes \mathbb{Z}[1/p]]$ in degrees $-1, 0$, which is quasi-isomorphic to G in degree zero (as ind-étale sheaves).

Assume that $\mathscr{P}$ is nilpotent. Then the complex $Z_{\mathrm{red}}(\mathscr{P})$ is acyclic because its value over $\mathrm{Spec}\, A$ is isomorphic to $[1 - V : P_{A_{\mathrm{red}}} \to P_{A_{\mathrm{red}}}]$, where V is a topologically nilpotent f^{-1}-linear homomorphism. Thus $Z(\mathscr{P})$ is quasi-isomorphic to $\hat{Z}(P)$. To $\mathscr{P}$ we associate a nilpotent display by the u_0-homomorphism of frames $\mathscr{D}_R \to \mathscr{W}_R$. By [Zink 2002, Theorem 81 and Corollary 89] there is a formal p-divisible group G over R associated to this display such that for each $A \in \mathcal{J}_R$ there is an exact sequence

$$0 \longrightarrow \hat{Q}(A) \xrightarrow{u_0 F_1 - 1} \hat{P}(A) \longrightarrow G(A) \longrightarrow 0;$$

this is the direct limit of the corresponding sequences for the finitely generated (nilpotent) subalgebras of $\mathcal{N}_A$. Since $u_0 \in W(\mathbb{Z}_p)$ maps to 1 in $W(\mathbb{F}_p)$, there is a unique $c \in W(\mathbb{Z}_p)$ which maps to 1 in $W(\mathbb{F}_p)$ such that $u_0 = c f(c^{-1})$, namely $c = u_0 f(u_0) f^2(u_0) \cdots$. Multiplication by c in both components defines an isomorphism of complexes

$$[\hat{Q}(A) \xrightarrow{F_1 - 1} \hat{P}(A)] \cong [\hat{Q}(A) \xrightarrow{u_0 F_1 - 1} \hat{P}(A)]$$

It follows that $Z'(\mathscr{P}) \simeq \hat{Z}'(\mathscr{P})$ is quasi-isomorphic to G in degree zero. $\square$

Remark 9.5. Recall that $\mathscr{D}_R = (\mathbb{W}(R), \mathbb{I}_R, f, \mathbb{f}_1)$ is viewed as a Dieudonné display over R. We have $\mathrm{BT}_R(\mathscr{D}_R) = \mu_{p^\infty}$ by [Zink 2002, (211)].

Lemma 9.6. *Let $R \to R'$ be a homomorphism of admissible rings which are local of dimension zero. For each Dieudonné display $\mathscr{P}$ over R there is a natural isomorphism*

$$\mathrm{BT}_R(\mathscr{P})_{R'} \cong \mathrm{BT}_{R'}(\mathscr{P}_{R'}).$$

Proof. If the residue field of R' is an algebraic extension of k, every ring in $\mathcal{J}_{R'}$ lies in $\mathcal{J}_R$, and the assertion follows directly from the construction of BT_R. In general, let $\mathcal{E}_R$ be the category of all R-algebras which are admissible rings, and let $\mathcal{E}_S$ be the category of affine S-schemes $\mathrm{Spec}\,A$ with $A \in \mathcal{E}_R$, endowed with the ind-étale topology. The complexes of presheaves $Z(\mathscr{P})$ and $Z'(\mathscr{P})$ on $\mathcal{J}_S$ defined in (9-1) and (9-2) extend to complexes of presheaves on $\mathcal{E}_S$ defined by the same formulas. The proof of Lemma 9.2 shows that for an ind-étale ring homomorphism $A \to B$ with $A \in \mathcal{E}_R$ we have $B \in \mathcal{E}_R$ as well. Using this, the proof of Proposition 9.4 shows that the ind-étale cohomology sheaves of $Z'(\mathscr{P})$ on $\mathcal{E}_S$ vanish outside degree zero, and $H^0(Z'(\mathscr{P}))$ is naturally isomorphic to $\mathrm{BT}_R(\mathscr{P})$ as a sheaf on $\mathcal{E}_S$. Since every ring in $\mathcal{E}_{R'}$ lies in $\mathcal{E}_R$, the lemma follows as in the first case. $\qquad\square$

Proposition 9.7. *The functor BT_R is an equivalence of exact categories which is a quasi-inverse of the functor Φ_R.*

Proof. By Section 3A we may assume that R is a local Artin ring. Since p-divisible groups and Dieudonné displays over k have universal deformation rings which are power series rings over $W(k)$, once the functor BT_R is defined, in order to show that it is an equivalence of categories it suffices to consider the cases $R = k$ and $R = k[\varepsilon]$. In particular, if $p = 2$, we may assume that $pR = 0$, so that the results of [Zink 2001a] and [Lau 2009] can be applied. The category $\mathcal{C}_R$ used in [Lau 2009] is the category of all $A \in \mathcal{J}_R$ such that $\mathcal{N}_A$ is nilpotent. Since this subcategory is stable under ind-étale extensions, it does not make a difference whether the functor BT_R is defined in terms of $\mathcal{C}_R$ or $\mathcal{J}_R$. Thus BT_R is an equivalence by [Lau 2009, Theorem 1.7], which relies on the equivalence proved in [Zink 2001a]. It is easily verified that $\mathrm{BT}_R(\Phi_R(\mathbb{Q}_p/\mathbb{Z}_p))$ is isomorphic to $\mathbb{Q}_p/\mathbb{Z}_p$. Thus BT_R is a quasi-inverse of Φ_R by Lemmas 8.2 and 9.6. It is easily verified that the functors BT_R and Φ_R preserve short exact sequences. $\qquad\square$

Appendix: PD envelopes of regular immersions

This section provides a reference for the flatness of the divided power envelope of a regular immersion, which is used in the proof of Lemma 1.13. Let us recall regular immersions following [SGA 1971, VII]. For a ring A, a projective A-module M of finite type, and a linear map $f : M \to A$, one defines the Koszul complex

$$K_*(A, f) = [\cdots \to \Lambda^2 M \to \Lambda^1 M \to A]$$

with differential given by $x_1 \wedge \cdots \wedge x_n \mapsto \sum (-1)^{i+1} f(x_i) x_1 \wedge \cdots \hat{x}_i \cdots \wedge x_n$. Let $I = f(M) \subseteq A$. One calls f regular if the augmentation $K_*(A, f) \to A/I$ is a quasi-isomorphism. If $x_1, \ldots, x_r$ is a regular sequence in A and $f : A^r \to A$ is given by $f(a) = \sum a_i x_i$, then f is regular in the previous sense. For a ring homomorphism $A \to A'$, let $f' : M' \to A'$ be the scalar extension of f, and let

$I' = f'(M')$. If both f and f' are regular, then $\operatorname{Tor}_i^A(A/I, A') = 0$ for $i \geq 1$ and thus $I' = I \otimes_A A'$. A closed immersion of schemes $Y \to X$ is called regular if locally in X it takes the form $\operatorname{Spec} A/I \to \operatorname{Spec} A$, where $I = f(M)$ for a regular homomorphism $f : M \to A$.

Proposition A.1. *Let S be a scheme and $i : Y \to X$ be a regular closed immersion of flat S-schemes. Then the divided power envelope $\mathscr{D}_X(Y)$ is flat over S.*

Under additional hypotheses, this is proved in [Berthelot et al. 1982, Lemme 2.3.3]. We use the following description of the divided power polynomial algebra:

Lemma A.2. *For a ring R, let $A_0 = R[T_1, \ldots, T_n]$, and let $B_0 = R\langle T_1, \ldots, T_n \rangle$ be the divided power envelope of $I_0 = (T_1, \ldots, T_n) \subseteq A_0$. Then one can write $B_0 = \varinjlim_r M_{0,r}$ as an A_0-module, the direct limit taken over $r \in \mathbb{N}$ ordered multiplicatively, such that there are exact sequences of A_0-modules*

$$0 \longrightarrow J_{0,r} \longrightarrow M_{0,r} \longrightarrow N_{0,r} \longrightarrow 0$$

with $J_{0,r} = (T_1^r, \ldots, T_n^r)$ and where $N_{0,r}$ has a finite filtration with quotients isomorphic to $A_0/I_0 = R$.

Proof. The assertion is stable under base change in R, so we may take $R = \mathbb{Z}$. Then B_0 is the A_0-subalgebra of $A_0 \otimes \mathbb{Q}$ generated by all $T_i^m/m!$. Let $M_{0,r} = B_0 \cap r^{-1} A_0$ inside $A_0 \otimes \mathbb{Q}$. Then $r^{-1} J_{0,r}$ is contained in $M_{0,r}$, and the quotient $N_{0,r}$ coincides with the image of $M_{0,r}$ in $(A_0/J_{0,r}) \otimes \mathbb{Q}$. Any maximal filtration of the latter by monomial ideals gives the required filtration of $N_{0,r}$. $\qquad \square$

Proof of Proposition A.1. We may assume that $S = \operatorname{Spec} R$, $X = \operatorname{Spec} A$ and $Y = \operatorname{Spec} A/I$, where I is the image of a regular map $f : A^r \to A$. We have $f(a) = \sum a_i x_i$ for a sequence $x_1, \ldots, x_r$ in A. Let $A_0 = \mathbb{Z}[T_1, \ldots, T_n]$ and $M_0 = A_0^n$ with $f_0 : M_0 \to A_0$ given by $a \mapsto \sum a_i T_i$. Let $I_0 = f_0(M_0)$. We consider the homomorphism $A_0 \to A$ defined by $T_i \mapsto x_i$. Let $B = \mathscr{D}_A(I)$ and $B_0 = \mathscr{D}_{A_0}(I_0)$ be the divided power envelopes. Since f and f_0 are regular, we have $I = I_0 \otimes_A A_0$. As in [Berthelot 1974, (3.4.8)] it follows that $B = B_0 \otimes_{A_0} A$. Using Lemma A.2, we get $B = \varinjlim M_r$ with $M_r = M_{0,r} \otimes_{A_0} A$. Moreover, since $\operatorname{Tor}_1^{A_0}(A_0/I_0, A) = 0$, we obtain exact sequences of A-modules

$$0 \longrightarrow J_r \longrightarrow M_r \longrightarrow N_r \longrightarrow 0$$

with $J_r = J_{0,r} \otimes_{A_0} A$ and $N_r = N_{0,r} \otimes_{A_0} A$, and we obtain filtrations of N_r with quotients isomorphic to A/I. Similarly there are exact sequences of A-modules

$$0 \longrightarrow J_r \longrightarrow A \longrightarrow A/J_r \longrightarrow 0$$

and filtrations of A/J_r with quotients isomorphic to A/I. Since A and A/I are flat over R, it follows that J_r and M_r and B are flat over R. $\qquad \square$

We will use the following example of regular immersions.

Lemma A.3. *For a ring R and a projective R-module T of finite type we consider the complete symmetric algebra $A = R[\![T]\!] = \prod_{n\geq 0} \mathrm{Sym}_R^n(T)$ and $M = T \otimes_R A$. Then the homomorphism $f : M \to A$ given by $t \otimes a \mapsto ta$ is regular.*

Proof. The complex $K_*(M, f)$ is the direct product over $m \geq 0$ of complexes $K_*^{(m)}$ with $K_n^{(m)} = \Lambda^n T \otimes_R \mathrm{Sym}^{m-n}(T)$, using the convention $\mathrm{Sym}^r(T) = 0$ for $r < 0$. Since the complexes $K_*^{(m)}$ are compatible with base change in R, the general case can be reduced to the case where T is free. Then an R-basis of T is a regular sequence in A, and the assertion follows. $\square$

Acknowledgements

References

[Berthelot 1974] P. Berthelot, *Cohomologie cristalline des schémas de caractéristique $p > 0$*, Lecture Notes in Mathematics **407**, Springer, New York, 1974. MR 52 #5676 Zbl 0298.14012

[Berthelot et al. 1982] P. Berthelot, L. Breen, and W. Messing, *Théorie de Dieudonné cristalline, II*, Lecture Notes in Mathematics **930**, Springer, Berlin, 1982. MR 85k:14023 Zbl 0516.14015

[Bourbaki 1983] N. Bourbaki, *Algèbre commutative: Chapitre 8: Dimension; Chapitre 9: Anneaux locaux noethériens complets*, Masson, Paris, 1983. MR 86j:13001 Zbl 0579.13001

[Breuil 1998] C. Breuil, "Schémas en groupes et corps des normes", unpublished manuscript, 1998, Available at http://www.math.u-psud.fr/~breuil/PUBLICATIONS/groupesnormes.pdf.

