Aims

The *Seminario Matematico* is a society of members of mathematics-related departments of the University and the Politecnico in Turin. Its scope is to promote study and research in all fields of mathematics and its applications, and to contribute to the diffusion of mathematical culture.

The *Rendiconti* is the official journal of the Seminario Matematico. It publishes papers and invited lectures. Papers should provide original contributions to research, while invited lectures will have a tutorial or overview character. Other information and the contents of recent issues are available at `http://seminariomatematico.dm.unito.it/rendiconti/`

Instructions to Authors

Authors should submit an electronic copy (preferably in the journal's style, see below) via

e-mail: `rend_sem_mat@unito.it`

English is the preferred language, but papers in Italian or French may also be submitted. The paper should incorporate the name and affiliation of each author, postal and e-mail addresses, a brief abstract, and MSC classification numbers. Authors should inform the executive editor as soon as possible in the event of any change of address.

The final decision concerning acceptance of a paper is taken by the editorial board, based upon the evaluation of an external referee. Papers are typically processed within two weeks of arrival, and referees are asked to pass on their reports within two months.

Authors should prepare their manuscripts in LaTeX, with a separate BibTeX file whenever possible. To avoid delays in publication, it is essential to provide a LaTeX version once a paper has been accepted. The appropriate style file and instructions can be downloaded from the journal's website. The editorial board reserves the right to refuse publication if the file does not meet these requirements.

Reprints are not normally offered.

RENDICONTI DEL SEMINARIO MATEMATICO-UNIVERSITÀ E POLITECNICO DI TORINO

Università e Politecnico di Torino

CONTENTS

RENDICONTI DEL SEMINARIO MATEMATICO-UNIVERSITÀ
E POLITECNICO DI TORINO

Autorizzazione del Tribunale di Torino N. 2962 del 6.VI.1980

DIRETTORE RESPONSABILE

Alberto Collino

QUESTO FASCICOLO È PRODOTTO CON IL CONTRIBUTO DI:
UNIVERSITÀ DEGLI STUDI DI TORINO
POLITECNICO DI TORINO

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 5 – 12

S. Barbero, U. Cerruti, N. Murru

ON POLYNOMIAL SOLUTIONS OF THE DIOPHANTINE EQUATION $(X+Y-1)^2 = WXY$

Abstract. In this paper we consider a particular class of polynomials arising from the solutions of the Diophantine equation $(x+y-1)^2 = wxy$. We highlight some interesting aspects, describing their relationship with many iportant integer sequences and pointing out their connection with Dickson and Chebyshev polynomials. We also study their coefficients finding a new identity involving Catalan numbers and proving that they are a Riordan array.

1. A class of polynomials related to integer sequences, Dickson and Chebyshev polynomials

In [1], the authors solved the Diophantine equation

$$(x+y-1)^2 = wxy, \tag{1}$$

where w is a given positive integer and x, y are unknown numbers, whose values are to be sought in the set of positive integers.
In particular, (x,y) is a solution of the Diophantine equation (1) if and only if $(x,y) = (u_{m+1}(w), u_m(w))$, for a given $m \in \mathbb{N}$, where $(u_n(w))_{n=0}^{+\infty}$ is the following linear recurrent sequence:

$$\begin{cases} u_0(w) = 0, \quad u_1(w) = 1, \quad u_2(w) = w \\ u_n(w) = (w-1)u_{n-1}(w) - (w-1)u_{n-2}(w) + u_{n-3}(w) \quad \forall n \geq 3. \end{cases} \tag{2}$$

This polynomial sequence is very interesting. Indeed, for several values of w, the polynomial sequence $(u_n(w))$ coincides with some well–known and studied integer sequences. For example, for $w = 4$, $(u_n(4)) = n^2$, that is the sequence A000290 in OEIS [7]. When $w = 5$, $(u_n(5))$ is the sequence of the alternate Lucas numbers minus 2 (see sequence A004146 in OEIS). If $w = 9$, $(u_n(9)) = F_{2n}^2$, where (F_n) is the sequence of the Fibonacci numbers. For $w = 4, \ldots, 20$, the sequence $(u_n(w))$ appears in OEIS [7]. In Table 1, we summarize sequences $u_n(w)$ for different values of w.

In the following, we prove that polynomials $u_n(w)$ are related to some well–known and studied polynomials like Chebyshev polynomials of the first and second kind, respectively $T_n(x)$ and $U_n(x)$ (see, e.g., [5]), and Dickson polynomials $D_n(x)$ and $E_n(x) = U_n\left(\frac{x}{2}\right)$ (see, e.g., [3]).
Here we define $T_n(x)$ and $U_n(x)$ as the n–th element of the linear recurrent sequence $(T_n(x))_{n=0}^{+\infty}$ and $(U_n(x))_{n=0}^{+\infty}$ with characteristic polynomial $t^2 - 2xt + 1$ and initial conditions $T_0(x) = 1$, $T_1(x) = x$ and $U_0(x) = 1$, $U_1(x) = 2x$, respectively.

w	$(u_n(w))_{n=0}^{+\infty}$	OEIS reference
4	$0,1,4,9,16,25,...$	A000290=$(n^2)_{n=0}^{+\infty}$,
5	$0,1,5,16,45,121,...$	A004146=Alternate Lucas numbers - 2
6	$0,1,6,25,96,361,...$	A092184
7	$0,1,7,36,175,841,...$	A054493 (shifted by one)
8	$0,1,8,49,288,1681,...$	A001108
9	$0,1,9,64,441,3025,...$	A049684=F_{2n}^2 (F_n Fibonacci numbers)
10	$0,1,10,81,640,5041,...,$	A095004 (shifted by one)
11	$0,1,11,100,891,7921,...,$	A098296
12	$0,1,12,121,1200,11881,...$	A098297
13	$0,1,13,144,1573,17161,...$	A098298
14	$0,1,14,169,2016,24025,...$	A098299
15	$0,1,15,196,2535,32761,...$	A098300
16	$0,1,16,225,3136,43681,...$	A098301
17	$0,1,17,256,3825,57121,...$	A098302
18	$0,1,18,289,4608,73441,...$	A098303
19	$0,1,19,324,5491,93025,...$	A098304
20	$0,1,20,361,6480,116281,...$	A049683=$(L_{6n}-2)/16$ (L_n Lucas numbers)

Table 1.1: Sequence $u_n(w)$ for different values of w

We recall that Dickson polynomials are defined as follows:

$$D_n(x)=\sum_{i=0}^{\lfloor n/2\rfloor}\frac{n}{n-i}\binom{n-i}{i}(-1)^i x^{n-2i}$$

and

$$E_n(x)=\sum_{i=0}^{\lfloor n/2\rfloor}\binom{n-i}{i}(-1)^i x^{n-2i}.$$

We also recall that for Dickson polynomials the following identities hold

$$D_n\left(x+x^{-1}\right)=x^n+x^{-n},\quad E_n\left(x+x^{-1}\right)=\frac{x^{n+1}-x^{-(n+1)}}{x-x^{-1}} \tag{3}$$

THEOREM 1. *We have*

$$u_n(w)=\frac{D_n(w-2)-2}{w-4}=2\frac{T_n(\frac{w-2}{2})-1}{w-4},\quad \forall n\geq 0 \tag{4}$$

and in particular for all $n\geq 1$

$$u_{2n}(w)=wE_{n-1}^2(w-2)=wU_{n-1}^2\left(\frac{w-2}{2}\right) \tag{5}$$

$$\begin{aligned} u_{2n-1}(w)&=(E_{n-1}(w-2)+E_{n-2}(w-2))^2=\\ &=\left(U_{n-1}\left(\frac{w-2}{2}\right)+U_{n-2}\left(\frac{w-2}{2}\right)\right)^2 \end{aligned} \tag{6}$$

Proof. The recurrence relation described in (2) clearly shows that the characteristic polynomial of $(u_n(w))$ is

$$x^3-(w-1)x^2+(w-1)x-1=(x-1)(x^2-(w-2)x+1)$$

whose zeros are $x_1=1$ and $x_{2,3}=\frac{w-2\pm\sqrt{w^2-4w}}{2}$. If we set $x_2=\zeta$ we easily observe that $x_3=\zeta^{-1}$ so that $\zeta+\zeta^{-1}=w-2$ and $\zeta-\zeta^{-1}=\sqrt{w^2-4w}$. Moreover, using the initial conditions in (2), with standard tecniques we find the following closed form for every element of $(u_n(w))$

$$u_n(w)=\frac{\zeta^n+\zeta^{-n}-2}{w-4}=\frac{\zeta^n+\zeta^{-n}-2}{\zeta+\zeta^{-1}-2} \tag{7}$$

. Thanks to the first identity in (3) it is straightforward to observe that

$$u_n(w)=\frac{D_n(\zeta+\zeta^{-1})-2}{w-4}=\frac{D_n(w-2)-2}{w-4}. \tag{8}$$

Since $x^2-(w-2)x+1$ is the characteristic polynomial of the sequence $(T_n\left(\frac{w-2}{2}\right))$, with roots $x_2=\zeta$ and $x_3=\zeta^{-1}$, and the initial conditions are $T_0\left(\frac{w-2}{2}\right)=1$, $T_1\left(\frac{w-2}{2}\right)=\frac{w-2}{2}$ we obtain

$$T_n\left(\frac{w-2}{2}\right)=\frac{\zeta^n+\zeta^{-n}}{2}=\frac{D_n(\zeta+\zeta^{-1})}{2}=\frac{D_n(w-2)}{2} \tag{9}$$

Thus substituting (9) in (8) we prove equality (4). Now considering the equality (7) and the second identity in (3) we have

$$u_{2n}(w)=\frac{\zeta^{2n}+\zeta^{-2n}-2}{\zeta+\zeta^{-1}-2}=\frac{(\zeta^n-\zeta^{-n})^2}{(\zeta-\zeta^{-1})^2}\frac{(\zeta-\zeta^{-1})^2}{\zeta+\zeta^{-1}-2}=w(E_{n-1}(w-2))^2,$$

which proves (5), and

$$u_{2n-1}(w)=\frac{\zeta^{2n-1}+\zeta^{-2n+1}-2}{\zeta+\zeta^{-1}-2}=\frac{(\zeta^{2n-1}+\zeta^{-2n+1}-2)(\zeta+\zeta^{-1}+2)}{(\zeta-\zeta^{-1})^2} \tag{10}$$

where we use the identity

$$(\zeta-\zeta^{-1})^2=w(w-4)=(\zeta+\zeta^{-1}+2)(\zeta+\zeta^{-1}-2).$$

An easy calculation shows that

$$(\zeta^{2n-1}+\zeta^{-2n+1}-2)(\zeta+\zeta^{-1}+2)=\left(\zeta^n-\zeta^{-n}+\zeta^{n-1}-\zeta^{-(n-1)}\right)^2$$

and substituting in (10) we find

$$\begin{aligned} u_{2n-1}(w) &= \frac{\left(\zeta^n-\zeta^{-n}+\zeta^{n-1}-\zeta^{-(n-1)}\right)^2}{(\zeta-\zeta^{-1})^2}= \\ &= \left(\frac{\zeta^n-\zeta^{-n}}{\zeta-\zeta^{-1}}+\frac{\zeta^{n-1}-\zeta^{-(n-1)}}{\zeta-\zeta^{-1}}\right)^2= \\ &= (E_{n-1}(w-2)+E_{n-2}(w-2))^2, \end{aligned}$$

proving (6). □

As a consequence of (4) we highlight the following relation, where we posed $\frac{w-2}{2} = x$

$$T_n(x) = 2D_n(2x) = u_n(2x+2)\cdot(x-1)+1 \tag{11}$$

The coefficients of polynomials $u_n(w)$ are particularly interesting and we explicitly determine them in the following

THEOREM 2. *For any integer $n \geq 1$, we have*

$$u_n(w) = \sum_{k=0}^{n} d_n(k) w^k,$$

where

$$d_n(k) = \sum_{i=0}^{n-k-1} (-1)^i \binom{i+2k}{2k}, \quad \forall 0 \leq k < n$$

and $d_n(n) = 0$.

Proof. The theorem can be proved by induction. For $n = 1$, we have $u_1(w) = 1$ and $d_1(0)w^0 + d_1(1)w = 1$. Similarly, it is straightforward to check the theorem when $n = 2$ and $n = 3$.
Now, let us suppose that the thesis holds for any integer less or equal than n, for a given integer n. We have

$$u_{n+1}(w) = (w-1)u_n(w) - (w-1)u_{n-1}(w) + u_{n-2}(w) =$$

$$= (w-1)\sum_{k=0}^{n} d_n(k)w^k - (w-1)\sum_{k=0}^{n-1} d_{n-1}(k)w^k + \sum_{k=0}^{n-2} d_{n-2}(k)w^k.$$

Observing that

$$d_n(k) = d_{n-1}(k) + (-1)^{n-k-1}\binom{n+k-1}{2k}$$

we obtain

$$\begin{aligned} u_{n+1}(w) &= (w-1)\sum_{k=0}^{n} d_n(k)w^k - (w-1)\sum_{k=0}^{n-1}\left(d_n(k) - (-1)^{n-k-1}\binom{n+k-1}{2k}\right)w^k + \\ &\quad + \sum_{k=0}^{n-2} d_{n-2}(k)w^k = \\ &= (w-1)\sum_{k=0}^{n-1}(-1)^{n-k-1}\binom{n+k-1}{2k}w^k + \sum_{k=0}^{n-2}\left(d_{n+1}(k) - (-1)^{n-k}\binom{n+k}{2k} + \right. \\ &\quad \left. -(-1)^{n-k-1}\binom{n+k-1}{2k} - (-1)^{n-k-2}\binom{n+k-2}{2k}\right)w^k. \end{aligned}$$

Thus we have to prove that

$$(12)\quad (w-1)\sum_{k=0}^{n-1}(-1)^{n-k-1}\binom{n+k-1}{2k}w^k$$
$$+\sum_{k=0}^{n-2}\left((-1)^{n-k-1}\binom{n+k}{2k}-(-1)^{n-k-1}\binom{n+k-1}{2k}-(-1)^{n-k-2}\binom{n+k-2}{2k}\right)w^k$$
$$-w^n+2(n-1)w^{n-1}=0$$

in order to prove that

$$u_{n+1}(w)=\sum_{k=0}^{n+1}d_{n+1}(k)w^k.$$

The left member of equation (12) is equal to

$$\sum_{k=0}^{n-3}(-1)^{n-k-1}\binom{n+k-1}{2k}w^{k+1}-\sum_{k=0}^{n-2}(-1)^{n-k-1}\binom{n+k-1}{2k}w^k+$$
$$+\sum_{k=0}^{n-2}\left((-1)^{n-k-1}\binom{n+k}{2k}-(-1)^{n-k-1}\binom{n+k-1}{2k}-(-1)^{n-k-2}\binom{n+k-2}{2k}\right)w^k=$$
$$=\sum_{k=1}^{n-2}(-1)^{n-k}\left(\binom{n+k-2}{2k-2}+2\binom{n+k-1}{2k}-\binom{n+k}{2k}-\binom{n+k-2}{2k}\right)w^k$$

and using the property of binomial coefficients

$$\binom{n}{k}=\binom{n-1}{k}+\binom{n-1}{k-1}$$

it is easy to check that

$$\binom{n+k-2}{2k-2}+2\binom{n+k-1}{2k}-\binom{n+k}{2k}-\binom{n+k-2}{2k}=0.$$

□

Thanks to previous theorems and relation (11) we find the following expression for Chebyshev polynomials

$$T_n(x)=1+(x-1)\sum_{k=0}^{n}d_n(k)(2x+2)^k,\quad \forall n\geq 1,$$

and an analogous one for Dickson polynomials

$$D_n(x)=\frac{1}{4}\left(2+(x-2)\sum_{k=0}^{n}d_n(k)(x+2)^k\right),\quad \forall n\geq 1.$$

In the following section, we see that coefficients $d_n(k)$ allow us to determine a new identity for Catalan numbers and they can be used to obtain a Riordan array.

2. Catalan numbers and Riordan array

Catalan numbers are very famous and interesting, deeply studied for their significance in combinatorics. In the beautiful book of Stanley [8] many combinatorial interpretations and identities involving Catalan numbers can be found. We whish to point out another new identity involving Catalan numbers and the coefficients $d_n(k)$ studied in the previous section.

THEOREM 3. *For any positive integer n, we have*

$$\sum_{k=0}^{n} d_n(k)C_k = 1,$$

where $(C_k)_{k=0}^{+\infty}$ *is the sequence of the Catalan numbers (A000108 in OEIS)*

Proof. Since

$$\int_{-1}^{1} \frac{T_n(x)}{\sqrt{1-x^2}}dx = 0,$$

by Theorem 1, we have

$$\int_{-1}^{1} \frac{u_n(2x+2)(x-1)+1}{\sqrt{1-x^2}}dx = 0.$$

Posing $y = 2x+2$, we obtain

$$\int_0^4 \left(\frac{u_n(y)(y-4)+1}{2}\right) \frac{1}{\sqrt{y(4-y)}}dy = 0$$

and consequenlty

$$\int_0^4 \frac{u_n(y)(y-4)}{2\sqrt{y(4-y)}}dy = -\pi,$$

$$\sum_{k=0}^{n} \int_0^4 \frac{d_n(k)y^k(4-y)}{\sqrt{y(4-y)}}dy = 2\pi.$$

Moreover, it is well–known that

$$\int_0^4 \frac{y^k(4-y)}{\sqrt{y(4-y)}} = 2\pi C_k,$$

thus

$$\sum_{k=0}^{n} d_n(k)C_k = 1.$$

□

Catalan numbers can be arranged in order to define a Riordan array. We recall that a Riordan array is an infinite lower triangular matrix, where the k–th column is a sequence having ordinary generating function of the form $f(x)g(x)^k$, see [6]. Catalan numbers are used to generate a particular Riordan array defined by $f(x) = \dfrac{1-\sqrt{1-4x}}{2x}$ and $g(x) = \dfrac{1-\sqrt{1-4x}}{2}$, see [4]. Thus, considering the previous relation between Catalan numbers and the coefficients of polynomials $u_n(w)$, we can suppose that also $d_n(k)$ may generate a Riordan array. Indeed, in the following theorem, we prove that the sequence $(d_n(k))_{n=0}^{+\infty}$ define a Riordan array where $f(x) = \dfrac{x}{1-x^2}$ and $g(x) = \dfrac{x}{(1+x)^2}$.

THEOREM 4. *Given an integer k the ordinary generating function of the sequence $(d_n(k))_{n=0}^{+\infty}$ is*

$$\frac{x}{1-x^2} \cdot \frac{x^k}{(1+x)^{2k}}$$

Proof. The ordinary generating function of the sequence $(d_n(k))_{n=0}^{+\infty}$ is

$$\sum_{n=0}^{+\infty} d_n(k)x^n = \sum_{n=k+1}^{+\infty} \sum_{i=0}^{n-k-1} (-1)^i \binom{i+2k}{2k} x^n,$$

where in the right member the first sum starts from $k+1$, since for $n < k+1$ the coefficients $d_n(k)$ are not defined. If we pose $n-k-1 = m$, the ordinary generating function becomes

$$\sum_{m=0}^{+\infty} \sum_{i=0}^{m} (-1)^i \binom{i+2k}{2k} x^{m+k+1} = x^{k+1} \sum_{m=0}^{+\infty} \sum_{i=0}^{m} (-1)^i \binom{i+2k}{2k} x^m =$$

$$= x^{k+1} \sum_{i=0}^{+\infty} (-1)^i \binom{i+2k}{2k} x^i \sum_{m=i}^{+\infty} x^{m-i} = x^{k+1} \sum_{i=0}^{+\infty} \binom{i+2k}{2k} (-x)^i \sum_{h=0}^{+\infty} x^h.$$

Considering that

$$\frac{1}{(1-z)^{n+1}} = \sum_{i=0}^{+\infty} \binom{i+n}{n} z^i,$$

(see, e.g., [2] pag. 199) we finally have that the ordinary generating function is

$$\frac{x^{k+1}}{1-x} \cdot \frac{1}{(1-(-x))^{2k+1}} = \frac{x}{1-x^2} \cdot \frac{x^k}{(1+x)^{2k}}.$$

□

Thus the following matrix is a Riordan array

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & \cdots \\ 1 & -2 & 1 & 0 & 0 & \cdots \\ 0 & 4 & -4 & 1 & 0 & \cdots \\ 1 & -6 & 11 & -6 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where the k–th column is the sequence $(d_n(k))$.

References

[1] M. Abrate, S. Barbero, U. Cerruti, N. Murru, *Polynomial sequences on quadratic curves*, Submitted to Integers: The Electronic Journal of Combinatorial Number Theory, 2014.

[2] R. Graham, D. Knuth, O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, Reading, Massachusetts: Addison-Wesley, 1994

[3] R. Lidl, G. L. Mullen, G. Turnwald, *Dickson polynomals*, Harlow, Essex, England : Longman Scientific and Technical; New York : Copublished in the United States with John Wiley and Sons, 1993.

[4] A. Luzon, D. Merlini, M. A. Moron, R. Sprugnoli, *Identities induced by Riordan arrays*, Linear Algebra and its Applications, Vol. **436**, 631–647, 2012.

[5] T. J. Rivlin, *The Chebyshev polynomials*, John Wiley and Sons, 1974.

[6] L. W. Shapiro, S. Getu, W. J. Woan, L. Woodson, *The Riordan group*, Discrete Applied Mathematics, Vol. **34**, 229–239, 1991

[7] N. J. A. Sloane, The On–Line Encyclopedia of Integer Sequences, Published electronically at http://www.research.att.com/njas/sequences (2010).

[8] R. P. Stanley, *Enumerative Combinatorics*, Vol. **2**, Cambridge University Press, 1997.

AMS Subject Classification: 11D09, 11B83

S. Barbero, U. Cerruti, N. Murru

Lavoro pervenuto in redazione il 09.06.2019.

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 13 – 19

F. Battistoni

DISCRIMINANTS OF NUMBER FIELDS AND SURJECTIVITY OF TRACE HOMOMORPHISM ON RINGS OF INTEGERS

Abstract. In this note we give a brief survey of the most elementary criteria used to determine the surjectivity of the trace operator on the ring of integers of a number field K. Furthermore, we introduce an easy to state yet unknown surjectivity criterion depending only on the prime factorization of the degree n of K and on the squarefree part of the discriminant d_K.

1. Preliminaries and trace homomorphism

Let K be a number field of degree $n \in \mathbb{N}$ over the field $\mathbb{Q}$ of rational numbers. It is known that, for $n > 1$, there is not a canonical way to embed K in the field $\mathbb{C}$ of complex numbers; nonetheless, the field K admits exactly n embeddings $\sigma_1, \dots, \sigma_n : K \to \mathbb{C}$.
By the Primitive Element Theorem (Theorem 5.1 of [6]) we know that any number field K has the form $\mathbb{Q}(\alpha)$ for some algebraic number $\alpha \in K$, and if $p(x) \in \mathbb{Z}[x]$ is the minimum polynomial of α, then there is a bijection between the embeddings $\sigma_1, \dots, \sigma_n$ of K and the complex roots of $p(x)$.

Given $\beta \in K$, define its **trace** as the number $\mathrm{Tr}(\beta) := \sum_{i=1}^{n} \sigma_i(\beta)$. By its very definition, the trace is an algebraic number which is invariant for the action of the n embeddings of K, and thus it is a rational number. This allows to define the trace function

$$\mathrm{Tr} : K \to \mathbb{Q}$$

which is immediately seen to be an homomorphism of $\mathbb{Q}$-vector spaces.

If $K = \mathbb{Q}(\alpha)$ and $p(x) := x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n \in \mathbb{Q}[x]$ is the minimum polynomial of α, then $\mathrm{Tr}(\alpha) = -a_1$; this follows immediately from the fact that $p(x)$ splits as $\prod_{i=1}^{n}(x - \sigma_i(\alpha))$ in any algebraic closure of K (Corollary 3.12 of [3]).

Let O_K be the ring of integers of K, i.e. the subring of the algebraic integers contained in K. If $\alpha \in O_K$, then not only $\mathrm{Tr}(\alpha)$ is a rational number but it is also an algebraic integer, and so $\mathrm{Tr}(\alpha) \in \mathbb{Z}$. The restricted map

$$\mathrm{Tr} : O_K \to \mathbb{Z}$$

is an homomorphism of abelian groups.

The ring of integers O_K satisfies the following two important properties:

- Any non-zero ideal $I \subset O_K$ can be written in a unique way as a finite product of prime ideals of O_K (Theorem 3.14, Chapter I of [2]);

– If K has degree n, then $(O_K,+)$ is a free abelian group of rank n (Theorem 1, Chapter I of [4]).

2. Surjectivity of trace operator

Given a number field K of degree n, it is very easy to see that the trace map is a surjective homomorphism: in fact, $\mathrm{Tr}(1) = n$ and so, given $a/b \in \mathbb{Q}$, the element $a/(nb)$ is such that $\mathrm{Tr}(a/(nb)) = a/(nb) \cdot \mathrm{Tr}(1) = a/b$.
Actually, this proves that considering the subfield $\mathbb{Q} \subset K$ is enough to yield a surjection.

Does this surjectivity hold also for the restricted map $\mathrm{Tr} : O_K \to \mathbb{Z}$? Surely the trick of dividing by the degree n of the number field no longer works, because given $\alpha \in O_K$ the element α/n may not be in O_K.
In fact, it is very easy to provide an example of number field for which the trace restricted to the ring of integers is not surjective: consider the field $K = \mathbb{Q}(\sqrt{2})$, which has minimum polynomial $p(x) := x^2 - 2$. The ring of integers O_K is then equal to $\mathbb{Z}[\sqrt{2}]$ (Propostions 1.32 and 1.33, Chapter II of [1]), i.e. any algebraic integer in K has the form $a + b\sqrt{2}$ with $a, b \in \mathbb{Z}$. Being $\mathrm{Tr}(m) = 2m$ for any $m \in \mathbb{Z}$ and $\mathrm{Tr}(\sqrt{2}) = 0$ because of $p(x)$, then the trace of any element of O_K is an even rational integer, and so the restricted trace map is not surjective.
The above considerations imply that the restricted trace is not surjective for any quadratic number field $\mathbb{Q}(\sqrt{d})$ with $d \in \mathbb{Z}$ squarefree and $d \equiv 2,3 \bmod 4$ (this last assumption is needed to ensure that the ring of integers is equal to $\mathbb{Z}[\sqrt{d}]$).

One could wonder if there exist any criteria, different from explicitly studying the trace map, to determine whether the restricted trace homomorphism is surjective.
A first try comes from looking at the minimum polynomial of the number field.

PROPOSITION 1. *Let K be a number field with minimum polynomial $p(x) := x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n \in \mathbb{Z}[x]$. If $a_1 = \pm 1$, then the trace map $\mathrm{Tr} : O_K \to \mathbb{Z}$ is surjective.*

Proof. $p(x)$ being a monic irreducible polynomial with integer coefficients, there exists $\alpha \in O_K$ root of $p(x)$ such that $\mathrm{Tr}(\alpha) = -a_1 = \mp 1$. Then, for every $m \in \mathbb{Z}$ it is $m = \mathrm{Tr}(m\alpha)$ or $m = \mathrm{Tr}(-m\alpha)$ depending on the sign of $\mathrm{Tr}(\alpha)$. □

What can be said for number fields K which are defined by polynomials with coefficient $a_1 \neq \pm 1$ and such that it seems not possible to produce elements $\alpha \in O_K$ with $\mathrm{Tr}(\alpha) = \pm 1$ by hands only? One can get further information thanks to the concept of ramification, which is naturally related to the trace homomorphism: this is the subject of the next section.

3. Discriminants and ramification

Let K be a number field and let O_K be its ring of integers. Given a prime number $p \in \mathbb{Z}$, the ideal pO_K is not necessarily prime but has a factorization

$$pO_K = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_r^{e_r}$$

where the $\mathfrak{p}_i$'s are prime ideals in O_K and $e_i \in \mathbb{N}$ for every $i = 1, \ldots, r$. The prime number p is said to be **ramified in** K if $e_i > 1$ for some index i.

It is a classical problem in Number Theory to detect the prime numbers ramifying in a number field K: its solution depends mainly on the folllowing concepts.
Let $\alpha_1 \ldots, \alpha_n \in O_K$ be independent $\mathbb{Z}$-generators of O_K as abelian group. The **discriminant of** K is defined as

$$d_K := (\det(\sigma_i(\alpha_j))_{i,j=1}^n)^2 = \det(\mathrm{Tr}(\alpha_i\alpha_j))_{i,j=1}^n.$$

One gets $d_K \in \mathbb{Z}$ because of the last equality, and it is obvious from the definition that the value of d_K does not change by considering a new system $\beta_1, \ldots, \beta_n$ of $\mathbb{Z}$-independent generators for O_K.
The importance of the discriminant for the study of the ramified primes relies in the following proposition:

PROPOSITION 2. *A prime number p ramifies in K if and only if p divides d_K.*

Proof. See Corollary III.2.12 of [8]. □

The prime numbers may ramify with different behaviours: the following distinction will be useful to provide criteria for the study of the restricted trace homomorphism.
Let p be a rational prime number ramifying in K and let $pO_K = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_r^{e_r}$ be its prime ideal factorization in O_K. Then p is said to be **wildly ramified** if there exists $i \in \{1, \ldots, r\}$ such that p divides e_i; otherwise p is said to be **tamely ramified**.
A number field K is said to be **tame** if every ramified prime number is tamely ramified, otherwise K is said to be **wild**.

The last tool needed is the concept of different ideal.
Consider the set $\hat{O}_K := \{\alpha \in K\colon \mathrm{Tr}(\alpha \cdot O_K) \subset \mathbb{Z}\}$. The set $\mathcal{D}_K := \{\beta \in K\colon \beta \cdot \hat{O}_K \subset O_K\}$ is called the **different ideal of** K (or simply the different of K); it is an abelian group with respect to the sum.

LEMMA 1. *The different $\mathcal{D}_K$ satifies the following properties:*

- *$\mathcal{D}_K$ is an ideal of O_K;*
- *If p wildly ramifies in K, $pO_K = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_r^{e_r}$ and there exists $i \in \{1, \ldots, r\}$ such that p divides e_i, then $\mathfrak{p}_i$ is a factor of $\mathcal{D}_K$ with exponent at least e_i;*

- *If p tamely ramifies in K and $pO_K = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_r^{e_r}$, then for any $i \in \{1, \dots, r\}$ the number $e_i - 1$ is the exact exponent of the prime $\mathfrak{p}_i$ as factor of $\mathcal{D}_K$;*

- *The size of the quotient ring $O_K/\mathcal{D}_K$ is equal to $|d_K|$.*

Proof. These results are all proved in Section 4.2 of [7]. □

The distinction between tame and wild number fields and the concept of different ideal have proved to be important in determining the surjectivity of the trace homomorphism restricted to the ring of integers.

THEOREM 1. *Let K be a tame number field. Then* $\mathrm{Tr} : O_K \to \mathbb{Z}$ *is surjective.*

Proof. See Corollary 5, Section 4.2 of [7]. The different ideal has a main role in the setting of the proof. □

One can get an interesting Corollary, from which the surjectivity of the restricted trace can be recovered by looking only at the factorization of the discriminant.

COROLLARY 1. *Let K be a number field with squarefree discriminant d_K. Then* $\mathrm{Tr} : O_K \to \mathbb{Z}$ *is surjective.*

Proof. If $d_K = \pm p_1 \cdots p_r$ is squarefree, then $\mathcal{D}_K = \mathfrak{p}_1 \cdots \mathfrak{p}_r$ where the size of every quotient ring $O_K/\mathfrak{p}_i$ is equal to p_i. This implies that, for any fixed factor p_i of the discriminant, $\mathfrak{p}_i$ is the unique factor of $p_i O_K$ which has exponent greater than 1, and the value of this exponent is precisely equal to 2. Thus, any odd p_i is tamely ramified.
If 2 divides d_K and $\mathfrak{Q}$ is the factor of $2O_K$ dividing $\mathcal{D}_K$, then either 2 wildly ramifies with the exponent of $\mathfrak{Q}$ being 1, or 2 tamely ramifies with the exponent of $\mathfrak{Q}$ being 2, and both these options are absurd.
Thus K is a tame number field, and from Theorem 1 the surjectivity on the trace over the ring of integers follows. □

4. A weaker discriminant criterion

Theorem 1 of the previous section proves the surjectivity of the restricted trace for a wide class of number fields, and it also yields a good sufficient criterion depending only on the factorization of the discriminant d_K.
The goal of this section is to present a simple, yet new, criterion for the surjectivity which not only relies on the factorization of d_K, but has the advantage to give a positive answer also for some wild number fields.

THEOREM 2. *Let K be a number field of degree n and assume that, for every prime number p dividing n, the number p^2 does not divide d_K. Then* $\mathrm{Tr} : O_K \to \mathbb{Z}$ *is surjective.*

Proof. Let $T_0(K) := \{\alpha \in O_K \colon \mathrm{Tr}(O_K) = 0\}$ be the kernel of the restricted trace homomorphism. The structure theorem of free abelian groups (Theorem 7.3, Chapter I of [5]) implies that $T_0(K)$ is a free abelian group too, its rank being equal to $n-1$.
The set $\mathrm{Tr}(O_K)$ is an ideal in $\mathbb{Z}$. Let t be the positive generator of this ideal. Since $n = \mathrm{Tr}(1)$ one gets that t divides n.

The previous considerations imply that the ring of integers admits a decomposition $O_K = T_0(K) \oplus \mathbb{Z}\gamma$ as free abelian group, where $\gamma \in O_K$ is such that $\mathrm{Tr}(\gamma) = t$. Let $\alpha_1, \ldots, \alpha_{n-1}$ be a $\mathbb{Z}$-basis for $T_0(K)$: then $\alpha_1, \ldots, \alpha_{n-1}, \gamma$ is a $\mathbb{Z}$-basis for O_K and so the discriminant d_K can be computed by means of this basis.
Let M_K denote the matrix

$$\begin{pmatrix} \sigma_1(\alpha_1) & \cdots & \sigma_n(\alpha_1) \\ \cdots & \cdots & \cdots \\ \sigma_1(\alpha_{n-1}) & \cdots & \sigma_n(\alpha_{n-1}) \\ \sigma_1(\gamma) & \cdots & \sigma_n(\gamma) \end{pmatrix}.$$

Since its determinant does not change by replacing the last column with the sum of every other column, we get that

$$\det M_K = \det \begin{pmatrix} \sigma_1(\alpha_1) & \cdots & \sigma_{n-1}(\alpha_1) & \mathrm{Tr}(\alpha_1) \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_1(\alpha_{n-1}) & \cdots & \sigma_{n-1}(\alpha_{n-1}) & \mathrm{Tr}(\alpha_{n-1}) \\ \sigma_1(\gamma) & \cdots & \sigma_{n-1}(\gamma) & \mathrm{Tr}(\gamma) \end{pmatrix}$$

$$= \det \begin{pmatrix} \sigma_1(\alpha_1) & \cdots & \sigma_{n-1}(\alpha_1) & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_1(\alpha_{n-1}) & \cdots & \sigma_{n-1}(\alpha_{n-1}) & 0 \\ \sigma_1(\gamma) & \cdots & \sigma_{n-1}(\gamma) & t \end{pmatrix}.$$

Consider now the minor given by the first $n-1$ rows and the first $n-1$ columns:

$$N_K := \begin{pmatrix} \sigma_1(\alpha_1) & \cdots & \sigma_{n-1}(\alpha_1) \\ \cdots & \cdots & \cdots \\ \sigma_1(\alpha_{n-1}) & \cdots & \sigma_{n-1}(\alpha_{n-1}) \end{pmatrix}.$$

Applying any σ_i which is not the identity embedding on K, one sees that a column of N_K is now formed by elements $\sigma_n(\alpha_j)$ with $j = 1, \ldots, n-1$, while the other columns are permutations of the remaining columns. But for every $j \in \{1, \ldots, n-1\}$ it is $\sigma_n(\alpha_j) = -\sum_{i=1}^{n-1} \sigma_i(\alpha_j)$: this implies that $\det N_K$ is invariant for the action of the embeddings σ_i, up to a possible change of sign due to the permutation of the columns.
Thus it is enough to take the square of $\det N_K$ to get an algebraic integer invariant for any embeddings σ_i, i.e. a rational integer, and so $d_K = (\det M_K)^2 = (\det N_K)^2 \cdot t^2 = C \cdot t^2$, with $C \in \mathbb{Z}$. In other words, it is $d_K/t^2 \in \mathbb{Z}$.
Finally, from the above considerations and the fact that t divides n, the hypothesis of the Theorem force $t = 1$, and so the trace $\mathrm{Tr} : O_K \to \mathbb{Z}$ must be surjective. □

An example of wild number field for which the surjectivity of the restricted trace is not evident without Theorem 2 is given by the cubic field K defined by the polynomial x^3+x-6. In fact, its discriminant is equal to $-2^2 \cdot 61$, so the primes 61 and 2 both ramify. A computation with the computer algebra package PARI/GP [9] shows that the prime 2 wildly ramifies, thus the extension is not tame and Theorem 1 or Corollary 1 do not apply. However, $\mathrm{Tr} : O_K \to \mathbb{Z}$ is surjective, by Theorem 2.

Some final considerations arise looking back at the quadratic fields studied in Section 2: in fact, these are wild fields for which the trace on the ring of integers is not surjective. Moreover, their discriminant is always divided by $4 = 2^2$, and thus they do not satisfy the hypotheses needed for the sufficient criterion introduced by Theorem 2. This suggests a possible conjecture for the complete characterization of the surjectivity of the trace map on the ring of integers:
Given a number field K, then $\mathrm{Tr} : O_K \to \mathbb{Z}$ *is not surjective if and only if K is wild and does not satisfy the hypotheses of Theorem 2.*

References

[1] A. Fröhlich and M. J. Taylor. *Algebraic number theory*, volume 27 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1993.

[2] G. J. Janusz. *Algebraic number fields*, volume 7 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 1996.

[3] F. Jarvis. *Algebraic number theory*. Springer Undergraduate Mathematics Series. Springer, Cham, 2014.

[4] S. Lang. *Algebraic number theory*, volume 110 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1994.

[5] S. Lang. *Algebra, Revised Third Edition*. Springer, 2002.

[6] J. S. Milne. Fields and Galois Theory (v4.60), 2018. Available at www.jmilne.org/math/.

[7] W. Narkiewicz. *Elementary and analytic theory of algebraic numbers*. Springer Science & Business Media, 2013.

[8] J. Neukirch. *Algebraic number theory*, volume 322 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1999. Translated from the 1992 German original and with a note by Norbert Schappacher, With a foreword by G. Harder.

[9] The PARI Group, Univ. Bordeaux. *PARI/GP version* `2.11.0`, 2018. Available at http://pari.math.u-bordeaux.fr/.

AMS Subject Classification: 11R04, 11R29

Francesco BATTISTONI,
Dipartimento di Matematica, Università degli Studi di Milano
Via Saldini 50, 20133 Milano, Italy
e-mail: `francesco.battistoni@unimi.it`

Lavoro pervenuto in redazione il 26.04.2019.

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 21 – 25

D. Bazzanella and C. Sanna

LEAST COMMON MULTIPLE OF POLYNOMIAL SEQUENCES

Abstract. We collect some results and problems about the quantity

$$L_f(n) := \mathrm{lcm}(f(1), f(2), \ldots, f(n)),$$

where f is a polynomial with integer coefficients and lcm denotes the least common multiple.

1. Introduction

For each positive integer n, let us define

$$L(n) := \mathrm{lcm}(1, 2, \ldots, n),$$

that is, the lowest common multiple of the first n positive integers. It is not difficult to show that

$$\log L(n) = \psi(n) := \sum_{p \le n} \log p,$$

where ψ denotes the first Chebyshev function, and p runs over all primes numbers not exceeding n. Hence, bounds for $L(n)$ are directly related to bounds for $\psi(n)$ and, consequently, to estimates for the prime counting function $\pi(n)$. In particular, since the Prime Number Theorem is equivalent to $\psi(n) \sim n$ as $n \to +\infty$, we have

$$\log L(n) \sim n.$$

In 1936 Gelfond and Shnirelman, proposed a new elementary and clever method for deriving a lower bound for the prime counting function $\pi(x)$ (see Gelfond's editorial remarks in the 1944 edition of Chebyshev's Collected Works [15, pag. 287–288]). In 1982 the Gelfond-Shnirelman method was rediscovered and developed by Nair [16, 17]. Their method was actually based on estimating $L(n)$, and in its simplest form [16] it gives

$$n \log 2 \le \log L(n) \le n \log 4,$$

for every $n \ge 9$, which in turn implies

$$(\log 2 + o(1)) \frac{n}{\log n} \le \pi(n) \le (\log 4 + o(1)) \frac{n}{\log n},$$

after some manipulations. Later, it was proved [18] that the Gelfond-Shnirelman-Nair method can give lower bound in the form

$$\pi(n) \ge C \, \frac{n}{\log n},$$

only for constants C less than 0.87, which is quite far from what is expected by the Prime Number Theorem. (A possible way around this problem has been considered in [13, 14, 19].)

Moving from this initial connection with estimates for $\pi(n)$ and the Prime Number Theorem, several authors have considered bounds and asymptotic for the following generalization of $L(n)$ to polynomials. For every polynomial $f \in \mathbb{Z}[x]$, let us define

$$L_f(n) := \operatorname{lcm}(f(1), f(2), \ldots, f(n)).$$

In the next section we collect some results on $L_f(n)$.

2. Products of linear polynomials

Stenger [12] used the Prime Number Theorem for arithmetic progressions to show the following asymptotic estimate for linear polynomials:

THEOREM 1. *For any linear polynomial $f(x) = ax + b \in \mathbb{Z}[x]$, we have*

$$\log L_f(n) \sim n \frac{q}{\varphi(q)} \sum_{\substack{1 \le r \le q \\ (q,r)=1}} \frac{1}{r},$$

as $n \to +\infty$, where $q = a/(a,b)$ and φ denotes the Euler's totient function.

Hong, Qian, and Tan [6] extended this result to polynomials f which are the product of linear polynomials, showing that an asymptotic of the form $\log L_f(n) \sim A_f n$ holds as $n \to +\infty$, where $A_f > 0$ is a constant depending only on f.

Moreover, effective lower bounds for $L_f(n)$ when f is a linear polynomial have been proved by Hong and Feng [3], Hong and Kominers [4], Hong, Tan and Wu [7], Hong and Yang [8], and Oon [9],

3. Quadratic polynomials

Cilleruelo [2, Theorem 1] considered irreducible quadratic polynomials and proved the following result:

THEOREM 2. *For any irreducible quadratic polynomial with integer coefficients $f(x) = ax^2 + bx + c$, we have*

$$\log L_f(n) = n \log n + B_f n + o(n),$$

where

$$B_f := \gamma - 1 - 2\log 2 - \sum_p \frac{(d/p)\log p}{p-1} + \frac{1}{\varphi(q)} \sum_{\substack{1\le r\le q\\(r,q)=1}} \log\left(1+\frac{r}{q}\right)$$
$$+ \log a + \sum_{p\mid 2aD} \log p \left(\frac{1+(d/p)}{p-1} - \sum_{k\ge 1} \frac{s(f,p^k)}{p^k} \right),$$

and γ is the Euler–Mascheroni constant, $D = b^2 - 4ac = d\ell^2$, where d is a fundamental discriminant, (d/p) is the Kronecker symbol, $q = a/(a,b)$ and $s(f,p^k)$ is the number of solutions of $f(x) \equiv 0 \pmod{p^k}$.

Rué, Šarka, and Zumalacárregui [11, Theorem 1.1] provided a more precise error term for the particular polynomial $f(x) = x^2 + 1$,

THEOREM 3. *Let $f(x) = x^2 + 1$. For any $\theta < 4/9$ we have*

$$\log L_f(n) = n\log n + B_f n + O_\theta\left(\frac{n}{(\log n)^\theta}\right).$$

4. Higher degree polynomials

Regarding general irreducible polynomials, Cilleruelo [2] formulated the following conjecture.

CONJECTURE 1. If $f(x) \in \mathbb{Z}[x]$ is an irreducible polynomial of degree $d \ge 2$, then

$$\log L_f(n) \sim (d-1) n \log n,$$

as $n \to +\infty$.

Except for the result of Theorem 2, no other case of Conjecture 1 is known to date. It can be proved (see [10, p. 2]) that for any irreducible f of degree $d \ge 3$, we have

$$n\log n \ll \log L_f(n) \le (1+o(1))(d-1)\, n\log n.$$

Also, Rudnick and Zehavi [10, Theorem 1.2] proved the following result, which established Conjecture 1 for almost all shifts of a fixed polynomial, in a range of n depending on the range of shifts.

THEOREM 4. *Let $f(x) \in \mathbb{Z}[x]$ be a monic polynomial of degree $d \ge 3$. Then, as $T \to +\infty$, we have that for all $a \in \mathbb{Z}$ with $|a| \le T$, but a set of cardinality $o(T)$, it holds*

$$\log L_{f(x)-a}(n) \sim (d-1)\, n \log n$$

uniformly for $T^{1/(d-1)} < n < T/\log T$.

Regarding lower bounds for $L_f(n)$, Hong and Qian [5, Lemma 3.1] proved the following:

THEOREM 5. *Let $f(x) \in \mathbb{Z}[x]$ be a polynomial of degree $d \geq 1$ and with leading coefficient a_d. Then for all integers $1 \leq m \leq n$, we have*

$$\operatorname{lcm}(f(m), f(m+1), \ldots, f(n)) \geq \frac{1}{(n-m)!} \prod_{k=m}^{n} \left| \frac{f(k)}{a_d} \right|^{1/d}.$$

Shparlinski [1] suggested to study a bivariate version of $L_f(n)$, posing the following problem:

PROBLEM 1. *Given a polynomial $f \in \mathbb{Z}[x,y]$, obtain an asymptotic formula for*

$$\log \operatorname{lcm}\{f(m,n) : 1 \leq m,n \leq N\}$$

with a power saving in the error term.