[Grothendieck 1974] A. Grothendieck, *Groupes de Barsotti–Tate et cristaux de Dieudonné*, Les Presses de l'Université de Montréal, Montreal, Que., 1974. Séminaire de Mathématiques Supérieures, No. 45 (Été, 1970). MR 54 #5250 Zbl 0331.14021

[Illusie 1985] L. Illusie, "Déformations de groupes de Barsotti–Tate (d'après A. Grothendieck)", pp. 151–198 in *Séminaire sur les pinceaux arithmétiques: la conjecture de Mordell* (Paris, 1983/84), edited by L. Szpiro, Astérisque **127**, 1985. MR 801922 Zbl 1182.14050

[Katz 1981] N. Katz, "Serre–Tate local moduli", pp. 138–202 in *Surfaces algébriques (Orsay, 1976–78)*, edited by J. Giraud et al., Lecture Notes in Math. **868**, Springer, Berlin, 1981. MR 83k:14039b Zbl 0477.14007

[Kim 2012] W. Kim, "The classification of p-divisible groups over 2-adic discrete valuation rings", *Math. Res. Lett.* **19**:1 (2012), 121–141. MR 2923180 Zbl 1284.14056

[Kisin 2006] M. Kisin, "Crystalline representations and F-crystals", pp. 459–496 in *Algebraic geometry and number theory*, edited by V. Ginzburg, Progr. Math. **253**, Birkhäuser, Boston, 2006. MR 2007j:11163 Zbl 1184.11052

[Kisin 2009] M. Kisin, "Modularity of 2-adic Barsotti–Tate representations", *Invent. Math.* **178**:3 (2009), 587–634. MR 2010k:11089

[Lau 2008] E. Lau, "Displays and formal p-divisible groups", *Invent. Math.* **171**:3 (2008), 617–628. MR 2009j:14058 Zbl 1186.14048

[Lau 2009] E. Lau, "A duality theorem for Dieudonné displays", *Ann. Sci. Éc. Norm. Supér.* (4) **42**:2 (2009), 241–259. MR 2010d:14065 Zbl 1182.14051

[Lau 2010] E. Lau, "Frames and finite group schemes over complete regular local rings", *Doc. Math.* **15** (2010), 545–569. MR 2011g:14107 Zbl 1237.14053

[Lau 2012] E. Lau, "Displayed equations for Galois representations", preprint, 2012. arXiv 1012.4436

[Lau 2013] E. Lau, "Smoothness of the truncated display functor", *J. Amer. Math. Soc.* **26**:1 (2013), 129–165. MR 2983008 Zbl 1273.14040

[Liu 2013] T. Liu, "The correspondence between Barsotti–Tate groups and Kisin modules when $p = 2$", *J. Théor. Nombres Bordeaux* **25**:3 (2013), 661–676. MR 3179680 Zbl 06291371

[Mazur and Messing 1974] B. Mazur and W. Messing, *Universal extensions and one dimensional crystalline cohomology*, Lecture Notes in Mathematics **370**, Springer, Berlin, 1974. MR 51 #10350 Zbl 0301.14016

[Messing 1972] W. Messing, *The crystals associated to Barsotti–Tate groups: with applications to abelian schemes*, Lecture Notes in Mathematics **264**, Springer, Berlin, 1972. MR 50 #337 Zbl 0243.14013

[Schlessinger 1968] M. Schlessinger, "Functors of Artin rings", *Trans. Amer. Math. Soc.* **130** (1968), 208–222. MR 36 #184 Zbl 0167.49503

[SGA 1970] M. Demazure and A. Grothendieck (editors), *Schémas en groupes, I: Propriétés générales des schémas en groupes* (Séminaire de Géométrie Algébrique du Bois Marie 1962/64 = SGA 3 I), Lecture Notes in Math. **151**, Springer, Berlin, 1970. MR 43 #223a Zbl 0207.51401

[SGA 1971] P. Berthelot, A. Grothendieck, and L. Illusie (editors), *Théorie des intersections et théorème de Riemann–Roch* (Séminaire de Géométrie Algébrique du Bois Marie 1966/67 = SGA 6), Lecture Notes in Math. **225**, Springer, Berlin, 1971. MR 50 #7133 Zbl 0218.14001

[Vasiu and Zink 2010] A. Vasiu and T. Zink, "Breuil's classification of p-divisible groups over regular local rings of arbitrary dimension", pp. 461–479 in *Algebraic and arithmetic structures of moduli spaces* (Sapporo, Japan, 2007), edited by I. Nakamura and L. Weng, Adv. Stud. Pure Math. **58**, Math. Soc. Japan, Tokyo, 2010. MR 2012a:14101 Zbl 1210.14049

[Viehmann and Wedhorn 2013] E. Viehmann and T. Wedhorn, "Ekedahl–Oort and Newton strata for Shimura varieties of PEL type", *Math. Ann.* **356**:4 (2013), 1493–1550. MR 3072810 Zbl 06194422

[Zink 2001a] T. Zink, "A Dieudonné theory for p-divisible groups", pp. 139–160 in *Class field theory — its centenary and prospect* (Tokyo, 1998), edited by K. Miyake, Adv. Stud. Pure Math. **30**, Math. Soc. Japan, Tokyo, 2001. MR 2002h:14075 Zbl 1052.14048

[Zink 2001b] T. Zink, "Windows for displays of p-divisible groups", pp. 491–518 in *Moduli of abelian varieties* (Texel Island, 1999), edited by C. Faber et al., Progr. Math. **195**, Birkhäuser, Basel, 2001. MR 2002c:14073 Zbl 1099.14036

[Zink 2002] T. Zink, "The display of a formal p-divisible group", pp. 127–248 in *Cohomologies p-adiques et applications arithmétiques, I*, edited by P. Berthelot et al., Astérisque **278**, Société Mathématique de France, 2002. MR 2004b:14083 Zbl 1008.14008

elau@math.upb.de   *Institut für Mathematik, Universität Paderborn,*
D-33098 Paderborn, Germany

Finiteness of unramified deformation rings

Patrick B. Allen and Frank Calegari

We prove that the universal unramified deformation ring R^{unr} of a continuous Galois representation $\bar{\rho} : G_{F^+} \to \mathrm{GL}_n(k)$ (for a totally real field F^+ and finite field k) is finite over $\mathbb{O} = W(k)$ in many cases. We also prove (under similar hypotheses) that the universal deformation ring R^{univ} is finite over the local deformation ring R^{loc}.

Introduction

Let k be a finite field of characteristic p, and let $\mathbb{O} = W(k)$. Let F be a number field, and consider a continuous absolutely irreducible Galois representation

$$\bar{\rho} : G_F \to \mathrm{GL}_n(k),$$

where $G_F = \mathrm{Gal}(\bar{F}/F)$ for some fixed algebraic closure $\bar{F}$ of F. If $(A, \mathfrak{m})$ is a complete local $\mathbb{O}$-algebra with residue field k, then a deformation ρ of $\bar{\rho}$ to A unramified outside a finite set of primes S consists an equivalence class of homomorphisms

$$\rho : G_F \to \mathrm{GL}_n(A)$$

such that the composite of ρ with the projection $\mathrm{GL}_n(A) \to \mathrm{GL}_n(A/\mathfrak{m}) = \mathrm{GL}_n(k)$ is $\bar{\rho}$, and such that the extension of fields $F(\ker(\rho))$ over $F(\ker(\bar{\rho}))$ is unramified away from places above primes in S (see [Mazur 1997]). The nature of such deformations is quite different depending on whether S contains the primes above p or not. If S contains all the primes above p, we denote the universal deformation ring by R^{univ}; if S contains no primes above p, we denote the corresponding universal deformation ring by R^{unr}. According to the Fontaine–Mazur conjecture (see [Fontaine and Mazur 1995, Conjecture 5a]), any map $R^{\mathrm{unr}} \to \overline{\mathbb{Q}}_p$ gives rise to a deformation ρ of $\bar{\rho}$ with finite image. (This form of the conjecture is known as the unramified Fontaine–Mazur conjecture.) Boston's strengthening of this conjecture

[Boston 1999, Conjecture 2 and the subsequent corollary] is the claim that the
universal unramified deformation

$$\rho^{\text{unr}} : G_F \to \text{GL}_n(R^{\text{unr}})$$

has finite image. In contrast, the ring R^{univ} is typically of large dimension (see
§1.10 of [Mazur 1989]). A conjecture of Mazur predicts that the relative dimension
of R^{univ} over $\mathcal{O}$ is (in odd characteristic)

$$(1 + r_2) + (n^2 - 1)[F : \mathbb{Q}] - \sum_{v|\infty} \dim H^0(D_v, \text{ad}^0(\bar{\rho})),$$

where $\text{ad}^0(\bar{\rho})$ denotes (in any choice of basis) the trace zero matrices in $\text{Hom}(\bar{\rho}, \bar{\rho})$.
A choice of basis for the universal deformation makes R^{univ} an algebra over a local
deformation ring

$$R^{\text{loc}} = \widehat{\bigotimes}_{v|p} R_v^{\text{loc}},$$

where R_v^{loc} is the universal framed local deformation ring of $\bar{\rho}|D_v$ for $v|p$. The
R^{loc}-algebra structure may depend on the choice of basis, but it is canonical up
to automorphisms of R^{loc}. It is not true in general that $\text{Spec}(R^{\text{univ}}) \to \text{Spec}(R^{\text{loc}})$
is a closed immersion, even in the minimal case where S is only divisible by
the primes dividing p. A simple example to consider is the deformation ring
of any one-dimensional representation $\bar{\rho} : G_F \to k^\times$; the corresponding map
$\text{Spec}(R^{\text{univ}}) \to \text{Spec}(R^{\text{loc}})$ is a closed immersion if and only if the maximal
everywhere-unramified abelian p-extension of F in which p splits completely
is trivial. It is, however, reasonable to conjecture that this map is always a finite
morphism. Indeed, one heuristic justification for the Fontaine–Mazur conjecture is
to imagine that the generic fibers of the image of $\text{Spec}(R^{\text{univ}})$ and the locus of local
crystalline representations of a fixed weight are transverse, and to infer (from a
conjectural computation of dimensions) that the intersection is finite, and hence that
there are only finitely many global crystalline representations of a fixed weight (see
pp. 191–192 of [Fontaine and Mazur 1995]); this line of thinking at least presumes
that the global-to-local map is quasifinite.

We prove the following:

Theorem 1. *Let F^+ be a totally real field, and let $\bar{\rho} : G_{F^+} \to \text{GL}_n(k)$ be a continuous absolutely irreducible representation. Suppose that*:

(1) $p > 2$.

(2) $\text{ad}^0(\bar{\rho}|G_{F^+(\zeta_p)})$ *is absolutely irreducible and $p > 2n^2 - 1$, or, if $n = 2$ and $\bar{\rho}$ is totally odd, $\bar{\rho}|G_{F^+(\zeta_p)}$ has adequate image.*

Then R^{unr} is a finite $\mathcal{O}$-algebra, and R^{univ} is a finite R^{loc}-algebra.

The second condition holds, for example, when $\bar{\rho}$ has image containing $\mathrm{SL}_n(k)$ and p is greater than $2n^2 - 1$. The finiteness of R^{univ} over R^{loc} can be deduced from an appropriate "$R = T$" theorem, since one proves that the maximal reduced quotient of R^{univ} modulo an ideal of R^{loc} is isomorphic to a finite $\mathbb{O}$-algebra T. However, in dimension greater than 2, without a conjugate self-dual assumption, the current $R = T$ theorems are contingent on conjectural properties of the cohomology of arithmetic quotients (see Part 2 of [Calegari and Geraghty 2012]).

We shall deduce from Theorem 1 the following corollaries:

Corollary 2. *For any $\bar{\rho}$ satisfying the conditions of Theorem 1, Boston's strengthening of the unramified Fontaine–Mazur conjecture is equivalent to the unramified Fontaine–Mazur conjecture.*

Corollary 3. *Suppose that $\bar{\rho} : G_{F^+} \to \mathrm{GL}_2(k)$ satisfies the conditions of Theorem 1. Assume further that:*

(1) *$\bar{\rho}$ is totally odd.*

(2) *If $p = 5$ and $\bar{\rho}$ has projective image $\mathrm{PGL}_2(\mathbb{F}_5)$, then $[F^+(\zeta_5) : F^+] = 4$.*

Then Boston's conjecture holds: the representation $\rho^{\mathrm{unr}} : G_{F^+} \to \mathrm{GL}_2(R^{\mathrm{unr}})$ has finite image.

When $n = 2$, $p > 2$, $F = \mathbb{Q}$, and $\bar{\rho}$ is totally odd and unramified at p, R^{unr} can be identified with the ring of Hecke operators acting on a (not necessarily torsion-free) coherent cohomology group (see [Calegari and Geraghty 2012]).

Let $\mathscr{G}_n$ be the group scheme over $\mathbb{Z}$ that is the semidirect product

$$(\mathrm{GL}_n \times \mathrm{GL}_1) \rtimes \{1, J\} = \mathscr{G}_n^0 \rtimes \{1, J\},$$

where J acts on $\mathrm{GL}_n \times \mathrm{GL}_1$ by $J(g, \mu)J^{-1} = (\mu^t g^{-1}, \mu)$. Let $\nu : \mathscr{G}_n \to \mathrm{GL}_1$ be the character that sends (g, μ) to μ and J to -1. Let F be a CM field with maximal totally real subfield F^+, and let

$$\bar{r} : G_{F^+} \to \mathscr{G}_n(k)$$

be a continuous homomorphism with $\bar{r}^{-1}(\mathscr{G}_n^0(k)) = G_F$. If $(A, \mathfrak{m})$ is a complete local $\mathbb{O}$-algebra with residue field k, then a deformation r of $\bar{r}$ to A unramified outside a finite set of primes S consists of an equivalence class of homomorphisms

$$r : G_{F^+} \to \mathscr{G}_n(A)$$

such that the composite of r with the projection $\mathscr{G}_n(A) \to \mathscr{G}_n(A/\mathfrak{m}) = \mathscr{G}_n(k)$ is $\bar{r}$, and such that the extension of fields $F(\ker(r))$ over $F(\ker(\bar{r}))$ is unramified away from places above primes in S. We say two lifts are equivalent if they are conjugate by an element of $\mathrm{GL}_n(A)$ that reduces to the identity modulo $\mathfrak{m}$. If $\bar{r}$ is Schur (see Definition 2.1.6 of [Clozel et al. 2008]), then this deformation problem is

representable. By abuse of notation, we will again denote the universal deformation ring of $\bar{r}$ by R^{univ} if S contains all the primes above p, and by R^{unr} if S contains no primes above p. This shouldn't cause any confusion, as we shall be very explicit regarding which deformation problem we refer to. As with the GL_n-valued theory, for each $v|p$ in F^+, there is a universal framed deformation ring $R_v^\square$ which represents the lifts of $\bar{r}|D_v$, and a choice of lift in the equivalence class of the universal deformation of $\bar{r}$ makes R^{univ} an algebra over

$$R^{\mathrm{loc}} = \widehat{\bigotimes}_{v|p} R_v^{\mathrm{loc}}.$$

We shall deduce Theorem 1 from the following result.