5. Acknowledgements

C. Sanna is supported by a postdoctoral fellowship of INdAM and is a member of the INdAM group GNSAGA.

References

[1] CANDELA P., *Memorial to Javier Cilleruelo: A problem list*, INTEGERS **18** (2018), #A28.

[2] CILLERUELO J., *The least common multiple of a quadratic sequence*, Compos. Math. **147** (2011), 1129–1150.

[3] HONG S., FENG W., *Lower bounds for the least common multiple of finite arithmetic progressions*, C. R. Math. Acad. Sci. Paris **343** (2016), 695–698.

[4] HONG S., KOMINERS S. D., *Further improvements of lower bounds for the least common multiples of arithmetic progressions*, Proc. Amer. Math. Soc. **138** (2010), 809–813.

[5] HONG S., QIAN G., *Uniform lower bound for the least common multiple of a polynomial sequence*, C. R. Math. Acad. Sci. Paris **351** (2013), 781–785.

[6] HONG S., QIAN G., TAN Q., *The least common multiple of a sequence of products of linear polynomials*, Acta Math. Hungar. **135** (2012), 160–167.

[7] HONG S., TAN Q., WU R., *New lower bounds for the least common multiples of arithmetic progressions*, Chin. Ann. Math. Ser. B, 34B(6) (2013), 861–864.

[8] HONG S., YANG Y., *Improvements of lower bounds for the least common multiple of finite arithmetic progressions*, Proc. Amer. Math. Soc., **136** (2008), 4111–4114.

[9] OON S.-M., *Note on the lower bound of least common multiple*, Abstr. Appl. Anal., (2013) Article ID 218125.

[10] RUDNICK Z. AND ZEHAVI S., *On Cilleruelo's conjecture for the least common multiple of polynomial sequences*, ArXiv: `http://arxiv.org/abs/1902.01102v2`.

[11] RUÉ J., ŠARKA, P., ZUMALACÁRREGUI A., *On the error term of the logarithm of the lcm of a quadratic sequence*, J. Théor. Nombres Bordeaux **25** (2013), 457–470.

[12] BATEMAN P., KALB J., STENGER A., *A limit involving least common multiples*, Amer. Math. Monthly **109** (2002), 393–394.

[13] D. BAZZANELLA, *A note on integer polynomials with small integrals*, Acta Math. Hungar. **141** (2013), n. 4, 320–328.

[14] D. BAZZANELLA, *A note on integer polynomials with small integrals. II*, Acta Math. Hungar. **149** (2016), n. 1, 71–81.

[15] P. L. CHEBYSHEV, *Collected Works, Vol. 1, Theory of Numbers*, Akad. Nauk. SSSR, Moskow, 1944.

[16] M. NAIR, *On Chebyshev's-type inequalities for primes*, Amer. Math. Monthly **89** (1982), 126–129.

[17] M. NAIR, *A new method in elementary prime number theory*, J. Lond. Math. Soc. (2) **25** (1982), 385–391.

[18] I. E. PRITSKER, *Small polynomials with integer coefficients*, J. Anal. Math., **96** (2005), pp. 151–190.

[19] C. SANNA, *A factor of integer polynomials with minimal integrals*, J. Théor. Nombres Bordeaux **29** (2017), 637–646.

AMS Subject Classification: 11N32, 11N37

Danilo BAZZANELLA,
Department of Mathematical Sciences, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: `danilo.bazzanella@polito.it`

Carlo SANNA,
Department of Mathematics, Università di Genova
Via Dodecaneso 35, 16146 Genova, Italy
e-mail: `carlo.sanna.dev@gmail.com`

Lavoro pervenuto in redazione il 30.10.2019.

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 27 – 42

F. Caldarola

ON THE MAXIMAL FINITE IWASAWA SUBMODULE IN $\mathbb{Z}_P$-EXTENSIONS AND CAPITULATION OF IDEALS

Abstract. For $\mathbb{Z}_p$-extensions of a number fields the properties of stabilization and capitulation of ideal classes are of great interest and are also related to very important aspects and problems such as Greenberg's conjecture. In [3] these properties are deeply investigated from the point of view of the maximum finite submodule of the Iwasawa module and new invariants and parameters are introduced to give precise characterizations of these phenomena. In this article we will discuss some bounds that control the increment of the index that measures the capitulation delay in the tower and moreover we will prove how some results on the capitulation kernels in [3] have to be considered optimal. Finally, we will also give some further applications and examples that emphasize the cases of false (or failed) stabilization in this context.

1. Introduction.

Iwasawa's theory in the last half century has been one of the richest areas of research in number theory. In this paper we will consider some basic objects of the theory such as $\mathbb{Z}_p$-extensions and the Iwasawa module with particular emphasis on its maximum finite submodule, in relation to the problems of capitulation and stabilization, typical of this context.

Let p be a prime number, k a number field and K/k a $\mathbb{Z}_p$-extension of k. Let moreover $L = L(K)$ be the maximal abelian unramified pro-p extension of K (in a fixed algebraic closure of $\mathbb{Q}$) and we also pose $\Gamma := \mathrm{Gal}(K/k)$ and $X(K) := \mathrm{Gal}(L(K)/K)$. We denote by k_n the n-th layer of K/k and by $A_n = A(k_n)$ the p-part of the ideal class group of k_n. For any $m \geq n \geq 0$ we write $N_{m,n} : A_m \to A_n$ and $i_{n,m} : A_n \to A_m$ for the natural maps induced by the norm and the inclusion of ideals, and we also consider the limits

$$\varprojlim_n A_n \qquad \text{and} \qquad A = A(K) := \varinjlim_n A_n \tag{1}$$

obtained via the $N_{m,n}$ and the $i_{n,m}$ maps, respectively. By class field theory, the first limit in (1) is canonically isomorphic to $X(K) \simeq \varprojlim \mathrm{Gal}(L(k_n)/k_n)$, where $L(k_n)$ (or L_n for short) is the maximal abelian unramified p-extension of k_n. Let

$$\mathbb{Z}_p[\![\Gamma]\!] := \varprojlim_n \mathbb{Z}_p[\Gamma/\Gamma^{p^n}] \simeq \varprojlim_n \mathbb{Z}_p[\mathrm{Gal}(k_n/k)]$$

be the *Iwasawa algebra* (completed group ring) associated to K/k, which is isomorphic to the formal power series ring $\Lambda := \mathbb{Z}_p[[T]]$ via the noncanonical isomorphism

$$\mathbb{Z}_p[\![\Gamma]\!] \xrightarrow{\simeq} \mathbb{Z}_p[[T]], \quad \gamma \mapsto T+1, \tag{2}$$

where γ is a topological generator of Γ. Since $\mathbb{Z}_p[\![\Gamma]\!]$ acts in a natural way, via conjugation, on $X(K)$, then it becomes a Λ-module through the isomorphism given in (2). With this structure, $X(K)$ (also denoted by $X(K/k)$ or simply X) is usually referred to as the *Iwasawa module* of K/k and, although to have been extensively studied by a lot of authors in the last sixty years, many problems of crucial importance remain open today (the interested reader can see [10] for the state-of-the-art and/or [16, 19, 22, 27] for some introductory references on the matter).

A classical problem concerns *capitulation* of ideals and ideal classes going up along the intermediate fields of the tower K/k. More precisely, denoting by $H_{n,m}$, $n \leq m$, the kernel of $i_{n,m} : A_n \to A_m$, we say that $[\mathfrak{a}] \in A_n$ *capitulates* in A_m if $[\mathfrak{a}] \in H_{n,m}$ (or, equivalently, $\mathfrak{a}$ becomes principal in k_m). From the beginning of the theory the groups $H_{n,m}$ and $H_n := \bigcup_{m \geq n} H_{n,m} = \mathrm{Ker}(i_n : A_n \to A)$ have been related, for example, to the finiteness of the module $X(K)$ (see, e.g., [9, 11, 13, 14, 20, 23]), but the phenomenon remains of a wild nature in general. In [3] the authors provide a description of the $H_{n,m}$ (named *relative capitulation kernels*) and of the H_n (named *absolute capitulation kernels*) in terms of the maximal finite submodule $D = D(K/k)$ of $X(K/k)$, obtaining isomorphisms with quotients of suitable submodules of D, finding some formulas for their order, and investigating their properties of stabilization both for orders and for p-ranks. For these purposes, in particular, two new functions $h(n)$ and $\rho(n)$ are introduced: the former is more technical and is linked to the vanishing of some submodules of D (see Definition 4 for a precise definition), while the second measures when the last ideal in A_n capitulates going up along the tower K/k_n.

In this paper we want to deepen the researches concerning the functions $h(n)$ and $\rho(n)$ and, in particular, we want to find bounds that allow to control their growth and therefore the so-called *capitulation delay*, especially in the lower levels that are potentially much more irregular. We also construct some examples in which we compute the explicit values of the mentioned invariants and parameters and then, generalizing such constructions, we give a method useful to show how some central results of [3] can be said to be optimal and some of the previous bounds sharp. Finally, even some emerging evidences of false, or failed, stabilization are discussed.

As regards the organization of the paper, it can be divided into two parts comprising two sections each. In Section 2 we give an overview of few known classical results about stabilization and capitulation in $\mathbb{Z}_p$-extensions, instead in Section 3 we will describe some of the main developments contained in the recent work [3] and necessary for the sequel. The second part, consisting of Section 4 and 5, contains the original results we referred to above. In particular, Section 4 deals with the bounds for $h(m) - h(n)$ and $\rho(m) - \rho(n)$ $(m \geq n)$, instead Section 5 draws the method for the optimality of results and sharpness of the bounds.

Lastly a notational remark: by convenience we include also zero in the set $\mathbb{N}$ of natural numbers.

2. A brief overview of classical results.

The literature object of this section is placed from the beginning of the theory until the 1990s. We divide it into two subsections for stabilization and capitulation, respectively.

2.1. Stabilization.

The term "stabilization" has the expected obvious meaning: given a sequence of finite $\mathbb{Z}_p$-modules or p-groups $\{M_n\}_{n\in\mathbb{N}}$, we say that their orders stabilize at an index $q\in\mathbb{N}$ if $|M_n|=|M_q|$ for all $n\geq q$. In the same way, we say that their p-ranks[2] stabilize at $q'\in\mathbb{N}$ if $\mathrm{rk}_p(M_n)=\mathrm{rk}_p(M_{q'})$ for all $n\geq q'$.

Stabilization is quite natural in Iwasawa theory even if there are not many results in this direction. We have stabilization theorems for $\{|A_n|\}_{n\in\mathbb{N}}$ and for $\{\mathrm{rk}_p(A_n)\}_{n\in\mathbb{N}}$ but not much else. Usually Iwasawa modules tend to stabilize at the very first level in which they do not grow (i.e., if we have no growth from n to $n+1$, we are not going to have any growth at all from n on).

Let $n_0=n_0(K/k)$ be the minimal $n\geq 0$ such that every prime which ramifies in the extension K/k_n is totally ramified. References for stabilization and for the following theorems are [1, 8, 18].

THEOREM 1. ([8, Theorem 1(1)]) *If $|A_n|=|A_{n+1}|$ for some $n\geq n_0$, then $A_m\simeq A_n\simeq X$ for all $m\geq n$.*

THEOREM 2. ([8, Theorem 1(2)]) *If $\mathrm{rk}_p(A_n)=\mathrm{rk}_p(A_{n+1})$ for some $n\geq n_0$, then $\mathrm{rk}_p(A_m)=\mathrm{rk}_p(A_n)=\mathrm{rk}_p(X)$ for all $m\geq n$ (and hence $\mu(K/k)=0$).*

2.2. Capitulation.

For any finitely generated Λ-module X we have an exact sequence of Λ-modules

$$0\to D(X)\to X\xrightarrow{\varphi} E(X)\to B(X)\to 0 \tag{3}$$

where φ is a *pseudo-isomorphism*, $E(X)$ an elementary Λ-module and $D(X)$, $B(X)$ are finite (see [3, Section 2] or [5, Chapter VII], [22, Chapter V], [27, Chapter 13] for more details). $D(X)$ is the maximal finite submodule of X and when $X=X(K/k)$ for some $\mathbb{Z}_p$-extension K/k, we also call $D(K/k)$ the *maximal finite Iwasawa module* of K/k, and we shall often simply write D to denote it in the following. We moreover recall that X is said *pseudo-null* if $E(X)=0$ in (3), or equivalently if $X=D(X)$.

The capitulation kernels $H_{n,m}$ and H_n are very important in Iwasawa theory, for example because of the following proposition which links them to Greenberg's conjecture which predicts the finiteness, or equivalently the pseudo-nullity, of $X(k_{cyc}/k)$ whenever k is a totally real number field (and k_{cyc} its cyclotomic $\mathbb{Z}_p$-extension).

[2] For any finitely generated $\mathbb{Z}_p$-module A, we obviously define its *p-rank* as $\mathrm{rk}_p(A):=\dim_{\mathbb{F}_p}(A/pA)$.

PROPOSITION 1. ([11, Proposition 2]) *We have that* $\lambda(K/k) = \mu(K/k) = 0$ *if and only if* $H_n = A_n$ *for every* $n \geq 0$.

Let $s = s(K/k)$ be the number of ramified prime ideals in K/k_{n_0}. In the following theorem the statement is not exactly the original one appearing in [11], but it can be easily derived from it because the proof only uses the hypothesis $s(K/k) = 1$.

THEOREM 3. ([11, Theorem 1]) *Let* $n_0 = 0$ *and* $s(K/k) = 1$ *(i.e., there is only one prime in k which ramifies in K), then X is pseudo-null if and only if* $H_0 = A_0$.

A very important remark is that, in the previous theorem, the hypothesis $n_0 = 0$ can be easily suppressed. This has been showed by J. Minardi and we write down it separately.

COROLLARY 1. ([21, Proposition 1.B]) *Assume* $s = 1$*, then X is pseudo-null if and only if* $H_0 = A_0$.

COROLLARY 2. ([21, pag. 6, Corollary]) *If there is only one prime* $\mathfrak{p}$ *in k dividing p and the class of some power of* $\mathfrak{p}$ *generates the whole* A_0*, then* $X(K/k)$ *is pseudo-null for every* $\mathbb{Z}_p$*-extension* K/k.

The statement (a) of the next theorem provides a stronger result and it was proved by T. Fukuda in 1994 in a very elegant way. Indeed Theorem 4 (a) gives the layer k_m for which $X \simeq A_m$, but there is a price to pay: with this kind of proof the hypothesis $n_0 = 0$ acquires a crucial role and it cannot be removed anymore.

THEOREM 4. ([8, Theorem 2]) *Let* $s(K/k) = 1$ *and* $n_0(K/k) = 0$.

(a) If $H_{0,n} = A_0$ *for some* $n \geq 1$*, then* $|A_m| = |A_n| = |X|$ *for all* $m \geq n$.

(b) If $|A_{n+1}| = |A_n|$ *for some* $n \geq 0$ *and the exponent of* A_n *is* p^t*, then* $H_{n,n+t} = A_n$.

We conclude this section recalling

THEOREM 5. ([23, Theorem]) *Let* $n_0(K/k) = 0$*, then* $X(K/k)$ *is pseudo-null if and only if* $\ker N_{1,0} \subseteq H_1$.

Not much more was known about the $H_{n,m}$'s before [3], in particular regarding their orders and their stabilization properties. For example, Iwasawa himself proved that $H_{n,m}$ is bounded by $|D| \cdot |B(X)| \cdot |A_{n_0}|$ independently from n and m (see e.g. [16] or [22]), M. Ozaki showed in [23] a relation between the H_n's and the submodule D of X, and M. Grandet and J.F. Jaulent considered capitulation in the special case $\mu = 0$ in [9]. Other documents related to these themes are [2, 13, 14, 18, 20, 21, 24, 26], while [7], although referring to a very different context, also draws inspiration from the vertical structures of Iwasawa's theory.

In the next section we will briefly summarize the description of the capitulation

kernels provided in [3] by going deeper into the study of the module D.

3. Recent developments.

The results obtained in [3] can be grouped into two families: the first concerns the behaviour of the capitulation kernels (see Subsection 3.1), the second instead provides a series of new equivalent conditions, which often involve capitulation kernels, for the vanishing of the Iwasawa invariants μ and λ, and therefore for the finiteness of $X(K/k)$ (see Subsection 3.2).

3.1. The behaviour of the capitulation kernels.

Focusing the attention on the sequence of the absolute capitulation kernels

$$H_{n_0},\ H_{n_0+1},\ H_{n_0+2},\dots \tag{4}$$

and on the chain of the relative capitulation ones (at the level n)

$$H_{n,n+1} \subseteq H_{n,n+2} \subseteq H_{n,n+3} \subseteq \dots, \tag{5}$$

it is natural to ask, first of all, for growing and stabilization both for the sequence of the orders and for the sequence of the p-ranks arising from (4) and (5). To answer to these questions, [3] begins with the introduction of some Λ-submodules of D as in Definition 2. But first we need a further piece of notation because, recalling the isomorphism in (2) which will be read as an identification in the following, there are some elements in $\Lambda = \mathbb{Z}_p[[T]]$ which play an important role in the study of the class groups A_n.

DEFINITION 1. *For every $m \geq n \geq 0$, we set*

- $\omega_n := \gamma^{p^n} - 1 = (1+T)^{p^n} - 1$,
- $\nu_{n,m} := 1 + \gamma^{p^n} + \gamma^{2p^n} + \dots + \gamma^{p^m - p^n}$
  $= \frac{\omega_m}{\omega_n} = \frac{(1+T)^{p^m}-1}{(1+T)^{p^n}-1} = 1 + (1+T)^{p^n} + \dots + ((1+T)^{p^n})^{p^{m-n}-1}$.

For simplicity we write ν_n in place of $\nu_{0,n}$: hence $\nu_n = 1 + \gamma + \gamma^2 + \dots + \gamma^{p^n-1} = \frac{\omega_n}{\omega_0} = \frac{(1+T)^{p^n}-1}{T}$ and $\nu_{n,m} = \frac{\nu_m}{\nu_n}$ as well.

DEFINITION 2. *We put*

(a) for all $m \geq n \geq 0$, $D_{n,m} := \nu_{n,m} D$;

(b) for all $n \geq n_0$, $D_n := D \cap Y_n$.

It is important, also for future use, to notice that the D_n have a good behaviour with respect to the usual Iwasawa relations, in the sense of the following

LEMMA 1. *For all $m \geq n \geq n_0$, we have $\nu_{n,m} D_n = D_m = D_{n,m} \cap Y_m$.*

For a proof see [3, Lemma 3.2]. Then, two new invariants for K/k are introduced as follows.

DEFINITION 3. *We set*

(a) $r = r(K/k) := \min\{z \geq n_0 : D_z = 0\}$, *and*

(b) $\widetilde{r} = \widetilde{r}(K/k) := \min\{z \geq n_0 : D_z \subseteq pD\}$.

Using Lemma 1 and Nakayama's Lemma, it is immediate to see that $r(K/k)$ and $\widetilde{r}(K/k)$ are always finite (nonnegative) integers.

Now we provide an isomorphism for the kernels $H_{n,m}$ in terms of the finite module D which leads to the formulas of Corollary 3 (a) and (b). For the proofs of the following results see [3, Section 3].

THEOREM 6. *For all $m \geq n \geq n_0$ there are the following isomorphisms*

$$H_{n,m} \simeq \mathrm{Ker}\{\nu_{n,m} : D/D_n \longrightarrow D/D_m\} \tag{6}$$

and

$$H_n \simeq D + Y_n/Y_n \simeq D/D_n. \tag{7}$$

Moreover, if $m \geq n \geq r(K/k)$, $H_{n,m} \simeq D[p^{m-n}]$ (where $D[p^{m-n}]$ is the submodule of the p^{m-n}-torsion elements of D).

COROLLARY 3. *For all $m \geq n \geq n_0$ we have*

(a) $|H_{n,m}| = \dfrac{|D| \cdot |D_m|}{|D_n| \cdot |D_{n,m}|}$;

(b) $|H_{n,m}| = |D + Y_n/D_{n,m} + Y_m| \cdot \dfrac{|A_n|}{|A_m|}$;

(c) if $D \neq 0$ and $n \geq n_0$, then $i_n : A_n \to A$ is injective if and only if $n = n_0$ and D is contained in Y_{n_0}.

COROLLARY 4. *For any $\mathbb{Z}_p$-extension K/k, the following are equivalent:*

(a) X does not contain any nontrivial finite submodule;

(b) $H_{n_0+1} = 0$;

(c) $i_{n,m} : A_n \to A_m$ *are injective for all* $m \geq n \geq n_0$.

An example for (a) is provided by the minus part of the Iwasawa module for the $\mathbb{Z}_p$-cyclotomic extension of a CM field (see [27, Propositions 13.26 and 13.28]); similar results can be derived from [23]. The following corollary instead generalizes [8, Proposition].

COROLLARY 5. *Let K/k be a $\mathbb{Z}_p$-extension, assume that $A_n \neq 0$ and $i_{n,m}$ is injective for some $m > n \geq n_0$. Then $|A_m| \geq p^{m-n}|A_n|$.*

As customary for Iwasawa modules, the H_n's verify some stabilization results.

THEOREM 7. *Assume $n \geq n_0$:*

(a) if $|H_n|=|H_{n+1}|$, then $H_m \simeq H_n \simeq D$ for all $m \geq n$. In particular

$$|H_{n_0}| < |H_{n_0+1}| < \ldots < |H_r| = |H_{r+1}| = \ldots = |D|; \tag{8}$$

(b) if $\mathrm{rk}_p(H_n) = \mathrm{rk}_p(H_{n+1})$, then $\mathrm{rk}_p(H_m) = \mathrm{rk}_p(H_n) = \mathrm{rk}_p(D)$ for all $m \geq n$. In particular

$$\mathrm{rk}_p(H_{n_0}) < \mathrm{rk}_p(H_{n_0+1}) < \ldots < \mathrm{rk}_p(H_{\widetilde{r}}) = \mathrm{rk}_p(H_{\widetilde{r}+1}) = \ldots = \mathrm{rk}_p(D). \tag{9}$$

From Theorem 7, we have $r = \min\{z \geq n_0 : H_z = H_{z+1}\}$ and $\widetilde{r} = \min\{z \geq n_0 : \mathrm{rk}_p(H_z) = \mathrm{rk}_p(H_{z+1})\}$, so these two invariants indicate also the stabilization of orders and p-ranks of the H_n's. To study instead the stabilization of the $H_{n,m}$'s and the delay of capitulation, we need to define an *intrinsic parameter* $h(n)$ and an *extrinsic* one $\rho(n)$, as follows.

DEFINITION 4. *For any $n \geq 0$ we set*

(a) $h(n) := \min\{z \geq n : D_{n,z} = 0\}$;

(b) $\rho(n) := \min\{z \geq n : H_{n,z} = H_z\}$;

(c) $\widetilde{\rho}(n) := \min\{z \geq n : \mathrm{rk}_p(H_{n,z}) = \mathrm{rk}_p(H_n)\}$.

REMARK 1.

(i) If $n \geq n_0$, the previous results imply $\rho(n) = \min\{z \geq n : D_{n,z} = D_z\}$.

(ii) If $n \geq n_0$, then $n \leq \widetilde{\rho}(n) \leq \rho(n) \leq h(n)$.

PROPOSITION 2. *Let $\delta, \varepsilon \in \mathbb{N}$ such that $|D| = p^\delta$ and p^ε is the exponent of D (i.e., the minimum positive integer t for which $tD = 0$). Then*

(a) for every $n \geq 0$, we have $h(n) - n \leq \delta$ and, for every $n \geq \delta - 1$, $h(n) - n = \varepsilon$;

(b) $h(n) - n = \varepsilon$ holds, also, for all $n \geqslant r$.

In the future we will continue to use δ and ε with the same meaning as in the previous proposition. Now we observe how the relative capitulation kernels $H_{n,m}$ have a rather irregular and therefore more interesting behaviour, as shown by the following

THEOREM 8.

(a) If $n_0 \leq n < r$, then we have

$$\begin{aligned} 1 = |H_{n,n}| \leq |H_{n,n+1}| \leq |H_{n,n+2}| \leq \ldots \leq |H_{n,r}| \\ = |H_{n,r}| < |H_{n,r+1}| < |H_{n,r+2}| < \ldots < |H_{n,h(n)}| \\ = |H_{n,h(n)}| = |H_{n,h(n)+1}| = |H_{n,h(n)+2}| = \ldots = |D/D_n|. \end{aligned} \tag{10}$$

(b) If $n \geq r$, then $|H_{n,m}| = |D|/|D_{n,m}|$ for all $m \geq n$ and

$$1 = |H_{n,n}| < |H_{n,n+1}| < \ldots < |H_{n,h(n)}| = |H_{n,h(n)+1}| = \ldots = |D|.$$

The three-line layout used in (10) with the last elements of the first two lines repeated on the next line, should help to better visualize what happens. If, for instance, $h(n) = r$, then the middle row disappears and we could have the stabilization of the order of the $H_{n,m}$'s even before arriving at the invariant $r(K/k)$. If, instead, $h(n) \neq r$ and the first line of (10) shows signs of equality, then we are faced with a "false stabilization" (which is something unusual and therefore very interesting in Iwasawa theory) as we will also see in Example 1 of Section 5.[†]

We close this subsection by collecting, for future use, some useful facts and consequences of what we have seen so far in the following

PROPOSITION 3.

(a) *If $n \geq n_0$ and $h(n) \neq r$, then $\rho(n) = h(n)$. Moreover if $n \geq r$, then $\rho(n) = h(n) = n + \varepsilon$.*

(b) *For all $n \geq n_0$, $\rho(n) - n \leq \delta$.*

(c) *Let $D \neq 0$ and $r \geq \delta$. If $r > n_0 + 1$ or $r = n_0 + 1$ and $D \not\subseteq Y_{n_0}$, then $\rho(r-1) = r - 1 + \varepsilon$.*

We finally notice that for other cases of false stabilization in a rather different context (non-abelian Iwasawa theory), the interested reader can see [4, Subsection 2.1].

3.2. Equivalent conditions to $\mu(K/k) = 0$ and/or $\lambda(K/k) = 0$.

As recalled in Section 2, the following conditions, found by different authors in the last 30 years of the last century, are equivalent to the finiteness of the Iwasawa module $X(K/k)$:

(i) $A_n = A_{n+1}$ for some $n \geq n_0$;

(ii) $H_n = A_n$ for all $n \geq 0$;

[†] See moreover the discussion immediately after the proof of Proposition 5.

(iii) $H_n = A_n$ for some $n \geq n_0 + 1$;

(iv) $\mathrm{Ker}(N_{n,n-1}) \subseteq H_n$ for some $n \geq n_0 + 1$.

Note that condition (iv) is equivalent at all to Ozaki's Theorem 5[‡] and condition (iii) can be also viewed as a particular case of (iv) itself. Theorem 9 gives instead some new conditions: see [3, Section 4] for the proof and for other similar results.

THEOREM 9. *The conditions below are equivalent to $\mu(K/k) = \lambda(K/k) = 0$:*

(a) $\mathrm{Im}(i_{n,m}) = \mathrm{Im}(i_{n-1,m})$ *for some* $m \geq n \geq n_0 + 1$;

(b) $\mathrm{Ker}(N_{m,n}) = \mathrm{Ker}(N_{m,n-1})$ *for some* $m \geq n \geq n_0 + 1$;

(c) $\mathrm{rk}_p(H_n) = \mathrm{rk}_p(A_n) = \mathrm{rk}_p(A_{n+1})$ *for some* $n \geq n_0$.

Note in particular as condition (c) is rather unexpected. It seems in fact the first statement that relates, or better, interprets the finiteness of the Iwasawa module $X(K/k)$ as equality not of orders like in Theorem 1, but of p-ranks.

The sequence of p-ranks $\{\mathrm{rk}_p(A_n)\}_n$ is classically linked only to the μ-invariant (e.g., it is bounded if and only if $\mu = 0$, see [27, Proposition 13.23]) and has the property of stabilization expressed in Theorem 2. The following theorem, instead, provides new insights in this direction: see [3, Theorem 4.4] for the proof and also [3, Subsections 4.1–4.3] for other properties regarding p-ranks.

THEOREM 10. *The following conditions are equivalent to $\mu(K/k) = 0$:*

(a) $\mathrm{rk}_p(\mathrm{Ker}(N_{m,n})) = \mathrm{rk}_p(\mathrm{Ker}(N_{m+1,n}))$ *for some* $m \geq n \geq n_0$;

(b) $\mathrm{rk}_p(\mathrm{Ker}(N_{m,n})) = \mathrm{rk}_p(\mathrm{Ker}(N_{m,n-1}))$ *for some* $m \geq n \geq n_0 + 1$;

(c) $\mathrm{rk}_p(\mathrm{Coker}(i_{n,m})) = \mathrm{rk}_p(\mathrm{Coker}(i_{n,m+1}))$ *for some* $m \geq n \geq n_0$;

(d) $\mathrm{rk}_p(\mathrm{Coker}(i_{n,m})) = \mathrm{rk}_p(\mathrm{Coker}(i_{n-1,m}))$ *for some* $m \geq n \geq n_0 + 1$;

(e) $\mathrm{rk}_p(\mathrm{Coker}(i_{n,m})) = \mathrm{rk}_p(A_m)$ *for some* $m \geq n \geq n_0 + 1$.

4. Bounds for $\rho(m) - \rho(n)$.

By definition we have that the last ideal in H_n capitulates exactly in $A_{\rho(n)}$, hence $\rho(n) - n$ measures how much complete capitulation is delayed in the tower. We have already given estimates for $\rho(n) - n$ in Proposition 3, and now we are going to provide bounds for the rate of growth of the sequence $\{\rho(n)\}_{n\in\mathbb{N}}$, i.e., for $\rho(m) - \rho(n)$ when $m \geq n$. By Proposition 3 (a) we know that $\rho(n) = n + \varepsilon$ for any $n \geq r$, hence for any $m \geq n \geq r$ one has $\rho(m) - \rho(n) = m - n$. But since in Iwasawa theory explicit computations are usually

[‡]It is enough, in fact, to consider the extension K/k_n.

possible only at very low levels,[§] it is more important to find bounds that hold in general and in particular for the layers n between $n_0(K/k)$ and $r(K/k)$. Hence, considering such indices, to avoid trivialities we assume $X \neq 0$ (while $D = 0$ is permitted even if no proofs would be needed in that case).

We begin with some estimates on the growth of the parameter $h(n)$ defined for all $n \geq 0$. The easiest one follows from $h(m) \leq m + \delta = m + \log_p(|D|)$ (Proposition 2 (a)) which yields

$$h(m) - h(n) \leq \log_p(|D|) + m - n \tag{11}$$

(if $D \neq 0$ one can add -1 on the right side). The following results improve this bound and lead to an estimate for $\rho(m) - \rho(n)$. They can all be formulated (and proved) in terms of the D_n's (which provide sharper bounds), but in the main statements we prefer to use the A_n's which are more convenient for computations in explicit examples.

PROPOSITION 4. *For all $m > n \geq 0$ we have*

$$h(m) - h(n) \leq \log_p\left(\frac{|D|}{|D_{n,m}|}\right) + \max\{0, m - h(n)\}. \tag{12}$$

Proof. Using Lemma 1, if $h(n) \geq m$ then $\nu_{n,m} D_{m,h(n)} = \nu_{n,m}\nu_{m,h(n)} D = \nu_{n,h(n)} D = 0$. Hence $|D_{m,h(n)}| \leq \frac{|D|}{|D_{n,m}|}$. Now note that $\nu_{h(n),h(m)} D_{m,h(n)} = D_{m,h(m)} = 0$, so, by Nakayama's Lemma and the minimality of $h(m)$, we have $|D_{m,h(n)}| \geq p^{h(m)-h(n)}$. Therefore

$$h(m) - h(n) \leq \log_p\left(\frac{|D|}{|D_{n,m}|}\right).$$

But if $m > h(n)$, then $D_{n,m} = 0$ and we write $h(m) - h(n) = h(m) - m + (m - h(n))$. Thus we also have

$$h(m) - h(n) \leq \log_p(|D|) + m - h(n) = \log_p\left(\frac{|D|}{|D_{n,m}|}\right) + m - h(n),$$

and the thesis is proved. □

Note that if $h(n) \geq m$, then the last term in (12) (i.e., $\max\{0, m - h(n)\}$) disappears: this certainly happens, for example, when $m = n + 1$.

THEOREM 11. *For all $m \geq n \geq n_0$ we have*

(a) $$h(m) - h(n) \leq \log_p\left(\frac{|A_m|}{|A_n|}\right) + \log_p(|H_{n,m}|) + m - n;$$

(b) $$h(m) - h(n) \leq \log_p\left(\frac{|A_m|}{|A_n|}\right) + \sum_{i=1}^{m-n} \log_p(|H_{n+i-1,n+i}|).$$

[§] See, for example, the methods used in [18] to determine the Iwasawa module for the cyclotomic $\mathbb{Z}_3$-extension of a real quadratic field of conductor less than 10^4 (and not congruent to 1 mod 3): all the machinery uses the very first layers of an extension and results of stabilization similar to those seen in Subsection 2.1.

Proof. For the first inequality we use Proposition 4 and Corollary 3 (a), which yield

$$h(m)-h(n) \le \log_p\left(\frac{|D|}{|D_{n,m}|}\right)+m-n = \log_p\left(\frac{|D_n|}{|D_m|}\right)+\log_p(|H_{n,m}|)+m-n.$$

Moreover note that we can embed D_n/D_m into Y_n/Y_m and $|Y_n/Y_m| = |A_m|/|A_n|$.

For the inequality (b), if $D \neq 0$, take any $i \in \{1,\dots,m-n\}$ and use (12) to get

$$h(n+i)-h(n+i-1) \le \log_p\left(\frac{|D|}{|D_{n+i-1,n+i}|}\right)$$

(note that if $D \neq 0$ then $h(n+i-1) \ge n+i$). Summing up, one finds

$$h(m)-h(n) \le \sum_{i=1}^{m-n} \log_p\left(\frac{|D|}{|D_{n+i-1,n+i}|}\right) = \log_p\left(\frac{|D|^{m-n}}{\prod_{i=1}^{m-n}|D_{n+i-1,n+i}|}\right).$$

Using again Corollary 3 (a), we have

$$\begin{aligned}
\log_p\left(\frac{|D|^{m-n}}{\prod_{i=1}^{m-n}|D_{n+i-1,n+i}|}\right) &= \log_p\left(\prod_{i=1}^{m-n}|H_{n+i-1,n+i}|\cdot\frac{|D_{n+i-1}|}{|D_{n+i}|}\right)\\
&= \log_p\left(\frac{|D_n|}{|D_m|}\right)+\log_p\left(\prod_{i=1}^{m-n}|H_{n+i-1,n+i}|\right)\\
&= \log_p\left(\frac{|D_n|}{|D_m|}\right)+\sum_{i=1}^{m-n}\log_p(|H_{n+i-1,n+i}|),
\end{aligned}$$

and since we have already seen that $|D_n/D_m| \le |A_m|/|A_n|$, then the proof of (b) is complete. □

Now we use the previous results to achieve the bounds for $\rho(m)-\rho(n)$ when $m \ge n \ge n_0$. The easiest one, coming from (11), is

$$\rho(m)-\rho(n) \le \delta+m-n, \tag{13}$$

and it is easy to realize that this bound is actually reached in very special cases.

COROLLARY 6. *For all $m \ge n \ge n_0$, we have*

(a) $\rho(m)-\rho(n) \le \log_p\left(\frac{|D|}{|D_{n,m}|}\right)+\max\{0,m-h(n)\}+(h(n)-\rho(n));$

(b) $\rho(m)-\rho(n) \le \log_p\left(\frac{|A_m|}{|A_n|}\right)+\log_p(|H_{n,m}|)+m-n+\max\{0,r-\rho(n)\};$

(c) $\rho(m)-\rho(n) \le \log_p\left(\frac{|A_m|}{|A_n|}\right)+\sum_{i=1}^{m-n}\log_p(|H_{n+i-1,n+i}|)+\max\{0,r-\rho(n)\}.$

Proof. Just note that $\rho(m)-\rho(n) \leq h(m)-\rho(n) = h(m)-h(n)+h(n)-\rho(n)$ and use the bounds of Proposition 4 and Theorem 11. □

The purpose of Corollary 6 is to find bounds that are as independent as possible from the knowledge of the module D, as done in (b) and (c). Between these two last bounds, there is not one always better than the other, but it is more convenient to use one or the other depending on the case (this, for instance, can be easily seen with techniques similar to those of the next Section 5). On the contrary, instead, the comparison between other bounds is often clear as the following remark, by way of example, shows.

REMARK 2. The bound given in Corollary 6 (a) is always better or equal to the one given in (13). The proof is an easy computation for which we have only to distinguish the following four cases: (1) $n \geq r$, (2) $n < r$, $m \leq h(n)$ and $m \geq r$, (3) $n < r$ and $m \geq h(n)$, (4) $m < r$. We prove (1) as an example: we have $\log_p\left(\frac{|D|}{|D_{n,m}|}\right) \leq \delta$, $h(n) = \rho(n) = n+\varepsilon$ and $\max\{0, m-h(n)\} = \max\{0, m-n-\varepsilon\}$. Then

$$\begin{aligned}\log_p\left(\frac{|D|}{|D_{n,m}|}\right) + \max\{0, m-h(n)\} + h(n) - \rho(n) & \\ \leq \begin{cases} \delta + m - h(n) & \text{if } m \geq h(n) \\ \delta & \text{if } m < h(n) \end{cases} & \\ \leq \delta + m - n. &\end{aligned}$$

The proofs of the other cases are similar.

5. Optimal results.

In this section we first give two examples in which we compute explicitly the values of the parameters and the invariants defined in the previous sections. Then, improving the methodologies used to construct such examples, we will show in what sense some of the results previously obtained can be considered optimal. We will therefore consider, by way of example, one of the most articulate statements like that of Theorem 8, and we will demonstrate how it is optimal, that is, not improvable from the point of view of the algebra of Λ-modules and, assuming Assumption 1, also in an absolute sense.

In the next examples we will use a result proved by M. Ozaki in [25]: for every prime p and every finite Λ-module D there exists a totally real field k whose cyclotomic $\mathbb{Z}_p$-extension has Iwasawa module isomorphic to D (see in particular [25, Theorem 1]).

EXAMPLE 1. Taking $D \simeq \Lambda/(p^u, T)$, by [25, Theorem 1] there exists at least a field k and a $\mathbb{Z}_p$-extension K/k that provides $X(K/k) \simeq D$. Moreover there exists u_0

such that $0 \leq u_0 \leq u$ and $D_0 = p^{u_0}D$. An easy calculation shows that

$$|H_{n,m}| = \begin{cases} 1 & \text{if } 0 \leq n \leq m \leq u-u_0 \\ p^{m-u+u_0} & \text{if } n \leq u-u_0 \text{ and } u-u_0 < m \leq n+u \\ p^{m-n} & \text{if } n > u-u_0 \text{ and } n \leq m \leq n+u \\ p^u & \text{if } n > u-u_0 \text{ and } m > n+u\,. \end{cases}$$

Furthermore we can easily see that our invariants and parameters take the following values: $r(K/k) = u-u_0$, $\widetilde{r}(K/k) = 0$, and for any $n \geq 0$

$$h(n) = \rho(n) = n+u \quad \text{and} \quad \widetilde{\rho}(n) = \max\{n+1,\, u-u_0+1\}.$$

In particular, if $n \leq u-u_0$, the chain of inequalities in Theorem 8 (a) becomes

$$1 = |H_{n,n}| = \ldots = |H_{n,u-u_0}| < \ldots < |H_{n,n+u}| = |H_{n,n+u+1}| = \ldots = |H_n|,$$

and whenever $n < u-u_0 < n+u$, we face a phenomenon of false, or failed, stabilization.

EXAMPLE 2. Let $v \geq 1$ and $D \simeq \Lambda/(p, T^v)$. Thus, by using [25, Theorem 1], there exists a number field k whose cyclotomic $\mathbb{Z}_p$-extension provides $X(k_{cyc}/k) \simeq D$, $n_0(k_{cyc}/k) = 0$ and $Y_0(k_{cyc}/k) = 0$ (possibly starting with some large layer). By a short computation one can check that, for all $m \geq n \geq 0$, we have

$$|H_{n,m}| = \begin{cases} p^{p^m - p^n} & \text{if } m \leq \lfloor \log_p(v+p^n-1) \rfloor \\ p^v & \text{if } m > \lfloor \log_p(v+p^n-1) \rfloor \end{cases}$$

and

$$\mathrm{rk}_p(H_{n,m}) = \begin{cases} p^m - p^n & \text{if } m \leq \lfloor \log_p(v+p^n-1) \rfloor \\ v & \text{if } m > \lfloor \log_p(v+p^n-1) \rfloor \end{cases},$$

where $\lfloor a \rfloor$ is the floor of $a \in \mathbb{R}$. Our invariants and parameters take moreover the following values: $r(k_{cyc}/k) = 0$, $\widetilde{r}(k_{cyc}/k) = 0$, and for any $n \geq 0$

$$h(n) = \rho(n) = \lfloor \log_p(v+p^n-1) \rfloor + 1 \quad \text{and} \quad \widetilde{\rho}(n) = \lfloor \log_p(v+p^n-1) \rfloor + 1.$$

As seen in the previous examples, for an explicit computation of all our parameters we need to know, in addition to X, the module Y_0 as well. There are easy cases in which $Y_0 = X$ or $Y_0 = TX$, depending also on the class group of k and on the number (and behaviour) of the primes of k above p, but in general Ozaki's results give no information about them. Going deeper in this direction we can show in what sense results like Theorem 8 can be considered optimal; it is also convenient to state the following assumption that can be considered as a conjecture generalizing Ozaki's results and similar others.

ASSUMPTION 1. Let Γ be a (multiplicative) topological group isomorphic to $\mathbb{Z}_p$ and let $D_0 \subseteq D$ be two finite $\mathbb{Z}_p[\![\Gamma]\!]$-modules. Then there exists a number field k and a $\mathbb{Z}_p$-extension K/k with $n_0(K/k) = 0$, such that

– the Iwasawa module $X(K/k)$ is isomorphic to D and

– the submodule $Y_0(K/k)$ is isomorphic to D_0,

via the isomorphism induced by $\Gamma \simeq \mathrm{Gal}(K/k)$.

The following proposition provides a strategy (and the explicit modules) to closely analyze the thesis of Theorem 8.

PROPOSITION 5. *Assuming Assumption 7, then for any $h' \geq r' \geq 0$ and any finite sequence of nonnegative integers $\{t_i\}_{1\leq i\leq h'}$ such that*

$$t_{r'+1} \geq t_{r'+2} \geq \ldots \geq t_{h'} \geq 1, \tag{14}$$

there exist a number field k and a $\mathbb{Z}_p$-extension K/k such that

$$n_0(K/k) = 0, \;\; r(K/k) = r', \;\; h(0) = h', \;\; \frac{|H_{0,i}|}{|H_{0,i-1}|} = p^{t_i} \;\; \textit{for all } 1 \leq i \leq h'$$

and

$$|H_{0,j}| = p^{\sum_{i=1}^{h'} t_i} \;\; \textit{for every } j \geq h'.$$

Sketch of proof. Let

$$D = \bigoplus_{i=1}^{r'} (\Lambda/(p^i,T))^{t_i} \oplus \Lambda/(p^{r'},T) \oplus \bigoplus_{i=r'+1}^{h'-1} (\Lambda/(p^i,T))^{t_i - t_{i+1}} \oplus (\Lambda/(p^{h'},T))^{t_{h'}}$$

and

$$\begin{aligned} D_0 \;\; = \;\; & \bigoplus_{i=1}^{r'} ((p,T)/(p^i,T))^{t_i} \oplus \Lambda/(p^{r'},T) \\ & \oplus \bigoplus_{i=r'+1}^{h'-1} ((p^{i-r'},T)/(p^i,T))^{t_i - t_{i+1}} \oplus ((p^{h'-r'},T)/(p^{h'},T))^{t_{h'}}. \end{aligned}$$

There are four summands both in D and D_0: the first one influences the part $|H_{0,0}| \leq |H_{0,1}| \leq \ldots \leq |H_{0,r'}|$, the second one, which is the same for D and D_0, guarantees that $r(K/k) = r'$, the last two affect the part $|H_{0,r'}| < |H_{0,r'+1}| < \ldots < |H_{0,h'}|$. The checking of all the claims of the thesis is a matter of direct (but rather patient) calculation that we leave to the reader. □

Now, looking at Theorem 8 (a), the previous proposition implies that every possibility for the part $|H_{n,n}| \leq |H_{n,n+1}| \leq \ldots \leq |H_{n,r}|$ (in the case $n < r$) is realizable. In particular, if $|H_{n,q}|$ seems to stabilize at a certain index m for many subsequent layers $m+1$, $m+2$, …, $m+t$, that is, $|H_{n,m}| = |H_{n,m+1}| = \ldots = |H_{n,m+t}|$ for some $n \leq m < m+t \leq r$, then this does not guarantee a definitive stabilization, i.e., we can still have $|H_{n,m+t}| < |H_{n,m+t+1}|$.

We conclude by observing that, if $n \geq n_0$ and $h(n) = r+1$, then Proposition 5 could take the following form:

Every situation not explicitly prohibited by Theorem 8 is realizable.

The reason to require $h(n) = r + 1$ is just technical and not substantial, and it lies precisely in the hypothesis (14). For instance, the proof of a similar proposition for the case $t_{r'+1} < t_{r'+2} < \ldots < t_{h'}$ can be easily realized,[1] but it involves a more complicated (and not particularly enlightening) computation.

Lastly, we notice that similar techniques to those of Proposition 5 can be employed to show that many bounds, found in the previous section, are sharp as well and not improvable in the general case. For example, we observed that the bound (13) is reached in very special cases and we proved in Remark 2 that the bound of Corollary 6 (a) is always better or equal to the one in (13). Now, using the methods of Proposition 5 and analyzing a series of cases, we can show how the bound in Corollary 6 (a) itself can be reached case by case.

References

[1] BANDINI A., *A note on p-ranks of class groups in $\mathbb{Z}_p$-extensions*, JP J. Algebra Number Theory Appl. **9** (2007), 95–103.

[2] BANDINI A., *Greenberg's conjecture and capitulation in $\mathbb{Z}_p^d$-extensions*, J. Number Theory **122** (2007), 121–134.

[3] BANDINI A. AND CALDAROLA F., *Stabilization for Iwasawa modules in $\mathbb{Z}_p$-extensions*, Rend. Semin. Matem. Univ. Padova **136** (2016), 137–155. DOI: 10.4171/RSMUP/136-10

[4] BANDINI A. AND CALDAROLA F., *Stabilization in non-abelian Iwasawa theory*, Acta Arithmetica **169** 4 (2015), 319–329. DOI: 10.4064/aa169-4-2

[5] BOURBAKI N., *Commutative Algebra, Chapters 1-7*, Spinger-Verlag, Berlin Heidelberg 1989.

[6] CALDAROLA F., *Invariants and coinvariants of class groups in $\mathbb{Z}_p$-extensions and Greenberg's Conjecture*, Riv. Matem. Univ. Parma **7** 1 (2016), 181–192.

[7] CALDAROLA F., *The exact measures of the Sierpiński d-dimensional tetrahedron in connection with a Diophantine nonlinear system*, Comm. Nonlinear Science Numer. Simul. **63** (2018), 228–238. https://doi.org/10.10.16/j.cnsns.2018.02.026

[8] T. FUKUDA, *Remarks on $\mathbb{Z}_p$-extensions of Number Fields*, Proj. Japan Acad. Ser. A **70** (1994), 264–266.

[9] M. GRANDET AND J.F. JAULENT, *Sur la capitulation dans une $\mathbb{Z}_l$-extension*, J. reine angew. Math. **362** (1985), 213–217.

[10] R. GREENBERG, *Iwasawa Theory - Past and Present*, Adv. Studies in Pure Math. **30** (2001), 335–385.

[11] R. GREENBERG, *On the Iwasawa invariants of totally real number fields*, Amer. J. Math. **98** (1976), 263–284.

[12] R. GREENBERG, *The Iwasawa invariants of Γ-extensions of a fixed number field*, Amer. J. Math. **95** (1973), 204–214.

[13] K. IWASAWA, *A note on capitulation problem for number fields*, Proc. Japan Acad. Ser. A **65** (1989), 59–61.

[14] K. IWASAWA, *A note on capitulation problem for number fields II*, Proc. Japan Acad. Ser. A **65** (1989), 183–186.

[15] K. IWASAWA, *On the theory of cyclotomic fields*, J. Math. Soc. Japan **20** (1964), 42–82.

[16] K. IWASAWA, *On $\mathbb{Z}_l$-extensions of algebraic number fields*, Ann. of Math. (2) **98** (1973), 246–326.

[1] Note also that a particular case is given by Example 2.

[17] K. IWASAWA, *On* Γ*-extensions of algebraic number fields*, Bull. Amer. Math. Soc. **65** (1959), 183–226.

[18] J.S. KRAFT AND R. SCHOOF, *Computing Iwasawa modules of real quadratic number fields*, Compositio Math. **97** (1995), 135–155.

[19] S. LANG, *Cyclotomic fields I and II - Combined second edition*, Springer-Verlag, GTM **121**, New York 1990.

[20] M. LE FLOC'H, A. MOVAHHEDI AND T. NGUYEN QUANG DO, *On capitulation cokernels in Iwasawa Theory*, Amer. J. Math. **127** 4 (2000), 851–877.

[21] J. MINARDI, *Iwasawa Modules for* $\mathbb{Z}_p^d$*-extensions of algebraic number fields*, Ph.D. Thesis, University of Washington (1986).

[22] J. NEUKIRCH, A. SCHMIDT AND K. WINGBERG, *Cohomology of Number Fields - Second edition*, Springer-Verlag, GMW **323**, Berlin Heidelberg 2008.

[23] M. OZAKI, *A Note on the Capitulation in* $\mathbb{Z}_p$*-extensions*, Proc. Japan Acad. Ser. A **71** (1995), 218–219.

[24] M. OZAKI, *Construction of real abelian fields of degree* p *with* $\lambda_p = \mu_p = 0$, Int. J. Open Problem Compt. Math. **3.2** (2009), 342–351.

[25] M. OZAKI, *Construction of* $\mathbb{Z}_p$*-extensions with prescribed Iwasawa modules*, J. Math. Soc. Japan **56** 3 (2004), 787–801.

[26] M. OZAKI, *Non-abelian Iwasawa theory of* $\mathbb{Z}_p$*-extensions*, J. reine angew. Math. **602** (2007), 59–94.

[27] L.C. WASHINGTON, *Introduction to Cyclotomic Fields - Second edition*, Springer-Verlag, GTM **83**, New York 1997.

AMS Subject Classification: 11R23, 11R29

Fabio CALDAROLA,
Department of Mathematics and Computer Science, Università della Calabria
Cubo 31/B, Ponte Bucci, 87036 Arcavacata di Rende, ITALY
e-mail: caldarola@mat.unical.it

Lavoro pervenuto in redazione il 02.11.2019.

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 43 – 56

M. Ceria*, T. Mora†, M. Sala‡

ZECH TABLEAUX AS TOOLS FOR SPARSE DECODING

Abstract. Within the framework of Groebner-free Solving, we introduce the notion of Zech tableau as a tool for producing a linear error locator polynomials for cyclic codes.

1. Introduction

In the Late Nineties, the classical approach on BCH decoding based on Berlekamp's *key equation* was upsetted by the application of Gröbner bases to the problem; it appeared a series of papers which terminated with two different proposals: Orsini-Sala general error locator polynomial [30] and Augot *et al.* Newton-Based decoder [3]; both approaches payed not only the hard pre-computation of a Gröbner basis but (mainly) the density of their decoders.

A recent work-in-progress [7, 8, 9] reconsidered the same problem within the frame of *Gröbner-free solving*, explicitly expressed and sponsored in the book [26, Vol.3,40.12,41.15]; such approach aims to avoid the computation of a Gröbner basis of a (0-dimensional) ideal $J \subset \mathcal{P}$ in favour of combinatorial algorithms describing instead the structure of the algebra $\mathcal{P}/J$.

The consequence is a preprocessing which is quadratic (and a decoding which is linear) on the length of the code.

The approach requires to describe and produce a monomial basis of the syndrome algebra; such description forced us to introduce the notion of *Zech tableau* which is the argument of this note.

2. Notations

$\mathbb{F}$ denotes an arbitrary field, $\overline{\mathbb{F}}$ denotes its algebraic closure and $\mathbb{F}_q$ denotes a finite field of size q (so q is implicitly understood to be a power of a prime) and $\mathcal{P} := \mathbb{F}[X] := \mathbb{F}[x_1, \ldots, x_n]$ the polynomial ring over the field $\mathbb{F}$.

Let $\mathcal{T}$ be the set of terms in $\mathcal{P}$, *id est*

$$\mathcal{T} := \{x_1^{\alpha_1} \cdots x_n^{\alpha_n} : (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n\}.$$

If $t = x_1^{\gamma_1} \cdots x_n^{\gamma_n} \in \mathcal{T}$, then $\deg(t) = \sum_{i=1}^n \gamma_i$ is the *degree* of t and, for each $h \in \{1, ..., n\}$, $\deg_h(t) := \deg_{x_h}(t) := \gamma_h$ is the *h-degree* of t.

*Department of Computer Science, University of Milan.
†Department of Mathematics, University of Genoa.
‡Department of Mathematics, University of Trento.

A *semigroup ordering* $<$ on $\mathcal{T}$ is a total ordering such that

$$t_1 < t_2 \Rightarrow st_1 < st_2, \text{ for each } s,t_1,t_2 \in \mathcal{T}.$$

For each semigroup ordering $<$ on $\mathcal{T}$, we can represent a polynomial $f \in \mathcal{P}$ as a linear combination of terms arranged w.r.t. $<$, with coefficients in the base field $\mathbb{F}$:

$$f = \sum_{t\in\mathcal{T}} c_t t = \sum_{t\in\mathcal{T}} c(f,t)t = \sum_{i=1}^{s} c(f,t_i)t_i : c(f,t_i) \in \mathbb{F}\setminus\{0\}, t_i \in \mathcal{T}, t_1 > ... > t_s.$$

For each such f its *support* is $\mathrm{supp}(f) := \{\tau \in \mathcal{T} : c(f,t) \neq 0\}$, its *leading term* is the term $\mathbf{T}_<(f) := \max_<(\mathrm{supp}(f)) = t_1$, its *leading coefficient* is $\mathrm{lc}_<(f) := c(f,t_1)$ and its *leading monomial* is $\mathbf{M}_<(f) := \mathrm{lc}_<(f)\mathbf{T}_<(f) = c(f,t_1)t_1$. When $<$ is understood we will drop the subscript, as in $\mathbf{T}(f) = \mathbf{T}_<(f)$.

A *term ordering* is a semigroup ordering such that 1 is lower than every variable or, equivalently, such that it is a *well ordering*.
In all paper, we consider the *lexicographical ordering* induced by $x_1 < ... < x_n$, i.e:

$$x_1^{\gamma_1}\cdots x_n^{\gamma_n} <_{Lex} x_1^{\delta_1}\cdots x_n^{\delta_n} \Leftrightarrow \exists j \,|\, \gamma_j < \delta_j, \gamma_i = \delta_i, \forall i > j,$$

which is a term ordering. Since we do not consider any term ordering other than Lex, we drop the subscript and denote it by $<$ instead of $<_{Lex}$.

The assignement of a finite set of terms

$$\mathsf{G} := \{\tau_1,\ldots,\tau_\nu\} \subset \mathcal{T}, \tau_i = x_1^{a_1^{(i)}}\cdots x_n^{a_n^{(i)}}$$

defines a partition $\mathcal{T} = \mathsf{T} \sqcup \mathsf{N}$ of $\mathcal{T}$ in two parts:

- $\mathsf{T} := \{\tau\tau_i : \tau \in \mathcal{T}, 1 \le i \le \nu\}$ which is a *semigroup ideal*, *id est* a subset $\mathsf{T} \subset \mathcal{T}$ such that
$$\tau \in \mathsf{T}, t \in \mathcal{T} \implies t\tau \in \mathsf{T};$$
- the *normal set* $\mathsf{N} := \mathcal{T} \setminus \mathsf{T}$ which is an *order ideal*, *id est* a subset $\mathsf{N} \subset \mathcal{T}$ such that
$$\tau \in \mathsf{N}, t \in \mathcal{T}, t \,|\, \tau \implies t \in \mathsf{N},$$

For any set $F \subset \mathcal{P}$, write

- $\mathbf{T}\{F\} := \{\mathbf{T}(f) : f \in F\}$;
- $\mathbf{M}\{F\} := \{\mathbf{M}(f) : f \in F\}$;
- $\mathbf{T}(F) := \{\tau\mathbf{T}(f) : \tau \in \mathcal{T}, f \in F\}$, a semigroup ideal;
- $\mathbf{N}(F) := \mathcal{T} \setminus \mathbf{T}(F)$, an order ideal;
- $\mathbb{I}(F) = \langle F \rangle$ the ideal generated by F.

– $\mathbb{F}[\mathbf{N}(F)] := \mathrm{Span}_{\mathbb{F}}(\mathbf{N}(F))$.

Given an ideal $J \subset \mathcal{P}$, denote G the minimal set of generators of the semigroup ideal $\mathsf{T} := \mathbf{T}(J)$; we denote by $\mathsf{N} := \mathbf{N}(J) = \mathcal{T} \setminus \mathbf{T}(J)$ the order ideal introduced by the partition $\mathcal{T} = \mathbf{T}(J) \sqcup \mathbf{N}(J) = \mathsf{T} \sqcup \mathsf{N}$; N will be called the *Groebner escalier* of J.

Let $\mathbf{X} = \{P_1, ..., P_N\} \subset \mathbb{F}^n$ be a finite set of simple points

$$P_i := (a_{1,i}, ..., a_{n,i}),\ i = 1, ..., N.$$

We call

$$I(\mathbf{X}) := \{f \in \mathcal{P} :\ f(P_i) = 0, \forall i\},$$

the *ideal of points* of $\mathbf{X}$.
If we are interested in the *ordered set*, instead of its support $\mathbf{X}$, we denote it by $\underline{\mathbf{X}} = [P_1, ..., P_N]$.

For any (0-dimensional, radical) ideal $J \subset \mathcal{P}$ and any extension field E of $\mathbb{F}$, let $\mathcal{V}_E(J)$ be the (finite) rational points of J over E. We also write $\mathcal{V}(J) = \mathcal{V}_{\mathbb{F}}(J)$. We have the obvious duality between I and $\mathcal{V} = \mathcal{V}_{\mathbb{F}}$.

Definition 1. For an ideal $J \subset \mathcal{P}$, a finite set $G \subset J$ will be called a *Groebner basis* of J if $\mathbf{T}(G) = \mathbf{T}(J)$, that is, $\mathbf{T}\{G\} := \{\mathbf{T}(g) : g \in G\}$ generates $\mathbf{T}(J) = \mathbf{T}\{J\}$.

We give now a brief recap on Cerlienco-Mureddu algorithm, introduced in [11, 12, 13], which is the first combinatorial algorithm that, given a finite set of simple points $\mathbf{X} = \{P_1, ..., P_N\}$ computes the lexicographical Groebner escalier $\mathsf{N}(I(\mathbf{X}))$ for the ideal of points of $\mathbf{X}$.
In particular, in [11], they consider an *ordered* finite set of simple points in $\mathbf{k}^n$, $\underline{\mathbf{X}} = [P_1, ..., P_N]$, and prove that there is a one-to-one correspondence between $\underline{\mathbf{X}}$ and the terms of the lexicographical Groebner escalier of $I(\mathbf{X})$:

$$\Phi : \underline{\mathbf{X}} \to \underline{\mathsf{N}(I(\mathbf{X}))}$$

$$P_i \mapsto x_1^{\alpha_1^{(i)}} \cdots x_n^{\alpha_n^{(i)}}.$$

They find Φ using only combinatorics on the coordinates of the elements in $\mathbf{X}$. In particular, only comparisons among the coordinates of the points are needed. The algorithm is iterative on the points and recursive on the variables, thus it pays the price of a rather bad complexity: a straightforward implementation of the algorithm is proportional to n^2N^2. Another iterative algorithm [10] gives the same result by eliminating recursion and keeping iterativity on the points, via the introduction of a data structure (the Bar Code) that stores the information on the terms needed to perform the algorithm.

We conclude this section briefly recalling the standard notation on cyclic codes, needed to understand what follows.
Let C be an $[n,k,d]_q$ q-ary cyclic code with length n, dimension k and distance d. We denote by $g(x) \in \mathbb{F}_q[x]$ its *generator polynomial*, remarking that $\deg(g) = n - k$ and

$g \mid x^n - 1$. Let $\mathbb{F}_{q^m}$ be the splitting field of $x^n - 1$ over $\mathbb{F}_q$.
If a is a primitive n-th root of unity, the *complete defining set* of C is

$$S_C = \{j | g(a^j) = 0, 0 \leq j \leq n-1\}.$$

This set is completely partitioned in cyclotomic classes, so we can pick an element for each such class, getting a set $S \subset S_C$, uniquely identifying the code. This set S is a *primary defining set* of C.
If H is a parity-check matrix of C, $\mathbf{c}$ is a codeword (i.e. $\mathbf{c} \in C$), $\mathbf{e} \in (\mathbb{F}_q)^n$ an error vector and $\mathbf{v} = \mathbf{c} + \mathbf{e}$ a received vector, the vector $\mathbf{s} \in (\mathbb{F}_{q^m})^{n-k}$ such that its transpose $\mathbf{s^T}$ is $\mathbf{s^T} = H\mathbf{v^T}$ is called *syndrome vector*. We call *correctable syndrome* a syndrome vector corresponding to an error of weight $\mu \leq t$, where t is the *error correction capability* of the code, i.e. the maximal number of errors that the code can correct.

3. Cooper Philosophy

In 1990 Cooper [17, 18] suggested to use Gröbner basis computations in order to decode cyclic codes. Let C be a binary BCH code correcting up to t errors, $\bar{s} = (s_1, \ldots, s_{2t-1})$ be the syndrome vector associated to a received word. Cooper's idea consisted in interpreting the error locations $z_1, \ldots z_t$ of C as the roots of the syndrome equation system:

$$f_i := \sum_{j=1}^{t} z_j^{2i-1} - s_{2i-1} = 0, \ 1 \leq i \leq t,$$

and, consequently, the plain error locator polynomial as the monic generator $g(z_1)$ of the principal ideal

$$\left\{ \sum_{i=1}^{t} g_i f_i, g_i \in \mathbb{F}_2(s_1, \ldots, s_{2t-1})[z_1, \ldots, z_t] \right\} \bigcap \mathbb{F}_2(s_1, \ldots, s_{2t-1})[z_1],$$

which was computed via the elimination property of lexicographical Gröbner bases.

In a series of papers [14, 15, 16] Chen et al. improved and generalized Cooper's approach to decoding. In particular, for a q-ary $[n,k,d]$ cyclic code, with error correction capability t, they made the following alternative proposals:

1. denoting, for an error with weight μ, $z_1, \ldots, z_\mu$ the error locations, $y_1, \ldots, y_\mu$ the error values, $s_1, \ldots, s_{n-k} \in \mathbb{F}_{q^m}$ the associated syndromes, they interpreted [14] the coefficients of the plain error locator polynomial as the elementary symmetric functions σ_j and the syndromes as the *Waring functions*, $s_i = \sum_{j=1}^{\mu} y_j z_j^i$, and suggested to deduce the σ_j's from the (known) s_i's via a Gröbner basis computation of the ideal generated by the Newton identities; a similar idea was later developed in [2, 3].
2. They considered [15] the *syndrome variety*, namely the variety

$$V := \left\{ (s_1, \ldots, s_{n-k}, y_1, \ldots, y_t, z_1, \ldots, z_t) \in (\mathbb{F}_{q^m})^{n-k+2t} : s_i = \sum_{j=1}^{\mu} y_j e_j^i, 1 \leq i \leq n-k \right\}$$

and proposed to deduce via a Groebner basis pre-computation in

$$\mathbb{F}_q[x_1,\ldots,x_{n-k},y_1,\ldots,y_t,z_1,\ldots,z_t]$$

a series of polynomials $g_\mu(x_1,\ldots,x_{n-k},Z),\mu\leq t$ such that, for any error with weight μ and associated syndromes $s_1,\ldots,s_{n-k}\in\mathbb{F}_{q^m}$, $g_\mu(s_1,\ldots,s_{n-k},Z)$ in $\mathbb{F}_{q^m}[Z]$ is the plain error locator polynomial. This approach was improved in a series of paper [4, 22] culminating with [30] which, specializing Gianni-Kalkbrener Theorem [20, 21], stated in Theorem 6 below.

For a survey of this *Cooper Philosophy* see [29] and on Sala-Orsini locator [5].

4. Syndrome Variety and spurious roots

The notion of *syndrome variety* was formalized in [15] in its approach to decoding q-ary $[n,k,d]$ cyclic codes, with error correction capability t.

Definition 2. For such a cyclic code, the *syndrome variety* is the set of points $\mathbf{V}:=\left\{(s_1,\ldots,s_{n-k},y_1,\ldots,y_t,z_1,\ldots,z_t)\in(\mathbb{F}_{q^m})^{n-k+2t}:s_l=\sum_{j=1}^{\mu}y_jz_j^l,\,1\leq i\leq n-k\right\}$ where for an error $(s_1,\ldots,s_{n-k},y_1,\ldots,y_t,z_1,\ldots,z_t)\in\mathbf{V}$ with weight $\mu\leq t$ and

$$y_{\mu+1}=\cdots=y_t=0,\quad z_{\mu+1}=\cdots=z_t=0,$$

$z_1,\ldots,z_\mu$ represent the *error locations*, $y_1,\ldots,y_\mu$ the *error values*, $s_1,\ldots,s_{n-k}\in\mathbb{F}_{q^m}$ *the associated syndromes*.

Definition 3. For such a cyclic code, and $\mu\leq t$ the *plain error locator polynomial* is the polynomial $\prod_{j=1}^{\mu}(X-z_i)$

Definition 4. [15, 30] A point $(s_1,\ldots,s_{n-k},y_1,\ldots,y_t,z_1,\ldots,z_t)\in\mathbf{V}$ is said *spurious* if there are at least two values $z_i,z_j,1\leq i\neq j\leq\mu$, such that $z_i=z_j\neq 0$.

Denote $\mathbf{V}_{OS}\subset\mathbf{V}$ the set of the non-spurious points of the syndrome variety and consider the polynomial set

$$\mathcal{F}_{OS}=\{f_i,h_j,\chi_i,\lambda_j,\mathsf{p}_{l\tilde{l}},\,1\leq l<\tilde{l}\leq t,1\leq i\leq n-k,1\leq j\leq t\}\;\subset\mathcal{P},$$

where

$$f_i:=\sum_{l=1}^{t}y_lz_l^j-x_i,\;\;\mathsf{p}_{l\tilde{l}}:=z_{\tilde{l}}z_l\frac{z_l^n-z_{\tilde{l}}^n}{z_l-z_{\tilde{l}}},$$
$$h_j:=z_j^{n+1}-z_j,\;\;\lambda_j:=y_j^{q-1}-1,\;\;\chi_i:=x_i^{q^m}-x_i.$$

Theorem 5. *[30] It holds* $\mathbb{I}(\mathcal{F}_{OS})=I(\mathbf{V}_{OS})$.

5. General error locator polynomial

Let G be the reduced Gröbner basis of $\mathbb{I}(\mathcal{F}_{OS}) = I(\mathbf{V}_{OS})$ w.r.t. the lex ordering with $x_1 < \cdots < x_{n-k} < z_t < \cdots < z_1 < y_1 < \cdots < y_t$ and let us denote, for each $\iota \leq t$ and each $\ell \in \mathbb{N}$

$$G_\iota := G \cap \mathbb{F}_q[x_1, \ldots, x_{n-k}, z_t, \cdots, z_\iota] \text{ and } G_{\iota\ell} := \{g \in G_\iota \setminus G_{\iota+1} : \deg_{x_\iota}(g) = \ell\}.$$

Moreover, we enumerate each $G_{\iota\ell}$ as

$$G_{\iota\ell} := \{g_{\iota\ell 1}, \ldots, g_{\iota\ell j_{\iota\ell}}\}, \mathbf{T}(g_{\iota\ell 1}) < \cdots < \mathbf{T}(g_{\iota\ell j_{\iota\ell}}).$$

Theorem 6. *[30] With the present notation we have*

1. $G \cap \mathbb{F}_q[x_1, ..., x_{n-k}, z_1, \ldots, z_t] = \cup_{i=1}^{t} G_i$;
2. $G_i = \sqcup_{\delta=1}^{i} G_{i\delta}$ *and* $G_{i\delta} \neq \emptyset$, $1 \leq i \leq t$, $1 \leq \delta \leq i$;
3. $G_{ii} = \{g_{ii1}\}$, $1 \leq i \leq t$, *i.e. exactly one polynomial exists with degree i w.r.t. the variable* z_i *in* G_i;
4. $\mathbf{T}(g_{ii1}) = z_i^i, \quad \mathrm{lc}(g_{ii1}) = 1$;
5. *if* $1 \leq i \leq t$ *and* $1 \leq \delta \leq i-1$, *then* $\forall g \in G_{i\delta}, z_1 \mid g$.

Definition 7. [30] The unique polynomial

$$g_{tt1} = z_t^t + \sum_{l=1}^{t} a_{t-l}(s_1, \ldots, s_{n-k}) z_t^{t-l}$$

with degree t w.r.t. the variable z_t in G_t, which is labelled the *general error locator polynomial*, is such that the following properties are equivalent for each syndrome vector $s = (s_1, \ldots, s_{n-k}) \in (\mathbb{F}_{q^m})^{n-k}$ corresponding to an error with weight bounded by t:

- there are exactly $\mu \leq t$ errors $\zeta_1, \ldots \zeta_\mu$;
- $a_{t-l}(s_1, \ldots, s_{n-k}) = 0$ for $l > \mu$ and $a_{t-\mu}(s_1, \ldots, s_{n-k}) \neq 0$;
- $g_{tt1}(s_1, \ldots, s_{n-k}, z_t) = z^{t-\mu} \prod_{j=1}^{\mu} (z - \zeta_i)$.

This means that the general error locator polynomial g_{tt1} is the monic polynomial in $\mathbb{F}_q[x_1, ..., x_{n-k}, z]$ which satisfies the following property:

> *given a syndrome vector* $s = (s_1, \ldots, s_{n-k}) \in (\mathbb{F}_{q^m})^{n-k}$ *corresponding to an error with weight* $\mu \leq t$, *then its* t *roots are the* μ *error locations plus zero counted with multiplicity* $t - \mu$.

Theorem 8 ([30])**.** *Every cyclic code possesses a general error locator polynomial.*

6. Degroebnerizing Error Correcting Codes (1)

Recently the same problem has been reconsidered in a group of papers [7, 9, 8] within the frame of *Groebner-free Solving* [23, 28, 23, 27], explicitly expressed and sponsored in the book [26, Vol.3,40.12,41.15]; such approach aims to avoid the computation of a Gröbner bases of a (0-dimensional) ideal $J \subset \mathcal{P}$ in favour of combinatorial algorithms, describing instead the structure of the algebra $\mathcal{P}/J$.

In particular, given the syndrome variety[*]

$$\mathsf{Z} = \left\{(c+d, c^3+d^3, c, d), c, d \in \mathbb{F}_{2^m}^*, c \neq d\right\}$$

of a BCH $[2^m-1, 2]$-code C over $\mathbb{F}_{2^m}$, and denoted $I(\mathsf{Z})$ the ideal of points of Z, [7] is able with good complexity to produce, via Cerlienco-Mureddu Algorithm [11, 12, 13] and Lazard Theorem [19], the set $\mathbf{N} := \mathbf{N}(I(\mathsf{Z}))$ and proves that the related Gröbner basis has the shape

$$G = (x_1^n - 1, g_2, z_2 + z_1 + x_1, g_4)$$

where (see [30])

$$g_2 = \frac{x_2^{\frac{n+1}{2}} - x_1^{\frac{n+1}{2}}}{x_2 - x_1} = x_2^{\frac{n-1}{2}} + \sum_{i=1}^{\frac{n-1}{2}} \binom{\frac{n-1}{2}}{i} x_1^i x_2^{\frac{n-1}{2}-i}$$

and $g_4 = z_1^2 - \sum_{t \in \mathbf{N}} c_t t$ is Sala-Orsini general error locator polynomial.

Such result allowed [7] to remark (applying Marinari-Mora Theorem [25, 1, 6]) that, for decoding, it is sufficient to compute a particular polynomial – the *half error locator polynomial* (HELP) – that is, a polynomial of the form

$$h(x_1, x_2, z_1) := z_1 - \sum_{t \in \mathbf{H}} c_t t \text{ where } \mathbf{H} := \{x_1^i x_2^j, 0 \leq i < n, 0 \leq j < \frac{n-1}{2}\}$$

which satisfies

$$h(c(1+a^{2j+1}), c^3(1+a^{3(2j+1)}), z_1) = z_1 - c, \text{ for each } c \in \mathbb{F}_{2^m}^*, 0 \leq j < \frac{n-1}{2},$$

the other error location ca^{2j+1} been computable via the polynomial $z_2 + z_1 + x_1 \in G$ as $z_2 := x_1 - z_1 = (c + ca^{2j+1}) - c = ca^{2j+1}$.

In other words, once the HELP $h(x_1, x_2, z_1)$ is known, in order to decode a received vector $\mathbf{v}$, one should:

1. compute the syndrome vector $\mathbf{s} = (s_1, s_2)$ from $\mathbf{v}$;
2. evaluate the HELP in $\mathbf{s}$, namely compute $h(s_1, s_2, z_1)$;

[*] We remark that the variables y_i, corresponding to the error values (see Definition 2) will be omitted in this paper, because talking about error values in a binary code is completely useless. Therefore $s_1 = x_1 = c + d$, $s_2 = x_2 = c^3 + d^3$ represent the two syndromes and $z_1 = c$, $z_2 = d$ represent the error locations.

3. find the unique root of $h(s_1,s_2,z_1)$ in z_1, i.e. $z_1 = c$; being an element of $\mathbb{F}_{2^m}$, it can be expressed either as $c = a^i$, $i \in \{1,...,n\}$, in terms of a fixed primitive n-th root of unity $a \in \mathbb{F}_{2^m}$ or as $c = 0$;
4. evaluate the polynomial $z_2 + z_1 + x_1$ in $(\mathbf{s}, c)$, namely compute $z_2 + c + s_1$;
5. solve $z_2 + c + s_1 = 0$, getting $z_2 = c + s_2 =: d$; being an element of $\mathbb{F}_{2^m}$, it can be expressed either as $d = a^j$, $j \in \{1,...,n\}$, again in terms of a, or as $d = 0$;
6. c, d are the two error locations, so that, if they are different from zero, they identify the position of an error. For $c = a^i, d = a^j \neq 0$, two errors occurred, exactly in positions i, j. Flipping the bits in that positions, we recover the correct sent codeword. If some of c, d are zero, it means that less than two errors occurred.

The HELP can be easily obtained with good complexity via Lundqvist interpolation formula [23] on the set of points

$$\{(c + ca^{2j+1}, c^3 + c^3 a^{3(2j+1)}, c), c \in \mathbb{F}_{2^m}^*, 0 \leq j < \frac{n-1}{2}\}.$$

Experimental showed that in that setting HELP has a very sparse formula, which has been proved in [7]:

$$h(x_1,x_2,z_1) = z_1 + \sum_{i=1}^{\frac{n-1}{2}} a_i x_1^{(4-3i) \mod n} x_2^{(i-1) \mod \frac{n-1}{2}}$$

where the unknown coefficients can be deduced by Lundqvist interpolation on the set of points

$$\{(1 + a^{2j+1}, 1 + a^{3(2j+1)}, 1), 0 \leq j < \frac{n-1}{2}\}$$

and on the monomials $\{x_1^{(4-3i) \mod n} x_2^{(i-1) \mod \frac{n+1}{2}}, 1 \leq i < \frac{n+1}{2}\}$.

Knowing the structure of the lexicographical Groebner escalier associated to the syndrome variety is a crucial step, in order to find the HELP and efficiently decode a binary cyclic code.
This suggested [9] to consider a binary cyclic code C over $GF(2^m)$, with length $n \mid 2^m - 1$ and *primary* defining set $S_C = \{1, l\}$. Thus it denoted by

- a a primitive $(2^m - 1)^{\text{th}}$ root of unity so that $\mathbb{F}_{2^m} = \mathbb{Z}_2[a]$, $\alpha := \frac{2^m-1}{n}$ and
- $b := a^\alpha$ a primitive n^{th} root of unity,
- $\mathcal{R}_u := \{e \in \mathbb{F}_{2^m} : e^n = 1\}$
- $\mathcal{S}_n := \mathcal{R}_u \sqcup \{0\}$;

considered the following sets of points

$$\mathsf{Z}_2 := \{(c + d, c^l + d^l, c, d), c, d \in \mathcal{R}_u, c \neq d\}, \#\mathsf{Z}_2^\times = n^2 - n;$$

$\mathsf{Z}_+ := \{(c+d, c^l+d^l, c, d), c, d \in \mathcal{S}_n, c \neq d\}, \#\mathsf{Z}_+^{\times} = n^2+n,$

$\mathsf{Z}_{ns} := \{(c+d, c^l+d^l, c, d), c, d \in \mathcal{S}_n\} \setminus \{(0,0,c,c), c \in \mathcal{R}_a\}, \#\mathsf{Z}_{ns}^{\times} = n^2+n+1,$

$\mathsf{Z}_e := \{(c+d, c^l+d^l, c, d), c, d \in \mathcal{S}_n\}, \#\mathsf{Z}_e^{\times} = (n+1)^2,$

and denoted, for $* \in \{e, ns, +, 2\}$,

- $J_* := I(\mathsf{Z}_*)$,
- $\mathsf{N}_* := \mathbf{N}(J_*)$ the Gröbner escalier of J_* w.r.t. the lex ordering with $x_1 < x_2 < z_1 < z_2$ and
- $\Phi_* : \mathsf{Z}_* \to \mathsf{N}_*$ a Cerlienco-Mureddu correspondence.

Then it assumed to know

(a). the structure of the order ideal N_2, $\#\mathsf{N}_2 = n^2 - n$, i.e. a minimal basis $\{t_1, \ldots, t_r\}, t_i := x_1^{a_i} x_2^{b_i}$, of the monomial ideal $\mathcal{T} \setminus \mathsf{N}_2 = \mathbf{T}(\mathfrak{I}(\mathsf{Z}_2))$,

(b). a Cerlienco Mureddu Correspodence $\Phi_2 : \mathsf{N}_2 \to \mathsf{Z}_2$

and deduced with elementary arguments N_* and Φ_* for $* \in \{e, ns, +\}$.

7. Zech Tableaux

We observe that the parameters of a minimal basis $\mathsf{G} = \{t_1, \ldots, t_r\}, t_i := x_1^{\gamma_i} x_2^{\delta_i}$, of a monomial ideal

$$\mathsf{T} \subset \mathcal{T} = \{x_1^{\gamma} x_2^{\delta} : (\gamma, \delta) \in \mathbb{N}^2\}$$

satisfy relations

- $\gamma_1 > \gamma_2 > \cdots > \gamma_r$
- $\delta_1 < \delta_2 < \cdots < \delta_r$
- and T is 0-dimensional if and only if $\delta_1 = 0 = \gamma_r$.

Indeed, the γ_i can be ordered so that $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_r$
If $\gamma_i = \gamma_{i+1}$ and without loss of generality, $\delta_i \geq \delta_{i+1}$ then $t_{i+1} | t_i$, contradicting the minimality of G. Moreover, if $\delta_i \geq \delta_{i+1}$, $t_{i+1} | t_i$ contradicting the minimality of G.
The corresponding escalier $\mathsf{N} = \mathcal{T}/\mathsf{T}$, in the zerodimensional case, is

$$\mathsf{N} := \bigsqcup_{i=1}^{r-1} \{x_1^{\gamma} x_2^{\delta} : 0 \leq \gamma < \gamma_i, 0 \leq \delta < \delta_{i+1}\}.$$

Definition 9. Consider the field $\mathbb{F}_{2^m} = \mathbb{Z}_2[a]$, a denoting a primitive $(2^m-1)^{\text{th}}$ root of unity; for a value $n \mid (2^m-1)$ we denote $\alpha := \frac{2^m-1}{n}$ and $b := a^\alpha$ a primitive n^{th} root of unity.
Denote, for $i, 0 \le i < \alpha$, $Z_i := \{j, 1 \le j \le n : 1 + b^j = 1 + a^{j\alpha} \equiv a^{i \bmod n}\}$, set $z(i) = \#Z_i$; for any set $H \subset \{j, 1 \le j \le n\}$ we consider also the values $\zeta(i) = \#(H \cap Z_i)$.
The $(2^m-1, n; H)$-*Zech Tableau* is the assignment of

- an ordered sequence $S := [j_0, ..., j_{r-1}] \subset \{i, 0 \le i < \alpha\}$ which satisfies
  - $\zeta(j_0) \ge \ldots \ge \zeta(j_{r-1}) > 0$,
  - $\zeta(j) = 0$ for each $j \notin S$.
- the minimal basis $\mathsf{G} = \{t_1, \ldots, t_r\}, t_i := x_1^{a_i} x_2^{b_i}$, of the monomial ideal $\mathsf{T} = \mathcal{T} \setminus \mathsf{N}$ corresponding to the escalier

$$\mathsf{N} := \bigsqcup_{i=1}^{r-1} \{x_1^a x_2^b : 0 \le a < a_i, 0 \le b < b_{i+1}\}.$$

Example 10. Let us consider the values $n = 21, m = 6, \alpha = \frac{63}{21} = 3$ and $O := \{2i-1, 1 \le i \le 10\}$. Let the primary defining set of our code be $S = \{1, 3\}$. The three classes induced by the 21-st roots of unity are divided in this way:
$[0] = \{1 + a^{21}, 1 + a^{45}, 1 + a^9, 1 + a^{27}\}, [1] = \{1 + a^{51}, 1 + a^{15}, 1 + a^3\}$, $[2] = \{1 + a^{39}, 1 + a^{57}, 1 + a^{33})\}$, so that $\zeta(0) = 4 > \zeta(1) = \zeta(2) = 3$ and the $(63, 21; O)$-Zech Tableau is given by the sequence $[0, 1, 2]$ and by the minimal basis $\{x_1^3, x_1 x_2^3, x_2^4\}$.

Example 11. Let us consider the value $n = 35, m = 12, \alpha = \frac{4095}{35} = 117$ and $O := \{2i-1, 1 \le i \le 17\}$. Let the primary defining set of our code be $S = \{1, 3\}$. The 35-th roots of unity, namely the powers of a^{117}: $\mathcal{R}_{35} = \{a^{117}, a^{234}, ..., a^{3978}, a^{4095} = 1\}$
The 117 classes induced by the 35-st roots of unity are divided in this way:
$[0] = \{1 + a^{2925}, 1 + a^{585}, (1 + a^{1755})\}$ and for each $u' \in \mathcal{R}_{35} \setminus \{a^{2925}, a^{585}, a^{1755}, a^{4095}\}, \{1 + u'\} = [k]$ where $1 + u = a^k, k \equiv u \mod 117$.

The $(4095, 35; O)$-Zech Tableau is given by the sequence

$$[0, 113, 106, 78, 116, 29, 58, 115, 53, 39, 73, 85, 95, 101, 109]$$

with $\zeta(0) = 3$, $\zeta(113) = \zeta(106) = \zeta(78) = \zeta(116) = \zeta(29) = \zeta(58) = \zeta(115) = \zeta(53) = \zeta(39) = \zeta(73) = \zeta(85) = \zeta(95) = \zeta(101) = \zeta(109) = 1$ and the minimal basis $\{x_1^{15}, x_1 x_2, x_2^3\}$.

8. Degroebnerizing Error Correcting Codes (2)

In this section, we deal with the case of codes such that $n \mid 2^m - 1$ with primary defining set $S_C = \{1, l\}$.
The escalier's shape is far from being trivial, and Zech tableaux will be used to study the escalier's shape.

Experiments showed that, for binary cyclic codes C over $GF(2^m)$, with length $n \mid 2^m-1$ and *primary* defining set $S_C = \{1,l\}$, the $(2^m-1,n;O)$-Zech Tableaux – with $O := \{2i-1, 1 \le i \le \frac{n-1}{2}\}$ – describe the structure of Z_2 thus making effective the results of [9] and allowing to extend those of [7]. In particular [8] reports (and proves) the following result. Still denoting a a primitive $(2^m-1)^{\text{th}}$ root of unity, $\alpha := \frac{2^m-1}{n}$ and $b := a^\alpha$ a primitive n^{th} root of unity, we consider the $(2^m-1,n;O)$-Zech Tableaux with

- ordered sequence $S := [j_0,...,j_{r-1}] \subset \{i, 0 \le i < \alpha\}$,
- minimal basis $\mathsf{G} = \{t_1,\ldots,t_r\}$, $t_i := x_1^{a_i}x_2^{b_i}$,

and let us enumerate

- each Z_{j_i} as $Z_{j_i} = [\beta_{i1},\ldots,\beta_{i\zeta(i)}]$

Then it holds.

(A). the minimal basis of $\mathbf{T}(J_2)$ is $\mathsf{G}_2 = \{\tau_1,\ldots,\tau_r\}, \tau_i := x_1^{na_i}x_2^{b_i}$, so that

(B). $\mathsf{N}_2 := \bigsqcup_{i=1}^{r-1}\{x_1^a x_2^b : 0 \le a < na_i, 0 \le b < b_{i+1}\}$ correlated to Z_2 via

(C). the Cerlienco-Mureddu correspondence

$$\Phi_2\left(b^\ell(1+b^{\beta_{i\imath}}), b^{l\ell}(1+b^{l\beta_{i\imath}}), b^\ell, b^{\ell+\beta_{i\imath}}\right) = (x_1^{(i-1)+\ell}x_2^1).$$

(D). Also in this more general frame the HELP has still a very sparse formula:

$$h(z_1,x_1,x_2) = z_1 - \sum_{j=0}^{\alpha-1} x_1^{nj+1} \sum_{i=0}^{\zeta(i)-1} a_{ji}(x_1^{-l}x_2)^i,$$

(E). where the unknown coefficient can be deduced by Lundqvist interpolation on the set of points

$$\left\{\left((1+b^{\beta_{i\imath}}), (1+b^{l\beta_{i\imath}}), 1\right)\right\}.$$

Example 10 (cont.). We have

$$\mathsf{N}_2 = \{1,x_1,...,x_1^{62},x_2,x_1x_2,...,x_1^{62}x_2,x_2^2,x_1x_2^2,...,x_1^{62}x_2^2,x_2^3,x_1x_2^3,...,x_1^{20}x_2^3\}$$

corresponding to $\mathsf{G}_2 = \{x_1^{63}, x_1^{21}x_2^3, x_2^4\}$ and HELP

$$z_1 + a^{47}x_1^{13}x_2^3 + a^{33}x_1^{58}x_2^2 + a^{47}x_1^{37}x_2^2 + a^{12}x_1^{16}x_2^2 + a^{41}x_1^{61}x_2 + a^{32}x_1^{40}x_2 + a^{47}x_1^{19}x_2 + a^{27}x_1^{43} + a^{42}x_1^{22} + a^9x_1$$

Example 11 (cont.). We have

$$\mathsf{N}_2 = \{1,x_1,...,x_1^{524},x_2,x_1x_2,...,x_1^{34}x_2,x_2^2,x_1x_2^2,...,x_1^{34}x_2^2\}$$

corresponding to $\mathsf{G}_2 = \{x_1^{525}, x_1^{35}x_2, x_2^3\}$ and HELP
$z_1 + a^{3510}x_1^{30}x_2^2a^{2340}x_1^{33}x_2 + a^{3381}x_1^{491} + a^{1140}x_1^{456} + a^{608}x_1^{421} + a^{56}x_1^{386} + a^{3477}x_1^{351} + a^{2238}x_1^{316} + a^{3445}x_1^{281} + a^{3709}x_1^{246} + a^{2260}x_1^{211} + a^{3761}x_1^{176} + a^{510}x_1^{141} + a^{400}x_1^{106} + a^{1044}x_1^{71} + a^{141}x_1^{36} + a^{1663}x_1$

References

[1] M.E. Alonso, M.G. Marinari, T Mora, *The Big Mother of All the Dualities, II: Macaulay Bases*, *Appl. Algebra Engrg. Comm. Comput.* **17** (2006) 409–451.

[2] D. Augot, M. Bardet, J.C. Faugere, Efficient decoding of (binary) cyclic codes above the correction capacity of the code using Gröbner bases, *Proc. IEEE Int. Symp. Information Theory 2003*, (2003) .

[3] D. Augot, M. Bardet, J.C. Faugere, On formulas for decoding binary cyclic codes, *Proc. IEEE Int. Symp. Information Theory 2007*, (2007) .

[4] M. Caboara, T. Mora The Chen-Reed-Helleseth-Truong Decoding Algorithm and the Gianni-Kalkbrenner Gröbner Shape Theorem, *Appl. Algebra Engrg. Comm. Comput.*, **13** (2002)

[5] F. Caruso, E. Orsini, C. Tinnirello and M. Sala *On the shape of the general error locator polynomial for cyclic codes IEEE Trans. Inform. Theory* 63.6 (2017): 3641-3657.

[6] M. Ceria, *A proof of the "Axis of Evil theorem" for distinct points*, *Rend. Semin. Mat. Univ. Politec. Torino*, Vol. 72 No. 3-4, pp. 213-233 (2014)

[7] M. Ceria, T. Mora, M. Sala, *HELP: a sparse error locator polynomial for BCH codes*, submitted.

[8] M. Ceria, *Half error locator polynomials for efficient decoding of binary cyclic codes*, in preparation.

[9] M. Ceria, *Macaulay, Lazard and the Syndrome Variety*, arxiv preprint arXiv:1910.13189 [math.CO].

[10] M. Ceria, T. Mora, *Combinatorics of ideals of points: a Cerlienco-Mureddu-like approach for an iterative lex game*, arxiv preprint arXiv:1805.09165.

[11] L. Cerlienco, M. Mureddu, *Algoritmi combinatori per l'interpolazione polinomiale in dimensione* ≥ 2, *Sém . Lothar. Combin*, B34e (1995).

[12] L. Cerlienco, M. Mureddu, *From algebraic sets to monomial linear bases by means of combinatorial algorithms*, *Discrete Math.* **139**, (1995) 73-87.

[13] L. Cerlienco, M. Mureddu, *Multivariate Interpolation and Standard Bases for Macaulay Modules*, *J. Algebra* **251** (2002), 686-726.

[14] X. Chen, I. S. Reed, T. Helleseth, K. Truong, Use of Gröbner Bases to Decode Binary Cyclic Codes up to the True Minimum Distance, *IEEE Trans. Inform. Theory*, **40** (1994) , 1654–1661.

[15] X. Chen, I. S. Reed, T. Helleseth, K. Truong, General Principles for the Algebraic Decoding of Cyclic Codes, *IEEE Trans. Inform. Theory*, **40** (1994) , 1661–1663.

[16] X. Chen, I. S. Reed, T. Helleseth, K. Truong, Algebraic decoding of cyclic codes: A polynomial Ideal Point of View, *Contemp. Math.*, **168** (1994), 15–22

[17] A.B. III Cooper, Direct solution of BCH decoding equations, In E. Arikan (Ed.) *Comm. Control and Signal Processing*, 281–286, Elsevier (1990)

[18] A.B. III Cooper, Finding BCH error locator polynomials in one step *Electron. Letters*, **27** (1991) 2090–2091

[19] D. Lazard, *Ideal Bases and Polynomial Decomposition: Case of Two Variables*, J. Symbolic Comput. **1** (1985), 261–270

[20] P. Gianni, Properties of Gröbner bases under specialization, step *Lecture Notes in Comput. Sci.*, **378** 293–297, (1991)

[21] M. Kalkbrenner, Solving systems of algebraic equations using Gröbner bases, step *Lecture Notes in Comput. Sci.*, **378** 282–292, (1991)

[22] P. Loustaunau, E.V. York, On the decoding of cyclic codes using Gröbner bases, *Appl. Algebra Engrg. Comm. Comput.*, **8** (1997) 469–483.