Theorem 4. *Let F be a CM field with maximal totally real subfield F^+. Let S denote a finite set of places of F^+ not containing any $v|p$, and let $\bar{r} : G_{F^+} \to \mathcal{G}_n(k)$ be a continuous homomorphism with $\bar{r}^{-1}(\mathcal{G}_n^0(k)) = G_F$ and such that $v \circ \bar{r}(c_v) = -1$ for each choice of complex conjugation c_v. Assume that $p \geq 2(n+1)$, that the image of $\bar{r}|G_{F(\zeta_p)}$ is adequate, and that $\zeta_p \notin F$. Let R^{unr} be the universal deformation ring of $\bar{r}$ unramified outside S, and let R^{univ} be the universal deformation ring of $\bar{r}$ unramified outside S and all primes $v|p$. Then R^{unr} is a finite $\mathbb{O}$-algebra, and R^{univ} is a finite R^{loc}-algebra.*

It turns out that the proof of this theorem is almost an immediate consequence of the finiteness results of [Thorne 2012] for ordinary deformation rings. The only required subtlety is to understand the relationship between the local ordinary deformation ring $R_{\Lambda_K}^{\triangle, ar}$ constructed in [Geraghty 2010] and the unramified local deformation ring R^{un}.

1. Some local deformation rings

Recall k is a finite field of characteristic p, and $\mathbb{O} = W(k)$. Let K be a finite extension of $\mathbb{Q}_p$ and let $G_K = \mathrm{Gal}(\bar{K}/K)$. Fix a continuous unramified representation

$$\bar{\rho} : G_K \to \mathrm{GL}_n(k)$$

and let $R^\square$ be its universal framed deformation ring. Let R^{un} be the quotient of $R^\square$ corresponding to unramified lifts.

Lemma 5. *The ring R^{un} is isomorphic to a power series ring over $\mathbb{O}$ in n^2 variables. In particular, it is reduced and its $\overline{\mathbb{Q}}_p$-points are Zariski dense in $\mathrm{Spec}(R^{\mathrm{un}})$.*

Proof. Fixing a choice of lift $g \in \mathrm{GL}_n(\mathbb{O})$ of $\bar{\rho}(\mathrm{Frob})$, it is easy to see that the lift to $\mathbb{O}[\![\{x_{ij}\}_{1 \leq i, j \leq n}]\!]$ given by $\mathrm{Frob} \mapsto g(I + (x_{ij}))$ is the universal framed deformation. $\square$

Let I_K^{ab} be the inertia subgroup of the abelianization of G_K, and let $I_K^{ab}(p)$ be its maximal pro-p quotient. Let $\Lambda_K = \mathbb{O}[\![(I_K^{ab}(p))^n]\!]$ and let $\psi = (\psi_1, \dots, \psi_n)$ be the universal n-tuple of characters $\psi_i : I_K \to \Lambda_K^\times$. Set $R_{\Lambda_K}^\square = R^\square \widehat{\otimes}_\mathbb{O} \Lambda_K$.

We briefly recall the construction of the universal ordinary deformation ring $R_{\Lambda_K}^\triangle$ by Geraghty (see §3 of [ibid.]). Let $\mathscr{F}$ be the flag variety over $\mathbb{O}$ whose S-points, for any $\mathbb{O}$-scheme S, is the set of increasing filtrations $0 = F_0 \subset F_1 \subset \cdots \subset F_n = \mathbb{O}_S^n$ of $\mathbb{O}_S^n$ by locally free submodules with $\operatorname{rank}(F_i) = i$ for each $i = 1, \dots, n$. Lemma 3.1.2 of [ibid.] shows that the subfunctor of

$$R_{\Lambda_K}^\square \otimes_\mathbb{O} \mathscr{F}$$

corresponding to pairs $(\rho, \{F_i\})$ such that

- $\{F_i\}$ is stabilized by ρ, and
- the action of I_K on F_i/F_{i-1} is given by the pushforward of ψ_i,

is represented by a closed subscheme $\mathscr{G}$. He then defines $R_{\Lambda_K}^\triangle$ as the image of

$$R_{\Lambda_K}^\square \to \mathbb{O}_\mathscr{G}(\mathscr{G}[1/p]).$$

Since scheme-theoretic image commutes with flat base change, $R_{\Lambda_K}^\triangle[1/p]$ is the scheme-theoretic image of

$$\mathscr{G}[1/p] \to \operatorname{Spec}(R_{\Lambda_K}^\square[1/p]).$$

Since this map is proper, $\mathscr{G}[1/p]$ surjects onto $\operatorname{Spec}(R_{\Lambda_K}^\triangle[1/p])$. Because $\mathscr{G}$ is of finite type over $R_{\Lambda_K}^\triangle$, we deduce that any $\overline{\mathbb{Q}}_p$-point of $\operatorname{Spec}(R_{\Lambda_K}^\triangle[1/p])$ lifts to a $\overline{\mathbb{Q}}_p$-point of $\mathscr{G}[1/p]$. This proves the following.

Lemma 6. *Let $x \in \operatorname{Spec}(R_{\Lambda_K}^\square)(\overline{\mathbb{Q}}_p)$, and let (ρ_x, ψ_x) denote the pushforward via x of the universal framed deformation and n-tuple of characters of I_K. Then x factors through $R_{\Lambda_K}^\triangle[1/p]$ if and only if there is a full flag $0 = F_0 \subset F_1 \subset \cdots \subset F_n = \overline{\mathbb{Q}}_p^n$ stabilized by ρ_x such that the action of I_K on F_i/F_{i-1} is given by $\psi_{i,x}$ for each $i = 1, \dots, n$.*

If $\bar{\rho}$ is the trivial representation, then Geraghty defines a further quotient $R_{\Lambda_K}^{\triangle, ar}$ of $R_{\Lambda_K}^\triangle$ as follows. Let $Q_1, \dots, Q_m$ be the minimal primes of Λ_K. For each $j = 1, \dots, m$, let $\mathscr{G}_j = \mathscr{G} \otimes_{\Lambda_K} \Lambda_K/Q_j$. Let $W_j \subset \operatorname{Spec}(\Lambda_K/Q_j)$ be the closed subscheme defined by $\psi_r = \epsilon_p \psi_s$ for some $1 \leq r < s \leq n$, and let U_j be the complement of W_j. Geraghty shows (see §3.4 of [ibid.]) that there is a unique irreducible component $\mathscr{G}_j^{ar}$ of $\mathscr{G}_j$ lying above U_j. We then set $\mathscr{G}^{ar} = \bigcup_{1 \leq j \leq m} \mathscr{G}_j^{ar}$ and define $R_{\Lambda_K}^{\triangle, ar}$ to be the image of

$$R_{\Lambda_K}^\triangle \to \mathbb{O}_{\mathscr{G}^{ar}}(\mathscr{G}^{ar}[1/p]).$$

The construction of $R_{\Lambda_K}^{\triangle, ar}$ together with Lemma 6 yields the following.

Lemma 7. *Assume that $\bar{\rho}$ is trivial. Let $x \in \operatorname{Spec}(R_{\Lambda_K}^{\square})(\overline{\mathbb{Q}}_p)$, and let (ρ_x, ψ_x) denote the pushforward via x of the universal framed deformation and n-tuple of characters of I_K. Assume that there is a full flag $0 = F_0 \subset F_1 \subset \cdots \subset F_n = \overline{\mathbb{Q}}_p^n$ stabilized by ρ_x such that the action of I_K on F_i/F_{i-1} is given by $\psi_{i,x}$ for each $i = 1, \ldots, n$. If $\psi_{i,x} \neq \epsilon_p \psi_{j,x}$ for any $i < j$, then x factors through $R_{\Lambda_K}^{\triangle, ar}$.*

Remark 8. If $[K : \mathbb{Q}_p] > \frac{1}{2}n(n-1)+1$ and $\bar{\rho}$ is trivial (which, for our applications, we could assume), then Thorne proves that $R_{\Lambda_K}^{\triangle, ar} = R_{\Lambda_K}^{\triangle}$ (see Corollary 3.12 of [Thorne 2014]).

There is a natural map $\Lambda_K \to R^{\mathrm{un}}$ given by modding out by the augmentation ideal $\mathfrak{a}$ of Λ_K. We thus have a natural surjection

$$R_{\Lambda_K}^{\square} \to R^{\mathrm{un}}.$$

Proposition 9. *The surjection $R_{\Lambda_K}^{\square} \to R^{\mathrm{un}}$ factors through $R_{\Lambda_K}^{\triangle}$. If $\bar{\rho}$ is trivial, then it further factors through $R_{\Lambda_K}^{\triangle, ar}$.*

Proof. The image of an unramified representation is the topological closure of the image of Frobenius. Since any element of $\mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ is conjugate to an upper triangular matrix, that the image of any unramified representation into $\mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ fixes a full flag for which the action of inertia on the corresponding quotients is trivial. It follows that the projection from $R_{\Lambda_K}^{\square}$ to any $\overline{\mathbb{Q}}_p$-point of R^{un} factors through $R_{\Lambda_K}^{\triangle}$ by Lemma 6 and, if $\bar{\rho}$ is trivial, through $R_{\Lambda_K}^{\triangle, ar}$, by Lemma 7. The result then follows from the fact that R^{un} is reduced and its $\overline{\mathbb{Q}}_p$-points are Zariski dense, by Lemma 5. $\qquad\square$

2. Proof of Theorem 4

We first prove the statement concerning R^{unr} over $\mathbb{O}$. Take a representation $\bar{r}$ as in Theorem 4. For each $v|p$ in F^+, let F_v^+ be the completion of F^+ at v and let $\Lambda_v = \Lambda_{F_v^+}$ with $\Lambda_{F_v^+}$ as in Section 1. Let $\Lambda = \widehat{\bigotimes}_{v|p, \mathbb{O}} \Lambda_v$.

We note that, using Lemma 1.2.2 of [Barnet-Lamb et al. 2014], we are free to make any base change disjoint from the fixed field of $\ker(\bar{r})$. After a base change, we may assume that $\bar{r}$ is everywhere unramified, and that $\bar{r}|D_v$ is trivial for all $v|p$ as well as any finite set of auxiliary primes. In particular, after a suitable base change, we may restrict ourselves to considering deformation rings which are unipotent at some finite set of auxiliary primes $v \in S$ (which corresponds to the local deformation condition $R_{\bar{v}}^1$ of [Thorne 2012, §8]). By Proposition 3.3.1 of [Barnet-Lamb et al. 2014], we may assume, after a further base change, that $\bar{r}$ lifts to a minimal crystalline ordinary modular representation (this is where we use the assumption that $p \geq 2(n+1)$). From Corollary 8.7 of [Thorne 2012], we deduce that the corresponding ordinary deformation ring $R_{\mathscr{G}}$ is finite over Λ. If we can

show that R^{unr} is a quotient of $R_{\mathcal{G}} \otimes \Lambda/\mathfrak{a}$, where $\mathfrak{a}$ is the augmentation ideal of Λ, then the result follows immediately by Nakayama, since $\Lambda/\mathfrak{a} = \mathcal{O}$. By definition, the local condition at $v|p$ for $R_{\mathcal{G}}$ is determined by the ordinary deformation ring $R_{\Lambda_v}^{\triangle,ar}$. By Proposition 9, the ring R^{un} is a quotient of $R_{\Lambda_v}^{\triangle,ar}/\mathfrak{a}$. Hence R^{unr} is a quotient of $R_{\mathcal{G}} \otimes \Lambda/\mathfrak{a}$ and we are done.

The finiteness of R^{univ} over R^{loc} then follows from the finiteness of R^{un} over $\mathcal{O}$ and Nakayama. Indeed, let $R^{\mathrm{split}} = R^{\mathrm{univ}} \otimes_{R^{\mathrm{loc}}} k$ and let r^{split} be the specialization of the universal deformation to R^{split}. Then $r^{\mathrm{split}}|D_v \cong \bar{r}|D_v$ for any $v|p$ in F^+, so the quotient $R^{\mathrm{univ}} \to R^{\mathrm{split}}$ factors through $R^{\mathrm{un}} \otimes_{\mathcal{O}} k$.

3. Some corollaries

3.1. *Proof of Theorem 1.* Let $\bar{\rho}$ satisfy the statement of Theorem 1. Consider $\mathrm{ad}^0(\bar{\rho})$ restricted to a suitable quadratic CM extension F/F^+. Since $p \nmid n$, the representation $\mathrm{ad}^0(\bar{\rho})$ is a direct summand of $\bar{\rho}^c \otimes \bar{\rho}^* = \bar{\rho} \otimes \bar{\rho}^*$ and is conjugate self-dual. The assumption of irreducibility together with the inequality $p > 2n^2 - 1$ imply that $\mathrm{ad}^0(\bar{\rho})$ is adequate by Theorem A.9 of [Thorne 2012]. If n is even, then $\mathrm{ad}^0(\bar{\rho})$ has odd dimension and so is automatically totally odd. If n is odd, then $\mathrm{ad}^0(\bar{\rho})$ is orthogonal (the conjugate self-duality is realized by the trace pairing, which is symmetric) and exactly self-dual (up to trivial twist) and so has trivial multiplier, which means that it is also totally odd. Both uses of totally odd refer to the properties of the multiplier character rather than the determinant of complex conjugation, and are the exact sign conditions required for automorphy lifting theorems for unitary groups (that is, totally odd means U-odd rather than GL-odd in the notation of [Calegari 2010]; see also §2.1 of [Barnet-Lamb et al. 2014]). Hence $\mathrm{ad}^0(\bar{\rho})|G_F$ extends to a homomorphism (see Lemma 2.1.1 of [Clozel et al. 2008])

$$\bar{r} : G_{F^+} \to \mathcal{G}_{n^2-1}(k),$$

which we fix, satisfying the conditions of Theorem 4. On the other hand, any deformation of $\bar{\rho}$ gives rise to a deformation of $\bar{r}$ in the natural way. By Yoneda's lemma, there is a corresponding morphism $R^{\mathrm{unr}}(\bar{r}) \to R^{\mathrm{unr}}(\bar{\rho})$. It suffices to prove this is finite. By Nakayama's lemma, this reduces to showing that the only deformations ρ of $\bar{\rho}$ to k-algebras such that $\mathrm{ad}^0(\rho)|G_F \cong \mathrm{ad}^0(\bar{\rho})|G_F$ are finite. The kernel of such a deformation must be contained in the maximal abelian pro-p extension of $F(\ker(\bar{\rho}))$ unramified outside S, which is finite by class field theory. As in the final paragraph of the proof of Theorem 4, the finiteness of R^{unr} implies the finiteness of R^{split} and hence that R^{univ} is a finite R^{loc}-algebra.