[23] S. Lundqvist, *Vector space bases associated to vanishing ideals of points*, J.Pure *Appl. Algebra* **214** (2010), 309-321.

[24] M.G. Marinari, T. Mora, H.M. Moeller, *Gröbner bases of ideals defined by functionals with an application to ideals of projective points*, *Appl. Algebra Engrg. Comm. Comput.* **4** (1993), 103-145.

[25] M.G. Marinari, T. Mora, *A remark on a remark by Macaulay or Enhancing Lazard Structural Theorem*, *Bull. of the Iranian Math. Soc.* **29** n 1 (2003), 103–145;

[26] T. Mora, *Solving Polynomial Equation Systems* 4 Vols., Cambridge University Press, I (2003), II (2005), III (2015), IV (2016)

[27] T. Mora, *An FGLM-like algorithm for computing the radical of a zero-dimensional ideal. J. Algebra Appl.*, **17**(01) (2018).

[28] B. Mourrain, *A New Criterion for Normal Form Algorithms* In: Fossorier M., Imai H., Lin S., Poli A. (eds) Applied Algebra, Algebraic Algorithms and Error-Correcting Codes. AAECC 1999. *Lecture Notes in Comput. Sci.*, **1719**. Springer, Berlin, Heidelberg (1999)

[29] E. Orsini, T. Mora,. *Decoding cyclic codes: the Cooper Philosophy*. in M.Sala et al., *Groebner Bases, Coding, and Cryptography*. Springer (2009), 62–92

[30] E. Orsini, M. Sala, *Correcting errors and erasures via the syndrome variety*, it J. Pure Appl. Algebra, **200** (2005), 191–226.

AMS Subject Classification: 05E40, 14G50, 11T71

M. Ceria, T. Mora, M. Sala

Lavoro pervenuto in redazione il 31.07.2019.

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 57 – 82

G. Coppola

RECENT RESULTS ON RAMANUJAN EXPANSIONS WITH APPLICATIONS TO CORRELATIONS

Abstract. This is a survey on some very recent results about Ramanujan expansions and their applications to the representation of shifted convolution sums of two arithmetical functions.

1. Some basic properties of the Ramanujan expansions

The so-called *Ramanujan sum* is [22]

$$c_q(n) \stackrel{def}{=} \sum_{\substack{j\le q\\(j,q)=1}} \cos(2\pi jn/q),\quad q\in\mathbb{N}, n\in\mathbb{Z}.$$

Hereafter, $d=(a,b)$ means that d is the greatest common divisor of a and b.

Note that $|c_q(n)|\le c_q(0)=\varphi(q)\stackrel{def}{=}\#\{j\in\mathbb{N}: j\le q \text{ and } (j,q)=1\}$ for all $q\in\mathbb{N}$, $n\in\mathbb{Z}$. Moreover, the arithmetic function $n\in\mathbb{N}\mapsto c_q(n)$ is periodic with period q. Throughout the paper, sometimes without further references, we apply other properties of the Ramanujan sums, quoted from [11], [13], [21] and summarized in the next proposition.

PROPOSITION 1. *Let μ be the Möbius function [24] and let $\omega(q)$ denote the number of the prime factors of $q\in\mathbb{N}$.*

① $c_q(n)=\sum_{\substack{d|q\\d|n}} d\mu(q/d)=\varphi(q)\dfrac{\mu\big(q/(q,n)\big)}{\varphi\big(q/(q,n)\big)}$, *for all* $q\in\mathbb{N}, n\in\mathbb{Z}$.

② $|c_q(n)|\le(q,n)$, *for all* $q,n\in\mathbb{N}$.

③ $\sum_{d|q}c_d(n)=q\mathbf{1}_{q|n}$, *for all* $q\in\mathbb{N}, n\in\mathbb{Z}$,

where $\mathbf{1}_{q|n}$ is the characteristic function of $\{n\in\mathbb{Z}:\ n\equiv 0\ (q)\}$.

④ *(Delange inequality)* $\sum_{d|q}|c_d(n)|\le n\sum_{d|q}\mu(d)^2=n2^{\omega(q)}$, *for all* $q,n\in\mathbb{N}$.

⑤ *If* $\ell,q\in\mathbb{N}$, $k\in\mathbb{Z}$, *then* $\lim_{x\to\infty}\dfrac{1}{x}\sum_{n\le x}c_\ell(n)c_q(n+k)=\begin{cases} c_\ell(k) & \text{if } \ell=q,\\ 0 & \text{otherwise.}\end{cases}$

REMARK 1. We will write $n\equiv m\ (q)$ to abbreviate $n\equiv m\ (\mathrm{mod}\, q)$. We denote the characteristic function of any set $U\cap\mathbb{Z}$, with $U\subseteq\mathbb{R}$, by $\mathbf{1}_U$. Equivalently, such a function is denoted by $\mathbf{1}_{\wp}$, as in ③ , provided that $\wp$ is a characteristic property of $U\cap\mathbb{Z}$. Note that the first equality in ① shows that $c_q(n)\in\mathbb{Z}$ for all $q\in\mathbb{N}, n\in\mathbb{Z}$.

DEFINITION 1. *Let* $\mathcal{A} \overset{def}{=} \{f : \mathbb{N} \to \mathbb{C}\}$ *be the set of all the arithmetic functions. The* **Ramanujan series** *associated to* $g \in \mathcal{A}$ *is the series*

$$\mathcal{R}_g(n) \overset{def}{=} \sum_{q=1}^{\infty} g(q) c_q(n), \quad \textit{for all } n \in \mathbb{N}.$$

We say that $f \in \mathcal{A}$ *admits a* **Ramanujan expansion** *if there exists* $g \in \mathcal{A}$ *such that*

$$f = \mathcal{R}_g.$$

For this, $g(q)$ *is the so-called q-th* **Ramanujan coefficient** *of the expansion* $f = \mathcal{R}_g$.

All classical expansions have this form (compare [22], [21], [23]).
However, we include the possibility that $g(q)$ also depends on n. As an example, take the following expansion of $0-$function, $\mathbf{0}(n) \overset{def}{=} 0$, $\forall n \in \mathbb{N}$ (recall $c_1(n) = 1$, $\forall n \in \mathbb{Z}$) :

$$\mathbf{0}(n) = 2c_1(n) + \left(2 \cdot \mathbf{1}_{n \neq 0(3)} - \mathbf{1}_{n \equiv 0(3)}\right) c_3(n), \quad \forall n \in \mathbb{N}. \tag{1}$$

In case $g(q)$ doesn't depends on n, as n varies in $\mathbb{N}$, we call $f = \mathcal{R}_g$ a **pure** expansion.

Given any $f \in \mathcal{A}$ (except $f = \mathbf{0}$, see Remark 4), we don't know if it has a pure expansion or not. (As far as we know no such a result is in the literature.)

However, all $f \in \mathcal{A}$ have at least one $n-$pointwise convergent $f(n) = \mathcal{R}_g(n)$, as $n \in \mathbb{N}$ (and it's finite, see Remark 3).

Consequently, we introduce

$$\langle f \rangle \overset{def}{=} \{g \in \mathcal{A} : \ f = \mathcal{R}_g\} \quad \text{the } \textit{Ramanujan } \textbf{cloud} \text{ of } f,$$

which is always non-empty, and its "pure part" (that can be empty)

$$\langle f \rangle_* \overset{def}{=} \{g \in \langle f \rangle : \ \mathcal{R}_g \text{ is pure}\}.$$

Assume that $\langle f \rangle_* \neq \emptyset$. We call the expansion $f = \mathcal{R}_g$ **completely uniform** if it's pure and the convergence of $f(n) = \mathcal{R}_g(n)$ is uniform w.r.t. n, as n varies in $\mathbb{N}$. We write

$$\langle f \rangle_{**} \overset{def}{=} \{g \in \langle f \rangle : \ \mathcal{R}_g \text{ is completely uniform}\}.$$

Let us write $f =_{\#} \mathcal{R}_g$ to mean that $\mathcal{R}_g$ is a finite sum. Assuming that f and g are not the identically zero functions, such a finite expansion, if pure, can be written in the form

$$f(n) = \mathcal{R}_g(n) = \sum_{q \leq Q} g(q) c_q(n), \qquad \forall n \in \mathbb{N},$$

where $Q \overset{def}{=} \max\{q \in \mathbb{N} : g(q) \neq 0\}$ is the so-called **length** of $\mathcal{R}_g(n)$. (Compare Remark 3 about finite expansions.) We indicate $\langle f \rangle_{\#} \overset{def}{=} \{g \in \langle f \rangle : \ \mathcal{R}_g \text{ is finite}\}$.

Henceforth, $\mathcal{R}$*-expansion* and $\mathcal{R}$*-coefficient* abbreviate *Ramanujan expansion* and *Ramanujan coefficient*, respectively. Sometimes, we'll abbreviate $\mathcal{R}-$cloud for the Ramanujan cloud.

REMARK 2. The reader is cautioned that some authors refer to Ramanujan series as *Fourier-Ramanujan* series ([15], [23]). Moreover, in the literature $\mathcal{R}$*-expansion* is often synonymous of Ramanujan series. Here we explicitly point out that by definition a $\mathcal{R}$-expansion is a convergent Ramanujan series taken as a representation of its sum (compare Remark 4 for its non-uniqueness). Further, it should be plain that in the present context Ramanujan sum cannot be synonymous of finite $\mathcal{R}$-expansion.

REMARK 3. A celebrated theorem of Hildebrand [18] ensures that for every $f \in \mathcal{A}$ and $n \in \mathbb{N}$ there exist $Q(n) \in \mathbb{N}$ and $h(q,n) \in \mathbb{C}$, with $1 \leq q \leq Q(n)$, such that

$$f(n) = \sum_{q=1}^{Q(n)} h(q,n) c_q(n) \qquad \forall n \in \mathbb{N}. \tag{2}$$

See also [23] for the proof, where the coefficients $h(q,n)$ are recursively defined. In other words, Hildebrand's result yields that $\langle f \rangle_{\#} \neq \emptyset$ for all $f \in \mathcal{A}$, implying that all $\mathcal{R}-$clouds are non-empty; however, it leaves open the possibilty that some $\langle f \rangle_{*} = \emptyset$. Note that for every $f \in \mathcal{A}$ there are always expansions of the form (2), where the dependence of $Q(n)$ and the coefficients $h(q,n)$ on n is effective. Indeed, besides the trivial choices of $Q(n) = 1$ and $h(1,n) = f(n)$, the expansion (2) holds also by taking

$$Q(n) = n, \quad h(q,n) = \sum_{\substack{d \leq n \\ d \equiv 0 (q)}} \frac{(f * \mu)(d)}{d},$$

where $*$ denotes the Dirichlet product [24]. (See the last line of the proof of the Wintner-Delange formula in Proposition 2 below.) The latter case yields the so-called *standard finite* $\mathcal{R}$*-expansion* of f, $\forall f \in \mathcal{A}$. This, in turn, proves (2) immediately.

REMARK 4. Since the first appearance of the Ramanujan series, it was clear at once that the $\mathcal{R}$-expansion of a given arithmetical function is very far from being unique. Indeed, besides the trivial fact that the identically zero function $\mathbf{0}$ belongs to $\langle \mathbf{0} \rangle$, non-trivial $\mathcal{R}$-expansions of $\mathbf{0}$ were found by Ramanujan himself [22] and Hardy [16], respectively as

$$\mathbf{0}(n) = \sum_{q=1}^{\infty} R_0(q) c_q(n), \qquad \text{where} \quad R_0(q) \stackrel{def}{=} \frac{1}{q}, \tag{3}$$

$$\mathbf{0}(n) = \sum_{q=1}^{\infty} H_0(q) c_q(n), \qquad \text{where} \quad H_0(q) \stackrel{def}{=} \frac{1}{\varphi(q)}. \tag{4}$$

Further samples of Ramanujan expansions are found in [23], [19], [21]. Moreover, it is plain that $\alpha \langle \mathbf{0} \rangle \subseteq \langle \mathbf{0} \rangle$ for all $\alpha \in \mathbb{C}$. Furthermore, for any $g \in \langle f \rangle$ one has

$$g + \langle \mathbf{0} \rangle \stackrel{def}{=} \{h \in \mathcal{A} : \ h = g + k \text{ for some } k \in \langle \mathbf{0} \rangle\} \subseteq \langle f \rangle.$$

Together with the aforementioned Hildebrand's result, this implies that the set $\langle f\rangle$ is infinite for any $f\in\mathcal{A}$. Namely, all $\mathcal{R}-$clouds contain infinitely many expansions each. Further, they are convex sets : $\alpha g_1+(1-\alpha)g_2\in\langle f\rangle$, $\forall g_1,g_2\in\langle f\rangle$ and $\forall\alpha\in\mathbb{C}$.
Some of the recent results concern the problem of the unique representation of the expansion $f=\mathcal{R}_g$, namely the search for suitable requirements on g that would yield uniqueness of such expansion. These results are discussed in §3.

REMARK 5. The convergence of a $\mathcal{R}$-expansion needs not be absolute. Indeed, since $|c_q(n)|=\mu(q)^2$ for $(q,n)=1$ (see Proposition 1), then for any fixed integer n we can write

$$\sum_{q=1}^{\infty}\frac{|c_q(n)|}{q}\geq\sum_{\substack{q=1\\(q,n)=1}}^{\infty}\frac{\mu(q)^2}{q}\geq\sum_p\frac{1}{p}-\sum_{p|n}\frac{1}{p}.$$

Hereafter, the letter p, with or without subscripts, is devoted to prime numbers. Thus, the absolute divergence of the series (3) follows from the well-known divergence of the series of prime numbers reciprocals.

2. The Wintner coefficients and the Carmichael coefficients

DEFINITION 2. *The Eratosthenes transform of f is $f'\in\mathcal{A}$ such that $f=f'*\mathbf{1}$, i.e.*

$$f(n)=\sum_{d|n}f'(d),\quad\forall n\in\mathbb{N}.$$

In particular, if for every $n\in\mathbb{N}$ and for some $Q\in\mathbb{N}$ independent of n one has

$$f(n)=\sum_{\substack{d\leq Q\\d|n}}f'(d),$$

then f is said to be a truncated divisor sum of range Q. We also say that f' is the Q-truncated Eratosthenes transform of f. The set of the truncated divisor sums of range Q is denoted by $\mathcal{A}_Q$.

REMARK 6. Henceforth, *E-transform* means *Eratosthenes transform*. Note that we have already abbreviated Wintner's terminology, where f' is used to be called the *Eratosthenes-Möbius transform* of f (see [23], [25]), being plain that $f'=f*\mu$ by the Möbius inversion formula [24]. We also refer to f as the *inverse E*-transform of f'. Finally, by definition the truncated divisor sum of range Q *associated* to $f=f'*\mathbf{1}$ is

$$f_Q(n)\stackrel{def}{=}\sum_{\substack{d\leq Q\\d|n}}f'(d),\quad\forall n\in\mathbb{N}.$$

In other words, the *E*-transform of f_Q is f' in $[1,Q]\cap\mathbb{N}$ and 0 otherwise. In §3, the set $\mathcal{A}_Q$ is characterized in terms of some peculiar $\mathcal{R}$-expansions (see Theorem 3).

DEFINITION 3. *Let f' be the E-transform of $f \in \mathcal{A}$.*

If $\displaystyle\sum_{\substack{d=1\\ d\equiv 0(q)}}^{\infty} \frac{f'(d)}{d}$ *converges, then its sum $\mathcal{W}_f(q)$ is the q-th* **Wintner coefficient** *of f.*

If $\mathcal{M}(f\cdot c_q) \stackrel{def}{=} \lim_{x\to\infty} \frac{1}{x}\sum_{n\le x} f(n)c_q(n)$ *exists and is finite, then* $\mathcal{C}_f(q) \stackrel{def}{=} \dfrac{\mathcal{M}(f\cdot c_q)}{\varphi(q)}$ *is the q-th* **Carmichael coefficient** *of f.*

REMARK 7. If it exists and is finite, then $\mathcal{M}(f\cdot c_q)$ is the so-called *mean value* of $f\cdot c_q$. Namely, $\mathcal{M}(f)=\mathcal{M}(f\cdot c_1)=\mathcal{C}_f(1)=\lim_{x\to\infty}\frac{1}{x}\sum_{n\le x} f(n)$ is the mean value of f.

The next proposition summarizes three classical results. In the first part we spot *Wintner's criterion* [23], a sufficient condition for the mere existence of both Wintner and Carmichael coefficients, which turn out to be equal. The second part of the proposition is the Wintner-Delange theorem [13], that provides with a sufficient condition for a given $f\in\mathcal{A}$ to be such that $\langle f\rangle_* \neq \emptyset$. In particular, such a theorem reveals that the Wintner coefficients (or equivalently the Carmichael ones in view of Wintner's criterion) are instances of $\mathcal{R}$-coefficients for f. The third part is Lucht's theorem [19] that gives a deep link between the $\mathcal{R}$-expansion of a function and its E-transform. In §3 we present a new result yielding the converse of Lucht's theorem. Such a result is a key argument for the problem of the unique $\mathcal{R}$-expansion (see Remark 4). In what follows, for any $f,g\in\mathcal{A}$ with g real and non-negative, the notation $f(n)\ll g(n)$, equivalent to $f(n)=O(g(n))$, means that there exist $n_0\in\mathbb{N}$ and a real number $C>0$ such that $|f(n)|\le Cg(n)$ for all $n>n_0$. The implicit constant C might depend on other variables, in which case they are displayed as subscripts in the symbols $\ll$ or O.

PROPOSITION 2. *Let f' be the E-transform of $f\in\mathcal{A}$.*

① *Wintner's criterion. If* $\displaystyle\sum_{d=1}^{\infty}\frac{f'(d)}{d}$ *converges absolutely, then $\mathcal{W}_f(q)$ and $\mathcal{C}_f(q)$ exist for all $q\in\mathbb{N}$. Moreover, one has $\mathcal{W}_f=\mathcal{C}_f$.*

② *The Wintner-Delange formula. If* $\displaystyle\sum_{d=1}^{\infty}2^{\omega(d)}\frac{f'(d)}{d}$ *converges absolutely, then the function $\mathcal{W}_f=\mathcal{C}_f$ belongs to $\langle f\rangle_*$:*

$$f(n)=\sum_{q=1}^{\infty}\mathcal{W}_f(q)c_q(n)=\sum_{q=1}^{\infty}\mathcal{C}_f(q)c_q(n),\quad \forall n\in\mathbb{N}.$$

③ *Lucht's theorem. If* $\displaystyle\sum_{\substack{q=1\\ q\equiv 0(d)}}^{\infty} g(q)\mu(q/d)$ *converges for every $d\in\mathbb{N}$, then $g\in\langle f\rangle$,*

where f is the inverse E-transform of $f'(d) \overset{def}{=} d \sum\limits_{\substack{q=1 \\ q\equiv 0\,(d)}}^{\infty} g(q)\mu(q/d)$, *i.e.*

$$f(n) = \sum_{d|n} f'(d) = \sum_{d|n} d \sum_{\substack{q=1 \\ q\equiv 0\,(d)}}^{\infty} g(q)\mu(q/d) = \mathcal{R}_g(n) \quad \forall n \in \mathbb{N}.$$

Proof. ① Clearly, the existence of $\mathcal{W}_f(q)$ for all $q \in \mathbb{N}$ is a straightforward consequence of the hypothesis. Thus, we have to show that for any fixed $q \in \mathbb{N}$ one has

$$\lim_{x\to\infty} \frac{1}{x} \sum_{n\le x} f(n)c_q(n) = \varphi(q)\mathcal{W}_f(q).$$

To this end, after recalling that $[\beta]$ and $\|\beta\|$ denote respectively the integer part of $\beta \in \mathbb{R}$ and the distance of β from the nearest integer, let us write

$$\begin{aligned}
\sum_{n\le x} f(n)c_q(n) &= \sum_{n\le x} c_q(n) \sum_{d|n} f'(d) = \sum_{d\le x} f'(d) \sum_{m\le x/d} c_q(dm) \\
&= \sum_{d\le x} f'(d) \sum_{\substack{j\le q \\ (j,q)=1}} \sum_{m\le x/d} e_q(jdm) \\
&= \varphi(q) \sum_{\substack{d\le x \\ d\equiv 0\,(q)}} f'(d)\left[\frac{x}{d}\right] + O\Big(\sum_{\substack{d\le x \\ d\not\equiv 0\,(q)}} |f'(d)| \sum_{\substack{j\le q \\ (j,q)=1}} \left\| \frac{jd}{q} \right\|^{-1} \Big),
\end{aligned}$$

where we have applied the well-known inequality (see [12], Ch.26)

$$\sum_{m\le x} e(m\beta) \ll \min(x, \|\beta\|^{-1}), \quad \forall x \ge 1, \forall \beta \in \mathbb{R}.$$

Since the O-term vanishes for $q = 1$, we can assume $q > 1$ henceforth. We see that

$$\begin{aligned}
\frac{1}{x} \sum_{n\le x} f(n)c_q(n) = \varphi(q) \sum_{\substack{d\le x \\ d\equiv 0\,(q)}} \frac{f'(d)}{d} &+ O\Big(\frac{\varphi(q)}{x} \sum_{d\le x} |f'(d)| \Big) \\
&+ O\Big(\frac{1}{x} \sum_{\substack{d\le x \\ d\not\equiv 0\,(q)}} |f'(d)| \sum_{\substack{j\le q \\ (j,q)=1}} \left\| \frac{jd}{q} \right\|^{-1} \Big).
\end{aligned}$$

Now, let us write

$$\sum_{\substack{d\le x \\ d\not\equiv 0\,(q)}} |f'(d)| \sum_{\substack{j\le q \\ (j,q)=1}} \left\| \frac{jd}{q} \right\|^{-1} \le \sum_{r<q} \sum_{\substack{d\le x \\ (d,q)=r}} |f'(d)| \sum_{\substack{j\le q \\ j\not\equiv 0\,(q/r)}} \left\| \frac{jd/r}{q/r} \right\|^{-1}$$

and note that, since $(d,q) = r$ yields $(d/r, q/r) = 1$, it turns out that (compare [20],§3.2)

$$\sum_{\substack{j\le q \\ j\not\equiv 0\,(q/r)}} \left\| \frac{jd/r}{q/r} \right\|^{-1} \le r \sum_{j'<q/r} \left\| \frac{j'}{q/r} \right\|^{-1} \ll q \sum_{j'<q/r} \frac{1}{j'} \ll q\log q, \quad \forall r|q, r \ne q.$$

Thus, we get

$$\frac{1}{x}\sum_{n\le x} f(n)c_q(n) = \varphi(q)\sum_{\substack{d\le x\\ d\equiv 0\,(q)}}\frac{f'(d)}{d} + O_q\Big(\frac{1}{x}\sum_{d\le x}|f'(d)|\Big). \tag{5}$$

Hence, the desired conclusion follows once it is shown that the O-term goes to zero as $x\to\infty$. To this end, by partial summation [24] we write

$$\sum_{d\le x}|f'(d)| = \sum_{d\le x}\frac{|f'(d)|}{d} + \int_1^x\Big(\sum_{d\le x}\frac{|f'(d)|}{d} - \sum_{d\le t}\frac{|f'(d)|}{d}\Big)\,\mathrm{d}t.$$

Note that by hypothesis, for any fixed real number $\varepsilon>0$, there exists $x_\varepsilon < x$ such that

$$\Big|\sum_{d\le x}\frac{|f'(d)|}{d} - \sum_{d\le t}\frac{|f'(d)|}{d}\Big| < \varepsilon,\ \forall t\in(x_\varepsilon, x).$$

Consequently, for all $x > x_\varepsilon$ one has

$$\begin{aligned}\frac{1}{x}\sum_{d\le x}|f'(d)| &< \frac{1}{x}\sum_{d\le x}\frac{|f'(d)|}{d} + \frac{1}{x}\int_1^{x_\varepsilon}\Big|\sum_{d\le x}\frac{|f'(d)|}{d} - \sum_{d\le t}\frac{|f'(d)|}{d}\Big|\,\mathrm{d}t + \varepsilon\\ &\le \frac{1+2x_\varepsilon}{x}\sum_{d=1}^{\infty}\frac{|f'(d)|}{d} + \varepsilon.\end{aligned}$$

② It is plain that the hypothesis and ① yield that $\mathcal{W}_f = \mathcal{C}_f$. From ④ of Prop. 1 we get

$$\begin{aligned}\sum_{q\le x}|\mathcal{W}_f(q)c_q(n)| &\le \sum_{q\le x}|c_q(n)|\sum_{d\equiv 0\,(q)}\frac{|f'(d)|}{d} = \sum_{d=1}^{\infty}\frac{|f'(d)|}{d}\sum_{\substack{q|d\\ q\le x}}|c_q(n)|\\ &\le n\sum_{d=1}^{\infty}\frac{2^{\omega(d)}}{d}\,\big|f'(d)\big|,\end{aligned}$$

where the double series on d and q converges absolutely because the latter series converges by hypothesis. (In particular, $\mathcal{R}_{\mathcal{W}_f}(n)$ is absolutely convergent for any fixed n.) Hence, we can exchange d and q summations, and apply ③ of Proposition 1 so that

$$\mathcal{R}_{\mathcal{W}_f}(n) = \sum_{q=1}^{\infty}\mathcal{W}_f(q)c_q(n) = \sum_{d=1}^{\infty}\frac{f'(d)}{d}\sum_{q|d}c_q(n) = \sum_{d=1}^{\infty}f'(d)\mathbf{1}_{d|n} = f(n),\ \ \forall n\in\mathbb{N}.$$

③ For $x\ge n$, from ① of Proposition 1 we get

$$\sum_{q\le x}g(q)c_q(n) = \sum_{d|n}d\sum_{\substack{q\le x\\ q\equiv 0\,(d)}}g(q)\mu(q/d),$$

yielding the conclusion immediately. The proposition is completely proved. □

REMARK 8. We underline that the absolute convergence of $\mathcal{R}_{\mathcal{W}_f}$ alone does not suffice to conclude that $f = \mathcal{R}_{\mathcal{W}_f}$. For example, if $\mathcal{R}_{\mathcal{W}_f}(n)$ is a finite sum, i.e. there exists $Q \in \mathbb{N}$ such that $\mathcal{W}_f(q) = 0$ for all $q > Q$, then obviously its convergence is absolute. However, the argument used to prove that $f = \mathcal{R}_{\mathcal{W}_f}$ in ② is no longer helpful because ③ of Proposition 1 cannot apply to

$$\mathcal{R}_{\mathcal{W}_f}(n) = \sum_{d=1}^{\infty} \frac{f'(d)}{d} \sum_{\substack{q \le Q \\ q|d}} c_q(n).$$

The reader should compare this case with Remark 14 after Theorem 3 below.

REMARK 9. We give many small new results, now.
Note that $|c_q(n)| \le \varphi(q)$ yields

$$\frac{1}{\varphi(q)} \left| \frac{1}{x} \sum_{n \le x} f(n) c_q(n) \right| \le \frac{1}{x} \sum_{n \le x} |f(n)|.$$

Hence, if $\mathcal{C}_f(q)$ and the mean value of $|f|$ exist, then $|\mathcal{C}_f(q)| \le \mathcal{M}(|f|) = \mathcal{C}_{|f|}(1)$. As a consequence, if $\mathcal{C}_f(q)$ exists for all $q \in \mathbb{N}$, then $\mathcal{M}(|f|) = 0$ implies $\mathcal{C}_f = \mathbf{0}$. In particular, this means that *a non-negative real function $f \neq \mathbf{0}$ with a null mean value does not admit its Carmichael coefficients as $\mathcal{R}$-coefficients, i.e. $\mathcal{C}_f \notin \langle f \rangle$*. Samples of such functions are the characteristic functions of subsets of $\mathbb{N}$ with zero density. Indeed, recalling that the density of $B \subseteq \mathbb{N}$ is

$$\delta(B) \stackrel{def}{=} \lim_{x \to \infty} \frac{\#\{n \in B : n \le x\}}{x} \in [0,1]$$

(provided that such a limit exists), this can be equivalently written as

$$\delta(B) = \lim_{x \to \infty} \frac{1}{x} \sum_{n \le x} \mathbf{1}_B(n) = \mathcal{C}_B(1),$$

where the first Carmichael coefficient of $\mathbf{1}_B$ is shortly denoted $\mathcal{C}_B(1)$. In particular, the set of prime numbers has zero density (esp., from the prime number theorem).
Further, it is well-known [T] that the inverse E-transform of the Liouville function, i.e.,

$$\lambda(p_1^{\alpha_1} p_2^{\alpha_2} \dots p_t^{\alpha_t}) \stackrel{def}{=} (-1)^{\alpha_1 + \alpha_2 + \dots + \alpha_t},$$

is the characteristic function of the set S of the square numbers, that has plainly zero density, i.e. $\mathcal{C}_S = \mathbf{0}$; a theorem of Landau and von Mangoldt states that $\sum_{d=1}^{\infty} \frac{\lambda(d)}{d} = 0$ is equivalent to the prime number theorem. Thus, being λ completely multiplicative, we see that $\sum_{\substack{d=1 \\ d \equiv 0 \,(q)}}^{\infty} \frac{\lambda(d)}{d} = \frac{\lambda(q)}{q} \sum_{d=1}^{\infty} \frac{\lambda(d)}{d} = 0$ for all $q \in \mathbb{N}$, i.e. $\mathcal{W}_S = \mathbf{0} = \mathcal{C}_S$.

On the other hand, $|\mathbf{1}'_S| = |\lambda| = \mathbf{1}$ doesn't satisfy Wintner's criterion hypothesis.

More in general, if $f \in \mathcal{A}$ is such that f' is completely multiplicative (c.m.), then

$$\mathcal{W}_f(q) = \sum_{d \equiv 0\,(q)} \frac{f'(d)}{d} = \frac{f'(q)}{q} \sum_{m=1}^{\infty} \frac{f'(m)}{m} = \frac{f'(q)}{q}\,\mathcal{W}_f(1), \quad \forall q \in \mathbb{N}.$$

Consequently,

$$f' \text{ c.m.}, \quad \mathcal{W}_f(1) = 0 \implies \mathcal{W}_f = \mathbf{0}$$

and also

$$f' \text{ c.m.}, \quad \mathcal{W}_f(1) \neq 0 \text{ and } \mathcal{W}_f(q) = 0, \forall q > Q \implies f'(q) = 0, \forall q > Q.$$

Furthermore, it is easily seen that (notice that here f' is not necessarily c.m.)

$$f' \geq 0 \text{ and } \mathcal{W}_f(q) = 0, \forall q > Q \implies f'(q) = 0, \quad \forall q > Q.$$

These properties suggest the following

Conjecture: *If $f \in \mathcal{A}$ is such that $\mathcal{W}_f(1) \neq 0$, then*

$$\mathcal{W}_f(q) = 0, \quad \forall q > Q \implies f'(q) = 0, \quad \forall q > Q.$$

In §4 it is shown how such a conjecture might replace the Delange hypothesis on the series $\sum_{d=1}^{\infty} 2^{\omega(d)} f'(d)/d$ within Proposition 2 in pursuing the Wintner-Delange formula for the shifted convolution sums.

REMARK 10. Formula (5) reveals that if $\mathcal{C}_{|f'|}(1) = 0$, i.e.

$$\sum_{d \leq x} |f'(d)| = o(x), \quad \text{as } x \to \infty, \tag{6}$$

then $\mathcal{C}_f(q)$ exists if and only if $\mathcal{W}_f(q)$ does. Further, if this is the case, then $\mathcal{C}_f(q) = \mathcal{W}_f(q)$. From the proof of Wintner's criterion it transpires that the absolute convergence of $\sum_{d=1}^{\infty} \frac{f'(d)}{d}$ implies (6), which alone however does not yield the existence of the Wintner coefficients; on the other hand, by taking $f'(d) = 1/\log(d+1)$ it is easily seen that the converse of such an implication is not true. Moreover, by taking f' as the Liouville function λ, it is plain that (6) does not hold, while the characteristic function of square numbers $\mathbf{1}_S = \lambda * \mathbf{1} = f' * \mathbf{1} = f$, say, satisfies the hypotheses of the next proposition, that is a result of Delange (see the theorem and remark 1.5 in [14]).

PROPOSITION 3. *Let $f \in \mathcal{A}$ and $q \in \mathbb{N}$ be such that $\sum_{n \leq x} |f(n)| = O(x)$ and $\mathcal{C}_f(d)$ exists for all $d|q$. Then $\mathcal{C}_f(q) = \mathcal{W}_f(q)$.*

In particular, by taking $q=1$, this result yields that if $\sum_{n\le x}|f(n)| = O(x)$ and there exists the mean value $\mathcal{M}(f) = C_f(1)$, then

$$\mathcal{M}(f) = \sum_{d=1}^{\infty} \frac{f'(d)}{d}.$$

3. Uniqueness for Ramanujan coefficients

Here we quote from [2] the next theorem, that provides with a kind of converse of Lucht's theorem (see ③ of Proposition 2).

THEOREM 1. *Let $f \in \mathcal{A}$ be such that $\langle f\rangle_* \neq \emptyset$.*
① *For any given $g \in \langle f\rangle_*$ the E-transform of f is*

$$f' : d \in \mathbb{N} \to f'(d) = d \sum_{\substack{q=1\\ q\equiv 0\,(d)}}^{\infty} g(q)\mu(q/d).$$

② *If $g \in \langle f\rangle_*$ is such that*

$$\sum_{q=1}^{\infty} 2^{\omega(q)} g(q) \quad \textit{converges absolutely}, \tag{7}$$

then $g = \mathcal{W}_f$.

Proof. ① We can exchange the sums in

$$\sum_{d|n} d \sum_{\substack{q=1\\ q\equiv 0\,(d)}}^{\infty} g(q)\mu(q/d)$$

because g does not depend on n by hypothesis. Thus, from ① of Proposition 1 for $x \ge n$ we get that

$$\sum_{d|n} d \sum_{\substack{q\le x\\ q\equiv 0\,(d)}} g(q)\mu(q/d) = \sum_{q\le x} g(q) c_q(n).$$

As $x \to \infty$, it follows that

$$\sum_{d|n} d \sum_{\substack{q=1\\ q\equiv 0\,(d)}}^{\infty} g(q)\mu(q/d) = \mathcal{R}_g(n) = f(n),$$

yielding that the E-transform of f is the claimed f'.
② From ① one has that

$$\mathcal{W}_f(q) = \sum_{\substack{d=1\\ d\equiv 0\,(q)}}^{\infty} \frac{f'(d)}{d} = \sum_{\substack{d=1\\ d\equiv 0\,(q)}}^{\infty} \sum_{k=1}^{\infty} \mu(k) g(dk) \quad \forall q \in \mathbb{N}.$$

Now, by using the well-known property [24]

$$\sum_{k|n}\mu(k)=\begin{cases}1 & \text{if } n=1,\\ 0 & \text{otherwise,}\end{cases}$$

for any $q\in\mathbb{N}$ we can write

$$g(q)=\sum_{n=1}^{\infty}g(qn)\sum_{k|n}\mu(k),$$

that converges unconditionally because of (7). Indeed, since $\omega(qn)\geq\omega(n)$ yields $2^{\omega(qn)}\geq 2^{\omega(n)}$, one has

$$\begin{aligned}\sum_{n=1}^{\infty}|g(qn)|\sum_{k|n}\mu^2(k) &= \sum_{n=1}^{\infty}2^{\omega(n)}|g(qn)|\leq\sum_{n=1}^{\infty}2^{\omega(qn)}|g(qn)|\\ &= \sum_{\substack{m=1\\ m\equiv 0(q)}}^{\infty}2^{\omega(m)}|g(m)|\leq\sum_{m=1}^{\infty}2^{\omega(m)}|g(m)|.\end{aligned}$$

Therefore, we can exchange the sums to get

$$\begin{aligned}g(q)=\sum_{n=1}^{\infty}g(qn)\sum_{k|n}\mu(k) &= \sum_{k=1}^{\infty}\mu(k)\sum_{m=1}^{\infty}g(qmk)=\sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\mu(k)g(qmk)\\ &= \sum_{\substack{d=1\\ d\equiv 0(q)}}^{\infty}\sum_{k=1}^{\infty}\mu(k)g(dk)=\mathcal{W}_f(q).\end{aligned}$$

The theorem is completely proved. □

REMARK 11. Since *any* $g\in\langle f\rangle_*$ determines the E-transform of f, the previous theorem establishes the uniqueness of g *modulo* $\langle\mathbf{0}\rangle_*$.
We refer to ② of the previous theorem as the *Wintner-Delange uniqueness formula* and (7) is called the *"Dual" Delange condition*. In fact, from Theorem 1 it follows that $\{g\in\langle f\rangle_* : g \text{ satisfies (7)}\}$ is either the empty set or $\{\mathcal{W}_f\}$.

REMARK 12. While it is plain that $\langle\mathbf{0}\rangle_{**}\neq\emptyset$ (for $\mathbf{0}\in\langle\mathbf{0}\rangle_{**}$), it might be possible that $\langle f\rangle_{**}=\emptyset$ for some $f\in\mathcal{A}$. Indeed, recall that there is the possibility that the larger set $\langle f\rangle_*$ might be empty because the (finite) $\mathcal{R}$-expansions ensured by Hildebrand's theorem are not necessarily pure. The next theorem, quoted from a lemma of [11], gives another positive answer to the uniqueness question posed in Remark 4. (First one coming from Theorem 1.) In particular, it implies $\langle\mathbf{0}\rangle_{**}=\{\mathbf{0}\}$.

THEOREM 2. *For any $f\in\mathcal{A}$, either $\langle f\rangle_{**}$ is the empty set or $\{C_f\}$.*

Proof. We have to show that if $g\in\langle f\rangle_{**}$, then

$$g(\ell)=\frac{1}{\varphi(\ell)}\lim_{x\to\infty}\frac{1}{x}\sum_{h\leq x}f(h)c_\ell(h)\quad\forall\ell\in\mathbb{N}.$$

Let $\ell \in \mathbb{N}$ be fixed. Note that, from the uniform convergence of the $\mathcal{R}$-expansion $f = \mathcal{R}_g$, it follows that for every $\varepsilon > 0$ there exists $Q = Q(\varepsilon, \ell) > \ell$ such that

$$\Big| \sum_{q>Q} g(q) c_q(h) \Big| < \frac{\varepsilon}{\mathbf{d}(\ell)}, \quad \forall h \in \mathbb{N},$$

where $\mathbf{d}(\ell) \overset{def}{=} \sum_{t|\ell} 1$ is the number of positive divisors of ℓ. Since the expansion $f = \mathcal{R}_g$ is also pure, this entails

$$\frac{1}{x} \sum_{h \le x} f(h) c_\ell(h) = \sum_{q \le Q} g(q) \frac{1}{x} \sum_{h \le x} c_\ell(h) c_q(h) + \frac{1}{x} \sum_{h \le x} c_\ell(h) \sum_{q>Q} g(q) c_q(h).$$

Recalling ④ of Proposition 1, from which in particular one has that

$$\lim_{x \to \infty} \frac{1}{x} \sum_{h \le x} c_\ell(h)^2 = \varphi(\ell),$$

and applying $|c_\ell(h)| \le (\ell, h)$ (see ② of Proposition 1), we can write

$$\begin{aligned} \left| \frac{1}{\varphi(\ell)} \lim_{x \to \infty} \frac{1}{x} \sum_{h \le x} f(h) c_\ell(h) - \frac{1}{\varphi(\ell)} \sum_{q \le Q} g(q) \lim_{x \to \infty} \frac{1}{x} \sum_{h \le x} c_\ell(h) c_q(h) \right| &= \\ \left| \frac{1}{\varphi(\ell)} \lim_{x \to \infty} \frac{1}{x} \sum_{h \le x} f(h) c_\ell(h) - g(\ell) \right| &\le \\ \frac{\varepsilon}{\varphi(\ell) \mathbf{d}(\ell)} \lim_{x \to \infty} \frac{1}{x} \sum_{h \le x} (\ell, h). & \end{aligned}$$

Therefore, the conclusion follows once it is proved that

$$\lim_{x \to \infty} \frac{1}{x} \sum_{h \le x} (\ell, h) = \sum_{d | \ell} \frac{\varphi(d)}{d}.$$

To this end, we write

$$\begin{aligned} \frac{1}{x} \sum_{h \le x} (\ell, h) &= \frac{1}{x} \sum_{t|\ell} t \sum_{\substack{h' \le \frac{x}{t} \\ (h', \frac{\ell}{t}) = 1}} 1 = \frac{1}{x} \sum_{t|\ell} t \sum_{d \mid \frac{\ell}{t}} \mu(d) \left[\frac{x}{dt} \right] \\ &= \sum_{t|\ell} \sum_{d \mid \frac{\ell}{t}} \frac{\mu(d)}{d} + O\left(\frac{1}{x} \sum_{t|\ell} t \mathbf{d}(\ell/t) \right) \\ &= \sum_{t|\ell} \frac{\varphi(\ell/t)}{\ell/t} + o(1) = \sum_{d|\ell} \frac{\varphi(d)}{d} + o(1). \end{aligned}$$

The theorem is completely proved. □

REMARK 13. To emphasize the fact that the $\mathcal{R}$-coefficients of f are uniquely determined, we set

$$\widehat{f} \stackrel{def}{=} \mathcal{C}_f;$$

or also, in the hypotheses of Theorem 1,

$$\widehat{f} \stackrel{def}{=} \mathcal{W}_f.$$

More generally, we write $\widehat{f} \stackrel{def}{=} g$ even if $\langle f \rangle = g + \langle \mathbf{0} \rangle$ or $\langle f \rangle_* = g + \langle \mathbf{0} \rangle_*$.

The next theorem yields the uniqueness of the $\mathcal{R}$-coefficients for pure and finite $\mathcal{R}$-expansions (see the following remark).

THEOREM 3. *$f \in \mathcal{A}_Q \Leftrightarrow \exists g \in \langle f \rangle_* \cap \langle f \rangle_\#$ such that $f =_\# \mathcal{R}_g$ has length at most Q.*

Proof. Let f' be the Q-truncated E-transform of f. By ③ of Proposition 1 we see that

$$f(n) = \sum_{\substack{d \le Q \\ d|n}} f'(d) = \sum_{d \le Q} \frac{f'(d)}{d} \sum_{q|d} c_q(n) = \sum_{q \le Q} \mathcal{W}_f(q) c_q(n),$$

where

$$\mathcal{W}_f(q) \stackrel{def}{=} \begin{cases} \displaystyle\sum_{\substack{d \le Q \\ d \equiv 0\,(q)}} \frac{f'(d)}{d} & \text{if } q \le Q, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

is a *Q-truncated* Wintner coefficient, say. It is plain that $\mathcal{W}_f \in \langle f \rangle_* \cap \langle f \rangle_\#$.
Vice versa, let $g \in \langle f \rangle_* \cap \langle f \rangle_\#$ be such that the length of the expansion $f(n) =_\# \mathcal{R}_g(n)$ is at most Q for all $n \in \mathbb{N}$. By applying ① of Proposition 1 we write

$$f(n) = \sum_{q \le Q} g(q) c_q(n) = \sum_{d|n} d \sum_{\substack{q \le Q \\ q \equiv 0\,(d)}} g(q) \mu(q/d) = \sum_{\substack{d \le Q \\ d|n}} f'(d),$$

where we have set

$$f'(d) \stackrel{def}{=} \begin{cases} d \displaystyle\sum_{\substack{q \le Q \\ q \equiv 0\,(d)}} g(q) \mu(q/d) & \text{if } d \le Q, \\ 0 & \text{otherwise.} \end{cases}$$

The theorem is completely proved. □

REMARK 14. Theorems 2 and 3 imply that $\langle f \rangle_* \cap \langle f \rangle_\# = \{\mathcal{C}_f\} = \{\mathcal{W}_f\}$ with $\mathcal{W}_f$ defined as in (8). In particular, note that

$$\frac{Q}{2} < q \le Q \Longrightarrow \widehat{f}(q) = \mathcal{W}_f(q) = \frac{f'(q)}{q}$$

(for $q > Q/2$, the conditions $d \leq Q$, $q|d$ hold simultaneously if and only if $d = q$). Also, see that if we assume the conjecture in Remark 9 for $f \in \mathcal{A}$ with $\mathcal{W}_f(1) \neq 0$, from Theorems 2 and 3 we get

$$\mathcal{W}_f(q) = 0 \ \forall q > Q \Longrightarrow \ f \in \mathcal{A}_Q \Longrightarrow \ f =_{\#} \mathcal{R}_{\widehat{f}},$$

where it turns out that $\widehat{f} = \mathcal{C}_f = \mathcal{W}_f$ is given by (8).

Finally, we underline the fact that the above proof provides with an explicit method to express a truncated divisor sum as a finite $\mathcal{R}$-expansion, and vice versa.

4. Ramanujan expansions of shifted convolution sums

DEFINITION 4. *The* **correlation** *(or shifted convolution sum) of $f, g \in \mathcal{A}$ is*

$$C_{f,g}(N,a) \stackrel{def}{=} \sum_{n \leq N} f(n)g(n+a).$$

Since without loss of generality one can assume that $f(N)g(N+a) \neq 0$, the number N is the length of such a correlation. Here $a \in \mathbb{N}$ is the **shift**.