If $n = 2$ and $\bar{\rho}$ is totally odd, we may work directly with $\bar{\rho}$. We first use Corollary 1.7 of [Taylor 2002] to conclude that $\bar{\rho}$ is potentially modular and Theorem A of [Barnet-Lamb et al. 2013] to assume it is potentially ordinarily modular. Then,

restricting $\bar{\rho}$ to a suitable CM field F, the proof is exactly as in the proof of Theorem 4 (without the appeal to Proposition 3.3.1 of [Barnet-Lamb et al. 2014]).

3.2. *Proof of Corollary 2.* This follows immediately from Theorem 1 and the following proposition.

Proposition 10. *Let F be a number field and let $\bar{\rho} : G_F \to \mathrm{GL}_n(k)$ be continuous and absolutely irreducible. Then*

$$\rho^{\mathrm{unr}} : G_F \to \mathrm{GL}_n(R^{\mathrm{unr}})$$

has finite image if and only if the following two properties hold:

(1) *R^{unr} is finite over $\mathbb{O}$;*

(2) *for any minimal prime $\mathfrak{p}$ of $R^{\mathrm{unr}}[1/p]$, the induced representation*

$$G_F \to \mathrm{GL}_n(R^{\mathrm{unr}}[1/p]/\mathfrak{p})$$

   has finite image.

Proof. If ρ^{unr} has finite image, then (2) is clearly satisfied, and (1) follows from Théorème 2 of [Carayol 1994], which shows that R^{unr} is generated over $\mathbb{O}$ by traces.

Now assume (1) and (2), and let E be the fraction field of $\mathbb{O}$. Since R^{unr} is a finite $\mathbb{O}$-algebra, the map $R^{\mathrm{unr}} \to R^{\mathrm{unr}}[1/p]$ has finite kernel. Hence it suffices to prove that the map

$$\rho : G_F \to \mathrm{GL}_n(R^{\mathrm{unr}}[1/p])$$

has finite image, assuming (2). Since R^{unr} is finite over $\mathbb{O}$, the ring $R^{\mathrm{unr}}[1/p]$ is a semilocal ring which is a direct sum of Artinian E-algebras A with residue field H for some finite $[H : E] < \infty$. In particular, the representation ρ breaks up into a finite direct sum of representations to such groups $\mathrm{GL}_n(A)$. If $A = H$, then assumption (2) implies that the image of such a representation is finite. If $A \neq H$, then A admits a surjective map to $H[\epsilon]/\epsilon^2$. In particular, there exists an unramified deformation

$$\rho : G_F \to \mathrm{GL}_n(H[\epsilon]/\epsilon^2).$$

By assumption (2) again, the corresponding residual representation with image in $\mathrm{GL}_n(H)$ is finite, and is given by some representation V on which G_F acts through a finite group. Moreover, ρ is then given by some nontrivial extension

$$0 \to V \to W \to V \to 0.$$

Consider the restriction of this representation to a finite extension L/F such that G_L acts trivially on V. Then the action of G_L on W factors through an unramified $\mathbb{Z}_p$-extension, which must be trivial by class field theory. It follows that the action of G_L on W is trivial, and hence that the extension W is trivial, a contradiction. $\square$

3.3. *Proof of Corollary 3.* By Theorem 0.2 of [Pilloni and Stroh 2013] (see also [Kassaei 2013]), one knows the unramified Fontaine–Mazur conjecture for $\bar{\rho}$ under the given hypothesis, hence the result follows from Corollary 2.

Acknowledgements

References

[Barnet-Lamb et al. 2013] T. Barnet-Lamb, T. Gee, and D. Geraghty, "Congruences between Hilbert modular forms: constructing ordinary lifts, II", *Math. Res. Lett.* **20**:1 (2013), 67–72. MR 3126722 Zbl 06255950

[Barnet-Lamb et al. 2014] T. Barnet-Lamb, T. Gee, D. Geraghty, and R. Taylor, "Potential automorphy and change of weight", *Ann. of Math.* (2) **179**:2 (2014), 501–609. MR 3152941 Zbl 06284344

[Boston 1999] N. Boston, "Some cases of the Fontaine–Mazur conjecture, II", *J. Number Theory* **75**:2 (1999), 161–169. MR 2000b:11124 Zbl 0928.11050

[Calegari 2010] F. Calegari, "Even Galois representations", lecture notes, Institut Henri Poincaré, 2010, http://www.math.northwestern.edu/∼fcale/papers/FontaineTalk-Adjusted.pdf.

[Calegari and Geraghty 2012] F. Calegari and D. Geraghty, "Modularity lifting beyond the Taylor–Wiles method", preprint, 2012, http://arxiv.org/abs/1207.4224. arXiv 1207.4224

[Carayol 1994] H. Carayol, "Formes modulaires et représentations galoisiennes à valeurs dans un anneau local complet", pp. 213–237 in *p-adic monodromy and the Birch and Swinnerton-Dyer conjecture* (Boston, MA, 1991), edited by B. Mazur and G. Stevens, Contemp. Math. **165**, Amer. Math. Soc., Providence, RI, 1994. MR 95i:11059 Zbl 0812.11036

[Clozel et al. 2008] L. Clozel, M. Harris, and R. Taylor, "Automorphy for some *l*-adic lifts of automorphic mod *l* Galois representations", *Publ. Math. Inst. Hautes Études Sci.* **108** (2008), 1–181. MR 2010j:11082 Zbl 1169.11020

[Fontaine and Mazur 1995] J.-M. Fontaine and B. Mazur, "Geometric Galois representations", pp. 41–78 in *Elliptic curves, modular forms, & Fermat's last theorem* (Hong Kong, 1993), edited by J. Coates and S.-T. Yau, Ser. Number Theory **1**, Int. Press, Cambridge, MA, 1995. MR 96h:11049 Zbl 0839.14011

[Geraghty 2010] D. Geraghty, "Modularity lifting theorems for ordinary Galois representations", preprint, 2010, https://www2.bc.edu/david-geraghty/files/oml.pdf.

[Kassaei 2013] P. L. Kassaei, "Modularity lifting in parallel weight one", *J. Amer. Math. Soc.* **26**:1 (2013), 199–225. MR 2983010 Zbl 1296.11052

[Mazur 1989] B. Mazur, "Deforming Galois representations", pp. 385–437 in *Galois groups over* **Q** (Berkeley, CA, 1987), edited by Y. Ihara et al., Math. Sci. Res. Inst. Publ. **16**, Springer, New York, 1989. MR 90k:11057 Zbl 0714.11076

[Mazur 1997] B. Mazur, "An introduction to the deformation theory of Galois representations", pp. 243–311 in *Modular forms and Fermat's last theorem* (Boston, MA, 1995), edited by G. Cornell et al., Springer, New York, 1997. MR 1638481 Zbl 0901.11015

[Pilloni and Stroh 2013] V. Pilloni and B. Stroh, "Surconvergence, ramification et modularité", preprint, 2013, http://perso.ens-lyon.fr/vincent.pilloni/Artinfinal.pdf.

[Taylor 2002] R. Taylor, "Remarks on a conjecture of Fontaine and Mazur", *J. Inst. Math. Jussieu* **1**:1 (2002), 125–143. MR 2004c:11082 Zbl 1047.11051

[Thorne 2012] J. Thorne, "On the automorphy of *l*-adic Galois representations with small residual image", *J. Inst. Math. Jussieu* **11**:4 (2012), 855–920. Appendix by R. Guralnick, F. Herzig, R. Taylor and J. Thorne. MR 2979825 Zbl 1269.11054

[Thorne 2014] J. Thorne, "Automorphy lifting for residually reducible l-adic Galois representations", *J. Amer. Math. Soc.* (online publication June 2014).

pballen@math.northwestern.edu   *Mathematics Department, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208, United States*

fcale@math.northwestern.edu   *Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208, United States*

msp

On direct images of pluricanonical bundles

Mihnea Popa and Christian Schnell

We show that techniques inspired by Kollár and Viehweg's study of weak positivity, combined with vanishing for log-canonical pairs, lead to new generation and vanishing results for direct images of pluricanonical bundles. We formulate the strongest such results as Fujita conjecture-type statements, which are then shown to govern a range of fundamental properties of direct images of pluricanonical and pluriadjoint line bundles, like effective vanishing theorems, weak positivity, or generic vanishing.

1. Introduction

The purpose of this paper is twofold: on the one hand we show that techniques inspired by Kollár and Viehweg's study of weak positivity, combined with vanishing theorems for log-canonical pairs, lead to new consequences regarding generation and vanishing properties for direct images of pluricanonical bundles. On the other hand, we formulate the strongest such results as Fujita conjecture-type statements, which are then shown to govern a range of fundamental properties of direct images of pluricanonical and pluriadjoint line bundles, like effective vanishing theorems, weak positivity, or generic vanishing.

Vanishing, regularity, and Fujita-type statements. All varieties we consider in this paper are defined over an algebraically closed field of characteristic zero. Recall to begin with the following celebrated conjecture.

Conjecture 1.1 (Fujita). *If X is a smooth projective variety of dimension n, and L is an ample line bundle on X, then $\omega_X \otimes L^{\otimes l}$ is globally generated for $l \geq n+1$.*

It is well known that Fujita's conjecture holds in the case when L is ample and globally generated, based on Kodaira vanishing and the theory of Castelnuovo–Mumford regularity, and that this can be extended to the relative setting as follows:

Proposition 1.2 (Kollár). *Let $f : X \to Y$ be a morphism of projective varieties, with X smooth and Y of dimension n. If L is an ample and globally generated line bundle on Y, then*

$$R^i f_* \omega_X \otimes L^{\otimes n+1}$$

is 0-regular, and therefore globally generated for all i.

Recall that a sheaf $\mathcal{F}$ on Y is 0-*regular* with respect to an ample and globally generated line bundle L if

$$H^i(Y, \mathcal{F} \otimes L^{\otimes -i}) = 0 \quad \text{for all } i > 0.$$

The Castelnuovo–Mumford Lemma says that every 0-regular sheaf is globally generated (see, e.g., [Lazarsfeld 2004a, Theorem 1.8.3]); the proposition is then a consequence of Kollár's vanishing theorem, recalled as Theorem 2.2 below.

An extension of Fujita's general conjecture to the relative case was formulated by Kawamata [1982, Conjecture 1.3], and proved in dimension up to four; the statement is that Proposition 1.2 should remain true for any L ample, at least as long as the branch locus of the morphism f is a divisor with simple normal crossings support (when the sheaves $R^i f_* \omega_X$ are locally free [Kollár 1986]). However, at least for $i = 0$, we propose the following unconditional extension of Conjecture 1.1:

Conjecture 1.3. *Let $f : X \to Y$ be a morphism of smooth projective varieties, with Y of dimension n, and let L be an ample line bundle on Y. Then, for every $k \geq 1$, the sheaf*

$$f_* \omega_X^{\otimes k} \otimes L^{\otimes l}$$

is globally generated for $l \geq k(n+1)$.

Our main result in this direction is a proof of a stronger version of Conjecture 1.3 in the case of ample and globally generated line bundles, generalizing Proposition 1.2 for $i = 0$ to arbitrary powers.

Theorem 1.4. *Let $f : X \to Y$ be a morphism of projective varieties, with X smooth and Y of dimension n. If L is an ample and globally generated line bundle on Y, and $k \geq 1$ an integer, then*

$$f_* \omega_X^{\otimes k} \otimes L^{\otimes l}$$

is 0-regular, and therefore globally generated, for $l \geq k(n+1)$.

This follows in fact from a more general effective vanishing theorem for direct images of powers of canonical bundles, which is Kollár vanishing when $k = 1$; see Corollary 2.9. We also observe in Proposition 2.13 that just knowing the (klt version of the) Fujita-type Conjecture 1.3 for $k = 1$ would imply a similar vanishing theorem when L is only ample. Using related methods, we find analogous statements in the contexts of pluriadjoint bundles and of log-canonical pairs as well. We will call a *fibration* a surjective morphism whose general fiber is irreducible.

Variant 1.5. *Let $f : X \to Y$ be a fibration between projective varieties, with X smooth and Y of dimension n. Let M be a nef and f-big line bundle on X. If L is an ample and globally generated line bundle on Y, and $k \geq 1$ an integer, then*

$$f_* (\omega_X \otimes M)^{\otimes k} \otimes L^{\otimes l}$$

is 0-regular, and therefore globally generated, for $l \geq k(n+1)$.

Variant 1.6. *Let $f : X \to Y$ be a morphism of projective varieties, with X normal and Y of dimension n, and consider a log-canonical $\mathbb{R}$-pair (X, Δ) on X. Consider a line bundle B on X such that $B \sim_{\mathbb{R}} k(K_X + \Delta + f^*H)$ for some $k \geq 1$, where H is an ample $\mathbb{R}$-Cartier $\mathbb{R}$-divisor on Y. If L is an ample and globally generated line bundle on Y, and $k \geq 1$ an integer, then*

$$f_* B \otimes L^{\otimes l} \quad \text{with } l \geq k(n+1-t)$$

is 0-regular, and so globally generated, where $t := \sup\{s \in \mathbb{R} \mid H - sL \text{ is ample}\}$.[1]

All of these results are consequences of our main technical result, stated next. It can be seen both as an effective vanishing theorem for direct images of powers, and as a partial extension of Ambro–Fujino vanishing (recalled as Theorem 2.3 below) to arbitrary log-canonical pairs.