From (2) it follows that

$$C_{f,g}(N,a) = \sum_q h(a,q)c_q(a), \ \ \forall a \in \mathbb{N}, \tag{9}$$

for some $h(a,q) = h(a,q,f,g,N) \in \mathbb{C}$. We refer to (9) as the *shift $\mathcal{R}$-expansion* of the correlation $C_{f,g}$. On the other hand, denoting by f', g' the E-transforms of f, g, respectively, we see that

$$C_{f,g}(N,a) = \sum_{n \leq N} \sum_{d|n} f'(d) \sum_{q|n+a} g'(q), \tag{10}$$

where observe that the conditions $n \leq N$ and $d|n$ yield $d \leq N$ in the second sum, while $n \leq N$ and $q|n+a$ yield $q \leq N+a$ in the third sum. In other words, within their correlation of length N, the functions f and g can be replaced respectively by the truncated divisor sums associated to f and g, of range respectively N and $N+a$, i.e.

$$f_N(n) \stackrel{def}{=} \sum_{\substack{d \leq N \\ d|n}} f'(d), \qquad g_{N+a}(n+a) \stackrel{def}{=} \sum_{\substack{q \leq N+a \\ q|n+a}} g'(q).$$

These functions admit pure (w.r.t n and $n+a$, respectively) and finite $\mathcal{R}$-expansions because of Theorem 3 (see also Remark 14). However, by plugging such expansions into (10) we can only get a finite expansion as

$$C_{f,g}(N,a) = \sum_{d \leq N} \sum_{q \leq N+a} \widehat{f_N}(d)\widehat{g_{N+a}}(q) \sum_{n \leq N} c_d(n)c_q(n+a).$$

Evidently, the latter cannot be considered as a shift $\mathcal{R}$-expansion of the form (9). On the other hand, by combining ④ of Proposition 1 with the Wintner-Delange formula of Proposition 2, the $\mathcal{R}$-expansions of the above truncated divisor sums can help in finding a pure and finite $\mathcal{R}$-expansion, which well approximates $C_{f,g}(N,a)$, provided that this correlation is *fair* and both functions f and g satisfy the *Ramanujan Conjecture* (see [10], [11]), accordingly to the following definitions.

DEFINITION 5. *The correlation $C_{f,g}(N,a)$ of $f,g\in\mathcal{A}$ is fair if it depends on a only because of the argument $n+a$ of g.*

DEFINITION 6. *We say that $f\in\mathcal{A}$ satisfies the Ramanujan Conjecture (or equivalently that f is essentially bounded) if $f(n)\ll_\varepsilon n^\varepsilon$ for any real number $\varepsilon>0$. In this case, we also write $f\ll\ll 1$. We denote the set of the essentially bounded arithmetic functions by $\mathcal{A}^\varepsilon$.*

For example, the correlation $C_{f,g}(N,a)$ is not fair if the support of f or g depends on a (see [11] for a specific example). In particular, note that the expression (10) is not fair in that the support of g' does depend on a. If $f,g\in\mathcal{A}^\varepsilon$ (consequently, also $f',g'\in\mathcal{A}^\varepsilon$), then somehow we can get rid of such a nuisance by writing

$$\begin{aligned} C_{f,g}(N,a) &= \sum_{n\le N}\sum_{d|n} f'(d)\sum_{\substack{q\le N\\ q|n+a}} g'(q)+\sum_{n\le N}\sum_{d|n} f'(d)\sum_{\substack{N<q\le N+a\\ q|n+a}} g'(q)\\ &= \sum_{n\le N}\sum_{d|n} f'(d)\sum_{\substack{q\le N\\ q|n+a}} g'(q)+\sum_{N<q\le N+a} g'(q)\sum_{\substack{n\le N\\ n\equiv -a\,(q)}}\sum_{d|n} f'(d). \end{aligned}$$

Since for $q>N$ one has that $\#\{n\le N : n\equiv -a\ (q)\}\le 1$, the second sum on the right hand side is $\ll aN^\varepsilon \max_{N<q\le N+a}|g'(q)|\max_{n\le N}|f'(n)|$. In all, we have for $f,g\in\mathcal{A}^\varepsilon$ that

$$C_{f,g}(N,a)=C_{f,g_N}(N,a)+O_\varepsilon(N^\varepsilon(N+a)^\varepsilon a).$$

Thus, we are reduced to deal with the correlation of the truncated divisor sums f_N,g_N of the same range N. For this reason we'll assume that $g\in\mathcal{A}_N$ (and, concerning $C_{f,g}(N,a)$, the hypothesis $f\in\mathcal{A}$ is equivalent to $f\in\mathcal{A}_N$). Using the finite $\mathcal{R}$-expansion of length at most N for g (see Theorem 3), namely $g(m)=\sum_{q\le N}\widehat{g}(q)c_q(m)$,

$$C_{f,g}(N,a)=\sum_{q\le N}\widehat{g}(q)\sum_{n\le N} f(n)c_q(n+a),\quad \forall a\in\mathbb{N}. \tag{11}$$

Note that $C_{f,g}(N,a)$ is fair, provided that neither $\widehat{g}$ nor f depends on a. By using ⑤ of Proposition 1 we calculate its Carmichael coefficients (we set $\mathcal{C}_C=\mathcal{C}_{C_{f,g}}$ for brevity):

$$\begin{aligned} \mathcal{C}_C(N,\ell) &= \frac{1}{\varphi(\ell)}\sum_{q\le N}\widehat{g}(q)\sum_{n\le N} f(n)\lim_{x\to\infty}\frac{1}{x}\sum_{a\le x} c_q(n+a)c_\ell(a)\\ &= \begin{cases} \dfrac{\widehat{g}(\ell)}{\varphi(\ell)}\displaystyle\sum_{n\le N} f(n)c_\ell(n) & \text{if } \ell\le N,\\ 0 & \text{if } \ell>N. \end{cases} \end{aligned} \tag{12}$$

Assuming that the E-transform $C'_{f,g}$ satisfies the Delange hypothesis (see Prop. 2), i.e.

$$\sum_d \frac{2^{\omega(d)} C'_{f,g}(N,d)}{d} \quad \text{converges absolutely,} \tag{13}$$

Proposition 2, Theorems 2 and 3 yield the so-called *Ramanujan exact explicit formula* (REEF)

$$C_{f,g}(N,a) = \sum_{q\le N} \mathcal{C}_C(N,q)c_q(a) = \sum_{q\le N}\left(\frac{\widehat{g}(q)}{\varphi(q)}\sum_{n\le N} f(n)c_q(n)\right)c_q(a), \ \ \forall a\in\mathbb{N}.$$

Without having (13) at our disposal we can proceed as it follows. Let us write

$$C_{f,g}(N,a) = \sum_{\substack{d\le N\\ d|a}} C'_{f,g}(N,d) + \sum_{\substack{d>N\\ d|a}} C'_{f,g}(N,d) = \sum\nolimits_I(a) + \sum\nolimits_{II}(a)\,, \text{ say,}$$

where clearly $\sum_{II}(a) = 0$, unless $a > N$. Since $\sum_I(a)$ is a truncated divisor sum, from Theorem 3 we get

$$C_{f,g}(N,a) = \sum_{q\le N} \mathcal{W}_C(N,q)c_q(a) + \sum\nolimits_{II}(a)$$

with

$$\mathcal{W}_C(N,q) \overset{def}{=} \begin{cases} \displaystyle\sum_{\substack{h\le N\\ h\equiv 0(q)}} \frac{C'_{f,g}(N,h)}{h} & \text{if } q\le N,\\ 0 & \text{otherwise.}\end{cases}$$

Therefore, calculating the Carmichael coefficients of

$$\sum_{q\le N} \mathcal{W}_C(N,q)c_q(a) = \begin{cases} C_{f,g}(N,a) - \sum\nolimits_{II}(a) & \text{if } a>N,\\ C_{f,g}(N,a) & \text{if } a\le N,\end{cases}$$

from (12) it follows that (recall the discussion on the *uniqueness* in Remark 14)

$$\begin{aligned}\mathcal{W}_C(N,q) = \ & \frac{1}{\varphi(q)}\lim_{x\to\infty}\frac{1}{x}\sum_{m\le N} C_{f,g}(N,m)c_q(m) + \\ & \frac{1}{\varphi(q)}\lim_{x\to\infty}\frac{1}{x}\sum_{N<m\le x}\Big(C_{f,g}(N,m) - \sum_{\substack{d>N\\ d|m}} C'_{f,g}(N,d)\Big)c_q(m)\\ = \ & \mathcal{C}_C(N,q) - \mathcal{L}(N,q), \ \ \forall q\in\mathbb{N},\end{aligned}$$

where

$$\mathcal{L}(N,q) = \mathcal{L}(f,g,N,q) \overset{def}{=} \frac{1}{\varphi(q)}\lim_{x\to\infty}\frac{1}{x}\sum_{N<m\le x} c_q(m)\sum_{\substack{d>N\\ d|m}} C'_{f,g}(N,d), \ \ \forall q\in\mathbb{N}.$$

In particular, $\mathcal{C}_C(N,q) = \mathcal{L}(N,q)$, $\forall q > N$, because $\mathcal{W}_C(N,q) = 0$, $\forall q > N$.

Therefore, under the only hypothesis that the correlation $C_{f,g}$ is fair, we obtain

$$\begin{aligned} C_{f,g}(N,a) &= \sum_{q\le N}\big(\mathcal{C}_C(N,q)-\mathcal{L}(N,q)\big)c_q(a)+\sum\nolimits_{II}(a) \\ &= \sum_{q\le N}\left(\frac{\widehat{g}(q)}{\varphi(q)}\sum_{n\le N}f(n)c_q(n)-\mathcal{L}(N,q)\right)c_q(a)+\sum\nolimits_{II}(a),\ \ \forall a\in\mathbb{N}, \end{aligned}$$

where

$$\sum\nolimits_{II}(a)\overset{def}{=}\begin{cases}\displaystyle\sum_{\substack{d>N\\ d|a}}C'_{f,g}(N,d) & \text{if } a>N,\\ 0 & \text{if } a\le N.\end{cases}$$

Hence, the following theorem and corollary are proved.

THEOREM 4. *If $f\in\mathcal{A}$ and $g\in\mathcal{A}_N$ are such that $C_{f,g}(N,a)$ is fair, then*

$$C_{f,g}(N,a)=\sum_{q\le N}\left(\frac{\widehat{g}(q)}{\varphi(q)}\sum_{n\le N}f(n)c_q(n)-\mathcal{L}(N,q)\right)c_q(a)+\sum\nolimits_{II}(a),$$

where $\sum_{II}(a)$ is defined above,

$$\mathcal{L}(N,q)=\mathcal{L}(f,g,N,q)\overset{def}{=}\frac{1}{\varphi(q)}\lim_{x\to\infty}\frac{1}{x}\sum_{N<m\le x}c_q(m)\sum_{\substack{d>N\\ d|m}}C'_{f,g}(N,d),$$

and $C'_{f,g}$ is the E-transform of $C_{f,g}$.

In particular, if $\displaystyle\sum_{d=1}^{\infty}\frac{2^{\omega(d)}}{d}|C'_{f,g}(N,d)|$ converges, then for all $a\in\mathbb{N}$ one has the REEF:

$$C_{f,g}(N,a)=\sum_{q\le N}\widehat{C_{f,g}}(N,q)c_q(a),\ \ \textit{with}\ \ \widehat{C_{f,g}}(N,q)=\frac{\widehat{g}(q)}{\varphi(q)}\sum_{n\le N}f(n)c_q(n).$$

COROLLARY 1. *Let $f,g\in\mathcal{A}^{\varepsilon}$. If $C_{f,g_N}(N,a)$ is fair and such that the series $\displaystyle\sum_{d=1}^{\infty}\frac{2^{\omega(d)}}{d}|C'_{f,g_N}(N,d)|$ converges, then for all $a\in\mathbb{N}$ one has*

$$C_{f,g}(N,a)=\sum_{q\le N}\left(\frac{\widehat{g_N}(q)}{\varphi(q)}\sum_{n\le N}f(n)c_q(n)\right)c_q(a)+O_{\varepsilon}(N^{\varepsilon}(N+a)^{\varepsilon}a).$$

REMARK 15. Given $f\in\mathcal{A}$ and $g\in\mathcal{A}_N$, assuming that $\displaystyle\sum_{d=1}^{\infty}\frac{C'_{f,g}(N,d)}{d}$ converges absolutely, from Wintner's criterion it follows that $C_{f,g}(N,a)$ has both Carmichael and Wintner coefficients with $\mathcal{C}_C(N,q)=\mathcal{W}_C(N,q)$ for all $q\in\mathbb{N}$. In particular, (12) yields that $\mathcal{W}_C(N,q)=0$ for all $q>N$. From this, if we further assume the conjecture formulated in Remark 9, we get that $C'_{f,g}(N,d)=0$ for all $d>N$.

Besides the consequence that the above series reduces to the finite sum of length at most N, such a conjecture yields the REEF without the Delange hypothesis (13). In other words, such a conjecture is an alternative way to get the REEF of Theorem 4.

REMARK 16. Assume that $f \in \mathcal{A}$ and $g \in \mathcal{A}_Q$, with $Q \leq N$, are such that $C_{f,g}(N,a)$ is fair. From (11) it follows that $C_{f,g}(N,a)$ is periodic with respect to a, which implies that it is a bounded arithmetic function of a. Together with (12), this reveals that $C_{f,g}(N,a)$ satisfies the hypotheses of Proposition 3, so that its Carmichael coefficients coincide with its Wintner ones.

Now, let us quote here the main result of [11].

THEOREM 5. *Let $f \in \mathcal{A}^{\varepsilon}$ and $g \in \mathcal{A}_N \cap \mathcal{A}^{\varepsilon}$ be such that $C_{f,g}(N,a)$ is fair for all $a \in \mathbb{N}$ and admits the shift $\mathcal{R}$-expansion (9). The following propositions are equivalent:*

① *The shift $\mathcal{R}$-expansion (9) is completely uniform, i.e. it is pure, with $h(a,q) = h(q)$, and converges uniformly with respect to a.*
② *The coefficients of (9) are the Carmichael coefficients of $C_{f,g}(N,a)$, i.e.*

$$h(a,q) = h(q) = \frac{1}{\varphi(q)} \lim_{x\to\infty} \frac{1}{x} \sum_{n\leq x} C_{f,g}(N,n) c_q(n).$$

③ *The coefficients of (9) are the REEF coefficients of $C_{f,g}(N,a)$, i.e.*

$$h(a,q) = h(q) = \widehat{C_{f,g}}(N,q) = \frac{\widehat{g}(q)}{\varphi(q)} \sum_{n\leq N} f(n) c_q(n).$$

④ *The shift $\mathcal{R}$-expansion (9) is finite and pure.*

We underline the latter equivalence between the condition $\langle C_{f,g}\rangle_* \cap \langle C_{f,g}\rangle_{\#} \neq \emptyset$ and the REEF. It highlights the fundamental role of Theorem 3 in this new approach to the $\mathcal{R}$-expansions of the correlations. Moreover, Corollary 1 has to be compared to the following result that can be proved in a similar fashion of Corollary 1 in [11].

THEOREM 6. *Let $g \in \mathcal{A}^{\varepsilon}$ and $f \in \mathcal{A}_D \cap \mathcal{A}^{\varepsilon}$, with $D < N^{1-\delta}$ for some $\delta \in (0,1)$. If $C_{f,g_N}(N,a)$ is fair and such that $\sum_{d=1}^{\infty} \frac{2^{\omega(d)}}{d} |C'_{f,g_N}(N,d)|$ converges, then uniformly for all $a \in \mathbb{N}$ one has*

$$C_{f,g}(N,a) = \mathfrak{S}_{f,g}(a) N + O\left(N^{1-\delta}\right) + O_{\varepsilon}\left(N^{\varepsilon}(N+a)^{\varepsilon} a\right),$$

where $\mathfrak{S}_{f,g}$ is the so-called **singular sum** *defined as*

$$\mathfrak{S}_{f,g}(a) \stackrel{def}{=} \sum_{q\leq N} \hat{f}(q)\hat{g}(q) c_q(a), \quad \forall a \in \mathbb{N}.$$

The elements of $\mathcal{A}_D \cap \mathcal{A}^{\varepsilon}$ are known as *sieve functions* (of range D). We refer the reader to [4]-[9] for further deepenings about such a class of functions.

We conclude the present section by quoting a result from [1] on the convolution sum of the von Mangoldt function Λ. Indeed, in [1] Corollary 1 is applied by taking $f = g = \Lambda$, that clearly belongs to $\mathcal{A}^{\varepsilon}$. From the well-known property [24]

$$\Lambda(n) = -\sum_{d|n} \mu(d)\log d,$$

it follows that its E-transform is $\Lambda'(n) = -\mu(n)\log n$. Further, for the N-truncated divisor sum of Λ we have (see Theorem 3)

$$\Lambda_N(n) = -\sum_{\substack{d\le N\\ d|n}} \mu(d)\log d = \sum_{q\le N} \widehat{\Lambda_N}(q)c_q(n),$$

with

$$\widehat{\Lambda_N}(q) \stackrel{def}{=} -\sum_{\substack{d\le N\\ d\equiv 0(q)}} \frac{\mu(d)\log d}{d} \ll \frac{L^2}{q},$$

where we have set $L \stackrel{def}{=} \log N$. Therefore, since $C_{\Lambda,\Lambda_N}(N,a)$ is fair, by assuming that $\sum_{d=1}^{\infty} \frac{2^{\omega(d)}}{d} C'_{\Lambda,\Lambda_N}(N,d)$ converges absolutely, Corollary 1 yields

$$C_{\Lambda,\Lambda}(N,a) = \sum_{q\le N}\left(\frac{\widehat{\Lambda_N}(q)}{\varphi(q)}\sum_{n\le N}\Lambda(n)c_q(n)\right)c_q(a) + O_{\varepsilon}(N^{\varepsilon}(N+a)^{\varepsilon}a).$$

The result established in [1] shows that the Delange hypotesis for $C_{\Lambda,\Lambda_N}(N,2k)$ yields the Hardy-Littlewood conjecture for the $2k$-twin primes [17]. Here it is stated as a further corollary of Theorem 4.

COROLLARY 2. *Let $k \in \mathbb{N}$ be such that $0 < k < N^{1-\delta}$, with $\delta \in (0,1/2)$ fixed. If $\sum_{d=1}^{\infty} \frac{2^{\omega(d)}}{d}|C'_{\Lambda,\Lambda_N}(N,d)|$ converges, then*

$$C_{\Lambda,\Lambda}(N,2k) = \mathfrak{S}_{\Lambda,\Lambda}(2k)N + O(Ne^{-c\sqrt{\log N}}),$$

where $c > 0$ is an absolute constant and

$$\mathfrak{S}_{\Lambda,\Lambda}(2k) \stackrel{def}{=} \sum_{q=1}^{\infty}\frac{\mu^2(q)}{\varphi^2(q)}c_q(2k) = 2\prod_{p|k}\left(1+\frac{1}{p-1}\right)\prod_{(p,2k)=1}\left(1-\frac{1}{(p-1)^2}\right).$$

5. Ramanujan expansions and smooth numbers

In the present section we resume some results of [3], where it is showed that all essentially bounded functions, with E-transform supported on smooth numbers, admit a unique $\mathcal{R}$-expansion with coefficients satisfying the "Dual" Delange condition (7).

DEFINITION 7. *Let $Q \geq 2$ be an integer.*

The set of the Q-smooth positive integers is $\mathcal{S} = \mathcal{S}(Q) \stackrel{def}{=} \{n \in \mathbb{N} : \ p|n \Rightarrow p \leq Q\} \cup \{1\}$.

The set of the Q-sifted positive integers is $\mathcal{T} = \mathcal{T}(Q) \stackrel{def}{=} \{n \in \mathbb{N} : \ p \leq Q \Rightarrow p \not| n\}$.

REMARK 17. Note that $\mathcal{S} \cap \mathcal{T} = \{1\}$ and $(n,m) = 1$ for all $n \in \mathcal{S}, m \in \mathcal{T}$. Further, any $n \in \mathcal{S}$ can be written as $n = p_1^{v_1} \cdots p_r^{v_r}$, for some integers $v_j \geq 0$, $j = 1, \ldots, r = \pi(Q)$, where $\pi(Q) = \#\{p \leq Q : \ p \text{ prime}\}$ and $2 = p_1, p_2, \ldots, p_r$ are all the consecutive prime numbers $\leq Q$. Thus, for any real number $x > 1$ and for all $\varepsilon > 0$ one has

$$\#\mathcal{S} \cap [1,x] \leq \sum_{\substack{n \in \mathcal{S} \\ n \leq x}} \frac{x^\varepsilon}{n^\varepsilon} \ll x^\varepsilon \sum_{v_1=0}^{\infty} \cdots \sum_{v_r=0}^{\infty} \frac{1}{\frac{\varepsilon v_1}{p_1}} \cdots \frac{1}{p_r^{\varepsilon v_r}} = x^\varepsilon \prod_{p \leq Q} \frac{1}{1-p^{-\varepsilon}} \ll_{\varepsilon, Q} x^\varepsilon.$$

Similarly, if $\varepsilon \in (0,1)$, we see that

$$\sum_{m \in \mathcal{S}} \frac{1}{m^{1-\varepsilon}} \ll_{\varepsilon, Q} 1.$$

Moreover, from the Legendre formula applied to $\#\mathcal{T} \cap [1,x] = \#\{n \leq x : \ (n, P_Q) = 1\}$, where $P_Q = \prod_{p \leq Q} p$, it follows that

$$\#\mathcal{T} \cap [1,x] = \sum_{d | P_Q} \mu(d) \left[\frac{x}{d}\right] = x \sum_{d | P_Q} \frac{\mu(d)}{d} + O\left(\sum_{d | P_Q} |\mu(d)|\right) = x \prod_{p \leq Q} \left(1 - \frac{1}{p}\right) + O_Q(1).$$

DEFINITION 8. *Let $Q \geq 2$ be an integer. The Q-smooth restriction of $f \in \mathcal{A}$ is the arithmetic function defined as*

$$f_{\mathcal{S}}(n) \stackrel{def}{=} \sum_{\substack{d | n \\ d \in \mathcal{S}}} f'(d), \ \ \forall n \in \mathbb{N},$$

where f' is the E-trasform of f.

REMARK 18. It is plain that $f_{\mathcal{S}}$ is the inverse E-trasform of $f' \cdot \mathbf{1}_{\mathcal{S}}$, where $\mathbf{1}_{\mathcal{S}}$ is the characteristic function of $\mathcal{S}$. Also note that $f_{\mathcal{S}}(n) = f(n)$ for all $n \in \mathcal{S}$. Further, one has

$$f_{\mathcal{S}}(n) = \sum_{t \in \mathcal{S}_n} f(t), \ \text{ where } \mathcal{S}_n \stackrel{def}{=} \{t \in \mathcal{S} : \ t | n \text{ and } n/t \in \mathcal{T}\}.$$

Indeed, this is trivially true for $n = 1$.

If $n \geq 2$, recall that $f' = f * \mu$ and use the (complete) multiplicativity of $\mathbf{1}_{\mathcal{S}}$, to get it:

$$\begin{aligned} f_{\mathcal{S}}(n) &= \sum_{\substack{d | n \\ d \in \mathcal{S}}} \sum_{t | d} f(t) \mu\left(\frac{d}{t}\right) = \sum_{\substack{t \in \mathcal{S} \\ t | n}} f(t) \sum_{\substack{k \in \mathcal{S} \\ k | \frac{n}{t}}} \mu(k) \\ &= \sum_{\substack{t \in \mathcal{S} \\ t | n}} f(t) \sum_{k | \frac{n}{t}} \mu(k) \mathbf{1}_{\mathcal{S}}(k) = \sum_{\substack{t \in \mathcal{S} \\ t | n}} f(t) \prod_{p | \frac{n}{t}} \left(1 - \mathbf{1}_{\mathcal{S}}(p)\right) = \sum_{\substack{t \in \mathcal{S} \\ t | n}} f(t) \mathbf{1}_{\mathcal{T}}(n/t). \end{aligned}$$

LEMMA 1. *Given any integer $Q \geq 2$, let us consider the set $\mathcal{S} = \mathcal{S}(Q)$ of the Q-smooth positive integers. For any $f \in \mathcal{A}^{\varepsilon}$, with $f' = f * \mu$, one has*

$$\textcircled{1} \qquad \sum_{t\in\mathcal{S}} \frac{|f(t)c_q(t)|}{t} \ll_{\varepsilon,q,Q} 1 \quad \textit{for all} \quad q \in \mathbb{N}$$

$$\textcircled{2} \qquad \sum_{t\in\mathcal{S}} \frac{|f'(t)|}{t} \ll_{\varepsilon,Q} 1 \qquad \textit{and} \qquad \sum_{t\in\mathcal{S}} \frac{2^{\omega(t)}|f'(t)|}{t} \ll_{\varepsilon,Q} 1.$$

Proof. Without loss of generality, we can assume that $\varepsilon \in (0,1)$.
① Using ② of Prop. 1 and the inequality $\sum_{m\in\mathcal{S}} m^{\varepsilon-1} \ll_{\varepsilon,Q} 1$ (see Remark 17), we get

$$\sum_{t\in\mathcal{S}} \frac{|f(t)c_q(t)|}{t} \ll_{\varepsilon} \sum_{t\in\mathcal{S}} (q,t)t^{\varepsilon-1} \ll_{\varepsilon} \sum_{\substack{d\in\mathcal{S}\\ d|q}} d \sum_{\substack{t\in\mathcal{S}\\ t\equiv 0\,(d)}} t^{\varepsilon-1} \ll_{\varepsilon} \sum_{\substack{d\in\mathcal{S}\\ d|q}} d^{\varepsilon} \sum_{m\in\mathcal{S}} m^{\varepsilon-1} \ll_{\varepsilon,q,Q} 1.$$

② Recalling that $f' \in \mathcal{A}^{\varepsilon}$ and arguing as before, we see that

$$\sum_{t\in\mathcal{S}} \frac{|f'(t)|}{t} \ll_{\varepsilon} \sum_{t\in\mathcal{S}} t^{\varepsilon-1} \ll_{\varepsilon,Q} 1.$$

Since $2^{\omega(t)} \leq 2^{\pi(Q)}$ for all $t \in \mathcal{S}$, the second inequality follows from the first one. □

THEOREM 7. *Let $Q \geq 2$ be an integer. For any $f \in \mathcal{A}^{\varepsilon}$, let us consider the Q-smooth restriction $f_{\mathcal{S}}$, where $\mathcal{S} = \mathcal{S}(Q)$ is the set of the Q-smooth positive integers. The Carmichael coefficients of $f_{\mathcal{S}}$ and the Wintner ones coincide, both given by*

$$\widehat{f_{\mathcal{S}}}(q) = \begin{cases} \dfrac{\tilde{f}(q,\mathcal{S})}{\varphi(q)} \displaystyle\prod_{p\leq Q}\left(1-\frac{1}{p}\right) = \sum_{\substack{d\in\mathcal{S}\\ d\equiv 0\,(q)}} \frac{f'(d)}{d} & \textit{if } q \in \mathcal{S}, \\ 0 & \textit{otherwise,} \end{cases} \tag{14}$$

where we set

$$\tilde{f}(q,\mathcal{S}) \overset{def}{=} \sum_{t\in\mathcal{S}} \frac{f(t)c_q(t)}{t}.$$

Further, one has $\widehat{f_{\mathcal{S}}} \in \langle f_{\mathcal{S}} \rangle$, i.e.

$$f_{\mathcal{S}}(a) = \sum_{q\in\mathcal{S}} \widehat{f_{\mathcal{S}}}(q)c_q(a), \quad \forall a \in \mathbb{N}, \tag{15}$$

and $\widehat{f_{\mathcal{S}}}$ satisfies the "Dual" Delange condition (7).

Proof. Without loss of generality, we can assume that $\varepsilon \in (0,1/2)$. First, note that $\tilde{f}(q,\mathcal{S})$ is well-defined for all $q \in \mathbb{N}$ because of ① in Lemma 1. Then, recalling that $f' \cdot \mathbf{1}_{\mathcal{S}}$ is the E-trasform of $f_{\mathcal{S}}$, the second inequality in ② of Lemma 1 implies that the hypothesis of the Wintner-Delange formula holds for $f_{\mathcal{S}}$.

Therefore, it follows from ② of Proposition 2 that Carmichael coefficients of $f_{\mathcal{S}}$ equal Wintner ones and they are $\mathcal{R}$-coefficients for $f_{\mathcal{S}}$. Being such coefficients uniquely determined (see the next remark), we denote the qth coefficient by $\widehat{f_{\mathcal{S}}}(q)$. In particular, since it is plain that the conditions $d \in \mathcal{S}$ and $q|d$ imply that $q \in \mathcal{S}$, the qth Wintner coefficient of $f_{\mathcal{S}}$ is

$$\widehat{f_{\mathcal{S}}}(q) = \begin{cases} \displaystyle\sum_{d\equiv 0(q)} \frac{(f'\cdot \mathbf{1}_{\mathcal{S}})(d)}{d} = \sum_{\substack{d\in\mathcal{S}\\ d\equiv 0(q)}} \frac{f'(d)}{d} & \text{if } q \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, recalling that $\sum_{m\in\mathcal{S}} m^{\varepsilon-1} \ll_{\varepsilon,Q} 1$ (see Remark 17), we see that

$$\begin{aligned} \sum_{q=1}^{\infty} 2^{\omega(q)} |\widehat{f_{\mathcal{S}}}(q)| &= \sum_{q\in\mathcal{S}} 2^{\omega(q)} \Big| \sum_{\substack{d\in\mathcal{S}\\ d\equiv 0(q)}} \frac{f'(d)}{d} \Big| \\ &\le 2^{\pi(Q)} \sum_{q\in\mathcal{S}} \sum_{\substack{d\in\mathcal{S}\\ d\equiv 0(q)}} \frac{|f'(d)|}{d} \ll_{\varepsilon,Q} \sum_{q\in\mathcal{S}} q^{\varepsilon-1} \sum_{k\in\mathcal{S}} k^{\varepsilon-1} \ll_{\varepsilon,Q} 1, \end{aligned}$$

that is $\widehat{f_{\mathcal{S}}}$ satisfies (1). Thus, it remains to prove that for every $q \in \mathcal{S}$ the qth Carmichael coefficient of $f_{\mathcal{S}}$ is

$$\frac{1}{\varphi(q)} \prod_{p\le Q} \Big(1-\frac{1}{p}\Big) \sum_{t\in\mathcal{S}} \frac{f(t)c_q(t)}{t}.$$

From Remark 18, $f_{\mathcal{S}}(a) = \displaystyle\sum_{t\in\mathcal{S}_a} f(t)$ and $\mathcal{S}_a \overset{def}{=} \{t \in \mathcal{S} : \ t|a \text{ and } a/t \in \mathcal{T}\}$, whence

$$\sum_{a\le x} f_{\mathcal{S}}(a)c_q(a) = \sum_{\substack{t\in\mathcal{S}\\ t\le x}} f(t) \sum_{\substack{k\in\mathcal{T}\\ k\le x/t}} c_q(tk) = \sum_{\substack{t\in\mathcal{S}\\ t\le x}} f(t)c_q(t)\#\mathcal{T}\cap[1,x/t],$$

where we exchange sums and $c_q(tk) = c_q(t)$ follows from the conditions $q \in \mathcal{S}$, $k \in \mathcal{T}$, which yield $(q,k) = 1$. Now, since (see Remark 17)

$$\#\mathcal{T}\cap[1,x/t] = \frac{x}{t} \prod_{p\le Q} \Big(1-\frac{1}{p}\Big) + O_Q(1),$$

for every $q \in \mathcal{S}$ the qth Carmichael coefficient of $f_{\mathcal{S}}$ is given by

$$\begin{aligned} \mathcal{C}_{f_{\mathcal{S}}}(q) &= \frac{1}{\varphi(q)} \lim_{x\to\infty} \frac{1}{x} \sum_{a\le x} f_{\mathcal{S}}(a)c_q(a) \\ &= \frac{1}{\varphi(q)} \prod_{p\le Q} \Big(1-\frac{1}{p}\Big) \sum_{t\in\mathcal{S}} \frac{f(t)c_q(t)}{t} + \frac{1}{\varphi(q)} \lim_{x\to\infty} \sum_{\substack{t\in\mathcal{S}\\ t\le x}} f(t)c_q(t)O_Q(1/x). \end{aligned}$$

The conclusion follows once we show that the latter limit is 0. Recalling that $f \in \mathcal{A}^{\varepsilon}$, $\#\mathcal{S} \cap [1,x] \ll_{\varepsilon,Q} x^{\varepsilon}$ (see Remark 17), and applying ② of Proposition 1, we see that

$$\sum_{\substack{t\in\mathcal{S}\\ t\le x}} f(t)c_q(t)O_Q(1/x) \ll_{\varepsilon,Q} x^{\varepsilon-1}\sum_{\substack{t\in\mathcal{S}\\ t\le x}}(q,t) \ll_{\varepsilon,Q} x^{\varepsilon-1}\sum_{d|q} d\#\mathcal{S}\cap[1,x/d] \ll_{\varepsilon,q,Q} x^{2\varepsilon-1}.$$

The theorem is completely proved. □

REMARK 19. The previous theorem and Theorem 1 yield that if $g \in \langle f_{\mathcal{S}}\rangle_*$ satisfies the "Dual" Delange condition (7), then $g = \widehat{f_{\mathcal{S}}}$. In other words, the $\mathcal{R}$-coefficients of a smooth restriction of an essentially bounded function are uniquely determined.

But even more important, since $f_{\mathcal{S}}(a) = f(a)$ for all $a \in \mathcal{S}$, from (15) we obtain the $\mathcal{R}$-expansion for the restriction of f to $\mathcal{S}$:

$$f(a) = \sum_{q=1}^{\infty} \widehat{f_{\mathcal{S}}}(q)c_q(a), \quad \forall a \in \mathcal{S}, \tag{16}$$

with the coefficients $\widehat{f_{\mathcal{S}}}(q)$ defined by (14). In particular, given $f \in \mathcal{A}$ and $g \in \mathcal{A}_Q$, with $Q \le N$, a fair correlation $C_{f,g}(N,a)$ for all $a \in \mathbb{N}$, being bounded (see Remark 16), satisfies the hypotheses of the previous theorem. Thus, from (16) we get the unique $\mathcal{R}$-expansion

$$\begin{aligned} C_{f,g}(N,a) &= \sum_{q\in\mathcal{S}}\sum_{\substack{d\in\mathcal{S}\\ d\equiv 0(q)}} \frac{C'_{f,g}(N,d)}{d}c_q(a) \\ &= \prod_{p\le Q}\left(1-\frac{1}{p}\right)\sum_{q\in\mathcal{S}}\frac{c_q(a)}{\varphi(q)}\sum_{t\in\mathcal{S}}\frac{C_{f,g}(N,t)c_q(t)}{t}, \quad \forall a\in\mathcal{S}. \end{aligned}$$

(It is easily see that such an expansion holds also if $f,g \in \mathcal{A}^{\varepsilon}$.) On the other hand, in view of Theorem 4, the conditions $f \in \mathcal{A}$, $g \in \mathcal{A}_Q$, with $Q \le N$, and $C_{f,g}(N,a)$ fair, do not suffice to get the REEF for such a correlation on the Q-smooth positive integers, i.e.

$$C_{f,g}(N,a) = \sum_{\ell\le Q}\left(\frac{\widehat{g}(\ell)}{\varphi(\ell)}\sum_{n\le N} f(n)c_\ell(n)\right)c_\ell(a), \ \ \forall a\in\mathcal{S}.$$

In [3] it is provided the following counterexample. For a fixed $q_0 \in [3,Q]\cap\mathbb{N}$, let us take $n_0 \in [1,N]\cap\mathbb{N}$ with $n_0 \equiv -1\,(q_0)$ and define $f,g \in \mathcal{A}$ as $f(n) = \mathbf{1}_{\{n_0\}}(n)$, $g(n) = c_{q_0}(n)$, $\forall n \in \mathbb{N}$.

It is easily seen that $g \in \mathcal{A}_Q$ and $C_{f,g}(N,a)$ is fair. However, it turns out that

$$C_{f,g}(N,1) = \varphi(q_0) \neq \frac{\mu(q_0)^2}{\varphi(q_0)} = \sum_{\ell\le Q}\left(\frac{\widehat{g}(\ell)}{\varphi(\ell)}\sum_{n\le N} f(n)c_\ell(n)\right)c_\ell(a).$$

Acknowledgements. Theorem 4 was a work in progress while the Author gave a talk about it and other results during the Second Number Theory Meeting - Torino 2017. He wishes to thank the organizers of the meeting for the invitation. Also, he acknowledges the encouragement by Ram Murty to pursue his research on these topics. Last but not least, special thanks to Maurizio Laporta for a big improvement in the paper exposition.

References

[1] COPPOLA G., *An elementary property of correlations*, Hardy-Ramanujan J. **41** (2018), 68–76.

[2] COPPOLA G., *A map of Ramanujan expansions*, ArXiV:1712.02970v2.

[3] COPPOLA G., *A smooth shift approach for a Ramanujan expansion*, ArXiV: 1901.01584v3

[4] COPPOLA G. AND LAPORTA M., *Generations of correlation averages*, J. Numbers, **2014**, Article ID 140840, (2014), 1-13.

[5] COPPOLA G. AND LAPORTA M., *On the Correlations, Selberg integral and symmetry of sieve functions in short intervals, III*, Mosc. J. Comb. Number Theory, **6** (1), (2016), 3-24.

[6] COPPOLA G. AND LAPORTA M., *Sieve functions in arithmetic bands*, Hardy-Ramanujan J., **39**, (2016), 21-37.

[7] COPPOLA G. AND LAPORTA M., *Symmetry and short interval mean-squares*, Analytic number theory, On the occasion of the 80th anniversary of the birth of Anatolii Alekseevich Karatsuba, Tr. Mat. Inst. Steklova, **299**, MAIK Nauka/Interperiodica, Moscow, 2017, 62-85; Proc. Steklov Inst. Math., **299**, (2017), 56-77.

[8] COPPOLA G. AND LAPORTA M., *Sieve functions in arithmetic bands, II*, Indian J. Pure Appl. Math., **49** (2), (2018), 301-311.

[9] COPPOLA G. AND LAPORTA M., *Correlations and distribution of essentially bounded functions in short intervals*, Afr. Mat. **29**, (2018), 1245-1263.

[10] COPPOLA G., MURTY, M.RAM AND SAHA, B., *Finite Ramanujan expansions and shifted convolution sums of arithmetical functions*, J. Number Theory **174** (2017), 78–92.

[11] COPPOLA G. AND MURTY M.R., *Finite Ramanujan expansions and shifted convolution sums of arithmetical functions, II*, J. Number Theory, **185** (2018), 16-47.

[12] DAVENPORT H., *Multiplicative Number Theory*, Third Edition, GTM **74**, Springer, New York, 2000.

[13] DELANGE H., *On Ramanujan expansions of certain arithmetical functions*, Acta Arith., **31** (1976), 259–270.

[14] DELANGE H., *On a formula for almost-even arithmetical functions*, Illinois J. Math. **31** (1987), 24–35.

[15] GADIYAR H.G., MURTY M.R. AND PADMA R., *Fourier series and a theorem of Ingham*, Indian J. Pure Appl. Math., **45** (5), (2014), 691-706.

[16] HARDY G.H., *Note on Ramanujan's trigonometrical function $c_q(n)$ and certain series of arithmetical functions*, Proc. Cambridge Phil. Soc., **20** (1921), 263–271.

[17] HARDY G.H. AND LITTLEWOOD J.E., *Some problems of partitio numerorum. III: On the expression of a number as a sum of primes*, Acta Math., **44** (1923), 1–70.

[18] HILDEBRAND A., *Über die punktweise Konvergenz von Ramanujan-Entwicklungen zahlentheoretischer Funktionen*, Acta Arith., **44** (1984), 109–140.

[19] LUCHT L., *A Survey of Ramanujan expansions*, Int. J. Number Theory **6**, (2010), 1785–1799.

[20] MONTGOMERY H.L., *Ten Lectures on the Interface Between Analytic Number Theory and Harmonic Analysis*, CBMS Regional Conf. Ser. in Math. **84**, Amer. Math. Soc., Providence, RI, 1994.

[21] MURTY M.R., *Ramanujan series for arithmetical functions*, Hardy-Ramanujan J., **36** (2013), 21-33.

[22] RAMANUJAN S., *On certain trigonometrical sums and their application to the theory of numbers*, Transactions Cambr. Phil. Soc., **22** (1918), 259–276.

[23] SCHWARZ W. AND SPILKER J., *Arithmetical functions, (An introduction to elementary and analytic properties of arithmetic functions and to some of their almost-periodic properties).* London Mathematical Society Lecture Note Series, **184**, Cambridge University Press, Cambridge, 1994.

[24] TENENBAUM G., *Introduction to Analytic and Probabilistic Number Theory*, Cambridge Studies in Advanced Mathematics, **46**, Cambridge University Press, 1995.

[25] WINTNER A., *Eratosthenian averages*, Waverly Press, Baltimore, MD, 1943.

AMS Subject Classification: 11A25, 11K65, 11N37

Keywords: Ramanujan expansion, Erathostenes transform, shifted convolution sum

Giovanni COPPOLA
Dipartimento di Matematica e Applicazioni,
Universitá degli Studi di Napoli "Federico II",
Complesso di Monte S. Angelo-Via Cinthia
80126 Napoli (NA), ITALY
Home address: Via Partenio, 12 - 83100 Avellino
e-mail: giovanni.coppola@unina.it
Lavoro pervenuto in redazione il 16.07.2019.

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 83 – 101

M. Elia

CONTINUED FRACTIONS AND FACTORING

Abstract. Legendre found that the continued fraction expansion of $\sqrt{N}$ having odd period leads directly to an explicit representation of N as the sum of two squares. Similarly, it is shown here that the continued fraction expansion of $\sqrt{N}$ having even period directly produces a factor of a composite N. Shanks' infrastructural method is then revisited, and some consequences of its application to factorization by means of the continued fraction expansion of $\sqrt{N}$ are derived.

Mathematics Subject Classification (2010): 11A55, 11A51

1. Introduction

Continued fractions have always held great fascination, for both aesthetic reasons and practical purposes. Among the many clever properties of periodic continued fractions, Legendre found how to obtain the representation of an integer N as the sum of two squares, in his own words, "*sans aucun tâtonnement*" from the continued fraction expansion of $\sqrt{N}$ when the period is odd [11]. In particular, this property holds for any prime p congruent 1 modulo 4, [11, 16]. As a kind of counterpart to Legendre's finding, this paper shows how to obtain a factor of a composite N directly from the continued fraction expansion of $\sqrt{N}$ when the period is even. In particular, this is certainly possible when both prime factors of N are congruent 3 modulo 4.

Based on this result, derived from peculiar properties of continued fraction convergents, and on an adaptation of Shanks' infrastructural machinery, a factoring algorithm is proposed whose complexity depends on the accuracy of the evaluation of certain integrals of Dirichlet's. The paper is organized as follows. Section 2 summarizes the properties of the continued fraction expansion of $\sqrt{N}$. In Section 3, some new properties of the convergents are proved, and Shanks' infrastructural method is revisited and applied to a sequence of quadratic forms generated from the convergent of the continued fraction expansion of $\sqrt{N}$. Section 4 discusses the factorization of composite numbers N when the period of the continued fraction expansion of $\sqrt{N}$ is even. Lastly, Section 5 briefly reports some conclusions.

2. Preliminaries

A regular continued fraction is an expression of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}, \tag{1}$$

where a_0, a_1, $a_2,\ldots,a_i,\ldots$ is a sequence, possibly infinite, of positive integers. A convergent of a continued fraction is a sequence of fractions $\frac{A_m}{B_m}$, each of which is obtained by truncating the continued fraction at the m-th term. The fraction $\frac{A_m}{B_m}$ is called the m-th convergent [5, 8, 12]. The first few initial terms of the convergent of (1) are

$$\frac{A_0}{B_0}=\frac{a_0}{1},\ \frac{A_1}{B_1}=\frac{a_0a_1+1}{a_1},\ \frac{A_2}{B_2}=\frac{a_0a_1a_2+a_0+a_2}{a_1a_2+1},\ldots\quad .$$

Numerators and denominators of the m-th convergent satisfy the second-order recurrences

$$(2)\qquad \begin{cases} A_m=a_mA_{m-1}+A_{m-2} & ,\ A_0=a_0,\ A_1=a_0a_1+1 \\ B_m=a_mB_{m-1}+B_{m-2} & ,\ B_0=1,\ B_1=a_1 \end{cases} \quad ,\ \forall\, m\geq 2;$$

further, we have [5, p.85] the relationships

$$(3)\qquad A_mB_{m-1}-A_{m-1}B_m=(-1)^{m-1}$$

$$(4)\qquad A_mB_{m-2}-A_{m-2}B_m=(-1)^{m-2}a_m\ .$$

Equation (3) shows that numerator and denominator of the m-th convergent are relatively prime.

A continued fraction is said to be definitively periodic, with period τ, if, starting from a finite n_o, a fixed pattern a'_1, $a'_2,\ldots,a'_\tau$ repeats indefinitely. Lagrange showed that any definitively periodic continued fraction represents a positive number of the form $a+b\sqrt{N}$, $a,b\in\mathbb{Q}$, i.e. an element of $\mathbb{F}=\mathbb{Q}(\sqrt{N})$, and conversely that any such positive number is represented by a definitively periodic continued fraction [5, 16]. The maximal order of $\mathbb{F}$ is denoted $\mathfrak{O}_\mathbb{F}$. Let $\mathcal{G}(\mathbb{F}/\mathbb{Q})=\{\iota,\sigma\}$ be the Galois group of $\mathbb{F}$ over $\mathbb{Q}$, where ι denotes the group identity, and the action of the automorphism σ, called conjugation, is defined as $\sigma(a+b\sqrt{N})=a-b\sqrt{N}$. The field norm $N_\mathbb{F}(\mathfrak{a})$ of $\mathfrak{a}\in\mathbb{F}$ is defined to be $N_\mathbb{F}(\mathfrak{a})=\mathfrak{a}\sigma(\mathfrak{a})$.

In the continued fraction expansion of $\sqrt{N}$, the period of length τ begins immediately after the first term a_0, and consists of a palindromic part formed by $\tau-1$ terms $a_1,a_2,\ldots,a_2,a_1,$ followed by $2a_0$. Periodic continued fractions of this sort are conventionally written in the form

$$(5)\qquad \sqrt{N}=\left[a_0,\overline{a_1,a_2,\ldots,a_2,a_1,2a_0}\right]\quad ,$$

where the over-lined part is the period. Note that the period of the irrational $\frac{a_0+\sqrt{N}}{N-a_0^2}$ starts immediately without anti-period; in this case, the continued fraction is called purely periodic and is denoted $\left[\overline{a_1,a_2,\ldots,a_2,a_1,2a_0}\right]$.