Theorem 1.7. *Let $f : X \to Y$ be a morphism of projective varieties, with X normal and Y of dimension n, and consider a log-canonical $\mathbb{R}$-pair (X, Δ) on X. Consider a line bundle B on X such that $B \sim_{\mathbb{R}} k(K_X + \Delta + f^*H)$ for some $k \geq 1$, where H*

---

[1]This is of course a generalization of Theorem 1.4. We chose to state it separately in order to preserve the simplicity of the main point, as will be done a few times throughout the paper.

is an ample $\mathbb{R}$*-Cartier* $\mathbb{R}$*-divisor on* Y. *If* L *is an ample and globally generated line bundle on* Y, *then*

$$H^i(Y, f_*B \otimes L^{\otimes l}) = 0 \quad \textit{for all } i > 0 \textit{ and } l \geq (k-1)(n+1-t) - t + 1,$$

where $t := \sup\{s \in \mathbb{R} \mid H - sL \textit{ is ample}\}$.

The proof of this result relies on a variation of a method used by Viehweg in the study of weak positivity, and on the use of the Ambro–Fujino vanishing theorem. Shifting emphasis from weak positivity to vanishing turns out to lead to stronger statements, as was already pointed out by Kollár [1986, §3] in the case $k = 1$; his point of view, essentially based on regularity, is indeed a crucial ingredient in the applications.

One final note in this regard is that all the vanishing theorems used in the paper hold for higher direct images as well. At the moment we do not know, however, how to obtain statements similar to those above for higher direct images, for instance for $R^i f_*\omega_X^{\otimes k}$ with $i > 0$.

Applications. The Fujita-type statements in Theorem 1.4 and its variants turn out to govern a number of fundamental properties of direct images of pluricanonical and pluriadjoint bundles. Besides the vanishing statements discussed above, we sample a few here, and refer to the main body of the paper for full statements. To begin with, we deduce in Section 4 an effective version of Viehweg's weak positivity theorem for sheaves of the form $f_*\omega_{X/Y}^{\otimes k}$ for arbitrary $k \geq 1$, just as Kollár did in the case $k = 1$; we leave the rather technical statement, Theorem 4.2, for the main text. The same method applies to pluriadjoint bundles (see Theorem 4.4) and in this case even the noneffective weak positivity consequence stated below is new. The case $k = 1$ is again due to Kollár and Viehweg; see also [Höring 2010, Theorem 3.30] for a nice exposition.

Theorem 1.8. *If* $f : X \to Y$ *is a fibration between smooth projective varieties, and* M *is a nef and* f*-big line bundle on* X, *then* $f_*(\omega_{X/Y} \otimes M)^{\otimes k}$ *is weakly positive for every* $k \geq 1$.

With this result at hand, Viehweg's machinery for studying Iitaka's conjecture can be applied to deduce the adjoint bundle analogue of his result on the additivity of the Kodaira dimension over a base of general type.

Theorem 1.9. *Let* $f : X \to Y$ *be a fibration between smooth projective varieties, and let* M *be a nef and* f*-big line bundle on* X. *We denote by* F *the general fiber of* f, *and by* M_F *the restriction of* M *to* F. *Then:*

(i) *If* L *is an ample line bundle on* Y, *and* $k > 0$, *then*

$$\kappa((\omega_X \otimes M)^{\otimes k} \otimes f^*L) = \kappa(\omega_F \otimes M_F) + \dim Y.$$

(ii) *If Y is of general type*, *then*

$$\kappa(\omega_X \otimes M) = \kappa(\omega_F \otimes M_F) + \dim Y.$$

In a different direction, the method involved in the proof of Theorem 1.4 (more precisely Corollary 2.9) leads to a generic vanishing statement for pluricanonical bundles. Let $f : X \to A$ be a morphism from a smooth projective variety to an abelian variety. Hacon [2004] showed that the higher direct images $R^i f_* \omega_X$ satisfy generic vanishing, i.e., are GV-sheaves on A; see Definition 5.1. This refines the well-known generic vanishing theorem of Green and Lazarsfeld [1987], and is crucial in studying the birational geometry of irregular varieties. In Section 5 we deduce the following statement, which is somewhat surprising given our previous knowledge about the behavior of powers of ω_X.

Theorem 1.10. *If $f : X \to A$ is a morphism from a smooth projective variety to an abelian variety*, *then $f_* \omega_X^{\otimes k}$ is a GV-sheaf for every $k \geq 1$.*

We also present a self-contained proof of this theorem based on an effective result, Proposition 5.2, which is weaker than Corollary 2.9, but has a more elementary proof of independent interest. Theorem 1.10 leads in turn to vanishing and generation consequences that are stronger than those for morphisms to arbitrary varieties; see Corollary 5.4. Similar statements are given for log-canonical pairs and for adjoint bundles in Variants 5.5 and 5.6.

2. Vanishing and freeness for direct images of pluri-log-canonical bundles

In this section we address results related to Conjecture 1.3, via vanishing theorems for direct images of pluricanonical bundles. The most general result we prove is for log-canonical pairs; this is of interest from a different perspective as well, as it partially extends a vanishing theorem of Ambro and Fujino.

Motivation and background. To motivate the main technical result, recall that, given an ample line bundle L on a smooth projective variety of dimension n, Conjecture 1.1 implies that $\omega_X \otimes L^{\otimes n+1}$ is a nef line bundle; this is in fact follows unconditionally from the fundamental theorems of the minimal model program. As a consequence, Kodaira vanishing implies that for every $k \geq 1$ one has

$$H^i(X, \omega_X^{\otimes k} \otimes L^{\otimes k(n+1)-n}) = 0 \quad \text{for all } i > 0, \tag{2.1}$$

an effective vanishing theorem for powers of ω_X.

We will look for similar results for direct images. Recall first that for $k = 1$ there is a well-known analogue of Kodaira vanishing, for all higher direct images.

Theorem 2.2 (Kollár vanishing [Kollár 1986, Theorem 2.1]). *Let $f : X \to Y$ be a morphism of projective varieties*, *with X smooth. If L is an ample line bundle on Y,*

then

$$H^j(Y, R^i f_* \omega_X \otimes L) = 0 \quad \text{for all } i \text{ and all } j > 0.$$

Moreover, of great use for the minimal model program are extensions of vanishing and positivity theorems to the log situation, in particular to log-canonical pairs; see, e.g., [Fujino 2011; 2014b]. For instance, Kollár's theorem above has an extension to this situation due to Ambro and Fujino; see, e.g., [Ambro 2003, Theorem 3.2] and [Fujino 2011, Theorem 6.3] (where a relative version can be found as well).

Theorem 2.3 (Ambro and Fujino). *Let $f : X \to Y$ be a morphism between projective varieties, with X smooth and Y of dimension n. Let (X, Δ) be a log-canonical log-smooth $\mathbb{R}$-pair,[2] and consider a line bundle B on X such that $B \sim_{\mathbb{R}} K_X + \Delta + f^* H$, where H is an ample $\mathbb{R}$-Cartier $\mathbb{R}$-divisor on Y. Then*

$$H^j(Y, R^i f_* B) = 0 \quad \text{for all } i \text{ and all } j > 0.$$

Just as Proposition 1.2 follows from Theorem 2.2 via the Castelnuovo–Mumford Lemma, so Theorem 2.3 has the following consequence:

Lemma 2.4. *Under the hypotheses of Theorem 2.3, consider in addition an ample and globally generated line bundle L on Y. Then*

$$R^i f_* B \otimes L^{\otimes n}$$

is 0-regular, and therefore globally generated, for all i.

Main technical result. We will now prove our main vanishing theorem, which can be seen as an extension of both vanishing of type (2.1) for powers of canonical bundles, for L ample and globally generated, and of Ambro–Fujino vanishing (Theorem 2.3) to log-canonical pairs with arbitrary Cartier index.

Proof of Theorem 1.7. Step 1. We will first show that we can reduce to the case when X is smooth, Δ has simple normal crossings support, and the image of the adjunction morphism

$$f^* f_* B \longrightarrow B$$

is a line bundle. *A priori* the image is $\mathfrak{b} \otimes B$, where $\mathfrak{b}$ is the relative base ideal of B. We consider a birational modification

$$\mu : \widetilde{X} \longrightarrow X$$

which is a common log-resolution of $\mathfrak{b}$ and (X, Δ). On $\widetilde{X}$ we can write

$$K_{\widetilde{X}} - \mu^*(K_X + \Delta) = P - N,$$

---

[2]This means that Δ is an effective $\mathbb{R}$-divisor with simple normal crossings support, and with the coefficient of each component at most equal to 1.

where P and N are effective $\mathbb{R}$-divisors with simple normal crossings support, without common components, and such that P is exceptional and all coefficients in N are at most 1. We consider the line bundle

$$\widetilde{B} := \mu^* B \otimes \mathcal{O}_{\widetilde{X}}(k\lceil P \rceil).$$

Note that by definition we have

$$\widetilde{B} \sim_{\mathbb{R}} k(K_{\widetilde{X}} + N + \lceil P \rceil - P + \mu^* f^* H).$$

Since $\lceil P \rceil$ is μ-exceptional, we have $\mu_* \widetilde{B} \simeq B$ for all k. Moreover,

$$\Delta_{\widetilde{X}} := N + \lceil P \rceil - P$$

is log-canonical with simple normal crossings support on $\widetilde{X}$.

Going back to the original notation, we can thus assume that X is smooth and Δ has simple normal crossings support, and the image sheaf of the adjunction morphism is of the form $B \otimes \mathcal{O}_X(-E)$ for a divisor E such that $E + \Delta$ has simple normal crossings support.

Step 2. Now since L is ample, there is a smallest integer $m \geq 0$ such that $f_* B \otimes L^{\otimes m}$ is globally generated, and so using the adjunction morphism we have that $B \otimes \mathcal{O}_X(-E) \otimes f^* L^{\otimes m}$ is globally generated as well. We can then write

$$B \otimes f^* L^{\otimes m} \simeq \mathcal{O}_X(D + E),$$

where D is an irreducible smooth divisor, not contained in the support of $E + \Delta$, and such that $D + E + \Delta$ has simple normal crossings support. Rewriting this in divisor notation, we have

$$k(K_X + \Delta + f^* H) + m f^* L \sim_{\mathbb{R}} D + E,$$

and hence

$$(k-1)(K_X + \Delta + f^* H) \sim_{\mathbb{R}} \frac{k-1}{k} D + \frac{k-1}{k} E - \frac{k-1}{k} \cdot m f^* L. \qquad (2.5)$$

Note that Δ and E may have common components in their support, which may cause trouble later on; therefore their coefficients need to be adjusted conveniently. Let's start by writing

$$\Delta = \sum_{i=1}^{l} a_i D_i, \quad a_i \in \mathbb{R} \text{ with } 0 < a_i \leq 1,$$

and

$$E = \sum_{i=1}^{l} s_i D_i + E_1, \quad s_i \in \mathbb{N},$$

where the support of E_1 and that of Δ have no common components.

Observe now that for every effective Cartier divisor $E' \preceq E$ we have

$$f_*(B \otimes \mathcal{O}_X(-E')) \simeq f_*B. \tag{2.6}$$

Indeed, it is enough to have this for E itself; but this is the base locus of B relative to f, so by construction we have

$$f^*f_*B \to B \otimes \mathcal{O}_X(-E) \hookrightarrow B,$$

and so the isomorphism follows by pushing forward to get the commutative diagram

$$f_*B \xrightarrow{\quad} f_*(B \otimes \mathcal{O}_X(-E)) \xrightarrow{\quad} f_*B.$$
$$\overset{\text{id}}{\overbrace{\qquad\qquad\qquad\qquad\qquad}}$$

Define now

$$\gamma_i := a_i + \frac{k-1}{k} \cdot s_i \quad \text{for } i = 1, \ldots, l.$$

We claim that we can find for each i an integer b_i such that

$$0 \le \gamma_i - b_i \le 1 \quad \text{and} \quad 0 \le b_i \le s_i.$$

This is the same as $\gamma_i - 1 \le b_i \le \gamma_i$, while on the other hand, $\gamma_i < 1 + s_i$, so it is clear that such integers exist. We define

$$E' := \sum_{i=1}^{l} b_i D_i + \left\lfloor \frac{k-1}{k} E_1 \right\rfloor \le E,$$

and for this divisor (2.6) applies.

Step 3. Using (2.5), for any integer l we can now write

$$B - E' + lf^*L \sim_{\mathbb{R}} K_X + \Delta + \frac{k-1}{k}E - E' + \frac{k-1}{k}D + f^*\left(H + \left(l - \frac{k-1}{k} \cdot m\right)L\right).$$

We first note that the $\mathbb{R}$-divisor

$$H' := H + \left(l - \frac{k-1}{k} \cdot m\right)L$$

on Y is ample provided $l + t - ((k-1)/k)m > 0$. On the other hand, the effective $\mathbb{R}$-divisor with simple normal crossings support

$$\Delta' := \Delta + \frac{k-1}{k}E - E' + \frac{k-1}{k}D$$

on X is log-canonical. Indeed, the only coefficients that could cause trouble are those of the D_i. Note however that these are equal to $\gamma_i - b_i$, which are between 0

and 1 by our choice of b_i. Putting everything together, it means that on X (which is now smooth) we have written

$$B - E' + lf^*L \sim_{\mathbb{R}} K_X + \Delta' + f^*H',$$

where Δ' is log-canonical with simple normal crossings support, and H' is ample on Y. The pushforward of the left-hand side is $f_*B \otimes L^{\otimes l}$, while for the right-hand side we can now apply Theorem 2.3 to conclude that

$$H^i(Y, f_*B \otimes L^{\otimes l}) = 0 \quad \text{for all } i > 0 \text{ and } l > \frac{k-1}{k} \cdot m - t. \tag{2.7}$$

We therefore have that for every $l > ((k-1)/k)m - t + n$ the sheaf $f_*B \otimes L^{\otimes l}$ is 0-regular, hence globally generated. Given our minimal choice of m, we conclude that for the smallest integer l_0 which is greater than $((k-1)/k)m - t$ we have $m \leq l_0 + n$. This implies

$$m \leq l_0 + n \leq \frac{k-1}{k} \cdot m + n + 1 - t,$$

which is equivalent to $m \leq k(n+1-t)$, and in particular the vanishing in (2.7) holds for

$$l \geq (k-1)(n+1-t) - t + 1. \qquad \square$$

Note that the inequality $m \leq k(n+1-t)$ obtained above implies the statement of Variant 1.6. Just as with the statement of Theorem 1.7 compared to that of the Ambro–Fujino Theorem 2.3, one notes that, even for $k = 1$, Variant 1.6 is slightly more general than the case $i = 0$ in Lemma 2.4. This is not surprising, but particular to zeroth direct images, as after passing to a log-resolution of the log-canonical pair there is no need to appeal to local vanishing for higher direct images; the Ambro–Fujino theorem and the lemma cannot be stated in this form in the case $i > 0$.

Special cases. We spell out the most important special cases of Theorem 1.7. They are obtained by taking $H = L$ in the statement of Theorem 1.7, so that $t = 1$.

Corollary 2.8. *Let $f : X \to Y$ be a morphism of projective varieties, with X normal and Y of dimension n. Consider a log-canonical pair (X, Δ) and an integer $k > 0$ such that $k(K_X + \Delta)$ is Cartier. If L is an ample and globally generated line bundle on Y, then*

$$H^i(Y, f_*\mathcal{O}_X(k(K_X + \Delta)) \otimes L^{\otimes l}) = 0 \quad \text{for all } i > 0 \text{ and } l \geq k(n+1) - n.$$

In particular, we have an extension of (2.1) to direct images, and of Proposition 1.2 for $i = 0$ to arbitrary k:

Corollary 2.9. *Let* $f : X \to Y$ *be a morphism of projective varieties*, *with* X *smooth and* Y *of dimension* n. *If* L *is an ample and globally generated line bundle on* Y, *and* $k > 0$ *is an integer*, *then*

$$H^i(Y, f_*\omega_X^{\otimes k} \otimes L^{\otimes l}) = 0 \quad \text{for all } i > 0 \text{ and } l \geq k(n+1) - n.$$

Remark. Note that if we perform the proof of Theorem 1.7 only in the "classical" case considered in Corollary 2.9, the second step is unnecessary since $\Delta = 0$, while Δ' in the third step is klt. This means that one does not need to appeal to the Ambro–Fujino vanishing theorem, but rather to the klt version of Theorem 2.2, still due to Kollár; see for instance [Kollár 1995, Theorem 10.19].

The regularity statement in the introduction is an immediate consequence.

Proof of Theorem 1.4. We note that $k(n + 1) = k(n + 1) - n + n$, and apply the vanishing statement in Corollary 2.9 by successively subtracting n powers of L. $\square$

A rephrasing of Theorem 1.4 is a useful uniform global generation statement involving powers of relative canonical bundles.

Corollary 2.10. *Let* $f : X \to Y$ *be a morphism of smooth projective varieties*, *with* Y *of dimension* n. *If* L *is an ample and globally generated line bundle on* Y, $k \geq 1$ *an integer*, *and* $A := \omega_Y \otimes L^{\otimes n+1}$, *then*

$$f_*\omega_{X/Y}^{\otimes k} \otimes A^{\otimes k}$$

is globally generated.

Question. The arguments leading to Corollary 2.9, and more generally Theorem 1.7 and its applications, do not extend to higher direct images. It is natural to ask, however, whether the statements do hold for all $R^i f_*\omega_X^{\otimes k}$ and analogues, just as Theorem 2.2 and Theorem 2.3 do.

Example: the main conjecture over curves. We record one case when the main Fujita-type Conjecture 1.3 can be shown to hold, namely when the base of the morphism has dimension one. This is not hard to check, but it uses important special facts about vector bundles on curves.

Proposition 2.11. *Let* $f : X \to C$ *be a morphism of smooth projective varieties*, *with* C *a curve*, *and let* L *be an ample line bundle on* C. *Then*, *for every* $k \geq 1$, *the vector bundle*

$$f_*\omega_X^{\otimes k} \otimes L^{\otimes m}$$

is globally generated for $m \geq 2k$.

Proof. First, note that the sheaf in question is locally free (since C is a curve). We can rewrite it as

$$f_*\omega_X^{\otimes k} \otimes L^{\otimes m} \simeq f_*\omega_{X/C}^{\otimes k} \otimes \omega_C^{\otimes k} \otimes L^{\otimes m}.$$

Now Theorem 1 of [Kawamata 2002] says that $f_*\omega_{X/C}^{\otimes k}$ is a semipositive vector bundle on C, while

$$\deg \omega_C^{\otimes k} \otimes L^{\otimes m} \geq k(2g-2) + m \deg L \geq 2g,$$

with g the genus of C, as $\deg L > 0$. The statement then follows from the following general result. $\square$

Lemma 2.12. *Let E be a semipositive vector bundle and L a line bundle of degree at least $2g$ on a smooth projective curve C of genus g. Then $E \otimes L$ is globally generated.*

Proof. It is enough to show that, for every $p \in C$, one has

$$H^1(C, E \otimes L \otimes \mathcal{O}_C(-p)) = 0,$$

or equivalently, by Serre duality, that there are no nontrivial homomorphisms

$$E \longrightarrow \omega_C \otimes \mathcal{O}_C(p) \otimes L^{-1}.$$

But the semipositivity of E means precisely that it cannot have any quotient line bundle of negative degree. $\square$

Remark. For curves of genus at least 1, the argument in Proposition 2.11 shows, in fact, that $f_*\omega_X^{\otimes k} \otimes L^{\otimes 2}$ is always globally generated.

Relative Fujita conjecture and vanishing for ample line bundles. It is worth observing that it suffices to know Conjecture 1.3 and its variants for $k = 1$ in order to obtain vanishing theorems for twists by line bundles that are assumed to be just ample, and not necessarily globally generated. For simplicity we spell out only the case of pluricanonical bundles, i.e., the analogue of Corollary 2.9.

Proposition 2.13. *Assume that Conjecture 1.3 holds for $k = 1$.[3] Then for any morphism $f : X \to Y$ of smooth projective varieties with Y of dimension n, any ample line bundle L on Y, and any integer $k \geq 2$, one has*

$$H^i(Y, f_*\omega_X^{\otimes k} \otimes L^{\otimes l}) = 0 \quad \text{for all } i > 0 \text{ and } l \geq k(n+1) - n.$$

---

[3]Or, more precisely, its klt version: if Y is smooth of dimension n, L is ample on Y, and (X, Δ) is a klt pair such that $B = K_X + \Delta + \alpha f^*L$ is Cartier for some $\alpha \in \mathbb{R}$, then $f_*B \otimes L^{\otimes l}$ is globally generated for any $l + \alpha \geq n + 1$.

Proof. This is a corollary of the proof of Corollary 2.9. Indeed, the only time we used that L is globally generated and not just ample was to deduce the global generation of a sheaf of the form $f_* B \otimes L^{\otimes l}$ from its 0-regularity with respect to L, where B is $\mathbb{Q}$-linearly equivalent to something of the form $K_X + \Delta + \alpha f^* L$ with (X, Δ) klt and $\alpha \in \mathbb{Q}$; see also the remark on page 2282. The klt version of Conjecture 1.3 for $k = 1$ would then serve as a replacement. $\square$

A natural version of Conjecture 1.3 can be stated in the log-canonical case,[4] with the same effect regarding the result of Theorem 1.7, but this would take us far beyond what is currently known.

3. Vanishing and freeness for direct images of pluriadjoint bundles

We now switch our attention to direct images of powers of line bundles of the form $\omega_X \otimes M$, where M is a nef and relatively big line bundle. Recall first that Proposition 1.2 has the following analogue:

Proposition 3.1. *Let $f : X \to Y$ be a fibration between projective varieties, with X smooth and Y of dimension n. Consider a nef and f-big line bundle M on X, and (X, Δ) a klt pair with Δ an $\mathbb{R}$-divisor with simple normal crossings support. If B is a line bundle on X such that $B \sim_{\mathbb{R}} K_X + M + \Delta + f^* H$ for some ample $\mathbb{R}$-Cartier $\mathbb{R}$-divisor H on Y, then*

$$H^i(Y, f_* B) = 0 \quad \text{for all } i > 0.$$

In particular, if L is an ample and globally generated line bundle on Y, then

$$f_* B \otimes L^{\otimes n}$$

is 0-regular, and therefore globally generated.

Proof. We include the well-known proof for completeness, as it is usually given in the case $\Delta = 0$ (see, e.g., [Höring 2010, Lemma 3.28]). Note first that $M + f^* H$ continues to be a nef and f-big $\mathbb{R}$-divisor on X. The local version of the Kawamata–Viehweg vanishing theorem (see [Lazarsfeld 2004b, Remarks 9.1.22 and 9.1.23]) applies then to give

$$R^i f_* B = 0 \quad \text{for all } i > 0.$$

We conclude that it is enough to show

$$H^i(X, B) = 0 \quad \text{for all } i > 0.$$

This will follow from the global $\mathbb{R}$-version of Kawamata–Viehweg vanishing as soon as we show that $M + f^* H$ is in fact a big divisor. Since it is nef, it suffices

---

[4]In the absolute case, an Angehrn–Siu type statement has been obtained by Kollár [1997, Theorem 5.8] in the klt case, and further extended by Fujino [2010, Theorem 1.1] to the log-canonical setting.

to check that $(M + f^*H)^m > 0$, where $m = \dim X$. Now $(M + f^*H)^m$ is a linear combination with positive coefficients of terms of the form

$$M^s \cdot f^* H^{m-s},$$

which are all nonnegative. Moreover, since M is f-big, the term $M^{m-n} \cdot f^* H^n$ is strictly positive, which gives the conclusion. $\square$

We now prove an analogue of Corollary 2.9 in this context. Just as with Theorem 1.4, Variant 1.5 is its immediate consequence.

Theorem 3.2. *Let* $f : X \to Y$ *be a fibration between projective varieties, with* X *smooth and* Y *of dimension* n. *Let* M *be a nef and* f-*big line bundle on* X. *If* L *is an ample and globally generated line bundle on* Y, *and* $k \geq 1$ *an integer, then*

$$H^i(Y, f_*(\omega_X \otimes M)^{\otimes k} \otimes L^{\otimes l}) = 0 \quad \text{for all } i > 0 \text{ and } l \geq k(n+1) - n.$$

Proof. The strategy is similar to that of the proof of Theorem 1.7, so we will be brief in some of the steps. We consider the minimal $m \geq 0$ such that $f_*(\omega_X \otimes M)^{\otimes k} \otimes L^{\otimes m}$ is globally generated. Using the adjunction morphism

$$f^* f_*(\omega_X \otimes M)^{\otimes k} \to (\omega_X \otimes M)^{\otimes k},$$

after possibly blowing up, we can write

$$(\omega_X \otimes M)^{\otimes k} \otimes f^* L^{\otimes m} \simeq \mathcal{O}_X(D + E),$$

with D smooth and $D + E$ a divisor with simple normal crossings support. In divisor notation, we obtain

$$K_X + M \sim_{\mathbb{Q}} \frac{1}{k}D + \frac{1}{k}E - \frac{m}{k}f^*L. \tag{3.3}$$

For any integer $l \geq 0$, using (3.3) we can then write the equivalence

$$k(K_X + M) - \left\lfloor \frac{k-1}{k}E \right\rfloor + lf^*L$$

$$= K_X + M + (k-1)(K_X + M) - \left\lfloor \frac{k-1}{k}E \right\rfloor + lf^*L$$

$$\sim_{\mathbb{Q}} K_X + M + \Delta + \left(l - \frac{k-1}{k} \cdot m\right)f^*L,$$

where

$$\Delta = \frac{k-1}{k}D + \frac{k-1}{k}E - \left\lfloor \frac{k-1}{k}E \right\rfloor$$

is a boundary divisor with simple normal crossings support. Since E is the base divisor of $(\omega_X \otimes M)^{\otimes k}$ relative to f, just as in the proof of Theorem 1.7 it follows that

$$f_*\mathcal{O}_X\left(k(K_X + M) - \left\lfloor \frac{k-1}{k}E \right\rfloor + lf^*L\right) \simeq f_*(\omega_X \otimes M)^{\otimes k} \otimes L^{\otimes l}.$$

On the other hand, on the right-hand side we can apply Proposition 3.1, to deduce that

$$H^i(Y, f_*(\omega_X \otimes M)^{\otimes k} \otimes L^{\otimes l}) = 0 \quad \text{for all } i > 0 \text{ and } l > \frac{k-1}{k} \cdot m.$$

We conclude that $f_*(\omega_X \otimes M)^{\otimes k} \otimes L^{\otimes l}$ is globally generated for $l > ((k-1)/k)m+n$. Since m was chosen minimal, we conclude as in Theorem 1.7 that $m \leq k(n+1)$, and that vanishing holds for all $l \geq k(n+1) - n$. $\qquad\square$

Remark. Fujita's conjecture and all similar statements have more refined numerical versions, replacing $L^{\otimes n+1}$ by any ample line bundle A such that $A^{\dim V} \cdot V > (\dim V)^{\dim V}$ for any subvariety $V \subseteq X$. Similarly, the analogues of Conjecture 1.3 and Proposition 2.13 make sense replacing ω_X by $\omega_X \otimes M$ as well.

4. Effective weak positivity, and additivity of adjoint Iitaka dimension

Recall the following fundamental definition (see, e.g., [Viehweg 1983, §1]):

Definition 4.1. A torsion-free coherent sheaf $\mathscr{F}$ on a projective variety X is *weakly positive* on a nonempty open set $U \subseteq X$ if for every ample line bundle A on X and every $a \in \mathbb{N}$, the sheaf $S^{[ab]}\mathscr{F} \otimes A^{\otimes b}$ is generated by global sections at each point of U for b sufficiently large. (Here $S^{[p]}\mathscr{F}$ denotes the reflexive hull of the symmetric power $S^p\mathscr{F}$.) As noted in [Viehweg 1983, Remark 1.3], it is not hard to see that it is enough to check this definition for a fixed line bundle A.