Carr's book [2, p.70-71] gives a good collection of properties of the continued fraction expansion of $\sqrt{N}$, which are summarized in the following, with the addition of some properties taken from [5, 16, 13]:

1. Let c_n and r_n be the elements of two sequences of positive integers defined by the relation
$$\frac{\sqrt{N}+c_n}{r_n} = a_{n+1} + \frac{r_{n+1}}{\sqrt{N}+c_{n+1}}$$
with $c_0 = \lfloor\sqrt{N}\rfloor$, and $r_0 = N - a_0^2$; the elements of the sequence $a_1, a_2, \ldots, a_n \ldots$ are thus obtained as the integer parts of the left-side fraction
$$a_{n+1} = \left\lfloor \frac{\sqrt{N}+c_n}{r_n} \right\rfloor \quad . \tag{6}$$

2. Let $a_0 = \lfloor\sqrt{N}\rfloor$ be initially computed, and set $c_0 = a_0$, $r_0 = N - a_0^2$, then sequences $\{c_n\}_{n\geq 0}$ and $\{r_n\}_{n\geq 0}$ are produced by the recursions
(7)
$$a_{m+1} = \left\lfloor \frac{a_0 + c_m}{r_m} \right\rfloor \quad , \quad c_{m+1} = a_{m+1} r_m - c_m \quad , \quad r_{m+1} = \frac{N - c_{m+1}^2}{r_m} \quad .$$
These recursive equations, together with (6), allow us to compute the sequence $\{a_m\}_{m\geq 1}$ using rational arithmetical operations; however, the iterations may be stopped when $a_m = 2a_0$, having completed a period.

3. The n-th convergent to $\sqrt{N}$ can be recursively computed as
$$\frac{A_n}{B_n} = \frac{a_n A_{n-1} + A_{n-2}}{a_n B_{n-1} + B_{n-2}} \quad n \geq 1 \quad , \tag{8}$$
with initial conditions $A_{-1} = 1$, $B_{-1} = 0$, $A_0 = a_0$, and $B_0 = 1$.

4. The sequence of ratios $\frac{A_n}{B_n}$ assumes the limit value $\sqrt{N}$ as n goes to infinity, due to the inequality
$$\left| \frac{A_n}{B_n} - \sqrt{N} \right| \leq \frac{1}{B_n B_{n+1}} \quad ,$$
since A_n and B_n go to infinity along with n. Furthermore, $\frac{A_n}{B_n} < \sqrt{N}$, if n is even, and $\frac{A_n}{B_n} > \sqrt{N}$ if n is odd [8, p.132]. Therefore, any convergent of even index is smaller than any convergent of odd index.

5. The true value of $\sqrt{N}$ is the value which (8) becomes when the "approximated" quotient a_n, as defined in (6), is substituted with the complete quotient $\frac{\sqrt{N}+c_{n-1}}{r_{n-1}}$. This gives
$$\sqrt{N} = \frac{(\sqrt{N}+c_{n-1})A_{n-1} + r_{n-1}A_{n-2}}{(\sqrt{N}+c_{n-1})B_{n-1} + r_{n-1}B_{n-2}} \quad .$$

6. The value $c_0 = a_0$ is the greatest value that c_n may assume. No a_n or r_n can be greater than $2a_0$.
If $r_n = 1$ then $a_{n+1} = a_0$. For all n greater than 0, we have $a_0 - c_n < r_n$.

7. The first complete quotient that is repeated is $\frac{\sqrt{N}+c_0}{r_0}$, and a_1, r_0, and c_0 commence each cycle of repeated terms.

8. Through the first period (or cycle) of length τ, the elements $a_{\tau-j}$, $r_{\tau-j-2}$, and $c_{\tau-j-1}$ are respectively equal to a_j, r_j, and c_j.

9. The period length cannot be greater than $2a_0^2$. This bound is very loose and was tightened by Kraitchik [17, p.95], who showed that τ is upper bounded by

$$0.72\sqrt{N}\ln N \qquad N > 7 \ . \tag{9}$$

However, the period length has irregular behavior as a function of N, because it may assume any value from 1, when $N = M^2 + 1$, to values close to the order $O(\sqrt{N}\ln N)$ [16].

10. The element $\mathfrak{c}_m = A_m + B_m\sqrt{N} \in \mathfrak{O}_{\mathbb{F}}$ is associated to the m-th convergent.

Numerators and denominators of the convergents satisfy interesting relations [12, p.92-95]

$$\begin{cases} A_0A_{\tau-1} + A_{\tau-2} - NB_{\tau-1} = 0 \\ A_1A_{\tau-2} + A_0A_{\tau-3} - N(B_1B_{\tau-2} + B_0B_{\tau-3}) = 0 \\ A_jA_{\tau-j-1} + A_{j-1}A_{\tau-j-2} - N(B_jB_{\tau-j-1} + B_{j-1}B_{\tau-j-2}) = 0 \qquad 3 \le j \le \tau-3 \ . \end{cases} \tag{10}$$

Besides these properties, the following equations, [16, p.329-332], are used in the proofs:

$$\begin{cases} A_\tau = 2a_0A_{\tau-1} + A_{\tau-2} \\ B_\tau = 2a_0B_{\tau-1} + B_{\tau-2} \end{cases} \tag{11}$$

$$\begin{cases} A_\tau B_{\tau-1} - A_{\tau-1}B_\tau = (-1)^{\tau-1} \\ A_{\tau-1}B_{\tau-2} - A_{\tau-2}B_{\tau-1} = (-1)^{\tau-2} \\ A_\tau B_{\tau-2} - A_{\tau-2}B_\tau = 2a_0(-1)^\tau \end{cases} \tag{12}$$

$$\begin{cases} A_{\tau-2} = -a_0A_{\tau-1} + NB_{\tau-1} \\ B_{\tau-2} = A_{\tau-1} - a_0B_{\tau-1} \end{cases} \tag{13}$$

$$\begin{cases} A_\tau = a_0A_{\tau-1} + NB_{\tau-1} \\ B_\tau = A_{\tau-1} + a_0B_{\tau-1} \end{cases} \tag{14}$$

REMARK 1. The smallest positive solution of Pell's equation $x^2 - Ny^2 = (\pm 1)$ is $\mathfrak{c}_{\tau-1}$, whenever a solution exists. If $\{1, \sqrt{N}\}$ is an integral basis of $\mathbb{F}$, then $\mathfrak{c}_{\tau-1}$ coincides with the fundamental positive unit ε_0 of $\mathbb{F}$. If $\{1, \frac{1+\sqrt{N}}{2}\}$ is an integral basis of $\mathbb{F}$, then $\mathfrak{c}_{\tau-1}$ may be either ε_0 or ε_0^3. An easy way to check whether $\mathfrak{c}_{\tau-1} = \varepsilon_0^3$ is to solve in $\mathbb{Q}$ the equation $(x+y\sqrt{N})^3 = A_{\tau-1} + B_{\tau-1}\sqrt{N}$, which is equivalent to verifying

whether some solution of the following Diophantine equation is a rational number with 2 as denominator

$$64x^9 - 48A_{\tau-1}x^6 + (27NB_{\tau-1}^2 - 15A_{\tau-1}^2)x^3 - A_{\tau-1}^3 = 0 \quad .$$

If a rational solution x_o of this equation exists, the corresponding y_o can be computed as $y_o = \sqrt{\frac{x_o^2-1}{N}}$.

The following proposition describes how to move from one period to another.

PROPOSITION 1. *The sequence* $\{\mathfrak{c}_m\}_{m\geq 0}$ *satisfies the relation*

$$\mathfrak{c}_{m+k\tau} = \mathfrak{c}_m \mathfrak{c}_{\tau-1}^k \quad \forall\, m,k \in \mathbb{N} \quad . \tag{15}$$

Proof. The two dependencies, with respect to m and k, are disposed of separately. The claimed equality is trivial for $m = k = 0$, and fixing $k = 1$, equation (14) allows us to write $\mathfrak{c}_\tau = a_0\mathfrak{c}_{\tau-1} + \sqrt{N}\mathfrak{c}_{\tau-1} = (a_0 + \sqrt{N})\mathfrak{c}_{\tau-1} = (A_0 + B_0\sqrt{N})\mathfrak{c}_{\tau-1}$. Then, by the recurrences (2) and the periodicity of the a_is, we can write

$$\mathfrak{c}_{\tau+1} = a_1\mathfrak{c}_\tau + \mathfrak{c}_{\tau-1} = a_1(A_0 + B_0\sqrt{N})\mathfrak{c}_{\tau-1} + \mathfrak{c}_{\tau-1} = \mathfrak{c}_1\mathfrak{c}_{\tau-1} \quad .$$

Clearly, we can iterate by using the recurrences (2) and the symmetry of the a_is to obtain the relation $\mathfrak{c}_{\tau+m} = \mathfrak{c}_m\mathfrak{c}_{\tau-1}$, which shows that multiplication by $\mathfrak{c}_{\tau-1}$ is equivalent to a translation by τ. The conclusion is immediate by iterating on k. □

3. Convergents and quadratic forms

Let $\Delta_m = A_m^2 - NB_m^2$ denote the field norm of $\mathfrak{c}_m = A_m + \sqrt{N}B_m \in \mathfrak{O}_{\mathbb{F}}$. Several properties of convergents are better described considering, besides the sequence $\Delta = \{\Delta_m\}_{m\geq 0}$, a second sequence $\Omega = \{\Omega_m = A_mA_{m-1} - NB_mB_{m-1}\}_{m\geq 1}$. Using (3), the following relation can be shown

$$\Omega_{m+1}^2 - \Delta_m\Delta_{m+1} = N \quad \forall m \geq 0 \quad . \tag{16}$$

The elements of the sequences Δ and Ω satisfy the recurrent relations

$$\begin{cases} \Delta_{m+1} = a_{m+1}^2\Delta_m + 2a_{m+1}\Omega_m + \Delta_{m-1} \\ \Omega_{m+1} = \Omega_m + a_{m+1}\Delta_m \end{cases} \qquad m \geq 1 \tag{17}$$

with initial conditions $\Delta_0 = a_0^2 - N$, $\Delta_1 = (1 + a_0a_1)^2 - Na_1^2$ and $\Omega_1 = (1 + a_0a_1)a_0 - Na_1$. Using (17), it is immediate to see that $c_{m+1} = |\Omega_m|$ and $r_{m+1} = |\Delta_m|$.
Introducing the matrix

$$T(a_m) = \begin{bmatrix} a_m^2 & a_m & 1 \\ 2a_m & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad , \tag{18}$$

and defining the column vector $\Lambda_m = [\Delta_m, 2\Omega_m, \Delta_{m-1}]^T$, equations (17) can be written as

$$\Lambda_{m+1} = T(a_{m+1})\Lambda_m \quad \forall\, m \geq 1 \ . \tag{19}$$

Iterating this relation, we have

$$\Lambda_{m+n} = T(a_{m+n})T(a_{m+n-1})\cdots T(a_{m+2})T(a_{m+1})\Lambda_m = T_{(m,n)}\Lambda_m \quad \forall\, m,n \geq 1 \ , \tag{20}$$

where $T_{(m,n)} = \prod_{j=m+1}^{m+n} T(a_j)$ is a matrix that only depends on the sequence of coefficients a_t. Furthermore, from (16) we may derive the relation

$$\Omega_{m+1}^2 - \Omega_m^2 = \Delta_m(\Delta_{m+1} - \Delta_{m-1}) \ \ \forall\, m \geq 1,$$

which allows us to write equation (17) as

$$\begin{cases} \Delta_{m+1} = \Delta_{m-1} + a_{m+1}(\Omega_{m+1} + \Omega_m) \\ \Omega_{m+1} = \Omega_m + a_{m+1}\Delta_m \end{cases} \quad \forall\, m \geq 1 \ . \tag{21}$$

DEFINITION 1. *Let Υ be the sequence of quadratic forms $f_m(x,y) = \Delta_m x^2 + 2\Omega_m xy + \Delta_{m-1}y^2$, $m \geq 1$, defined by means of the sequences Δ and Ω.*

Note that it may sometimes be convenient to denote a quadratic form simply with the triple of coefficients, i.e. the 3-dimensional vector Λ_m; further, due to equation (16), all quadratic forms in Υ have the same discriminant $4N$.

REMARK 2. The absolute values of Δ_m and Ω_m are bounded as

$$|\Delta_m| < 2\frac{1}{a_{m+1}}\sqrt{N} \leq 2\sqrt{N} \qquad , \qquad |\Omega_m| < \sqrt{N} \quad \forall\, m \geq 1 \ .$$

The bound $2\sqrt{N}$ for Δ_m is well known, [8, Theorem 171, p.140], and can be slightly tightened considering the following chain of inequalities

$$|A_m^2 - NB_m^2| = B_m^2\left|\frac{A_m}{B_m} - \sqrt{N}\right|\left(\frac{A_m}{B_m} + \sqrt{N}\right) \leq \frac{B_m}{B_{m+1}}\left|\frac{A_m}{B_m} - \sqrt{N} + 2\sqrt{N}\right|$$

$$\leq \frac{B_m}{B_{m+1}}\left|\frac{A_m}{B_m} - \sqrt{N}\right| + 2\sqrt{N}\frac{B_m}{B_{m+1}} \leq \frac{1}{B_{m+1}^2} + 2\frac{B_m}{a_{m+1}B_m + B_{m-1}}\sqrt{N}$$

$$= 2\frac{1}{a_{m+1}}\sqrt{N} + \frac{1}{B_{m+1}^2} - 2\sqrt{N}\frac{B_{m-1}}{a_{m+1}(a_{m+1}B_m + B_{m-1})} < 2\frac{1}{a_{m+1}}\sqrt{N} \ .$$

The bound for $|\Omega_m|$ is an immediate consequence of equation (16), we have $\Delta_m\Delta_{m+1} < 0$ since the signs in the sequence Δ alternate; consequently

$$\Omega_m^2 = N + \Delta_m\Delta_{m+1} < N \ ,$$

thus taking the positive square root of both sides, the inequality $|\Omega_m| < \sqrt{N}$ is obtained.

3.1. Periodicity and Symmetry

The sequences Δ and Ω are periodic in the same way as the sequence of coefficients a_m, although their periods are even, and may be τ or 2τ depending on whether τ is even or odd. Further, within a period, there exist interesting symmetries.

THEOREM 1 (Periodicity of Δ). *Starting with $m = 1$, the sequence $\Delta = \{\Delta_m\}_{m\geq 0}$ is periodic with period τ or 2τ depending on whether τ is even or odd. The elements of the first block $\{\Delta_m\}_{m=0}^{\tau} \subset \Delta$ satisfy the symmetry relation $\Delta_m = (-1)^{\tau}\Delta_{\tau-m-2}$, $\forall\, 0 \leq m \leq \tau-2$.*

Proof. The period of the sequence Δ is τ or 2τ, as a consequence of equation (15), because the norm of $A_{\tau-1} + \sqrt{N}B_{\tau-1}$ is $(-1)^{\tau}$.
The symmetry of the sequence Δ within the τ elements of the first period follows from the relations

$$\left\{ \begin{array}{l} A_{\tau-m-2} = (-1)^{m-1}A_{\tau-1}A_m + (-1)^m NB_{\tau-1}B_m \\ B_{\tau-m-2} = (-1)^m A_{\tau-1}B_m + (-1)^{m-1}B_{\tau-1}A_m \end{array} \right. , \quad 0 \leq m \leq \tau-2 \, , \tag{22}$$

which are proved using the recurrences (2) together with (13) and (14) [16, p.329-330]; the transformation defined by (22) is identified by the matrix

$$M_{\tau-1} = \left[\begin{array}{cc} -A_{\tau-1} & NB_{\tau-1} \\ -B_{\tau-1} & A_{\tau-1} \end{array} \right] . \tag{23}$$

We have

$$\left\{ \begin{array}{lcl} A^2_{\tau-m-2} - NB^2_{\tau-m-2} & = & (A_{\tau-1}A_m - NB_{\tau-1}B_m)^2 - N(-A_{\tau-1}B_m + B_{\tau-1}A_m)^2 \\ & = & (A^2_m - NB^2_m)(A^2_{\tau-1} - NB^2_{\tau-1}) = (-1)^{\tau}(A^2_m - NB^2_m) \end{array} \right.$$

that is $\Delta_{\tau-m-2} = (-1)^{\tau}\Delta_m$. Actually, equation (22) can be written in the form

$$A_{\tau-m-2} + \sqrt{N}B_{\tau-m-2} = (-1)^{m-1}(A_{\tau-1} + \sqrt{N}B_{\tau-1})(A_m - \sqrt{N}B_m) \tag{24}$$

or more compactly as $\mathfrak{c}_{\tau-m-2} = (-1)^{m-1}\mathfrak{c}_{\tau-1}\sigma(\mathfrak{c}_m)$. □

THEOREM 2 (Periodicity of Ω). *The sequence $\Omega = \{\Omega_m\}_{m\geq 1}$ is periodic of period τ or 2τ depending on whether τ is even or odd. The elements of the first block $\{\Omega_m\}_{m=1}^{\tau} \subset \Omega$ satisfy the symmetry relation*
$\Omega_{\tau-m-1} = (-1)^{\tau+1}\Omega_m$, $\forall\, m \leq \tau-2$.

Proof. The periodicity of the sequence Ω follows from the property expressed by equation (15), noting that

$$\Omega_j = \frac{1}{2}\left((A_j + \sqrt{N}B_j)((A_{j-1} - \sqrt{N}B_{j-1}) + (A_j - \sqrt{N}B_j)((A_{j-1} + \sqrt{N}B_{j-1})\right) .$$

The symmetry property of the sequence Ω within a period follows from (22) in the same way as does that of the sequence Δ; we have

$$\begin{array}{rcl} A_{\tau-1-j}A_{(\tau-1)-j-1} - NB_{\tau-1-j}B_{(\tau-1)-j-1} = & & -(A_{\tau-1}A_j - NB_{\tau-1}B_j)(A_{\tau-1}A_{j-1} - NB_{\tau-1}B_{j-1}) \\ & & +N(A_{\tau-1}B_j - B_{\tau-1}A_j)(A_{\tau-1}B_{j-1} - B_{\tau-1}A_{j-1}) \\ & = & -(A^2_{\tau-1} - NB^2_{\tau-1})(A_jA_{j-1} - NB_jB_{j-1}) \end{array}$$

that is, $\Omega_{\tau-j-1} = (-1)^{\tau+1}\Omega_j$. □

The two quadratic forms $f_n(x,y) = \Delta_n x^2 + 2\Omega_n xy + \Delta_{n-1}y^2$ and $f_{\tau-1-n}(x,y) = \Delta_{n-1}x^2 - 2\Omega_n xy + \Delta_n y^2$ are associated respectively to the positions n and $\tau-1-n$, as a consequence of the symmetries of the sequences Δ and Ω shown by Theorems 1 and 2, within the first block of length τ in Υ. It should be noted that $f_m(x,y)$ and $f_{\tau-1-m}(x,y)$ are improperly equivalent.

Key matrix. Clearly, the column vectors Λ_m and $\Lambda_{\tau-m-1}$ are transformed one into the other by an involutory matrix J of determinant 1

$$\begin{bmatrix} \Delta_{m-1} \\ -2\Omega_m \\ \Delta_m \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta_m \\ 2\Omega_m \\ \Delta_{m-1} \end{bmatrix} .$$

Using the matrices $T(a_n)$ and equation (20), and applying to Λ_m the sequence of matrices $T(a_{m+1}), T(a_{m+2}), \ldots, T(a_{\tau-1-m})$ in reverse order, we obtain $\Lambda_{\tau-1-m}$

$$\Lambda_{\tau-1-m} = T(a_{\tau-1-m})\cdots T(a_{m+1})\Lambda_m \;\Rightarrow\; \Lambda_m = JT(a_{\tau-1-m})\cdots T(a_{m+1})\Lambda_m \quad . \tag{25}$$

Assuming τ is even, this equation implies that Λ_m is an eigenvector of eigenvalue 1 of the matrix

$$E_m = JT(a_{\tau-1-m})\cdots T(a_{m+1}) = JT(a_{m+1})T(a_m)\cdots T(a_{\frac{\tau}{2}-1})T(a_{\frac{\tau}{2}})T(a_{\frac{\tau}{2}+1})\cdots T(a_{m+1})$$

since $T(a_{\tau-1-n}) = T(a_{n+1})$ by the symmetry of the sequence $\{a_n\}_{n=1}^{\tau-1}$. Observing that $JT(a_m)J = T(a_m)^{-1}$ and $J^2 = I$, we have

$$\begin{aligned} E_m &= (JT(a_{n+2})J)(JT(a_{n+3})J)J\cdots\cdots(JT(a_{\frac{\tau}{2}-1})J)JT(a_{\frac{\tau}{2}})T(a_{\frac{\tau}{2}-1})\cdots T(a_{n+2}) \\ &= T(a_{n+2})^{-1}\cdots T(a_{\frac{\tau}{2}-1})^{-1}JT(a_{\frac{\tau}{2}})T(a_{\frac{\tau}{2}-1})\cdots T(a_{n+2}) \\ &= (T(a_{\frac{\tau}{2}-1})\cdots T(a_{n+2}))^{-1}JT(a_{\frac{\tau}{2}})(T(a_{\frac{\tau}{2}-1})\cdots T(a_{n+2})) \quad . \end{aligned} \tag{26}$$

It follows that the matrix E_m has the same characteristic polynomial $z^3 - z^2 - z + 1$ as $JT(a_{\frac{\tau}{2}})$, i.e. E_m has eigenvalue -1 with multiplicity 1, and eigenvalue 1 with geometric multiplicity 2.

Assuming τ is odd, the symmetries of the sequences $\{a_n\}_{n=1}^{\tau-1}$, $\{\Delta_n\}_{n=1}^{\tau-1}$, and $\{\Omega_n\}_{n=1}^{\tau-1}$, refer to an even number $\tau-1$ of terms, and equation (26) is written as

$$\begin{aligned} D_n &= (JT(a_{n+2})J)(JT(a_{n+3})J)J\cdots\cdots(JT(a_{\frac{\tau-3}{2}})J)JT(a_{\frac{\tau-3}{2}})\cdots T(a_{n+2}) \\ &= T(a_{n+2})^{-1}\cdots T(a_{\frac{\tau-3}{2}})^{-1}JT(a_{\frac{\tau-3}{2}})\cdots T(a_{n+2}) \\ &= (T(a_{\frac{\tau-3}{2}})\cdots T(a_{n+2}))^{-1}J(T(a_{\frac{\tau-3}{2}})\cdots T(a_{n+2})) \quad . \end{aligned} \tag{27}$$

It follows that the matrix D_n has the same characteristic polynomial $z^3 + z^2 - z - 1$ of J, i.e. D_n has eigenvalue 1 with multiplicity 1, and eigenvalue -1 with geometric multiplicity 2.

An example may clarify the method.

EXAMPLE 1. Consider the continued fraction expansion of $\sqrt{386}$, which has period $\tau = 12$

$$[[19],[1,1,1,4,1,18,1,4,1,1,1,38]]$$

Consider the vector $\Lambda_3 = [7,-30,-23]$, since $\tau - 1 - 3 = 8$ the vector Λ_8 by symmetry is $[-23,30,7]$, i.e. $\Lambda_8 = J\Lambda_3$. However, Λ_8 may be obtained by multiplying Λ_3 by a convenient sequence of matrices

$$T(a) = \begin{bmatrix} a^2 & a & 1 \\ 2a & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\Lambda_8 = T(4)T(1)T(18)T(1)T(4)\Lambda_3$$

Since $\Lambda_3 = J\Lambda_8$, we have the equation $\Lambda_3 = JT(4)T(1)T(18)T(1)T(4)\Lambda_3$, that is

$$\Lambda_3 = \begin{bmatrix} 9801 & 1980 & 400 \\ -97020 & -19601 & -3960 \\ 240100 & 48510 & 9801 \end{bmatrix} \Lambda_3 \quad \Rightarrow \quad \Lambda_3 = E_3\Lambda_3 \ ,$$

i.e. Λ_3 is an eigenvector of E_3 for the eigenvalue 1.
The characteristic polynomial of E_3 is found to be $Z^3 - Z^2 - Z + 1 = (Z+1)(Z-1)^2$ which is the same of the matrix $JT(a_6)$, with

$$T_{\frac{\tau}{2}} = T(18) = \begin{bmatrix} 324 & 18 & 1 \\ 36 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \ ;$$

note that $\frac{\tau}{2} = 6$, and in position 5 we find the vector $\Lambda_5 = [2,-36,-31]$ whose first entry gives the factor 2 of 386.

THEOREM 3. *The correspondence $m \leftrightarrow \Lambda_m$ is one-to-one for $1 \le m \le \tau$, i.e. all quadratic forms $f_m(x,y)$ within a period are distinct.*

Proof. The proof is by contradiction. Suppose, contrary to the theorem's claim, that $\Lambda_{n_1} = \Lambda_{n_2} = X$ for some $n_1 < n_2$, then equation (20) implies the existence of a matrix $P_{n_2 n_1} = \prod_{j=n_1+1}^{n_2} T(a_j)$ such that $\Lambda_{n_2} = P_{n_2 n_1}\Lambda_{n_1}$. Thus X must be an eigenvector, for the eigenvalue 1, of the non-negative (positive whenever $n_2 - n_1 \ge 2$) matrix $P_{n_2 n_1}$ which is the product of non-negative matrices.
If $n_2 = n_1 + 1$, it is direct to compute the characteristic polynomial $p(x)$ of $P_{n_2 n_1} = T(a_{n_1})$

$$p(x) = x^3 - (a_{n_1}^2 + 1)x^2 - (a_{n_1}^2 + 1)x + 1 \ ,$$

which is a 3-degree reciprocal polynomial which has a single root -1, and the remaining roots are certainly different from 1, because $a_{n_1} \neq 0$; thus, in this case, X cannot exist.

To prove in general that X does not exist, we observe that any $P_{n_2 n_1}$ has a reciprocal characteristic polynomial $q(x)$ of degree 3, because we have

$$q(x) = \det\left(\lambda I_3 - \prod_{j=n_1+1}^{n_2} T(a_j)\right) = \det\left(\lambda I_3 - J\prod_{j=n_1+1}^{n_2} T(a_j)J\right) = \det\left(\lambda I_3 - \prod_{j=n_1+1}^{n_2} T(a_j)^{-1}\right) ,$$

$$q(x) = \det\left(\lambda I_3 - \prod_{j=n_1+1}^{n_2} T(a_j)\right) = \det\left(\lambda I_3 - \left(\prod_{j=n_1+1}^{n_2} T(a_j)\right)^{-1}\right) ,$$

where the last equality is justified by [9, Theorem 1.3.20, p.53]. The reciprocal polynomial $q(x)$ has an eigenvalue equal to either -1 or 1. If the eigenvalue is -1, which occurs when $n_2 - n_1$ is odd, the eigenvector X does not exist. If the eigenvalue is 1, which occurs when $n_2 - n_1$ is even, there is a second eigenvector for the same eigenvalue, because we have

$$J\Lambda_{n_2} = JP_{n_2 n_1}\Lambda_{n_1} = JP_{n_2 n_1}J\cdot J\Lambda_{n_1} = \left(\prod_{j=n_1+1}^{n_2} T(a_j)\right)^{-1} J\Lambda_{n_1} \;\Rightarrow\; \left(\prod_{j=n_1+1}^{n_2} T(a_j)\right) J\Lambda_{n_2} = J\Lambda_{n_2}.$$

Then, X and JX should be distinct eigenvectors (because $\Omega_{n_2} \neq 0$ for every n_2) of the same eigenvalue 1 of multiplicity one, which is impossible.
In conclusion, the eigenvector X of eigenvalue 1 does not exist, so $m \leftrightarrow \Lambda_m^T$ is a one-to-one mapping within each period. □

3.2. Odd period

In [11, p.59-60], Legendre describes a constructive method for computing the representation of a positive (square-free) N as the sum of two squares, by means of the continued fraction expansion of $\sqrt{N}$. This result is stated as a theorem with a different proof from that of Legendre [11, p.60].

THEOREM 4. *Let N be a positive integer such that the continued fraction expansion of $\sqrt{N}$ has odd period τ. The representation of $N = x^2 + y^2$ is given by $x = \Delta_{\frac{\tau-1}{2}}$ and $y = \Omega_{\frac{\tau-1}{2}}$.*

PROOF. Since τ is odd, by the anti-symmetry in the sequence $\{\Delta_n\}_{n=0}^{\tau-2}$, we have $\Delta_{\frac{\tau-1}{2}-1} = -\Delta_{\frac{\tau-1}{2}}$, so that the quadratic form $\Delta_{\frac{\tau-1}{2}}X^2 + 2\Omega_{\frac{\tau-1}{2}}XY + \Delta_{\frac{\tau-3}{2}}Y^2$ has discriminant $4\Delta^2_{\frac{\tau-1}{2}} + 4\Omega^2_{\frac{\tau-1}{2}} = 4N$, which shows the assertion. □

3.3. Even period

Let N be a square-free composite integer such that the continued fraction of $\sqrt{N}$ has even period. We say that $\mathfrak{c}_{\tau-1} = A_{\tau-1} + B_{\tau-1}\sqrt{N}$ splits N whenever $A_{\tau-1} + 1$ and $A_{\tau-1} - 1$ are divisible by proper factors, say m_1 and m_2, of $N = m_1 m_2$, respectively.

LEMMA 1. *If the period τ of the continued fraction expansion of $\sqrt{N}$ is even, we have*

$$\Delta_\tau = \Delta_{\tau-2} \quad \textit{and} \quad \Omega_\tau = -\Omega_{\tau-1}$$

with $\Omega_{\tau-1} = -a_0$.

Proof. Since $\Delta_{\tau-1} = 1$, we have $\Omega_{\tau-1}^2 - \Delta_{\tau-2} = N$, thus $\Omega_{\tau-1} = -\sqrt{N+\Delta_{\tau-2}}$ because $\tau-1$ is odd. Considering the Taylor series around the origin for the square root, we have

$$\Omega_{\tau-1} = -\sqrt{N+\Delta_{\tau-2}} = -\sqrt{N}\left(1 - \frac{\Delta_{\tau-2}}{2N} + \frac{\Delta_{\tau-2}^2}{8N^2} + \cdots\right) = -\left\lfloor\sqrt{N}\right\rfloor = -a_0 \ .$$

Using equation (17) with $m = \tau - 1$ we have

$$\Delta_\tau = \Delta_{\tau-2} + a_\tau\left(2\Omega_{\tau-1} + a_\tau\Delta_{\tau-1}\right) = \Delta_{\tau-2} \ .$$

Thus, equation (21) finally gives $\Omega_\tau = -\Omega_{\tau-1}$. □

LEMMA 2. *Let τ be even, and define the integer $\gamma \in \mathfrak{O}_\mathbb{F}$ by the product*

$$\gamma = \prod_{m=0}^{\tau-1}\left(\sqrt{N} + (-1)^m\Omega_m\right) \ ,$$

then $\frac{\gamma}{\sigma(\gamma)} = \left(A_{\tau-1} + B_{\tau-1}\sqrt{N}\right)^2 = \mathfrak{c}_{\tau-1}^2$.

Proof. The norm of $\frac{\gamma}{\sigma(\gamma)}$ is patently 1, thus it remains to prove that $\frac{\gamma}{\sigma(\gamma)}$ lies in $\mathfrak{O}_\mathbb{F}$. We have

$$\frac{\gamma}{\sigma(\gamma)} = \prod_{m=0}^{\tau-1}\frac{\sqrt{N}+(-1)^m\Omega_m}{-\sqrt{N}+(-1)^m\Omega_m} = \prod_{m=0}^{\tau-1}\frac{(\sqrt{N}+(-1)^m\Omega_m)^2}{\Omega_m^2 - N} = \prod_{m=0}^{\tau-1}\frac{(\sqrt{N}+(-1)^m\Omega_m)^2}{\Delta_m\Delta_{m-1}} \ .$$

Observing that $\prod_{m=0}^{\tau-1}(\Delta_m\Delta_{m-1}) = \prod_{m=0}^{\tau-1}\Delta_m^2$ by the periodicity of the sequence $\{\Delta_m\}_m$, it follows that $\frac{\gamma}{\sigma(\gamma)}$ is a perfect square. Considering the following identity

$$\frac{\sqrt{N}+(-1)^m\Omega_m}{\Delta_m} = (-1)^m\frac{A_{m-1} - B_{m-1}\sqrt{N}}{A_m - B_m\sqrt{N}} \ ,$$

we have that the base of the square giving $\frac{\gamma}{\sigma(\gamma)}$ is

$$\prod_{m=0}^{\tau-1}\frac{(\sqrt{N}+(-1)^m\Omega_m)}{\Delta_m} = \prod_{m=0}^{\tau-1}(-1)^m\frac{A_{m-1} - B_{m-1}\sqrt{N}}{A_m - B_m\sqrt{N}} = (-1)^{\frac{\tau}{2}}\frac{A_{-1} - B_{-1}\sqrt{N}}{A_{\tau-1} - B_{\tau-1}\sqrt{N}} \ .$$

Now, $A_{-1} = 1$ and $B_{-1} = 0$ by definition, thus

$$\prod_{m=0}^{\tau-1}\frac{(\sqrt{N}+(-1)^m\Omega_m)}{\Delta_m} = (-1)^{\frac{\tau}{2}}\left(A_{\tau-1} + B_{\tau-1}\sqrt{N}\right) = (-1)^{\frac{\tau}{2}}\mathfrak{c}_{\tau-1} \ , \tag{28}$$

and in conclusion $\frac{\gamma}{\sigma(\gamma)} = \mathfrak{c}_{\tau-1}^2$, which shows the claimed property. □

The close connection between the continued fraction expansion of $\sqrt{N}$ and the factorization of N is proved using the matrix $M_{\tau-1}$ defined in equation (23). Note that the matrix $M_{\tau-1}$ is involutory, or neg-involutory, since its square is either plus or minus the identity matrix I_2, i.e. $M_{\tau-1}^2 = (-1)^\tau I_2$. If τ is even, the eigenvalues of matrix $M_{\tau-1}$ are ± 1, and $M_{\tau-1}$ is involutory. If τ is odd, the eigenvalues are $\pm i$, and $M_{\tau-1}$ is neg-involutory.

THEOREM 5. *If the period τ of the continued fraction expansion of $\sqrt{N}$ is even, the element $\mathfrak{c}_{\tau-1}$ in $\mathbb{Q}(\sqrt{N})$ splits $2N$, and a factor of $2N$ is located at positions $\frac{\tau-2}{2} + j\tau$, $j = 0, 1, \ldots$, in the sequence $\Delta = \{\mathfrak{c}_m \sigma(\mathfrak{c}_m)\}_{m \geq 1}$.*

Proof. It is sufficient to consider $j = 0$, due to the periodicity of Δ. Since τ is even, $M_{\tau-1}$ is involutory and has eigenvalues ± 1 with corresponding eigenvectors

$$X^{(h)} = \left[\frac{A_{\tau-1} - (-1)^h}{d}, \frac{B_{\tau-1}}{d} \right]^T \quad \text{with} \quad d = \gcd\{A_{\tau-1} - (-1)^h, B_{\tau-1}\} \quad h = 0, 1 \ .$$

Considering equation (22) written as

$$\begin{bmatrix} A_{\tau-j-2} \\ B_{\tau-j-2} \end{bmatrix} = (-1)^{j-1} M_{\tau-1} \begin{bmatrix} A_j \\ B_j \end{bmatrix} \ ,$$

we see that $Y^{(j)} = [A_j, B_j]^T$ is an eigenvector of $M_{\tau-1}$, of eigenvalue $(-1)^{j-1}$ if and only if j satisfies the condition $\tau - j - 2 = j$, that is $j = \frac{\tau-2}{2} = \tau_0$. From the comparison of $X^{(h)}$ and $Y^{(\tau_0)}$, we have

$$A_{\tau_0} = \frac{A_{\tau-1} - (-1)^{\tau_0 - 1}}{d} \qquad B_{\tau_0} = \frac{B_{\tau-1}}{d} \ , \tag{29}$$

where the equalities are fully motivated because $\gcd\{A_{\tau_0}, B_{\tau_0}\} = 1$. Direct computation yields

$$\Delta_{\tau_0} = \frac{(A_{\tau-1} - (-1)^{\tau_0-1})^2 - N B_{\tau-1}^2}{d^2} = 2\frac{(-1)^{\tau_0-1} A_{\tau-1} + 1}{d^2} \ , \tag{30}$$

which can be written as $A_{\tau_0}^2 - N B_{\tau_0}^2 = 2(-1)^{\tau_0 - 1} \frac{A_{\tau_0}}{d}$; dividing this equality by $2\frac{A_{\tau_0}}{d}$ we have

$$\frac{d A_{\tau_0}}{2} - N \frac{1}{\frac{2A_{\tau_0}}{d}} B_{\tau_0}^2 = (-1)^{\tau_0 - 1} \ .$$

Noting that $\gcd\{A_{\tau_0}, B_{\tau_0}\} = 1$, it follows that $\frac{2A_{\tau_0}}{d}$ is certainly a divisor of $2N$, i.e. $\Delta_{\tau_0} | 2N$. □

EXAMPLE 2. Consider $N = 3 \cdot 5 \cdot 7 \cdot 11 \cdot 19 = 21945$, the period of the continued fraction of $\sqrt{21945}$ is 10, and is fully shown in the following table for the sequences Δ and Ω

j	Δ_j	Ω_j
-1	1	
0	-41	148
1	64	-139
2	-129	117
3	16	-141
4	-21	147
5	16	-147
6	-129	141
7	64	-117
8	-41	139
9	1	-148
10	-41	148

In position $j = \frac{\tau-2}{2} = 4$ we find 21, a factor of N, as expected. The same factor 21 can be found by considering the fundamental unit $\varepsilon_9 = 3004586089 + 20282284\sqrt{21945}$, in fact we have $3004586089 - 1 = 2^3 \cdot (3 \cdot 7) \cdot 4229^2$, and the second factor $5 \cdot 11 \cdot 19$ may be obtained from $3004586089 + 1 = 2 \cdot (5 \cdot 11^3 \cdot 19) \cdot 109^2$.

In principle, in many cases the above Theorem 5 yields a factor of N; however there are examples in which only the factor 2 appears.

EXAMPLE 3. Let $N = 8527 \times 8537 = 72794999$ be a composite number. The period of $\sqrt{N}$ is $\tau = 3864$ and in position 1931 we do not find a factor of N but $\Delta_{1931} = 2$ which is a factor of $2N$.

It would be interesting to find a general condition that can discriminate the various situations, i.e. whether a factor of N is found or not. This objective can be achieved almost in full when $N = pq$ is the product of two primes, a case that cleverly shows the difficulty of the whole problem.

3.4. Factoring $N = pq$

When $N = pq$ is the product of two distinct primes, the analysis of section 3.3 may be further pursued, leading to the following remarkable property:

PROPOSITION 2. *If $p \equiv q \equiv 3 \bmod 4$, the fundamental unit ε_0 (or the cube ε_0^3) splits $N = pq$, then $\Delta_{\frac{\tau-2}{2}}$ is equal to $(q|p)p$, with $p < q$.*

This proposition is given without the proof, which uses units and splitting of primes in quadratic number fields (see [6, 4, 10]); further, the complete classification in terms of residues of p and q modulo 8, proved in [6],. is reported in Table 7.1 for easy reference.

4. Factorization

Gauss recognized that the factoring problem was important, although very difficult,

> *... Problema, numeros primos a compositis dignoscendi, hosque in factores suos primos resolvendi, ad gravissima ac utilissima totius arithmeticae pertinere, et geometrarum tum veterum tum recentiorum industriam ac sagacitatem occupavisse, tam notum est, ut de hac re copiose loqui superfluum foret. ...* C. F. GAUSS [*Disquisitiones Arithmeticae* ART. 329]

and, in spite of much effort, various different approaches, and the problem's increased importance due to the large number of cryptographic applications, no satisfactorily factoring method has yet been found.

Many factorizations make use of the regular continued fraction expansion of $\sqrt{N}$, combined with the idea of using quadratic forms [7, 13]. The infrastructure method, proposed by Shanks [15], considers the subset $\Psi = \{f_m(x,y)\}_{1\leq m\leq \tau-1}$ in the periodic sequence $\Upsilon = \{f_m(x,y)\}_{m\geq 1}^{\infty}$ of reduced principal quadratic forms. It should be remarked that the forms $f_m(x,y) = \Delta_m x^2 + 2\Omega_m xy + \Delta_{m-1} y^2$ in Υ are reduced following a different convention from that commonly adopted [1].

DEFINITION 2. *A real quadratic form $f(x,y) = ax^2 + 2bxy + cy^2$ of discriminant $4N$ is said to be reduced if, defining $\kappa = \min\{|a|,|c|\}$, b is the sole integer such that $\sqrt{N} - |b| < \kappa < \sqrt{N} + |b|$, with the sign of b chosen opposite to the sign of a.*

DEFINITION 3. *The distance between $f_{m+1}(x,y)$ and $f_m(x,y)$ is defined to be*

$$d(f_{m+1}, f_m) = \frac{1}{2}\ln\left(\frac{\sqrt{N} + (-1)^m \Omega_m}{\sqrt{N} - (-1)^m \Omega_m}\right) \quad . \tag{31}$$

The distance between two quadratic forms $f_m(x,y)$ and $f_n(x,y)$, with $m > n$, is defined to be the sum

$$d(f_m, f_n) = \sum_{j=n}^{m-1} d(f_{j+1}, f_j) \quad . \tag{32}$$

Taking the above definitions, Shanks showed that, by the Gauss composition law of quadratic forms with the same determinant, followed by reduction, the set Ψ equipped with the distance $d(f_{m+1}, f_m)$ modulo $R = \ln \mathfrak{c}_{\tau-1}$ resembles a cyclic group, with $f_{\tau-1}(x,y)$ playing the role of identity. Composition followed by reduction affords big steps (giant steps) within Ψ, thus two operators were further defined [3, p.259] to allow small steps (baby steps), precisely

1. One-step forward: The operator ρ^+ that transforms one reduced quadratic form into the next in the sequence Υ, is defined as

$$\rho^+([a,2b,c]) = [\frac{b_1^2 - N}{a}, 2b_1, a] \quad ,$$

where b_1 is $2b_1 = [2b \bmod (2a)] + 2ka$ with k chosen in such a way that $-|a| < b_1 < |a|$.

2. One-step backward: The operator ρ^- that transforms a reduced quadratic form into the immediately preceding quadratic form in the sequence Υ is defined as

$$\rho^-([a,2b,c]) = [c, 2b_1, \frac{b_1^2 - N}{c}] \ ,$$

where b_1 is $2b_1 = [-2b \bmod (2c)] + 2kc$ with k chosen such that $-|c| < b_1 < |c|$.

The infrastructure machinery was used to compute the fundamental unit, the regulator, and the class number [3], with complexity smaller than $O(\sqrt{N})$, although not of polynomial complexity in N. From a different perspective, by Theorem 5, in many cases a factor of N is exactly positioned in the middle of a period of the sequence Δ. Therefore, instead of trying to find special quadratic forms randomly located in Ψ (the principal genus), or some ambigue form in some non-principal genus, we may try to localize the position of some factor of N within a period whose length τ is unknown. Then, it is shown that, by extending the infrastructure machinery to the whole sequence Υ, some factors of N can be computed with a complexity substantially bounded by the complexity required to evaluate an integral of Dirichlet's at a given accuracy: the more precise the evaluation of the integral, the less complex the factorization; at the limit, it is of polynomial complexity; clearly, to be more accurate in the integral evaluation, greater complexity is required. To pursue this idea, we briefly review and adapt the previous definitions of the infrastructure components to the new task. Let us recall that the quadratic forms $f_m(x,y)$ are primitive, i.e. $\gcd\{\Delta_m, 2\Omega_m, \Delta_{m-1}\} = 1$, and at least one between $|\Delta_m|$ and $|\Delta_{m-1}|$ is less than $\sqrt{N}$ and $0 < |\Omega_m| < \sqrt{N}$. Further, since $\mathfrak{c}_{\tau-1}$ is either equal to the positive fundamental unit of $\mathbb{F} = \mathbb{Q}(\sqrt{N})$ or equal to its cube, the regulator of $\mathfrak{O}_\mathbb{F}$ is either $R_\mathbb{F} = \ln \mathfrak{c}_{\tau-1}$, or $R_\mathbb{F} = \frac{1}{3} \ln \mathfrak{c}_{\tau-1}$. The following observations are instrumental to motivate the procedure:

1. The sign of Δ_{m-1} is the same as that of Ω_m, which is opposite to that of Δ_m, thus in the sequence Υ the two triplets of signs $(-,+,+)$ and $(+,-,-)$ alternate.

2. The distance of $f_m(x,y)$ from the beginning of Υ is defined by referring to a properly selected hypothetical quadratic form, i.e. $f_0(x,y) = f_\tau(x,y) = f_0(x,y) = \Delta_0 x^2 - 2\sqrt{N - \Delta_0} xy + y^2$, which is located before $f_1(x,y)$, that is $d(f_m, f_0)$ is given by (32) if $m < \tau$, and by $d(f_m, f_0) = d(f_{m \bmod \tau}, f_0) + kR_\mathbb{F}$ if $k\tau \le m < (k+1)\tau$.

3. Let "$\bullet$" denote the form composition $f_m(x,y) \bullet f_n(x,y)$ in Υ, that is the Gauss composition [3] of $f_m(x,y)$ and $f_n(x,y)$ followed by a reduction performed with the minimum number of steps, ending with a reduced form whose triplet of signs is $(-,+,+)$ if m and n have the same parity, and $(+,-,-)$ otherwise. This distance defined by (31) holds in Υ with good approximation, and is compatible with the "$\bullet$" operation, that is we have

$$f_{\ell(m,n)}(x,y) = f_m(x,y) \bullet f_n(x,y) \Rightarrow d(f_{\ell(m,n)}, f_0) \approx d(f_m, f_0) + d(f_n, f_0) \ .$$

It is remarked that the error affecting this distance estimation is of order $O(\ln N)$ as shown by Schoof in [14].