Kollár [1986, §3] introduced an approach to proving the weak positivity of sheaves of the form $f_*\omega_{X/Y}$ based on his vanishing theorem for $f_*\omega_X$, which in particular gives effective statements. Here we first provide a complement to Kollár's result, using Theorem 1.4, in order to make this approach work for all $f_*\omega_{X/Y}^{\otimes k}$ with $k \geq 1$. Concretely, below is the analogue of [Kollár 1986, Theorem 3.5(i)]; the proof is very similar, and we only sketch it for convenience.

Theorem 4.2. *Let $f : X \to Y$ be a surjective morphism of smooth projective varieties, with generically reduced fibers in codimension one.[5] Let L be an ample and globally generated line bundle on Y, and $A = \omega_Y \otimes L^{\otimes n+1}$, where $n = \dim Y$. Then for every $s \geq 1$, the sheaf*

$$f_*(\omega_{X/Y}^{\otimes k})^{[\otimes s]} \otimes A^{\otimes k}$$

is globally generated over a fixed open set U containing the smooth locus of f; here $f_*(\omega_{X/Y}^{\otimes k})^{[\otimes s]}$ denotes the reflexive hull of $f_*(\omega_{X/Y}^{\otimes k})^{\otimes s}$.

---

[5]This means that there exists a closed subset $Z \subset Y$ of codimension at least two such that over $Y - Z$ the fibers of f are generically reduced. This condition is realized for instance if there is such a Z such that over $Y - Z$ the branch locus of f is smooth, and its preimage is a simple normal crossings divisor; see [Kollár 1986, Lemma 3.4].

Proof. As in [Viehweg 1983, §3] and in the proof of [Kollár 1986, Theorem 3.5], based on Viehweg's fiber product construction one can show that there is an open set $U \subset Y$, whose complement $Y - U$ has codimension at least two, over which there exists a morphism

$$\varphi : f_*^{(s)}(\omega_{X^{(s)}/Y}^{\otimes k}) \longrightarrow f_*(\omega_{X/Y}^{\otimes k})^{[\otimes s]}$$

which is an isomorphism over the smooth locus of f. Here $\mu : X^{(s)} \to X^s$ is a desingularization of the unique irreducible component X^s of the s-fold fiber product of X over Y which dominates Y; we have natural morphisms $f^s : X^s \to Y$ and $f^{(s)} = f^s \circ \mu : X^{(s)} \to Y$. The reason one can do this for any $k \geq 1$ is this: the hypothesis on the morphism implies that X^s is normal and Gorenstein over such a U (contained in the flat locus of f) with complement of small codimension; see also [Höring 2010, Lemma 3.12]. In particular, for every $k \geq 1$ there is a morphism

$$t : \mu_* \omega_{X^{(s)}/Y}^{\otimes k} \longrightarrow \omega_{X^s/Y}^{\otimes k}$$

which induces φ.

Now, without changing the notation, we can pass to a compactification of $X^{(s)}$, and the morphism φ extends to a morphism of sheaves on Y, since it is defined in codimension one and the sheaf on the right is reflexive. Corollary 2.10 says that

$$f_*^{(s)}(\omega_{X^{(s)}/Y}^{\otimes k}) \otimes A^{\otimes k}$$

is globally generated for all s and k, which implies that

$$f_*(\omega_{X/Y}^{\otimes k})^{[\otimes s]} \otimes A^{\otimes k}$$

is generated by global sections over the locus where φ is an isomorphism.   □

Corollary 4.3 [Viehweg 1983, Theorem III]. *If $f : X \to Y$ is a surjective morphism of smooth projective varieties, then $f_* \omega_{X/Y}^{\otimes k}$ is weakly positive for every $k \geq 1$.*

This follows in standard fashion from Theorem 4.2, by passing to semistable reduction along the lines of [Viehweg 1983, Lemma 3.2 and Proposition 6.1]. This was already noted by Kollár [1986, Corollary 3.7 and the preceding comments] in the case $k = 1$. As mentioned above, the theorem has the advantage of producing an effective bound, at least for sufficiently nice morphisms. We note also that Fujino [2014a] has used the argument above in order to deduce results on the semipositivity of direct images of pluricanonical bundles.

We now switch our attention to the context of direct images of adjoint line bundles of the form $\omega_X \otimes M$, where M is a nef and f-big line bundle for a fibration $f : X \to Y$. Given Theorem 3.2, we are now able to use the cohomological approach to weak positivity for higher powers of adjoint bundles as well. Concretely, Theorem 1.8

again follows via Viehweg's semistable reduction methods from the following analogue of the effective Theorem 4.2.

Theorem 4.4. *Let $f : X \to Y$ be a fibration between smooth projective varieties, with generically reduced fibers in codimension one. Let M be a nef and f-big line bundle on X, L an ample and globally generated line bundle on Y, and $A = \omega_Y \otimes L^{\otimes n+1}$ with $n = \dim Y$. Then*

$$f_*((\omega_{X/Y} \otimes M)^{\otimes k})^{[\otimes s]} \otimes A^{\otimes k}$$

is globally generated over a fixed nonempty open set U for any $s \geq 1$.

Proof. Using the notation in the proof of Theorem 4.2, over the same open subset $U \subset Y$ with complement of codimension at least two, one has a morphism which is generically an isomorphism:

$$\varphi : f_*^{(s)}((\omega_{X^{(s)}/Y} \otimes M^{(s)})^{\otimes k}) \longrightarrow f_*((\omega_{X/Y} \otimes M)^{\otimes k})^{[\otimes s]}. \quad (4.5)$$

Here $M^{(s)}$ is the line bundle on the desingularization $X^{(s)}$ defined inductively as

$$M^{(s)} := p_1^* M \otimes p_2^* M^{(s-1)},$$

with p_1 and p_2 the projections of $X^{(s)}$ to X and $X^{(s-1)}$, respectively. The morphism in (4.5) is obtained as a consequence of flatness and the projection formula; an excellent detailed discussion of the case $k = 1$, as well as of this whole circle of ideas, can be found in [Höring 2010, §3.D], in particular Lemma 3.15 and Lemma 3.24. The case $k > 1$ follows completely analogously, given the morphism t in the proof of Theorem 4.2.

Finally, Variant 1.5 immediately gives the analogue of Corollary 2.10 for twists by nef and relatively big line bundles, implying that $f_*^{(s)}((\omega_{X^{(s)}/Y} \otimes M^{(s)})^{\otimes k}) \otimes A^{\otimes k}$ is globally generated for all s and k. Combined with the reflexivity of the right-hand side, this leads to the desired conclusion. $\square$

We conclude by noting that Corollary 4.3 has a natural extension to the setting of log-canonical pairs; see [Campana 2004, §4], and also [Fujino 2014b, §6]. It is an interesting and delicate problem to obtain an analogue of Theorem 4.2 in this setting as well.

Subadditivity of Iitaka dimension for adjoint bundles. Theorem 1.8 allows us to make use of an argument developed by Viehweg in order to provide the analogue in the adjoint setting of [Viehweg 1983, Corollary IV] on the subadditivity of Kodaira dimension for fibrations with base of general type.

Proof of Theorem 1.9. Note that the $\leq$ inequalities are consequences of the easy addition formula; see [Mori 1987, Corollary 1.7]. The proof of the reverse inequalities closely follows the ideas of Viehweg [1983] based on the use of weak positivity,

as streamlined by Mori with the use of a result of Fujita; we include it below for completeness. Namely, we will apply the following lemma (but not directly for the line bundles on the left-hand side in (i) and (ii)).

Lemma 4.6 [Fujita 1977, Proposition 1; Mori 1987, Lemma 1.14]. *Let $f : X \to Y$ be a fibration with general fiber F, and N a line bundle on X. Then there exists a big line bundle L on Y and an integer $m > 0$ with $f^*L \hookrightarrow N^{\otimes m}$ if and only if*

$$\kappa(N) = \kappa(N_F) + \dim Y.$$

To make use of this, note first that, according to [Viehweg 1983, Lemma 7.3], there exists a smooth birational modification $\tau : Y' \to Y$ and a resolution X' of $X \times_Y Y'$ giving a commutative diagram

$$
\begin{array}{ccc}
X' & \xrightarrow{\tau'} & X \\
\downarrow{\scriptstyle f'} & & \downarrow{\scriptstyle f} \\
Y' & \xrightarrow{\tau} & Y
\end{array}
$$

with the property that every effective divisor B on X' that is exceptional for f' lies in the exceptional locus of τ'. Note that in this case $\tau'_* \omega_{X'}^{\otimes k}(kB) \simeq \omega_X^{\otimes k}$ for every $k \geq 0$. Also, τ'^*M is still nef and f'-big.

Fix now an ample line bundle L on Y, and consider the big line bundle $L' = \tau^*L$ on Y'. By Theorem 1.8 we have that for any $k > 0$ (which we can assume to be such that $f'_*(\omega_{X'/Y'} \otimes \tau'^*M)^{\otimes k} \neq 0$) there exists $b > 0$ such that

$$S^{[2b]} f'_*(\omega_{X'/Y'} \otimes \tau'^*M)^{\otimes k} \otimes L'^{\otimes b}$$

is generically globally generated. On the other hand, there exists an effective divisor B on X', exceptional for f', such that the reflexive hull of

$$f'_*(\omega_{X'/Y'} \otimes \tau'^*M)^{\otimes p}$$

is equal to

$$f'_*(\omega_{X'/Y'}(B) \otimes \tau'^*M)^{\otimes p}$$

for every $p \leq kb$. Using the nontrivial map induced by multiplication of sections on the fibers, we obtain that

$$f'_*(\omega_{X'/Y'}(B) \otimes \tau'^*M)^{\otimes 2kb} \otimes L'^{\otimes b}$$

has a nonzero section, and hence we obtain an inclusion

$$f'^*L'^{\otimes b} \hookrightarrow (\omega_{X'/Y'}(B) \otimes \tau'^*M)^{\otimes 2kb} \otimes f'^*L'^{\otimes 2b}.$$

According to Lemma 4.6, we obtain that

$$\kappa\big((\omega_{X'/Y'}(B) \otimes \tau'^*M)^{\otimes k} \otimes f'^*L'\big) = \kappa(\omega_{F'} \otimes (\tau'^*M)_{F'}) + \dim Y'$$
$$= \kappa(\omega_F \otimes M_F) + \dim Y,$$

where F' is the general fiber of f'.

To deduce (i), note that, as we have observed that $\tau'_* \omega_{X'}^{\otimes k}(kB) \simeq \omega_X^{\otimes k}$, we have

$$\tau'_*\big((\omega_{X'/Y'}(B) \otimes \tau'^*M)^{\otimes k} \otimes f'^*L'\big) \simeq (\omega_{X/Y} \otimes M)^{\otimes k} \otimes f^*L.$$

To deduce (ii), since Y' is of general type, recall that by Kodaira's lemma there exists an inclusion $L' \hookrightarrow \omega_{Y'}^{\otimes r}$ for some $r > 0$. This implies that

$$\kappa(\omega_X \otimes M) = \kappa(\omega_{X'}(B) \otimes \tau'^*M) \geq \kappa\big((\omega_{X'/Y'}(B) \otimes \tau'^*M)^{\otimes r} \otimes f'^*L'\big),$$

which is equal to $\kappa(\omega_F \otimes M_F) + \dim Y$ by the above. $\qquad\square$

5. Generic vanishing for direct images of pluricanonical bundles

We concentrate now on the case of morphisms $f : X \to A$, where X is a smooth projective variety and A is an abelian variety. We denote by P the normalized Poincaré bundle on the product $A \times \mathrm{Pic}^0(A)$, and by P_α its restriction to the slice $A \times \{\alpha\}$; this is of course just a different name for the point $\alpha \in \mathrm{Pic}^0(A)$.

Definition 5.1 [Pareschi and Popa 2011a, Definition 3.1]. A coherent sheaf $\mathscr{F}$ on X is called a *GV-sheaf* (with respect to the given morphism f) if it satisfies

$$\mathrm{codim}\{\alpha \in \mathrm{Pic}^0(A) \mid H^k(X, \mathscr{F} \otimes f^*P_\alpha) \neq 0\} \geq k$$

for every $k \geq 0$.

If f is generically finite, then by a special case of the generic vanishing theorem of Green and Lazarsfeld [1987], ω_X is a GV-sheaf. This was generalized by Hacon [2004] to the effect that for an arbitrary f the higher direct images $R^i f_* \omega_X$ are GV-sheaves on A for all i. On the other hand, there exist simple examples showing that even when f is generically finite, the powers $\omega_X^{\otimes k}$ with $k \geq 2$ are not necessarily GV-sheaves; see [Pareschi and Popa 2011a, Example 5.6]. Therefore Theorem 1.10 in the introduction is a quite surprising application of the methods in this paper.

Proof of Theorem 1.10. Let M be a very high power of an ample line bundle on $\widehat{A}$, and let $\varphi_M : \widehat{A} \to A$ be the isogeny induced by M. According to a criterion of Hacon [2004, Corollary 3.1], the assertion will be proved if we manage to show that

$$H^i(\widehat{A}, \varphi_M^* f_* \omega_X^{\otimes k} \otimes M) = 0 \quad \text{for all } i > 0.$$

Equivalently, we need to show that

$$H^i(\widehat{A}, g_*\omega_{X_1}^{\otimes k} \otimes M) = 0 \quad \text{for all } i > 0,$$

where $g : X_1 \to \widehat{A}$ is the base change of $f : X \to A$ via φ_M. We can, however, perform another base change $\mu : \widehat{A} \to \widehat{A}$ by a multiplication map of large degree, such that $\mu^* M \simeq L^{\otimes d}$, where L is an ample line bundle, which we can also assume to be globally generated, and d is arbitrarily large. The situation is summarized in the diagram

$$\begin{array}{ccccc}
X_2 & \longrightarrow & X_1 & \longrightarrow & X \\
\downarrow{\scriptstyle h} & & \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle f} \\
\widehat{A} & \xrightarrow{\ \mu\ } & \widehat{A} & \xrightarrow{\ \varphi_M\ } & A
\end{array}$$

It is then enough to show that

$$H^i(\widehat{A}, h_*\omega_{X_2}^{\otimes k} \otimes L^{\otimes d}) = 0 \quad \text{for all } i > 0.$$

Note that we cannot apply Serre vanishing here, as all of our constructions depend on the original choice of M. However, we can conclude if we know that there exists a bound $d = d(n, k)$, i.e., depending only on $n = \dim A$ and k, such that the vanishing in question holds for any morphism h.