4. Shanks [15] observed that, within the first period, the composition law "$\bullet$" induces a structure similar to a cyclic group for the addition of distances modulo the "regulator".

5. Between the elements of Υ the distance is nearly maintained by the giant steps, and is rigorously maintained by the baby steps.

THEOREM 6. *The distance $d(f_\tau, f_0)$ is exactly equal to $\ln \mathfrak{c}_{\tau-1}$, i.e. this distance $d(f_\tau, f_0)$ is either the regulator $R_\mathbb{F}$ or $3R_\mathbb{F}$. The distance $d(f_{\frac{\tau}{2}}, f_0)$ is exactly equal to $\frac{1}{2}d(f_\tau, f_0)$.*

Proof. The distance between f_τ and f_0 is the summation

$$d(f_\tau, f_0) = \sum_{j=0}^{\tau-1} d(f_{j+1}, f_j) = \sum_{j=0}^{\tau-1} \frac{1}{2} \ln\left(\sum_{j=0}^{\tau-1} \frac{\sqrt{N}+(-1)^j\Omega_j}{\sqrt{N}-(-1)^j\Omega_j}\right) = \frac{1}{2}\ln\left(\prod_{j=0}^{\tau-1} \frac{\sqrt{N}+(-1)^j\Omega_j}{\sqrt{N}-(-1)^j\Omega_j}\right) .$$

Recalling that $N-\Omega_j^2 = -\Delta_j\Delta_{j-1} > 0$, and taking into account the periodicity of the sequence Δ, the last expression can be written with rational denominator as

$$\frac{1}{2}\ln\left(\prod_{j=0}^{\tau-1} \frac{(\sqrt{N}+(-1)^j\Omega_j)^2}{-\Delta_j\Delta_{j-1}}\right) = \frac{1}{2}\ln\left(\prod_{j=0}^{\tau-1} \frac{(\sqrt{N}+(-1)^j\Omega_j)^2}{\Delta_j^2}\right) = \ln\left(\prod_{j=0}^{\tau-1} \frac{\sqrt{N}+(-1)^j\Omega_j}{(-1)^{j-1}\Delta_j}\right) .$$

The conclusion follows from Lemma 2, showing that the product $\prod_{j=0}^{\tau-1} \frac{\sqrt{N}+(-1)^j\Omega_j}{(-1)^{j-1}\Delta_j}$, which has field norm one and is an element of the order $\mathfrak{O}_\mathbb{F}$, is actually the unit $\mathfrak{c}_{\tau-1}$ by equation (28). The connection between $\ln \mathfrak{c}_{\tau-1}$ and the regulator is motivated by Remark 1.
The equality $d(f_{\frac{\tau}{2}}, f_0) = \frac{1}{2}d(f_\tau, f_0)$ is an immediate consequence of the symmetry of the sequence $f_m(x,y)$ within a period. □

Since Theorem 5 guarantees that, when τ is even, a factor of N is located in the positions $\frac{\tau-2}{2}+k\tau$ of the sequence Υ, Shanks' method allows us to find such a factor, if $\ln(\mathfrak{c}_{\tau-1})$, or an odd multiple of it, is exactly known. Now, a formula of Dirichlet's gives the product

$$h_\mathbb{F}R_\mathbb{F} = \frac{\sqrt{D}}{2}L(1,\chi) = -\sum_{n=1}^{\lfloor\frac{D-1}{2}\rfloor} \left(\frac{D}{n}\right) \ln\left(\sin\frac{n\pi}{D}\right) \tag{33}$$

where $R_\mathbb{F}$ is the regulator, $L(1,\chi)$ is a Dedekind L-function, $D = N$ if $N \equiv 1 \bmod 4$ or $D = 4N$ otherwise, and character χ is the Jacobi symbol in this case. If the product $h_\mathbb{F}R_\mathbb{F}$ is known exactly (computed), for example using equation (33), the distance from the beginning of the sequence where the quadratic form can be found $[1, 2\Omega_{\tau-1}, \Delta_{\tau-2}]$ is known. Since this distance is an integer multiple of the regulator, and our target is to find a quadratic form that is located in the middle of some period, then

1. if $h_{\mathbb{F}}$ is odd, a factor of N is found in the position at distance $\frac{h_{\mathbb{F}}R_{\mathbb{F}}}{2}$, or $3\frac{h_{\mathbb{F}}R_{\mathbb{F}}}{2}$, from the beginning;

2. If $h_{\mathbb{F}}$ is even, in a position at distance $\frac{h_{\mathbb{F}}R_{\mathbb{F}}}{2}$, or $3\frac{h_{\mathbb{F}}R_{\mathbb{F}}}{2}$ the quadratic form $[1, 2\Omega_{\tau-1}, \Delta_{\tau-2}]$ is found, (which reveals a posteriori that $h_{\mathbb{F}}$ is even); in this case, the procedure can be repeated with target the position at distance $\frac{h_{\mathbb{F}}R_{\mathbb{F}}}{4}$, or $3\frac{h_{\mathbb{F}}R_{\mathbb{F}}}{4}$; again, either a factor of N is found or $h_{\mathbb{F}}$ is found to be a multiple of 4. Clearly the process can be iterated ℓ times until $\frac{h_{\mathbb{F}}R_{\mathbb{F}}}{2^{\ell}}$ is an odd multiple of $R_{\mathbb{F}}$, and a factor of N is found.

When the factor m_1 of N is found, the second factor is $m_2 = \frac{N}{m_1}$, thus the procedure can be iterated to find all factors of N. Mimicking Shanks' infrastructure, giant steps are performed to get close to forms at distance $\frac{kR_{\mathbb{F}}}{2}$, or $3\frac{kR_{\mathbb{F}}}{2}$, for some $1 \leq k \leq h_{\mathbb{F}}$, then baby steps are performed to get the exact position.

5. Conclusions

It has been shown that the complexity of factoring a composite number $4N$ is upper bounded by the complexity of evaluating, at a certain degree of accuracy, the product $h_{\mathbb{F}}R_{\mathbb{F}}$, as defined by Dirichlet using the $L(1,\chi_N)$ function, and also that is not necessary to know $h_{\mathbb{F}}$ and $R_{\mathbb{F}}$ separately. The more precise the evaluation of the product $h_{\mathbb{F}}R_{\mathbb{F}}$, the less complex the factoring $2N$; if we are lucky, the complexity could be polynomial in N. It is an open problem to find which is the best compromise between the approximate evaluation of $h_{\mathbb{F}}R_{\mathbb{F}}$ and the computational complexity for obtaining such approximation. In this context, the following expression, taken from [3, p.262], may be useful for efficiently evaluating the product $h_{\mathbb{F}}R_{\mathbb{F}}$ as a function of N

$$h_{\mathbb{F}}R_{\mathbb{F}} = \frac{1}{2}\sum_{x\geq 1}\left(\frac{N}{x}\right)\left(\frac{\sqrt{N}}{x}\mathrm{erfc}\left(x\sqrt{\frac{\pi}{N}}\right) + E_1\left(\frac{\pi x^2}{N}\right)\right) \quad , \tag{34}$$

where the complementary error function $\mathrm{erfc}(x)$, and the exponential integral function $E_1(x)$, can be closely approximated [18, p.297-299]

$$\mathrm{erfc}(z) = \frac{2}{\sqrt{\pi}}\int_z^{\infty} e^{-t^2}dt = 1 - \mathrm{erf}(z) = 1 - \frac{2}{\sqrt{\pi}}\sum_{n=0}^{\infty}\frac{(-1)^n z^{2n+1}}{n!(2n+1)}$$

$$E_1(z) = \int_1^{\infty}\frac{e^{-tz}}{t}dt = -\gamma - ln(z) - \sum_{n=1}^{\infty}\frac{(-1)^n z^n}{n\cdot n!} \quad .$$

As a last observation, the arguably, a fast (how fast is open) algorithm for factoring is achievable by combining results of Dirichlet, Shanks, and the above observations, which were suggested by Legendre's finding that continued fractions permit the representation of primes as the sum of two squares explicitly computed.

Acknowledgement. The very useful and constructive suggestions of the unknown referee, who pointed out several misprints, incompleteness, and provided the Example 3 are gratefully acknowledged. I also thank Karan Khathuria and Simran Tinani, PhD students at the University of Zurich, for their careful reading of a preliminary version of the paper, and for pointing out several misprints, errors, and imprecisions.

References

[1] D.A. Buell, *Binary Quadratic Forms*, New York: Springer-Verlag, 1989.

[2] G.S. Carr, *Formulas and Theorems in Mathematics*, Chelsea: New York, 1970.

[3] H. Cohen, *A Course in Computational Algebraic Number Theory*, New York: Springer-Verlag, 1993.

[4] **H. Cohn**, *A dvanced in Number Theory*, New York: Dover, 1980.

[5] H. Davenport, *The Higher Arithmetic*, New York: Dover, 1983.

[6] M. Elia, Relative Densities of Ramified Primes in $\mathbb{Q}(\sqrt{pq})$, *International Mathematical Forum*, vol. 3, n.8, 2008, p.375-384.

[7] C.F. Gauss, *Disquisitiones Arithmeticae*, New York: Springer-Verlag, 1986.

[8] G.H. Hardy, E.M. Wright, *An Introduction to the Theory of Numbers*, Oxford: Clarendon Press, 1971.

[9] R. A. Horn, C. R. Johnson, *Matrix Analysis*, Cambridge: Cambridge Univ. Press, 1999.

[10] Hua Loo Keng, *Introduction to Number Theory*, New York: Springer, 1981.

[11] A-M. Legendre, *Essai sur la Théorie des Nombres*, Chez Courcier, Paris, 1808, reissued by Cambridge University Press, 2009.

[12] O. Perron, *Die Lehre von den Kettenbrüchen, BandI:Elementare Kettenbrüche*, Springer, 1977.

[13] H. Riesel, *Prime Numbers and Computer Methods for Factorization*, Boston: Birkhäuser, 1984.

[14] R. Schoof, Quadratic Fields and Factorization, *Computational methods in number theory*, Mathematical Centre Tracts 154, Amsterdam, (1982), p.235-286. (p 129 of .pdf file)

[15] D. Shanks, The infrastructure of a real quadratic field and its applications, *Proc. 1972 Number Theory Conference*, Boulder (1972), p.217-224. New York, 1985.

[16] W. Sierpinski, *Elementary Theory of Numbers*, North Holland, New York, 1988.

[17] J. Steuding, *Diophantine Analysis*, Chapman & Hall, New York, 2003.

[18] M. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions*, New York: Dover, 1968.

AMS Subject Classification: 11A55, 11A51

M. Elia

Lavoro pervenuto in redazione il 19.10.2019.

p mod 8	q mod 8	Split?	$(p\|q)$	$\Delta_{\tau/2-1}$	T mod 4
3	3	Yes	± 1	$-(p\|q)p$	$1+(p\|q)$
3	7	Yes	± 1	$-(p\|q)p$	$1+(p\|q)$
7	3	Yes	± 1	$-(p\|q)p$	$1+(p\|q)$
7	7	Yes	± 1	$-(p\|q)p$	$1+(p\|q)$
5	3	Yes	1	p	0
3	5	Yes	1	$-p$	2
5	3	Yes	-1	$2p$	0
3	5	Yes	-1	$-2p$	2
5	7	Yes	1	p	0
7	5	Yes	1	$-p$	2
5	7	Yes	-1	$-2p$	2
7	5	Yes	-1	$2p$	0
1	3	No	-1	-2	2
1	3	Yes	1	p	AND 0
1	3	No/Yes	1	$-2,-2p$	2
3	1	No	-1	-2	2
3	1	Yes	1	$2p$	AND 0
3	1	No/Yes	1	$-2,-p$	2
7	1	No	-1	2	0
7	1	No	1	2	AND 0
7	1	Yes	1	$-p,-2p$	2
1	7	No	-1	2	0
1	7	No/Yes	1	$2,p,2p$	0
5	1	No	-1		1,3
5	1	No	1		AND 1,3
5	1	Yes	1	$-p$	AND 2
5	1	Yes	1	p	AND 0
1	5	No	-1		1,3
1	5	No	1		AND 1,3
1	5	Yes	1	$-p$	AND 2
1	5	Yes	1	p	AND 0
5	5	No	-1		1,3
5	5	No	1		AND 1,3
5	5	Yes	1	$-p$	AND 2
5	5	Yes	1	p	AND 0
1	1	No	-1		1,3
1	1	No	1		AND 1,3
1	1	Yes	1	$-p$	AND 2
1	1	Yes	1	p	AND 0

Table 7.1: $p < q$

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 103 – 124

E. Tron

THE GREATEST COMMON DIVISOR OF LINEAR RECURRENCES

Abstract. We survey the existing theory on the greatest common divisor $\gcd(u_n, v_n)$ of two linear recurrence sequences $(u_n)_n$ and $(v_n)_n$, with focus on recent development in the case where one of the two sequences is polynomial.

1. The problem

A *linear recurrence sequence* (or just *linear recurrence* for short) is a sequence $(u_n)_{n\in\mathbb{N}}$ specified by giving values $u_0, \dots, u_{d-1}$ and the condition that, for $n \geq d$,

$$u_{n+1} = \sum_{i=1}^{d} a_i u_{n+1-i}$$

for fixed $a_1, \dots, a_d$ and $a_d \neq 0$; the integer d is taken to be the least one for which a linear relation of this form holds and is called the *order* of the recurrence. All our recurrences will be assumed for simplicity to have rational integer terms, although the reader should keep in mind that much of what we are going to state holds with little to no change when they are instead defined over the ring of integers of a number field. The characteristic polynomial of the recurrence is $P(X) := X^d - \sum_{i=1}^{d} a_i X^{d-i}$ and its discriminant is $\Delta_u := \Delta(P)$: accordingly, the recurrence is called *simple* if the distinct roots of P (which are also referred to as the roots of u), say $\alpha_1, \dots, \alpha_r \in \mathbb{C}^\times$, are simple, and *non-degenerate* if no ratio α_i/α_j of any two distinct roots of P is a root of unity. Any term of the sequence can be expressed as a generalized power sum

$$u_n = \sum_{i=1}^{r} Q_i(n)\alpha_i^n$$

where the Q_i are polynomials over $\mathbb{C}$ whose degree is less than the multiplicity of α_i. The basic theory of linear recurrences will be assumed throughout, and we shall not develop it here but instead point to a general reference work such as the one of Everest–van der Poorten–Shparlinski–Ward [21] for further detail.

The problem that we are interested in is as follows. Given are two linear recurrences $(u_n)_n$ and $(v_n)_n$; what can one say about the quantity

$$g_n := \gcd(u_n, v_n)?$$

This can be thought as measuring the "arithmetical proximity" of u and v, as the G.C.D. puts together, for all non-archimedean places, how much the sequences share

(termwise) at each place. We will be interested in the distribution of the values of such a sequence, for instance in the counting function

$$G(x,y) := \#\{n \le x : g_n \ge y\};$$

this is slightly more convenient than the version with reversed inequality sign since one expects g_n to be often relatively small.

The plan is as follows. In Section 2 we shall see how one can bound large values of g_n when u and v are simple, by use of Schmidt's Subspace Theorem, and then interpret those bounds as cases of Vojta's conjecture. In Section 3 we will on the other hand see that if one recurrence is fully non-simple (one root with maximal multiplicity), almost everything concerning large and small values and averages of g_n can be determined. In Section 4, we will hint at how to translate the statements when studying other objects, such as elliptic divisibility sequences and meromorphic functions. We shall adopt an expository layout, with a focus on results over proofs.

We shall suppose, in each section, that the recurrence u (or the recurrences u and v) is fixed once and for all, so that all Vinogradov symbols depend on u in addition to other parameters: hence, read O_u, o_u, $\ll_u$, C_u (or $O_{u,v}$ etc.) for O, o, $\ll$, C respectively, which is the same as saying $O_{d,a_1,\dots,a_d}$ etc. The same is understood to hold for the objects that are meant to stand in place of linear recurrences in Sections 2 and 4.

2. The case with both recurrences non-degenerate

Throughout this section, we assume that the recurrences u and v are simple, and that their roots generate together a torsion-free multiplicative group (in particular, u and v are non-degenerate). This assumption is convenient in that it simplifies the statements of the theorems in the next sub-section, and does not entail a loss of generality [13, Sect. 1].

2.1. The Subspace Theorem and S-units

We first, and mostly, examine large values of g_n. For instance, what can one say on the cases when it is as large as possible, that is equal to $\min(|u_n|,|v_n|)$? The answer is given by the classical Hadamard Quotient Theorem.

THEOREM 1 (Pourchet [60], van der Poorten [59]). *Suppose that v_n divides u_n for all n.*[*] *Then $(u_n/v_n)_n$ is a linear recurrence.*

We may also rephrase the conclusion by saying that v has to divide u in the ring of linear recurrences.

Spectacular progress on the problem came next from exploiting the Subspace Theorem of Schmidt (as generalized by Schlickewei, Evertse, ...) in an ingenious

[*] Except possibly for those n for which $v_n = 0$, but there is a finite number of them—cf. the Skolem–Mahler–Lech theorem.

way; for the theorem itself the reader may see for instance Schmidt [68], or Bilu [6] for applications. The generalization that is used most often in applications involves all places of a number field and is due to Schlickewei.

THEOREM 2. *Suppose that K is a number field and S a finite set of places containing the Archimedean ones, $n \geq 1$ an integer. For each $\nu \in S$, let $L_{\nu,1}, \ldots, L_{\nu,n}$ be linearly independent linear forms in n variables defined over K. Then, for every fixed $\varepsilon > 0$, the nonzero solutions of*

$$\prod_{\nu \in S} \prod_{i=1}^{n} |L_{\nu,i}(x)|_\nu < H(x)^{-\varepsilon},$$

with $x \in O_K^n$, lie in a finite union of proper subspaces of K^n.

First, a powerful improvement to the Hadamard Quotient Theorem was proved by Corvaja–Zannier [13]. If we only assume that the divisibility occurs for infinitely many n, then the quotient might not be a linear recurrence anymore, but it is almost so. The result is also remarkable for not requiring the so-called "dominant root condition", which had plagued many applications thus far.

THEOREM 3 (Corvaja–Zannier [13, Th. 1]). *Suppose that v_n divides u_n for infinitely many n. Then there is a polynomial $P(X) \in \mathbb{C}[X]$ such that both sequences $(P(n)u_n/v_n)_n$ and $(v_n/P(n))_n$ are linear recurrences.*

In quantitative form, they also prove that if $(u_n/v_n)_n$ is not a linear recurrence, then u_n/v_n can be an integer only for $o(x)$ values of $n \leq x$. This was made precise by Sanna [63], improving on a remark in Corvaja–Zannier [13, Cor. 2].

THEOREM 4 (Sanna [63, Th. 1.5], Corvaja–Zannier [13, Sect. 4]). *If $(u_n/v_n)_n$ is not a linear recurrence, then u_n/v_n can be an integer only for*

$$x \left(\frac{\log\log x}{\log x} \right)^C$$

values of $n \leq x$, for some explicit positive integer C. This is best possible up to a power of $\log\log x$.

The G.C.D. bounds were made quantitatively explicit in a series of works whose heart were more complex applications of the Subspace Theorem. We now consider sequences of the form $a^n - 1$ for simplicity. First, if $a = c^r$ and $b = c^s$, then the $\gcd(a^n - 1, b^n - 1)$ is as large as a power of $\min(a^n - 1, b^n - 1)$ for trivial reasons; we exclude this case by saying that a and b are multiplicatively independent. Apart from this case, the greatest common divisor is always smaller than any fixed power of the smallest of the two sequences.

THEOREM 5 (Bugeaud–Corvaja–Zannier [8, Th. 1]). *Let $a, b \geq 2$ be multiplicatively independent integers. Then for $n > n(\varepsilon)$,*

$$\gcd(a^n - 1, b^n - 1) < \exp(\varepsilon n).$$

If b is not a power of a, then $\gcd(a^n - 1, b^n - 1) \ll a^{n/2}$ for large n.

This is close to best possible. Bugeaud–Corvaja–Zannier [8, Rem. 2] observe, after Adleman–Pomerance-Rumely [1, Prop. 10], that there are infinitely many n's that achieve $\exp(n^{c/\log\log n})$ (though they do not make a conjecture for the true maximal order; for instance, could it be $\exp(n^{(1+o(1))\log\log\log n/\log\log n})$?).

The start of the proof is as follows. For a positive integer i, write

$$z_i(n) := \frac{b^{in} - 1}{a^n - 1} = \frac{c_{i,n}}{d_n}$$

where $c_{i,n}$, d_n are integers, and d_n is taken as the denominator of $z_1(n)$.

Observe that for a fixed integer m we have the approximation

$$\frac{1}{a^n - 1} = a^{-n}\frac{1}{1 - a^{-n}} = a^{-n}\sum_{r=0}^{\infty} a^{-rn} = \sum_{r=1}^{m} \frac{1}{a^{rn}} + O(a^{-(m+1)n}).$$

If we multiply this by $b^{in} - 1$ we get

$$\left| z_i(n) + \sum_{s=1}^{m} \frac{1}{a^{sn}} - \sum_{r=1}^{m} \left(\frac{b^i}{a^r}\right)^n \right| = O(b^{in} a^{-(m+1)n});$$

the key idea is to see the left-hand side of this as a linear form in the variables $z_i(n)$, b^{in}/a^{rn}, a^{-sn}, for various values of i: if it were the case that $d_n \leq a^{(1-\varepsilon)n}$ infinitely often, then such forms would be small too often and contradict Theorem 2.

Corvaja–Rudnick–Zannier [12] prove a matrix generalization of this in the setting of periods of toral automorphisms. If B is a square matrix over $\mathbb{Z}$, we write $\gcd(B)$ for the greatest common divisor of the entries of B.

THEOREM 6 (Corvaja–Rudnick–Zannier [12, Th. 2]). *Suppose that $\varepsilon > 0$ is fixed and A is a square matrix of rational integers. Under some conditions on the eigenvalues of A, we have*

$$\gcd(A^n - I) < \exp(\varepsilon n)$$

for all large n.

The Bugeaud–Corvaja–Zannier bound is recovered as a special case of this, for the diagonal matrix $A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$.

Fuchs [22], building on the work of Bugeaud–Corvaja–Zannier [8] and Hernández–Luca [35], further generalized the theorem as follows.

THEOREM 7 (Fuchs [22, Thms. 1 and 2]). *Suppose that u, v have only positive real roots, and that v does not divide u in the ring of linear recurrences. Then there is an explicit constant $C < 1$ such that for large n*

$$\gcd(u_n, v_n) < \min(|u_n|, |v_n|)^C.$$

If moreover all the roots of u, v are integer, u_n is of the form $ba^n + c$, and a is coprime to all the roots of v, then for sufficiently large n this can be strengthened to

$$\gcd(u_n, v_n) < \min(|u_n|, |v_n|)^\varepsilon.$$

A further generalization by Levin [43] concerns greatest common divisors of terms with distinct indices; we give a simplified version for the sake of exposition.

THEOREM 8 (Levin [43, Th. 1.11]). *Suppose that u, v are simple linear recurrences such that for each place ν of $\mathbb{Q}$ at least one of the roots α of u or v has $|\alpha|_\nu \geq 1$. If the inequality*

$$\gcd(u_n, v_m) < \exp(\varepsilon \max(m, n))$$

has infinitely many solutions (m, n), then all but finitely many of those solutions satisfy one of finitely many linear relations $(m, n) = (a_i k + b_i, c_i k + d_i)$ $(1 \leq i \leq t)$, where the linear recurrences $(u_{a_i n + b_i})_n$ and $(v_{c_i n + d_i})_n$ have a nontrivial common factor in the ring of linear recurrences for all i.

Another direction for generalizations starts from the observation that a^n is an S-unit for a finite S, so that theorems on terms of linear recurrences really are at their heart theorems concerning sums of S-units. Hence the following:

THEOREM 9 (Corvaja–Zannier [14, Th.], Hernández–Luca [35]). *Let $S \supseteq \{\infty\}$ be a finite set of rational primes and $\varepsilon > 0$ fixed. Then for all but finitely many multiplicatively independent S-units u, v we have*

$$\gcd(u - 1, v - 1) < \max(|u|, |v|)^\varepsilon.$$

Corvaja–Zannier also give further generalizations of these to $\gcd(F(u, v), G(u, v))$ [15] and versions in positive characteristic [17].

Further still, one can obtain bounds where u, v are just assumed to be "near" S-units–for instance, of the form $F(n)a^n$. This is the case in the following.

THEOREM 10 (Luca [46, Cor. 3.3]). *Let a, b be positive integers, and F_1, F_2, G_1, G_2 non-zero polynomials with integer coefficients, $\varepsilon > 0$ fixed. Then for all large m, n we have*

$$\gcd(F_1(n)a^n + G_1(n), F_2(n)b^n + G_2(n)) < \exp(\varepsilon n).$$

Grieve–Wang [31] combine the ideas of Levin and Luca to obtain a very general upper bound in the case of non-simple recurrences, by means of the moving form of the Subspace Theorem.

For more applications of the Subspace Theorem to linear recurrences we refer to Fuchs [23] and Corvaja–Zannier [18].

2.2. More over function fields

The problem of small values of $\gcd(u_n, v_n)$ is more obscure. Indeed, the following conjecture is open (and probably very difficult).

THEOREM 11 (Ailon–Rudnick [2, Conj. A]). *If a and b are multiplicatively independent integers, then there are infinitely many n for which*

$$\gcd(a^n - 1, b^n - 1) = \gcd(a - 1, b - 1).$$

Evidence for this was given by Silverman [75].

To avoid the obstacle, people have thus been trying to study what happens when a, b belong to other rings. The outcome can turn out to be far more satisfying.

THEOREM 12 (Ailon–Rudnick [2, Th. 1]). *If $F, G \in \mathbb{C}[x]$ are non-constant multiplicatively independent polynomials, then there is a polynomial $H \in \mathbb{C}[X]$ such that for any n*

$$\gcd(F^n - 1, G^n - 1) \text{ divides } H.$$

In particular, $\deg \gcd(F^n - 1, G^n - 1) \leq C_{F,G}$.

The idea of Ailon and Rudnick is very simple but relies crucially on a deep theorem of Ihara–Serre–Tate, which states that an irreducible curve in $\mathbb{C}^\times \times \mathbb{C}^\times$ can only contain finitely many points both of whose coordinates are roots of unity, unless it is defined by an equation of the form $X^m Y^n - \zeta = 0$ or $X^m - \zeta Y^n = 0$ with ζ a root of unity [84, Ch. 1.1]. Applying this to the curve $\{(F(t), G(t)) : t \in \mathbb{C}\}$ we find that $F(z)$ and $G(z)$ are simultaneously roots of unity for finitely many $z \in \mathbb{C}$.

Now, for any root t of $\gcd(F^n - 1, G^n - 1)$, both $F(t)$ and $G(t)$ must simultaneously be roots of unity, so there are only finitely many possible roots t for $\gcd(F^n - 1, G^n - 1)$. Moreover, since $F^n - 1 = \prod_{i=1}^{n-1}(F - \zeta_n^i)$ and the factors on the right-hand side are pairwise coprime, any $X - t$ can divide at most one of them with multiplicity at most $\deg F$, and the same for G. Hence, we may take $H(X) = \prod (X - t)^{\min(\deg F, \deg G)}$.

THEOREM 13 (Silverman [70, Th. 4]). *If $P, Q \in \mathbb{F}_q[x]$ are non-constant monic, then*

$$\deg \gcd(P^n - 1, Q^n - 1) \geq C_{P,Q} n.$$

for infinitely many n.[†]

Denis [20, Th. 1.1] gives lower bounds for the number of integers n for which $\deg \gcd(P^n - 1, Q^n - 1)$ is on the other hand bounded, and studies the analogous problem on Drinfeld modules. Cohen–Sonn generalize Silverman's theorem to the quantity $\gcd(\Phi_m(a^n), \Phi_m(b^n))$ with $(\Phi_m)_m$ the classical cyclotomic polynomials [10, Th. 2.1].

[†]Notice thus the trichotomy $\mathbb{Z}$–$\mathbb{C}[X]$–$\mathbb{F}_q[X]$ in the results, with profoundly different kinds of bounds in each case.

Corvaja–Zannier [17] give more generally estimates on the $\gcd(u-1,v-1)$ when u,v belong to a function field of positive characteristic, and derive a bound of Weil type from this.

For more applications, and an extensive development of such concepts in the area of unlikely intersections, see Zannier [84, Ch. 2]. We just mention in passing a nice application of these bounds due to Luca–Shparlinski [47], who exploit them to show that the groups $E(\mathbb{F}_{q^n})$ ($E/\mathbb{F}_q$ an ordinary elliptic curve) have a large cyclic factor; and a follow-up by Magagna [50, Th. 5] proving $\gcd(\#E_1(\mathbb{F}_{q^n}),\#E_2(\mathbb{F}_{q^n})) < \exp(\varepsilon n)$ if E,E' are ordinary and non-isogenous.

2.3. The geometric approach and Vojta's conjecture

We come back to the results of Section 2 to put them in a different light. The connection between G.C.D. bounds and Vojta's conjecture that we are going to see was first noticed by Silverman [72]. We recall here the statement of the conjecture in a form that suits applications; here we take the ambient variety X as fixed and fix a choice of height functions as well.

CONJECTURE 2 (Vojta). Let X/k be a smooth projective variety over a number field k and S a finite set of places of k, K_X a canonical divisor, A an ample divisor, D a divisor with normal crossings. Then for any $\varepsilon > 0$ there is a proper Zariski closed subset Z of X and a constant C such that, for all $P \in X(k)\setminus Z$, it holds that

$$\sum_{v\in S} \lambda_{D,v}(P) + h_{K_X}(P) \le \varepsilon h_A(P) + C.$$

This conjecture is very general and encompasses many open problems in Diophantine geometry. For our needs, the point is that *G.C.D. bounds and are essentially equivalent to cases of Vojta's conjecture*. We immediately state an instance of this.

THEOREM 14 (Silverman [72, Th. 1]). *We let $|x|'_S$ be the prime-to-S part of x, i.e. the largest divisor of x that is not divisible by any prime in S. Let S be a finite set of places, $F_1,\dots,F_t \in \mathbb{Z}[X_1,\dots,X_n]$ homogeneous polynomials such that their zero set V is a smooth variety in $\mathbb{P}^n$ which does not intersect any hyperplane $\{X_i = 0\}$; let $r := n - \dim V$. Assume Vojta's conjecture for $\mathbb{P}^n$ blown up along V and fix $\varepsilon > 0$. Then there is a homogeneous $G \in \mathbb{Z}[X_1,\dots,X_n]$ and a constant $\delta > 0$ such that for any $n+1$-tuple of coprime integers $x_0,\dots,x_n \in \mathbb{Z}$ either $G(x_0,\dots,x_n) = 0$ or*

$$\gcd(F_1(x_0,\dots,x_n),\dots,F_t(x_0,\dots,x_n)) \le \max(|x_0|,\dots,|x_n|)^{\varepsilon}(|x_0\cdots x_n|'_S)^{1/(r-1+\delta\varepsilon)}.$$

If we apply for instance this with $n=2$, $F_1 = X_1 - X_0$, $F_2 = X_2 - X_0$, this theorem says that outside a one-dimensional set we have

$$\gcd(x_1 - x_0, x_2 - x_0) \le \max(|x_0|,|x_1|,|x_2|)^{\varepsilon}(|x_0x_1x_2|'_S)^{1/(1+\delta\varepsilon)}.$$

If we specialize further to $x_0 = 1$ and x_1, x_2 S-units, this becomes

$$\gcd(x_1 - 1, x_2 - 1) \le \max(|x_1|,|x_2|)^{\varepsilon}$$

and we recover Theorem 9 (up to the exceptional set, which is however not hard to determine). As for the proof of Theorem 14 itself, it involves Vojta's conjecture with $X=\mathbb{P}^n$, $A=\{X_0=0\}$, $D=-\pi^*K_X=-\pi^*\sum_{i=0}^n\{X_i=0\}$ and π the blow-up of X along $\{F_1=\cdots=F_t=0\}$.

Instead of explaining the general proof, let us just see where the analogy starts from [72, Sect. 2]. One writes for $a,b\in\mathbb{Q}$

$$\log\gcd(a,b)=\sum_p\min(\nu_p(a),\nu_p(b))\log p=\sum_{\nu\in M_{\mathbb{Q}}^0}\min(\nu(a),\nu(b));$$

for general a, b in a number field we then define

$$\log\gcd(a,b):=\sum_{\nu\in M_k}\min(\nu^+(a),\nu^+(b)).$$

To bring heights into play, we note that ν^+ is the local height function on $\mathbb{P}^1(k)$ with respect to the divisor (0). We would like a similar height-theoretic interpretation for the function $\min(\nu^+(\cdot),\nu^+(\cdot))$, but here $(0,0)$ is not a divisor on $(\mathbb{P}^1(k))^2$. To try and make things work we then blow up the plane at this point, and it turns out that the height with respect to the exceptional divisor on this blow-up is in fact the logarithmic G.C.D. In general, the G.C.D. is to be interpreted as a height function with respect to a closed subscheme, following the definitions laid out by Silverman [69]. Again, the analogy is rich and complex and we will not illustrate it, but point to the ultimate reference for this—the landmark article by Silverman [72].

This important analogy has thence been used to prove various cases of Vojta's conjecture for blow-ups by mutuating the techniques that successfully apply for G.C.D. problems, namely the Subspace Theorem. Levin [43] proves some cases on toric varieties; Wang–Yasufuki [81] on Cohen-Macaulay varieties; Yasufuki [82] on $\mathbb{P}^n$, and links it with the *abc* conjecture; Yasufuki again [83] on rational surfaces; Grieve [30] on Fano toric varieties.

3. The case with one recurrence fully non-simple

It has been realized in recent times [3] that the case where one of the sequences is instead fully degenerate, and in particular a polynomial sequence, the distribution problem for $g_n=\gcd(P(n),u_n)$ offers a more approachable toy version of the general problem. For the time being, we shall take one of the sequences to be the identity sequence and the other one to be a simple[‡] linear recurrence, and study $g_n=\gcd(n,u_n)$; the stronger results will then follow from the fact that here we have complete control over the places that divide one of the two recurrences.

Firstly, the case of u a first-order recurrence is easily settled. For instance, large and small values are immediate to estimate [3, Sect.1], and the observation that $\gcd(n,p^n)=p^{v_p(n)}$ implies the following asymptotic for the moments.

[‡]There is no loss of generality in assuming the simplicity of u [3, Sect. 1].

THEOREM 15. *As* $x \to \infty$,

$$\sum_{n\le x} (\log \gcd(n, p^n))^k = x\left(1-\frac{1}{p}\right)(\log p)^k \sum_{m=0}^{\infty} \frac{m^k}{p^m} + O\left(\left(\frac{\log x}{\log p}\right)^{k+1}\right).$$

Moreover, if $k \ge 1$,

$$\sum_{n\le x} \gcd(n, p^n)^k = x^{k(1+o_p(1))}$$

as $x \to \infty$.[§]

The expression with $u_n = a^n$ for composite a is, however, not nearly as nice. At any rate, we shall henceforth assume that the order of the recurrence u is greater than 1.

3.1. Large values of $\gcd(n, u_n)$

We first look at large values of $\gcd(n, u_n)$. Remember that u is always a simple linear recurrence of order at least 2, and that it is fixed once and for all without further mention of it in all Vinogradov symbols.

Studies on this quantity mostly involved the naïve formulation "when does n divide u_n"—in our perspective, this is asking for which n's the $\gcd(n, u_n)$ equals n, i.e. is as large as it can possibly get. The early works were partial characterizations, usually in terms of a (more or less explicit) recursive tree structure which is however unsuited to quantitative estimates. Credit for this is to be given here to Jarden [39], Hoggatt–Bergum [36], André-Jeannin [4], Somer [78], Smyth [77], and Győry–Smyth [33].

The first major work was that of Alba González–Luca–Pomerance–Shparlinski [3], where they obtained good bounds for various cases according to how nice the recurrence is.

THEOREM 16 (Alba González–Luca–Pomerance–Shparlinski [3, Th. 1.1]). *If* u *is non-degenerate, then as* $x \to \infty$

$$\#\{n \le x : n \text{ divides } u_n\} \ll \frac{x}{\log x}.$$

An ingredient of the proof is again the Subspace Theorem 2, or rather a consequence of it due to Schlickewei, to bound the number of zeros of the recurrence modulo p, hence number of solutions modulo p of an exponential equation [67].

This is essentially best possible: if we consider for instance the recurrence $u_n = 2^n - 2$, then p always divides u_p, and the composite n's for which n divides u_n are pseudoprimes and hence [58, Th. 2] much fewer than odd primes, so that in this case $\#\{n \le x : \gcd(n, u_n) = n\} = (1+o(1))x/\log x$.

[§] In fact the sum admits an asymptotic of the form $\left(x/p^{\psi_{p,k}(\log x/\log p)}\right)^k$, where ψ is a bounded periodic function with an explicit description as well; but we are not concerned here with such higher order terms.

THEOREM 17 (Alba González–Luca–Pomerance–Shparlinski [3, Th. 1.2]). *If the recurrence u is a non-degenerate Lucas sequence then as $x \to \infty$*

$$\#\{n \le x : n \text{ divides } u_n\} \le x \exp\left(-(1+o(1))\sqrt{\log x \log\log x}\right). \tag{1}$$

THEOREM 18 (Alba González–Luca–Pomerance–Shparlinski [3, Thms. 1.3 and 1.4]). *Suppose that u is a non-degenerate Lucas sequence with characteristic polynomial $X^2 - a_1X - a_2$.*

If $a_2 = \pm 1$ then as $x \to \infty$

$$\#\{n \le x : n \text{ divides } u_n\} \ge x^{1/4+o(1)}. \tag{2}$$

If $a_2 \ne \pm 1$ but $\Delta_u \ne \pm 1$ then, as $x \to \infty$

$$\#\{n \le x : n \text{ divides } u_n\} \ge \exp\left(C(\log\log x)^2\right).$$

In fact, to show (2) they use an explicit construction of integers of the form $2s\prod_{p\le x} p$ with s as follows: every one if its prime factors q is greater than x and such that q^2-1 is x-friable (has only prime factors smaller than x). If the factorization of integers of the form q^2-1 is statistically the same as a typical integer of their size, a lower bound $x^{1+o(1)}$ in (2) holds.

The next step was that of Luca and the author [49], who showed that the upper bound (1) can be vastly improved, and gave an explicit structure theorem for such integers. Their result was for Fibonacci numbers and was generalized by Sanna [62] to any Lucas sequence, using the appropriate formulae for the p-adic valuation of Lucas sequences [61]. From now on Lucas sequences will be understood to be non-degenerate as degenerate ones pose no problem [62, Sect. 2].

THEOREM 19 (Sanna [62, Th. 1.2], Luca–Tron [49, Th. 1]). *If the recurrence u is a Lucas sequence, then*

$$\#\{n \le x : n \text{ divides } u_n\} \le x \exp\left(-\left(\frac{1}{2}+o(1)\right)\frac{\log x \log\log\log x}{\log\log x}\right).$$

The $1/2+o(1)$ factor is just an artifact of the methods [27, Th. 3]. In fact, based on this and on analogies [58, Sect. 4] with Carmichael numbers via Korselt's criterion, Luca–Tron conjecture the following.

CONJECTURE 3 (Luca–Tron [49, Sect. 1]). If the recurrence u is a Lucas sequence, then

$$\#\{n \le x : n \text{ divides } u_n\} = x \exp\left(-(1+o(1))\frac{\log x \log\log\log x}{\log\log x}\right).$$

It should be noted that numerical evidence supporting this conjecture is relatively poor [58, Sect. 5], but there is a very precise and interesting reason why [29].

The "workhorse" here is a structure theorem for such integers n which reads as follows. We let $z_u(n)$ to be the least positive integer m for which n divides a term u_m of the sequence, whenever it is defined.

LEMMA 1 (Luca–Tron [49, Th. 2], Sanna [62, Lemma 3.3]). *For any fixed k let $\mathcal{R}_k := \{n \in \mathbb{N} : n/z_u(n) = k\}$. If n is in $\mathcal{R}_k$, then it is of the form $\gamma(k)m$, where m is a positive integer all of whose prime factors divide $6\Delta_u k$, and $\gamma(k)$ an integer depending only on k.*

In other words, if the ratio $n/z_u(n)$ is prescribed, every integer n is the product of a fixed integer times an S-integer with controlled S. This can be proved using explicit formulas for the p-adic valuation of u_n [61] and then, taking any n that belongs to $\mathcal{R}_k$, inspect for which n' the integer nn' also belongs to $\mathcal{R}_k$. This is of course no use without being able to estimate $\gamma(k)$, and the little miracle here is the existence of a very neat expression for it.

LEMMA 2 (Luca–Tron [49, Th. 2], Sanna [62, Lemma 3.3], Leonetti). *For any k, $\gamma(k)$ is the least element in $\mathcal{R}_k$ and we have*

$$\gamma(k) = k \operatorname*{lcm}_{m \geq 1} z^{\circ m}(k).$$

One can indeed see that this is well defined; once we know this expression we can notice that indeed $\gamma(k) \in \mathcal{R}_k$ almost by construction. This kind of expression might be telling for someone working in dynamical systems, but a satisfying dynamical interpretation is still lacking.

The work of Luca–Tron and Sanna does in fact prove an upper bound for the counting function when one instead asks for $\gcd(n, u_n) \geq \alpha n$ with $0 \leq \alpha \leq 1$ fixed (and thus a bound on $G(x,y)$ in the range $y \gg x$). With some more work, the methods would imply the following uniform bound.

CONJECTURE 4. If $0 \leq \alpha \leq 1$ is fixed, then

$$\#\{n \leq x : \gcd(n, u_n) \geq \alpha n\} \leq x \exp\left(-\left(\frac{1}{2} + o_\alpha(1)\right) \frac{\log x \log\log\log x}{\log\log x}\right).$$

The conjecture for the correct order of magnitude is still the same, that the $1/2 + o(1)$ on the right-hand side is actually an $1 + o(1)$. The key here is that Lemma 11, as well as its proof, adapts almost word by word when instead of $n = bz(n)$, $b \in \mathbb{N}$ a fixed integer, one asks for $n = \beta z(n)$, $\beta \in \mathbb{Q}$ a fixed rational number.

We end the section by considering the more general case when one of the recurrences is fully non-simple but of possibly higher order, i.e. the G.C.D. has the form $\gcd(F(n), u_n)$ with F a non-constant polynomial with integer coefficients. In this case, using sieve methods Alba González–Luca–Pomerance–Shparlinski prove a sligthly worse upper bound.

THEOREM 20 (Alba González–Luca–Pomerance–Shparlinski [3, Sect. 7]). *If*

the recurrence u has order $d \geq 2$ and F is as above, as $x \to \infty$ it holds that

$$\#\{n \leq x : F(n) \text{ divides } u_n\} \ll_F \frac{x \log\log x}{\log x}.$$

3.2. Small values of $\gcd(n, u_n)$

After studying when n divides u_n, next is the "dual" problem of when n is coprime to u_n. We retain the notation and hypotheses of the previous section.

The first basic theorem is due to Sanna [65] and proves, under very general assumptions, that those n have an asymptotic density.

THEOREM 21 (Sanna [65, Th. 1.1]). *If u is non-degenerate, the set of integers n such that $\gcd(n, u_n) = 1$ has an asymptotic density. Such a density is positive, unless $(u_n/n)_n$ is also a linear recurrence, in which case this set is in fact finite.*

Next came the work of Sanna and the author [66], where it was shown that not only this generalizes to any fixed value of the G.C.D., but also that another little miracle occurs: there is a very explicit expression for the asymptotic density. For notational convenience set $\ell_u(m) := \operatorname{lcm}(m, z_u(m))$.

THEOREM 22 (Sanna–Tron [66, Thms. 1.3 and 1.4]). *Let u be a non-degenerate Lucas sequence with characteristic polynomial $X^2 - a_1X - a_2$. For any $k \in \mathbb{N}$, let $\mathcal{A}_k$ be the set of integers n such that $\gcd(n, u_n) = k$. Then $\mathcal{A}_k$ has an asymptotic density which is given by the absolutely convergent series*

$$\sum_{\gcd(d,a_2)=1} \frac{\mu(d)}{\ell_u(dk)}.$$

Such a density is positive if and only if $\mathcal{A}_k$ is not empty if and only if $\gcd(k, a_2) = 1$ and $k = \gcd(\ell_u(k), u_{\ell_u(k)})$.

The last part vindicates a conjecture made in another setting by Silverman [73, Q. 1]. The statement is moderately far-reaching: for instance, the integers n such that $\gcd(n, 2^n - 1) = k$ have an asymptotic density given by $\sum_{n \text{ odd}} 1/\operatorname{lcm}(kn, \operatorname{ord}_{kn}(2))$. However, a way of proving *a priori* the criterion for such a sum to be zero or not, or even just showing its non-negativity, directly without going through the related arithmetical problem, is not known to exist.

The heart of the proof is also the apparently least interesting part, to show that the expression is well defined. We record it separately to emphasize it.

LEMMA 3. *The series*

$$\sum_{\gcd(d,a_2)=1} \frac{1}{\ell_u(d)}$$

converges absolutely.

If in place of $\ell_u(d) = \operatorname{lcm}(d, z_u(d))$ we just had $d\, z_u(d)$ things would be much easier: the convergence of the sum $\sum_d 1/dz_u(d)$ has been known at least since the work of Romanoff in the '30s [53].

Once we know this, the expression for the density in Theorem 22 is straightforward to derive; let us do the case $k = 1$ and $a_2 = 1$, so that z_u is defined on all integers. If we set $\rho(n,d)$ to be the indicator function of "$d|u_n$" then

$$\#\mathcal{A}_1(x) = \sum_{n\le x}\prod_{p|n}(1-\rho(n,p)) = \sum_{n\le x}\sum_{d|n}\mu(d)\rho(n,d) = \sum_{d\le x}\mu(d)\sum_{m\le x/d}\rho(dm,d);$$

now, $\rho(dm,d) = 1$ is equivalent to m being divisible by $\ell_u(d)/d$, so the latter quantity is

$$\sum_{d\le x}\mu(d)\sum_{m\le x/d}1 = \sum_{d\le x}\mu(d)\left\lfloor\frac{x}{\ell(d)}\right\rfloor = x\left(\sum_{d\le x}\frac{\mu(d)}{\ell(d)}\right) - \sum_{d\le x}\mu(d)\left\{\frac{x}{\ell(d)}\right\}.$$

All we need to do now is to use that $\sum_{d>x}\frac{\mu(d)}{\ell(d)}$ is the tail of a convergent series, and split the latter sum into large and small d (say, at a cutoff of $x^{1/2}$) to recover Theorem 22.