At this stage we can of course apply Corollary 2.9, which allows us to take $d \geq k(n+1) - n$. We stress, however, that as long as we know that such a uniform bound for d exists, for this argument its precise shape does not matter. We therefore choose to present below a weaker but more elementary result that does not need vanishing theorems for $\mathbb{Q}$-divisors, making the argument self-contained.

Indeed, Proposition 5.2 below shows that there exists a morphism $\varphi : Z \to \widehat{A}$ with Z smooth projective, and $m \leq n + k$, such that $h_*\omega_{X_2}^{\otimes k} \otimes L^{\otimes m(k-1)}$ is a direct summand in $\varphi_*\omega_Z$. Applying Kollár vanishing (Theorem 2.2), we deduce that

$$H^i(\widehat{A}, h_*\omega_{X_2}^{\otimes k} \otimes L^{\otimes d}) = 0 \quad \text{for all } i > 0 \text{ and all } d \geq (n+k)(k-1) + 1,$$

which suffices to conclude the proof. $\qquad\square$

Proposition 5.2. *Let $f : X \to Y$ be a morphism of projective varieties, with X smooth and Y of dimension n. Let L be an ample and globally generated line bundle on Y, and $k \geq 1$ an integer. Then there exists a smooth projective variety Z with a morphism $\varphi : Z \to Y$, and an integer $0 \leq m \leq n+k$, such that $f_*\omega_X^{\otimes k} \otimes L^{\otimes m(k-1)}$ is a direct summand in $\varphi_*\omega_Z$.*

Proof. This is closer in spirit to the arguments towards weak positivity used in [Viehweg 1983, §5]. Note first that $f_*\omega_X^{\otimes k} \otimes L^{\otimes pk}$ is globally generated for some sufficiently large p. Denote by m the minimal $p \geq 0$ for which this is satisfied.

We are going to use a branched covering construction to show that $m \leq n + k$. First, consider the adjunction morphism

$$f^* f_* \omega_X^{\otimes k} \longrightarrow \omega_X^{\otimes k}.$$

After blowing up on X, if necessary, we can assume that the image sheaf is of the form $\omega_X^{\otimes k} \otimes \mathcal{O}_X(-E)$ for a divisor E with normal crossing support. As $f_* \omega_X^{\otimes k} \otimes L^{\otimes mk}$ is globally generated, we have that the line bundle

$$\omega_X^{\otimes k} \otimes f^* L^{\otimes mk} \otimes \mathcal{O}_X(-E)$$

is globally generated as well. It is therefore isomorphic to $\mathcal{O}_X(D)$, where D is an irreducible smooth divisor, not contained in the support of E, such that $D + E$ still has normal crossings. We have arranged that

$$(\omega_X \otimes f^* L^{\otimes m})^{\otimes k} \simeq \mathcal{O}_X(D + E),$$

and so we can take the associated covering of X branched along $D + E$ and resolve its singularities. This gives us a generically finite morphism $g : Z \to X$ of degree k, and we denote $\varphi = f \circ g : Z \to Y$.

Now by a well-known calculation of Esnault and Viehweg [Viehweg 1983, Lemma 2.3], the direct image $g_* \omega_Z$ contains the sheaf

$$\omega_X \otimes (\omega_X \otimes f^* L^{\otimes m})^{\otimes k-1} \otimes \mathcal{O}_X\left(-\left\lfloor \frac{k-1}{k}(D + E) \right\rfloor\right)$$

$$\simeq \omega_X^{\otimes k} \otimes f^* L^{\otimes m(k-1)} \otimes \mathcal{O}_X\left(-\left\lfloor \frac{k-1}{k} E \right\rfloor\right)$$

as a direct summand. If we now apply f_*, we find that

$$f_*\left(\omega_X^{\otimes k} \otimes \mathcal{O}_X\left(-\left\lfloor \frac{k-1}{k} E \right\rfloor\right)\right) \otimes L^{\otimes m(k-1)} \tag{5.3}$$

is a direct summand of $\varphi_* \omega_Z$. At this point we observe, as in the proof of Theorem 1.7, that, since E is the relative base locus of $\omega_X^{\otimes k}$, we have

$$f_*\left(\omega_X^{\otimes k} \otimes \mathcal{O}_X\left(-\left\lfloor \frac{k-1}{k} E \right\rfloor\right)\right) \simeq f_* \omega_X^{\otimes k}.$$

In other words, $f_* \omega_X^{\otimes k} \otimes L^{\otimes m(k-1)}$ is a direct summand in $\varphi_* \omega_Z$. Applying Proposition 1.2, we deduce in turn that $f_* \omega_X^{\otimes k} \otimes L^{\otimes m(k-1)+n+1}$ is globally generated. By our minimal choice of m, this is only possible if

$$m(k - 1) + n + 1 \geq (m - 1)k + 1,$$

which is equivalent to $m \leq n + k$. $\hspace{2cm} \square$

Remark. With slightly more clever choices, the integer m in Proposition 5.2 can be chosen to satisfy $m \leq n + 2$, but the effective vanishing consequence is still

weaker than that obtained in Corollary 2.9. Note also that one can show analogous results in the case of log-canonical pairs and of adjoint bundles, with only small additional technicalities.

Going back to the case when the base is an abelian variety, once we know generic vanishing the situation is in fact much better than what we obtained for morphisms to arbitrary varieties.

Corollary 5.4. *If* $f : X \to A$ *is a morphism from a smooth projective variety to an abelian variety, for every ample line bundle L on A and every $k \geq 1$ one has*:

  (i) $f_* \omega_X^{\otimes k}$ *is a nef sheaf on A.*

 (ii) $H^i(A, f_* \omega_X^{\otimes k} \otimes L) = 0$ *for all $i > 0$.*

(iii) $f_* \omega_X^{\otimes k} \otimes L^{\otimes 2}$ *is globally generated.*

Proof. For (i), note that every GV-sheaf is nef by [Pareschi and Popa 2011b, Theorem 4.1]. Part (ii) follows from the more general fact that the tensor product of a GV-sheaf with an IT_0 locally free sheaf is IT_0; see [ibid., Proposition 3.1]. Finally, (iii) follows from [Pareschi and Popa 2003, Theorem 2.4], as by (ii) $f_* \omega_X^{\otimes k} \otimes L$ is an M-regular sheaf on A. □

Question. It is again natural to ask whether, given a morphism $f : X \to A$, the higher direct images $R^i f_* \omega_X^{\otimes k}$ are GV-sheaves for all i.

The exact same method, with appropriate technical modifications, gives the following analogues for log-canonical pairs and pluriadjoint bundles, either based on Corollary 2.8 and Theorem 3.2, or on the analogues of Proposition 5.2; we will not repeat the argument.

Variant 5.5. *Let $f : X \to A$ be a morphism from a normal projective variety to an abelian variety. If (X, Δ) is a log-canonical pair and $k \geq 1$ is any integer such that $k(K_X + \Delta)$ is Cartier, then $f_* \mathcal{O}_X(k(K_X + \Delta))$ is a GV-sheaf for every $k \geq 1$.*

Variant 5.6. *Let $f : X \to A$ be a fibration between a smooth projective variety and an abelian variety, and M a nef and f-big line bundle on X. Then $f_*(\omega_X \otimes M)^{\otimes k}$ is a GV-sheaf for every $k \geq 1$.*

Acknowledgements

References

[Ambro 2003] F. Ambro, "Quasi-log varieties", *Tr. Mat. Inst. Steklova* **240** (2003), 220–239. In Russian; translated in *Proc. Steklov Inst. Math.* **240** (2003), 214–233. MR 2004f:14027 Zbl 1081.14021

[Campana 2004] F. Campana, "Orbifolds, special varieties and classification theory", *Ann. Inst. Fourier* (*Grenoble*) **54**:3 (2004), 499–630. MR 2006c:14013 Zbl 1062.14014

[Fujino 2010] O. Fujino, "Effective base point free theorem for log canonical pairs, II. Angehrn–Siu type theorems", *Michigan Math. J.* **59**:2 (2010), 303–312. MR 2011h:14016 Zbl 1201.14010

[Fujino 2011] O. Fujino, "Fundamental theorems for the log minimal model program", *Publ. Res. Inst. Math. Sci.* **47**:3 (2011), 727–789. MR 2012h:14031 Zbl 1234.14013

[Fujino 2014a] O. Fujino, "Direct images of pluricanonical divisors", preprint, 2014. arXiv 1409.7437

[Fujino 2014b] O. Fujino, "Notes on the weak positivity theorems", preprint, 2014. arXiv 1406.1834

[Fujita 1977] T. Fujita, "Some remarks on Kodaira dimensions of fiber spaces", *Proc. Japan Acad. Ser. A Math. Sci.* **53**:1 (1977), 28–30. MR 58 #1276 Zbl 0399.14024

[Green and Lazarsfeld 1987] M. Green and R. Lazarsfeld, "Deformation theory, generic vanishing theorems, and some conjectures of Enriques, Catanese and Beauville", *Invent. Math.* **90**:2 (1987), 389–407. MR 89b:32025 Zbl 0659.14007

[Hacon 2004] C. D. Hacon, "A derived category approach to generic vanishing", *J. Reine Angew. Math.* **575** (2004), 173–187. MR 2005m:14026 Zbl 1137.14012

[Höring 2010] A. Höring, "Positivity of direct image sheaves—a geometric point of view", *Enseign. Math.* (2) **56**:1-2 (2010), 87–142. MR 2011g:14022 Zbl 1203.14011

[Kawamata 1982] Y. Kawamata, "Kodaira dimension of algebraic fiber spaces over curves", *Invent. Math.* **66**:1 (1982), 57–71. MR 83h:14025 Zbl 0461.14004

[Kawamata 2002] Y. Kawamata, "On a relative version of Fujita's freeness conjecture", pp. 135–146 in *Complex geometry* (Göttingen, 2000), edited by I. Bauer et al., Springer, Berlin, 2002. MR 2003g:14056 Zbl 1058.14010

[Kollár 1986] J. Kollár, "Higher direct images of dualizing sheaves I", *Ann. of Math.* (2) **123**:1 (1986), 11–42. MR 87c:14038 Zbl 0598.14015

[Kollár 1995] J. Kollár, *Shafarevich maps and automorphic forms*, Princeton University Press, Princeton, NJ, 1995. MR 96i:14016 Zbl 0871.14015

[Kollár 1997] J. Kollár, "Singularities of pairs", pp. 221–287 in *Algebraic geometry* (Santa Cruz, 1995), edited by J. Kollár et al., Proc. Sympos. Pure Math. **62**, Amer. Math. Soc., Providence, RI, 1997. MR 99m:14033 Zbl 0905.14002

[Lazarsfeld 2004a] R. Lazarsfeld, *Positivity in algebraic geometry, I*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) **48**, Springer, Berlin, 2004. MR 2005k:14001a Zbl 1093.14501

[Lazarsfeld 2004b] R. Lazarsfeld, *Positivity in algebraic geometry, II*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) **49**, Springer, Berlin, 2004. MR 2005k:14001b Zbl 1093.14500

[Mori 1987] S. Mori, "Classification of higher-dimensional varieties", pp. 269–331 in *Algebraic geometry* (Bowdoin College, Brunswick, Maine, 1985), edited by S. J. Bloch, Proc. Sympos. Pure Math. **46**, Amer. Math. Soc., Providence, RI, 1987. MR 89a:14040 Zbl 0656.14022

[Pareschi and Popa 2003] G. Pareschi and M. Popa, "Regularity on abelian varieties I", *J. Amer. Math. Soc.* **16**:2 (2003), 285–302. MR 2004c:14086 Zbl 1022.14012

[Pareschi and Popa 2011a] G. Pareschi and M. Popa, "GV-sheaves, Fourier–Mukai transform, and generic vanishing", *Amer. J. Math.* **133**:1 (2011), 235–271. MR 2012e:14043 Zbl 1208.14015

[Pareschi and Popa 2011b] G. Pareschi and M. Popa, "Regularity on abelian varieties III: relationship with generic vanishing and applications", pp. 141–167 in *Grassmannians, moduli spaces and vector bundles* (Cambridge, MA, 2006), edited by D. A. Ellwood and E. Previato, Clay Math. Proc. **14**, Amer. Math. Soc., 2011. MR 2012h:14113 Zbl 1236.14020

[Viehweg 1983] E. Viehweg, "Weak positivity and the additivity of the Kodaira dimension for certain fibre spaces", pp. 329–353 in *Algebraic varieties and analytic varieties* (Tokyo, 1981), edited by S. Iitaka, Adv. Stud. Pure Math. **1**, North-Holland, Amsterdam, 1983. MR 85b:14041 Zbl 0513.14019

mpopa@math.northwestern.edu

Department of Mathematics, Northwestern University,
2033 Sheridan Road, Evanston, IL 60208, United States

cschnell@math.sunysb.edu   Department of Mathematics, Stony Brook University,
Stony Brook, NY 11794, United States

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the ANT website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *ANT* are usually in English, but articles written in other languages are welcome.

Length There is no a priori limit on the length of an *ANT* article, but *ANT* considers long articles only if the significance-to-length ratio is appropriate. Very long manuscripts might be more suitable elsewhere as a memoir instead of a journal article.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

Algebra & Number Theory

Volume 8   No. 9   2014