REMARK 1. In light of Theorem 22, the set of numbers k for which $\mathcal{A}_k$ is empty (or not) is itself of interest. Leonetti–Sanna [42] prove that there are at least $Cx/\log x$ and at most $o(x)$ integers k up to x for which $\mathcal{A}_k$ is not empty. Given that they only consider prime numbers in the lower bound, the true order of magnitude should be somewhat larger; are there, say, at least $x\log\log x/\log x$ such integers up to x?

Parallel to the previous sections, the problem with $\gcd(F(n),u_n)$ a fixed integer, where F is a polynomial with integer coefficients, has also been studied.

THEOREM 23 (Mastrostefano–Sanna [52, Th. 1.4]). *Suppose that F splits over $\mathbb{Q}$, and let k be a fixed integer. Then the set of integers n such that $\gcd(F(n),u_n) = k$ has an asymptotic density. If moreover u is non-degenerate and F does not have fixed divisors, then the set set of integers n such that $\gcd(F(n),u_n) = 1$ has zero asymptotic density if and only if it is finite.*

However, no nice expression for the density is presently known in cases other than $F(n) = n$.

3.3. Averages of $\gcd(n,u_n)$

The previous sections give quite satisfying answers to the problem of determining extreme values of $\gcd(n,u_n)$. If we inquire, however, about its average size, much less is known–let alone the distribution function $G(x,y)$ in general. We summarize here partial progress towards the solution.

If we allow for some more regular version of the G.C.D., say its logarithm $\log\gcd(n,u_n)$, the situation is already quite different.

THEOREM 24 (Sanna [64, Th. 1.1]). *Let u be a non-degenerate Lucas sequence. Then for any fixed positive integer k, as $x\to\infty$,*

$$\sum_{n\le x}(\log\gcd(n,u_n))^k = M_k x + O_k(x^{1-1/(3k+3)}).$$

Moreover, the constant M_k is given explicitly by an absolutely convergent series

$$M_k = \sum_{\gcd(d,a_2)=1}\frac{\rho_k(d)}{\ell_u(d)}$$

and ρ_k is a certain, explicitly defined, arithmetical function such that $\rho_k(m)\le(k\log m)^k$.

This implies directly a bound for the counting function.

COROLLARY 1 (Sanna [64, Cor. 1.3]).

$$G(x,y)\ll_{u,k}\frac{x}{(\log y)^k}.$$

The argument itself is not too different to what we have seen already in the previous section. Suppose for instance that $k=1$: we can write

$$\sum_{n\le x}\log\gcd(n,u_n)=\sum_{n\le x}\sum_{\ell_u(p^e)\mid n}\log p=\sum_{p^e}\log p\sum_{\substack{n\le x\\ \ell_u(p^e)\mid n}}1=\sum_{p^e}\log p\left\lfloor\frac{x}{\ell_u(p^e)}\right\rfloor$$

$$=:\sum_{\gcd(m,a_2)=1}\rho_1(m)\left\lfloor\frac{x}{\ell_u(m)}\right\rfloor=\left(\sum_{\gcd(m,a_2)=1}\frac{\rho_1(m)}{\ell_u(m)}\right)x-\sum_{\gcd(m,a_2)=1}\rho_1(m)\left\{\frac{x}{\ell_u(m)}\right\},$$

then argue as in Section 3.2; for larger k there is more combinatorial work involved, but again convergence of the relevant sum is the bulk of the proof.

Inspired by this work, Mastrostefano set out to find more on the moments themselves. Here is the upper bound that he obtained.

THEOREM 25 (Mastrostefano [51, Th. 1.3]). *Let u be a non-degenerate Lucas sequence. Then for any fixed positive integer k, as $x\to\infty$,*

$$\sum_{n\le x}\gcd(n,u_n)^k\le x^{k+1-(1+o_k(1))\sqrt{\log\log x/\log x}}.$$

The key to improving these estimates is the study of the tail of a series

$$\sum_{\substack{d>x\\ \gcd(d,a_2)=1}}\frac{1}{\ell_u(d)}:$$

Mastrostefano bounds it by $\exp\left(-(1/\sqrt{6}-\varepsilon+o_\varepsilon(1))\sqrt{\log x\log\log x}\right)$. We also get the following for the counting function.

COROLLARY 2 (Mastrostefano [51, Cor. 1.5]). *As $x\to\infty$,*

$$G(x,y)\le x^{2-(1+o(1))\sqrt{\log\log x/\log x}}/y.$$

The determination of the moments can be a subtle problem [51, Sect. 6]. However, it is not difficult to conjure up a simple heuristic: if we come back to numbers n such that $\gcd(n,u_n)=n$, there are conjecturally $x^{1-(1+o(1))\log\log\log x/\log\log x}$ of them up to x. If they were evenly spaced (which they are not, but they are at least well distributed) they would contribute at least

$$\sum_{n\le x/x^{(1+o(1))\log\log\log x/\log\log x}}\left(nx^{(1+o(1))\log\log\log x/\log\log x}\right)^k = x^{k+1-(1+o(1))\log\log\log x/\log\log x}$$

to the k-th moment. If we compound this with the *ansatz* that "most" of the mass of the moments comes from those n with large $\gcd(n,u_n)$—e.g. larger than βn, cf. Conjecture 4—we end up with the following conjecture.

CONJECTURE 5. If the recurrence u is a non-degenerate Lucas sequence, then as $x\to\infty$

$$\sum_{n\le x}\gcd(n,u_n)^k = x^{k+1-(1+o_k(1))\log\log\log x/\log\log x}.$$

As Mastrostefano kindly pointed out to me, this very argument, coupled with the input from Alba González–Luca–Pomerance–Shparlinski (cf. Theorem 18), immediately provides the following.

THEOREM 26. *If $a_2=\pm1$ then as $x\to\infty$*

$$\sum_{n\le x}\gcd(n,u_n)^k \ge x^{k+1/4+o_k(1)}.$$

It is maybe worth to point out the formal resemblance of Theorems 24 and 15 with work of Luca–Shparlinski [48, Th. 2]. They study sums of the form $\sum_{n\le x}f(u_n)^k$, where f is any arithmetic function satisfying certain stringent growth conditions, and they prove an estimate $M_{f,k}\,x+O_{f,k}(x(\log\log x)^k/\log x)$.

4. The problem in other settings

4.1. Elliptic divisibility sequences

The most straightforward adaptation of statements from Part 3 is in the setting of elliptic divisibility sequences—which by the way is an indicator that some properties have more to do with u_n being a divisibility sequence rather than a linear recurrence. We

recall that an elliptic divisibility sequence, call it u_n still, is defined by taking a non-torsion point $P \in E(\mathbb{Q})$ of an elliptic curve $E/\mathbb{Q}$ defined by a Weierstrass equation and then the reduced x-coordinates of its orbit $x_{[n]P} = v_n/u_n^2$.

The recursive structure theorems mentioned at the start of Section 3.1 have an elliptic version by Silverman–Stange [76]; the theorems for the distribution of $\gcd(n,u_n) = n$ are due to Gottschlich [28].

THEOREM 27 (Gottschlich [28, Th. 1.1]). *As $x \to \infty$, we have*

$$\#\{n \le x : n \text{ divides } u_n\} \ll_{E,P} x\,\frac{(\log\log x)^{5/3}(\log\log\log x)^{1/3}}{(\log x)^{4/3}}.$$

When E has complex multiplication, and for any E under the Lang–Trotter conjecture, he also obtains an upper bound

$$x\,\exp\left(-(1+o_{E,P}(1))\cdot\sqrt{\log x\log\log x/8}\right).$$

On the other hand, the analogy is even closer for the problem of $\gcd(n,u_n) = k$ constant. In this case, Kim [40] proved that a theorem *formally analogous* to Theorem 2.2 holds. Again, the delicate point is the convergence of the sum [40, App. A], while the proof itself is otherwise formally the same.

As an aside, we comment that the setting of elliptic curves gives a more transparent geometric interpretation which otherwise, in the case of linear recurrences, is to be found in the work of Cubre–Rouse [19] (after Lagarias [41]), solving a conjecture of Bruckman–Anderson [7] by means of the "torus trick" of Hasse–Ballot [5]. For a slightly different take on this, also see Silverman [72].

Finally, the Ailon–Rudnick theorem 1.2 as well is proved by Silverman for elliptic divisibility sequences over function fields (i.e. obtained from a curve $E/k(T)$) in case the j-invariant of the curve is k-rational [71, Th. 3]. Ghioca–Hsia–Tucker give a variant over any field of positive characteristic [25], Ostafe [56] for multivariate polynomials, Ghioca–Hsia–Tucker again [26] over elliptic curves, Ulmer–Urzúa [79] a result of similar flavor on unlikely intersections. Silverman [72] has a theorem analogous to Theorem 1.4 where a bound in the same form as Theorem 9 but for elliptic divisibility sequences is shown to be another consequence of Vojta's conjecture.

4.2. Nevanlinna theory

An extremely fruitful development in analogy with the greatest common divisors of recurrences is in Nevanlinna theory, where the quantities are replaced by their cousins in the setting of entire functions in the spirit of Vojta's celebrated dictionary between Nevanlinna theory and diophantine approximation [80]. Without developing the basics of Nevanlinna theory, we shall limit ourselves to mentioning the most relevant results.

The basic ideas involved in the correct analogy where introduced in the landmark work of Noguchi–Winkelmann–Yamanoi [55]. The article of Pastén–Wang [57]

is the most complete source of meromorphic counterparts to the arithmetic G.C.D. bounds, and we now introduce some of them.

For f a meromorphic function on $\mathbb{C}$ and $z \in \mathbb{C}$, we set $\nu_z^+(f) := \max(0, \mathrm{ord}_z(f))$ and $\nu_z^-(f) := -\min(0, \mathrm{ord}_z(f))$. We then define the characteristic function

$$T(f,r) := \frac{1}{2\pi}\int_0^{2\pi} \max(0, \log|f(re^{i\theta})|)d\theta + \sum_{0<|z|\le r} \nu_z^-(f)\log|r/z| + \nu_0^-(f)\log r.$$

The analogue for the G.C.D. is defined as follows: if

$$n(f,g,r) := \sum_{|z|\le r} \min(\nu_z^+(f), \nu_z^+(g)),$$

then the relevant counting function is

$$N(f,g,r) := \int_0^r \frac{n(f,g,t) - n(f,g,0)}{t}dt + n(f,g,0)\log r.$$

A sample of the many G.C.D. bounds that Pastén–Wang obtain in this setting are the following.

THEOREM 28 (Pastén–Wang [57, Th. 1.3]). *Let f, g be algebraically independent meromorphic functions and $\varepsilon > 0$. Then*

$$N(f^n - 1, g^n - 1, r) < \varepsilon \max(nT(f,r), nT(g,r))$$

for all r in a set of infinite Lebesgue measure.

THEOREM 29 (Pastén–Wang [57, Th. 1.5]). *Let f, g be multiplicatively independent entire functions without zeros, both of finite order, and $\varepsilon > 0$. Then for all large n, as $r \to \infty$ we have*

$$N(f^n - 1, g^n - 1, r) < \varepsilon \min(T(f^n, r), T(g^n, r)) + O(\log r).$$

They give many more theorems under various different hypotheses on the growth of the functions, and even general results for meromorphic functions over any complete algebraically closed field, so the reader is advised to read their introduction. For more on the general technical background, see Noguchi–Winkelmann [54].

This line of work spawned the following developments.

THEOREM 30 (Guo–Wang [32, Th. 1.1]). *Let f, g be algebraically independent meromorphic functions and $\varepsilon > 0$. Then for all large n, and for all r outside a set of finite Lebesgue measure,*

$$N(f^n - 1, g^n - 1, r) < (1/2 + \varepsilon)\max(T(f^n, r), T(g^n, r)).$$

THEOREM 31 (Levin–Wang [44, Cor. 1.6]). *Let f, g be multiplicatively independent meromorphic functions, and $\varepsilon > 0$. Then for all large n, as $r \to \infty$ (outside a set of finite Lebesgue measure), we have*

$$N(f^n - 1, g^n - 1, r) < \varepsilon \max(T(f^n, r), T(g^n, r)).$$

The Corvaja–Zannier version of the Hadamard Quotient Theorem has an analog for entire functions as well, due to Guo [34].

THEOREM 32 (Guo [34, Th. 1.2]). *Let $f_1,\dots,f_k,g_1,\dots,g_m$ be nonconstant entire functions such that $\max_i T(f_i,r) \asymp \max_j T(g_j,r)$ as $r\to\infty$. Set $F(n) := a_0 + a_1 f_1^n + \cdots + a_k f_k^n$, $G(n) := b_0 + b_1 g_1^n + \cdots + b_k g_k^n$ where the a_i and b_j are nonzero complex numbers. If $F(n)/G(n)$ is an entire function for infinitely many n, then the f_i, g_j are multiplicatively dependent (there is a product $f_1^{r_1}\cdots f_k^{r_k} g_1^{s_1}\cdots g_k^{s_k}$ which is a nonzero constant).*

For more work on G.C.D. bounds in Nevanlinna theory in the setting of holomorphic maps to semi-abelian varieties also see Liu–Yu [45]. Corvaja–Noguchi [11] prove another counterpart to the Corvaja–Zannier theorem [13].

4.3. Rational dynamical systems

Another domain of research which is rich in analogies with the problems that we have studied is that of rational dynamical systems [74], i.e. the study of the behavior of iterates of rational maps (which is itself linked to the domain of unlikely intersections [84, Ch. 3.4.7]). The links usually exploit Silverman's ideas in some way or another, and the powers of integers are replaced by n-fold iterates of polynomials.

Chen–Gassert–Stange [9] prove analogues of the structure theorems mentioned at the beginning of Section 3.1 and Gassert–Urbanski [24] study the divisibility by n of $F^{\circ n}(0)$, F a polynomial.

More interestingly, Hsia–Tucker [37] prove a "compositional" cousin to the Ailon–Rudnick theorem.

THEOREM 33 (Hsia–Tucker [37, Th. 4]). *Let $F,G\in\mathbb{C}[X]$ be compositionally independent polynomials, of degree greater than 1, and $C\in\mathbb{C}[X]$ another polynomial satisfying some extra conditions. Then there is a polynomial $H\in\mathbb{C}[X]$ such that, for all m,n,*

$$\gcd(F^{\circ m}-C, G^{\circ n}-C) \text{ divides } H.$$

A compositional analogue of the Bugeaud–Corvaja–Zannier bound is known as well; here, however, the substantial recourse to Silverman's method requires Vojta's conjecture in a form not yet proved in such generality. Assuming thus Vojta's conjecture, the theorem reads as follows.

THEOREM 34 (Huang [38, Th. A]). *Let $F,G\in\mathbb{Z}[X]$ be polynomials of the same degree $d=\deg F=\deg G\geq 2$, and $a,b,\alpha,\beta\in\mathbb{Z}$ integers. Under some genericity assumption, there is a constant $C>0$ such that for all n*

$$\gcd(F^{\circ n}(a)-\alpha, G^{\circ n}(b)-\beta) \leq C\exp(\varepsilon d^n).$$

In fact he proves more general versions for rational maps and also gives more in-depth characterizations in case the genericity assumption is not satisfied.

Acknowledgements

I am thankful to Yuri Bilu, Francesco Campagna, Pietro Corvaja, Luca Ghidelli, Paolo Leonetti, Daniele Mastrostefano, Carlo Sanna, Joe Silverman, Umberto Zannier, and the anonymous referee, for useful discussion and comments before and during the preparation of this work. I also thank the organizers of the *2nd Number Theory Meeting*, where my lecture constituted the early core of this survey.

References

[1] ADLEMAN L.M., POMERANCE C., AND RUMELY S., *On Distinguishing Prime Numbers from Composite Numbers*, Annals of Math. **117** 1 (1983), 173–206.

[2] AILON N. AND RUDNICK Z., *Torsion points on curves and common divisors of a^k-1 and b^k-1*, Acta Arith. **113** (2004), 31–38.

[3] ALBA GONZÁLEZ J.J., LUCA F., POMERANCE C., AND SHPARLINSKI I.E., *On numbers n dividing the nth term of a linear recurrence*, Proc. Edinb. Math. Soc. **55** 2 (2012), 271–289.

[4] ANDRÉ-JEANNIN R., *Divisibility of generalized Fibonacci and Lucas numbers by their subscripts*, Fibonacci Q. **29** 4 (1991), 364–366.

[5] BALLOT C., *Density of Prime Divisors of Linear Recurrences*, Mem. Am. Math. Soc. **551**, AMS, Providence 2005.

[6] BILU YU., *The Many Faces of the Subspace Theorem (after Adamczewski, Bugeaud, Corvaja, Zannier...)*, Séminaire Bourbaki n° 967, Astérisque **317** (2008), 1–38.

[7] BRUCKMAN P.S. AND ANDERSON P.G., *Conjectures on the Z-densities of the Fibonacci sequence*, Fibonacci Q. **36** 3 (1998), 263–271.

[8] BUGEAUD Y., CORVAJA P., AND ZANNIER U., *An upper bound for the G.C.D. of a^n-1 and b^n-1*, Math. Z. **243** (2003), 79–84.

[9] CHEN A.S., GASSERT T.A., AND STANGE K.E., *Index divisibility in dynamical sequences and cyclic orbits modulo p*, New York J. Math. **23** (2017), 1045–1063.

[10] COHEN J. AND SONN J., *A cyclotomic generalization of the sequence* $\gcd(a^n-1,b^n-1)$, J. Théor. Nombres Bordx. **27** 1 (2015), 53–65.

[11] CORVAJA P. AND NOGUCHI J., *A new unicity theorem and Erdős problem for polarized semi-abelian varieties*, Math. Ann. **353** 2 (2012), 439–464.

[12] CORVAJA P., RUDNICK Z., AND ZANNIER U., *A Lower Bound for Periods of Matrices*, Commun. Math. Phys. **252** (2004), 535–541.

[13] CORVAJA P. AND ZANNIER U., *Finiteness of integral values for the ratio of two linear recurrences*, Invent. Math. **149** 2 (2002), 431–451.

[14] CORVAJA P. AND ZANNIER U., *On the greatest prime factor of* $(ab+1)(ac+1)$, Proc. Am. Math. Soc. **131** 6 (2003), 1705–1709 .

[15] CORVAJA P. AND ZANNIER U., *A Lower Bound for the Height of a Rational Function at S-unit Points*, Monatsh. Math. **144** 3 (2005), 203–224.

[16] CORVAJA P. AND ZANNIER U., *Some cases of Vojtas Conjecture on integral points over function fields*, J. Algebr. Geom. **17** (2008), 295–333. Addendum: Asian J. Math. **14** (2010), 581–584.

[17] CORVAJA P. AND ZANNIER U., *Greatest common divisors of $u-1$, $v-1$ in positive characteristic and rational points on curves over finite fields*, JEMS **15** 5 (2013), 1927–1942.

[18] CORVAJA P. AND ZANNIER U., *Applications of Diophantine Approximation to Integral Points and Transcendence*, Camb. Tracts Math. **212**, Cambridge University Press, Cambridge 2018.

[19] CUBRE P. AND ROUSE J., *Divisibility properties of the Fibonacci entry point*, Proc. Am. Math. Soc. **142** 3 (2014), 3771–3785.

[20] DENIS L., *Facteurs communs et torsion en caractéristique non nulle* , J. Théor. Nombres Bordx. **23** 2 (2011), 347–352.

[21] EVEREST G., VAN DER POORTEN A., SHPARLINSKI I., AND WARD T., *Recurrence Sequences*, Math. Surv. Monogr. **104**, AMS, Providence 2003.

[22] FUCHS C., *An upper bound for the G.C.D. of two linear recurring sequences*, Math. Slovaca **53** 1 (2003), 21–42.

[23] FUCHS C., *Diophantine problems with linear recurrences via the Subspace Theorem*, Integers **5** 3 (2005), A08.

[24] GASSERT T.A. AND URBANSKI M.T., *Index divisibility in the orbit of 0 for integral polynomials*, `arXiv:1709.08751 [math.NT]` (2017).

[25] GHIOCA D., HSIA L.-C., AND TUCKER T.J., *On a variant of the AilonRudnick theorem in finite characteristic*, New York J. Math. **23** (2017), 213–225.

[26] GHIOCA D., HSIA L.-C., AND TUCKER T.J., *A variant of a theorem by AilonRudnick for elliptic curves*, Pac. J. Math. **295** 1 (2018), 1–15.

[27] GORDON D.M. AND POMERANCE C., *The distribution of Lucas and elliptic pseudoprimes*, Math. Comput. **57** (1991), 825–838.

[28] GOTTSCHLICH A., *On positive integers n dividing the nth term of an elliptic divisibility sequence*, New York J. Math. **18** (2012), 409–420.

[29] GRANVILLE A. AND POMERANCE C., *Two contradictory conjectures concerning Carmichael numbers*, Math. Comput. **71** (2001), 883–908.

[30] GRIEVE N., *Generalized GCD for toric Fano varieties*, `arXiv:1904.13188 [math.AG]` (2019).

[31] GRIEVE N. AND WANG J.T.-Y., *Greatest common divisors with moving targets and linear recurrence sequences*, `arXiv:1902.09109 [math.NT]` (2019).

[32] GUO J. AND WANG J.T.-Y., *Asymptotic gcd and divisible sequences for entire functions*, Trans. Am. Math. Soc. **37** 9 (2019), 6241–6256.

[33] GYŐRY K. AND SMYTH C., *The divisibility of $a^n - b^n$ by powers of n*, Integers **10** (2010), A27/319–334.

[34] GUO J., *The Quotient Problem for Entire Functions*, Can. Math. Bull. **62** 3 (2019), 479–489.

[35] HERNÁNDEZ S. AND LUCA F., *On the largest prime factor of $(ab+1)(ac+1)(bc+1)$*, Bol. Soc. Mat. Mex., III. Ser. **9** 2 (2003), 235–244 .

[36] HOGGATT V.E. JR. AND BERGUM G.E., *Divisibility and Congruence Relations*, Fibonacci Q. **12** 2 (1974), 189–195.

[37] HSIA L.C. AND TUCKER T.J., *Greatest common divisors of iterates of polynomials*, Algebra Number Theory **11** 6 (2017), 1437–1459.

[38] HUANG K., *Generalized Greatest Common Divisors for the Orbits under Rational Functions*, `arXiv:1702.03881 [math.NT]` (2017).

[39] JARDEN D., *Divisibility of terms by subscripts in Fibonaccis sequence and associate sequence*, Riveon Lematematika **13** (1959), 51–56.

[40] KIM S., *The density of the terms in an elliptic divisibility sequence having a fixed G.C.D. with their indices*, J. Number Theory, to appear. Appendix by M. R. Murty.

[41] LAGARIAS J.C., *The set of primes dividing the Lucas numbers has density 2/3*, Pac. J. Math. **118** 2 (1985), 449–461. Errata: Pac. J. Math. **162** 2 (1994), 393–396.

[42] LEONETTI P. AND SANNA C., *On the greatest common divisor of n and the nth Fibonacci number*, Rocky Mt. J. Math. **48** 4 (2018), 1191–1199.

[43] LEVIN A., *Greatest common divisors and Vojta's conjecture for blowups of algebraic tori*, Invent. Math. **215** 2 (2019), 493–533.

[44] LEVIN A. AND WANG J.T.-Y., *Greatest common divisors of analytic functions and Nevanlinna theory on algebraic tori*, J. Reine Angew. Math., to appear.

[45] LIU X. AND YU G., *Upper Bounds of GCD Counting Function for Holomorphic Maps*, J. Geom. Anal. **29** 2 (2019), 1032–1042.

[46] LUCA F., *On the Greatest Common Divisor of $u-1$ and $v-1$ with u and v Near S-units*, Monatsh. Math. **146** 3 (2005), 239–256.

[47] LUCA F. AND SHPARLINSKI I.E., *On the exponent of the group of points on elliptic curves in extension fields*, Int. Math. Res. Not. **23** (2005), 1391–1409.

[48] LUCA F. AND SHPARLINSKI I.E., *Arithmetic functions with linear recurrence sequences*, J. Number Theory **125** (2007), 459–472.

[49] LUCA F. AND TRON E., *The Distribution of Self-Fibonacci Divisors*, in *Advances in the Theory of Numbers*, Fields Inst. Commun. **77**, 149–158, Springer, New York 2015.

[50] MAGAGNA C., *A lower bound for the r-order of a matrix modulo N*, Monatsh. Math. **153** 1 (2008), 59–81.

[51] MASTROSTEFANO D., *An upper bound for the moments of a GCD related to Lucas sequences*, Rocky Mt. J. Math. **49** 3 (2019), 887–902.

[52] MASTROSTEFANO D. AND SANNA C., *On numbers n with polynomial image coprime with the nth term of a linear recurrence*, Bull. Aust. Math. Soc. **99** 1 (2019), 23–33.

[53] MURTY M.R., ROSEN M., AND SILVERMAN J.H., *Variations on a theme of Romanoff*, Int. J. Math. **7** 3 (1996), 373–391.

[54] NOGUCHI J. AND WINKELMANN J., *Nevanlinna Theory in Several Complex Variables and Diophantine Approximation*, Grundlehren Math. Wiss. **350**, Springer, Berlin 2014.

[55] NOGUCHI J., WINKELMANN J., AND YAMANOI K., *The second main theorem for holomorphic curves into semi-Abelian varieties*, Acta Math. **188** 1 (2002), 129–161.

[56] OSTAFE A., *On some extensions of the AilonRudnick theorem*, Monatsh. Math. **181** 2 (2016), 451–471.

[57] PASTEN H. AND WANG J.T.-Y., *GCD Bounds for Analytic Functions*, Int. Math. Res. Not. **2017** 1 (2017), 47–95.

[58] POMERANCE C., *On the Distribution of Pseudoprimes*, Math. Comput. **37** 156 (1981), 587–593.

[59] VAN DER POORTEN A.J., *Solution de la conjecture de Pisot sur le quotient de Hadamard de deux fractions rationnelles*, C. R. Acad. Sci. Paris Sér. I **306** (1988), 97–102.

[60] POURCHET Y., *Solution de la conjecture de Pisot sur le quotient de Hadamard de deux fractions rationnelles*, C. R. Acad. Sci. Paris Sér. I **288** (1979), A 1055–1057.

[61] SANNA C., *The p-adic valuation of Lucas sequences*, Fibonacci Q. **54** 2 (2016), 118–124.

[62] SANNA C., *On numbers n dividing the nth term of a Lucas sequence*, Int. J. Number Theory **13** 3 (2017), 725–734.

[63] SANNA C., *Distribution of integral values for the ratio of two linear recurrences*, J. Number Theory **180** (2017), 195–207.

[64] SANNA C., *The moments of the logarithm of a G.C.D. related to Lucas sequences*, J. Number Theory **191** (2018), 305–315.

[65] SANNA C., *On Numbers n Relatively Prime to the nth Term of a Linear Recurrence*, Bull. Malays. Math. Sci. Soc. **42** 2 (2019), 827–833.

[66] SANNA C. AND TRON E., *The density of numbers n having a prescribed G.C.D. with the nth Fibonacci number*, Indag. Math. **29** (2018), 972–980.

[67] SCHLICKEWEI H.P. AND SCHMIDT W.M., *The Number of Solutions of Polynomial-Exponential Equations*, Compos. Math. **120** 2 (2000), 193–225.

[68] SCHMIDT W.M., *Diophantine Approximation*, Lect. Notes Math. **785**, Springer, Berlin 1980.

[69] SILVERMAN J.H., *Arithmetic distance functions and height functions in Diophantine geometry*, Math. Ann. **279** (1987), 193–216.

[70] SILVERMAN J.H., *Common divisors of $a^n - 1$ and $b^n - 1$ over function fields*, New York J. Math. **10** (2004), 37–43.

[71] SILVERMAN J.H., *Common divisors of elliptic divisibility sequences over function fields*, Manuscr. Math. **114** 4 (2004), 431–446.

[72] SILVERMAN J.H., *Generalized greatest common divisors, divisibility sequences, and Vojta's conjecture for blowups*, Monatsh. Math. **145** 4 (2005), 333–350.

[73] SILVERMAN J.H., *Divisibility sequences and powers of algebraic integers*, Doc. Math. extra vol. (2007), 711–727.

[74] SILVERMAN J.H., *The Arithmetic of Dynamical Systems*, Grad. Texts Math. **241**, Springer–Verlag, New York 2007.

[75] SILVERMAN J.H., *The Greatest Common Divisor of $a^n - 1$ and $b^n - 1$ and the Ailon–Rudnick Conjecture*, Contemp. Math. **517**, 339–347, AMS, Providence 2010.

[76] SILVERMAN J.H. AND STANGE K.E., *Terms in elliptic divisibility sequences divisible by their indices*, Acta Arith. **146** 4 (2011), 355–378.

[77] SMYTH C., *The terms in Lucas sequences divisible by their indices*, J. Integer Seq. **13** (2010), 10.2.4.

[78] SOMER L., *Divisibility of terms in Lucas sequences by their subscripts*, Applications of Fibonacci Numbers **5**, 515–525, Kluwer Academic Publishers, Dordrecht 1992.

[79] ULMER D. AND URZÚA G., *Transversality of sections on elliptic surfaces with applications to elliptic divisibility sequences and geography of surfaces*, `arXiv:1908.02208 [math.AG]` (2019).

[80] VOJTA P., *Diophantine approximation and Nevanlinna theory*, in *Arithmetic Geometry*, Lect. Notes Math. **2009**, Springer, Berlin 2011.

[81] WANG J.T.-Y. AND YASUFUKI Y., *Greatest common divisors of integral points of numerically equivalent divisors*, `arXiv:1907.09324 [math.NT]` (2019).

[82] YASUFUKI Y., *Vojtas conjecture on blowups of $\mathbb{P}^n$, greatest common divisors, and the abc conjecture*, Monatsh. Math. **163** 2 (2011), 237–247.

[83] YASUFUKI Y., *Integral points and Vojta's conjecture on rational surfaces*, Trans. Am. Math. Soc. **364** (2012), 767–784.

[84] ZANNIER U., *Some Problems of Unlikely Intersections in Arithmetic and Geometry*, Ann. Math. Stud. **181**, Princeton University Press, Princeton 2012.

AMS Subject Classification: 11B37, 11J87

Emanuele TRON
Institut de Mathématiques de Bordeaux
351 cours de la Libération, 33405 Talence, FRANCE
e-mail: `emanuele.tron@math.u-bordeaux.fr`

Lavoro pervenuto in redazione il 02.10.2019.

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 125 – 141

Alessandro Gambini, Remis Tonon, Alessandro Zaccagnini,
with an addendum by Jacques Benatar and Alon Nishry

SIGNED HARMONIC SUMS OF INTEGERS WITH k DISTINCT PRIME FACTORS

Abstract. We give some theoretical and computational results on "random" harmonic sums with prime numbers, and more generally, for integers with a fixed number of prime factors.

Keywords: Egyptian fractions; harmonic numbers; harmonic sums.
2010 Mathematics Subject Classification: Primary 11D75, Secondary 11B99.

1. Introduction and general setting

It is well known that the harmonic series restricted to prime numbers diverges, as the harmonic series itself. This was first proved by Leonhard Euler in 1737 [7], and it is considered as a landmark in number theory. The proof relies on the fact that

$$\sum_{n=1}^{N} \frac{1}{n} = \log N + \gamma + O(1/N),$$

where $\gamma \simeq 0.577215\ldots$ is the Euler–Mascheroni constant. The corresponding result for primes is one of the formulae proved by Mertens, namely

$$\sum_{p \le N} \frac{1}{p} = \log\log N + A + O\left(\frac{1}{\log N}\right),$$

where $A \simeq 0.2614972\ldots$ is the Meissel–Mertens constant. It is also referred to as Hadamard–de la Vallée-Poussin constant that appears in Mertens' second theorem.

Recently, Bettin, Molteni and Sanna [2] studied the random harmonic series

$$X := \sum_{n=1}^{\infty} \frac{s_n}{n}, \tag{1}$$

where $s_1, s_2, \ldots$ are independent uniformly distributed random variables in $\{-1,+1\}$. Based on the previous work by Morrison [9, 10] and Schmuland [12], they proved the almost sure convergence of (1) to a density function g, getting lower and upper bounds of the minimum of the distance of a number $\tau \in \mathbb{R}$ to a partial sum $\sum_{n=1}^{N} s_n/n$. In 1976 Worley studied the same problem in terms of upper bound of (1) both in the case $\tau = 0$ (see [13]) and for a generic $\tau \in \mathbb{R}$ (see [14]); his approach is not probabilistic but he has achieved an upper bound comparable to that of [2]. For further references, see also Bleicher and Erdős [3, 4], where the authors treated the number of distinct subsums

of $\sum_1^N 1/n$, which corresponds to taking s_i independent uniformly distributed random variables in $\{0,1\}$. A more complete list of references can be found in [2].

The purpose of this paper is firstly to show that basically the same results hold for a general sequence of integers under some suitable, and not too restrictive, conditions; moreover, that a stronger result can be reached if we restrict to integers with exactly k distinct prime factors.

Although Bettin, Molteni and Sanna [2] treat both the lower bound and the upper bound, we are mainly interested in the upper bound using a probabilistic approach. As we will see, in the cases that we treat, we will not be able to say anything about the lower bound, except in terms of numerical computations.

We will use a consistent notation with the previous works by Bettin, Molteni and Sanna [1], [2], Crandall [6] and Schmuland [12].

1.1. General setting of the problem

We denote by $\mathbb{N}$ the set of positive integers. Let $(a_n)_{n\in\mathbb{N}}$ be a strictly decreasing sequence of positive real numbers such that

$$\lim_{n\to+\infty} a_n = 0 \qquad \text{and} \qquad \sum_{n\geq 1} a_n = +\infty. \tag{2}$$

Notice that

$$\sum_{n\geq 1} (-1)^n a_n$$

converges (not absolutely) by Leibniz's rule. Hence, by Riemann's theorem, given λ, $\Lambda \in [-\infty,+\infty]$ with $\lambda \leq \Lambda$, we can arrange the choice of the signs $s_n = s_n(\lambda,\Lambda) \in \{-1,1\}$, in such a way that

$$\liminf_{N\to+\infty} \sum_{n\leq N} s_n a_n = \lambda \qquad \text{and} \qquad \limsup_{N\to+\infty} \sum_{n\leq N} s_n a_n = \Lambda.$$

As we said above, we are mainly interested in prime numbers, so we introduce some further reasonable hypotheses on the sequence a_n: we assume that $b_n = a_n^{-1} \in \mathbb{N}$, so that b_n is strictly increasing, and that

$$n \leq b_n \leq nB(n), \tag{3}$$

where $B(n) = n^{\beta(n)}$, with β a real-valued decreasing function such that $\beta(n) = o(1)$. In order to prove Proposition 20 below, we will assume a more restrictive condition on β, that is

$$\beta(n) \leq \frac{1}{8\log\log n} \qquad \text{for sufficiently large } n. \tag{4}$$

Actually, this assumption is not strictly necessary and we will discuss this in Remark 25. Nevertheless, since the series $\sum a_n$ must diverge, this condition is not too restrictive, and besides it is satisfied by most of the interesting sequences, like arithmetic progressions, the one of primes, and primes in arithmetic progressions.

Let us introduce some more notation: we consider the set

$$\mathfrak{S}(N) = \Big\{ \sum_{n\le N} s_n a_n \colon s_n \in \{-1,1\} \text{ for } n \in \{1,\dots,N\} \Big\}, \tag{5}$$

and, for a given $\tau \in \mathbb{R}$, we set

$$\mathfrak{m}_N(\tau) = \min\big\{ |S_N - \tau| \colon S_N \in \mathfrak{S}(N) \big\}.$$

In other words, for a given $N \in \mathbb{N}$, the goal is to find the choice of signs such that $|S_N - \tau|$ attains its minimum value. Finally, we define the random variable

$$X_N := \sum_{n=1}^{N} s_n a_n,$$

where the signs s_n are taken uniformly and independently at random in $\{-1,1\}$. We will study its small scale distribution. With a slight abuse of notation, we denote by s_n both the signs in the definition (5) and the random variables in the definition above.

1.2. Results

For ease of comparison with the results in Bettin, Molteni and Sanna [2], we now state our main results in the following form, even though more precise versions of them are to be found within the paper.

Theorem 12. *Let β satisfy (4). Then there exists $C > 0$ such that for every $\tau \in \mathbb{R}$ we have*

$$\mathfrak{m}_N(\tau) < \exp(-C \log^2 N)$$

for all sufficiently large N depending on τ.

Theorem 13. *Let $(b_n)_{n\in\mathbb{N}}$ be the sequence of integers having exactly k distinct prime factors. Then, for every $\tau \in \mathbb{R}$ and for all sufficiently large N depending on τ, we have*

$$\mathfrak{m}_N(\tau) < \exp(-f(N)),$$

where f is any function satisfying

$$f(N) = o\left(N^{1/(2k+1)-\varepsilon} \right).$$

Remark 14. We emphasize the fact that the estimate obtained in Theorem 13 holds uniformly for every $\tau \in \mathbb{R}$ in any fixed compact set.

Corollary 15 (J. Benatar and A. Nishry)**.** *For any fixed $\tau \in \mathbb{R}$, $\varepsilon > 0$ and any sufficiently large N there exists a choice of signs $(s_n)_{n\le N} \in \{-1,1\}^N$, such that*

$$\left| \sum_{n\le N} \frac{s_n}{n} - \tau \right| \ll_{\tau,\varepsilon} \exp\left(-N^{1/3-\varepsilon} \right).$$

We collect some numerical results for $k = 1$ in Tables 1, 2 and 3. The sequence of Tables 1 and 2 appears in OEIS A332399: see [5].

Acknowledgements. We thank Sandro Bettin and Giuseppe Molteni for many conversations on the subject, and Mattia Cafferata for his help in computing the tables at the end of the present paper. We also warmly thank Jacques Benatar and Alon Nishry for their fruitful suggestions which improved our paper, for providing us references and for letting us include their proof of Corollary 15 in this paper. R. Tonon and A. Zaccagnini are members of the INdAM group GNSAGA, which partially funded their participation to the Second Symposium on Analytic Number Theory in Cetraro, where some of this work was done.

2. Lemmas

In this section we study some properties of the general sequence defined in (2), using the classical notation: $\mathbb{E}[X]$ denotes the expected value of a random variable X, $\mathbb{P}(E)$ the probability of an event E. For each continuous function with compact support $\Phi \in \mathcal{C}_c(\mathbb{R})$ we denote by $\widehat{\Phi}$ its Fourier transform defined as follows:

$$\widehat{\Phi}(x) := \int_{\mathbb{R}} \Phi(y)\, \mathrm{e}^{-2\pi \mathrm{i} xy}\, \mathrm{d}y.$$

We are actually interested in smooth functions, because the smoothness of the density of any random variable X is related to the decay at infinity of its characteristic function, defined precisely by its Fourier transform.

For each $N \in \mathbb{N} \cup \{\infty\}$, for any $x \in \mathbb{R}$ and for any sequence satisfying (2), we also define the product

$$\rho_N(x) := \prod_{n=1}^{N} \cos(\pi x a_n) \qquad \text{and} \qquad \rho(x) := \rho_\infty(x).$$

We begin with the following lemma, which is a more general version of Lemma 2.4 from [2].

Lemma 16. *We have*

$$\mathbb{E}[\Phi(X_N)] = \int_{\mathbb{R}} \widehat{\Phi}(x) \rho_N(2x)\, \mathrm{d}x$$

for all $\Phi \in \mathcal{C}_c^1(\mathbb{R})$.

Proof. By the definition of expected value we have

$$\mathbb{E}[\Phi(X_N)] = \frac{1}{2^N} \sum_{s_1,\dots,s_N \in \{-1,1\}} \Phi\left(\sum_{n=1}^{N} s_n a_n \right).$$

Using the inverse Fourier transform we get

$$\begin{aligned}\mathbb{E}[\Phi(X_N)] &= \frac{1}{2^N}\sum_{s_1,\dots,s_N\in\{-1,1\}}\int_{\mathbb{R}}\widehat{\Phi}(x)\exp\left(2\pi\mathrm{i}x\sum_{n=1}^{N}s_n a_n\right)\mathrm{d}x\\ &= \int_{\mathbb{R}}\widehat{\Phi}(x)\frac{1}{2^N}\sum_{s_1,\dots,s_N\in\{-1,1\}}\exp\left(2\pi\mathrm{i}x\sum_{n=1}^{N}s_n a_n\right)\mathrm{d}x.\end{aligned}$$

Exploiting the fact that $\mathrm{e}^{\mathrm{i}\alpha}+\mathrm{e}^{-\mathrm{i}\alpha}=2\cos(\alpha)$, we have

$$\sum_{s_1,\dots,s_N\in\{-1,1\}}\exp\left(2\pi ix\sum_{n=1}^{N}s_n a_n\right)=\frac{1}{2}\sum_{s_1,\dots,s_N\in\{-1,1\}}2\cos\left(2\pi x\sum_{n=1}^{N}s_n a_n\right).$$

Finally, taking advantage of Werner's trigonometric identities, we obtain

$$\mathbb{E}[\Phi(X_N)]=\int_{\mathbb{R}}\widehat{\Phi}(x)\rho_N(2x)\,\mathrm{d}x. \qquad \square$$

We will need also a generalisation of Lemma 2.5 from [2], which is the following

Lemma 17. *For all $N\in\mathbb{N}$ and $x\in[0,\sqrt{N}]$ we have*

$$\rho_N(x)=\rho(x)\left(1+O\left(x^2/N\right)\right).$$

Proof. We recall that a_n is defined as in (2) and satisfies (3). In particular $a_n=O(1/n)$, so that the same argument in the proof of Lemma 2.5 of [2] holds. $\square$

Let us now define, for every positive integer N and any real δ and x the set

$$\mathcal{S}\left(N,\delta,x,(a_n)_{n\geq 1}\right):=\{n\in\{1,\dots,N\}\colon \|xa_n\|\geq\delta\},$$

where $\|\cdot\|$ denotes the distance from the nearest integer. For brevity, we sometimes drop the dependence on the sequence $(a_n)_{n\geq 1}$.

Lemma 18. *For all $N\in\mathbb{N}$ and for all $x,\delta\geq 0$ we have*

$$|\rho_N(x)|\leq\exp\left(-\frac{\pi^2\delta^2}{2}\cdot\#\mathcal{S}(N,\delta,x)\right).$$

Proof. It is a straightforward consequence of the inequality

$$|\cos(\pi x)|\leq\exp\left(-\frac{\pi^2\|x\|^2}{2}\right). \qquad \square$$

Lemma 19. *For any $N\in\mathbb{N}$, $x\in\mathbb{R}$ and $0<\delta<1/2$ we have*

$$\frac{N}{2}-D(N,y(\delta),x)<\#\mathcal{S}(N,\delta,x)<N-D(N,y(\delta)/2,x),$$

where

$$D(N,y,x) = D(N,y,x,(b_n)_{n\geq 1}) := \sum_{x-y<m<x+y} \sum_{\substack{b_n \mid m \\ N/2\leq n\leq N}} 1$$

and $y(\delta) := \delta NB(N)$.

Proof. As in Lemma 3.3 of [2], we observe that

$$\frac{N}{2} - T(N,\delta,x) < \#\mathcal{S}(N,\delta,x) < N - T(N,\delta,x),$$

where

$$T(N,\delta,x) := \#\{n \in \mathbb{N}\cap[N/2,N]\colon \|xa_n\| < \delta\}.$$

Now, recalling that $a_n = 1/b_n$, we have

$$\begin{aligned} T(N,\delta,x) &= \#\{n \in \mathbb{N}\cap[N/2,N] : \exists \ell \in \mathbb{N},\ \ell - \delta < xa_n < \ell + \delta\} \\ &= \#\{n \in \mathbb{N}\cap[N/2,N] : \exists \ell \in \mathbb{N},\ x - \delta b_n < \ell b_n < x + \delta b_n\}. \end{aligned}$$

From our hypothesis (3) we know that $b_n \leq NB(N)$; then

$$\begin{aligned} T(N,\delta,x) &< \#\{n \in \mathbb{N}\cap[N/2,N]\colon \exists \ell \in \mathbb{Z},\ x - y(\delta) < \ell b_n < x + y(\delta)\} \\ &= D(N,y(\delta),x). \end{aligned}$$

This proves the lower bound; the upper bound follows with the same argument. □

Proposition 20. *Let A be a fixed positive constant and, for N sufficiently large,*

$$\beta(N) \leq \frac{1}{8\log\log N}.$$

Then there exists $C' > 0$ such that $|\rho_N(x)| < x^{-A}$ for all sufficiently large positive integers N and for all $x \in [N, \exp(C'(\log N)^2)]$.

Proof. The proof follows along the same lines as Proposition 3.2 of [2]: we take

$$\bar{\delta} = \frac{2\sqrt{2A\log x}}{\pi} N^{-1/2} \quad \text{and} \quad x \in \left[N, \exp\left(\frac{\pi^2 N}{32A}\right)\right),$$

so that $0 < \bar{\delta} < 1/2$ and $y(\bar{\delta}) = \bar{\delta}NB(N) < x$.

By Lemmas 18 and 19, if we show that $D(N,y(\bar{\delta}),x) < N/4$, then we get $|\rho_N(x)| < 1/x^A$. Considering that b_n is a sequence of positive integers, we use Rankin's

trick with $w \in (1/4, 1/2)$ and Ramanujan's result on $\sigma_{-s}(n)$ [11] to obtain

$$\begin{aligned}
D(N, y(\overline{\delta}), x) &< \frac{4}{\pi}\sqrt{2AN\log x}B(N)\cdot \max_{m\le 2x} \sum_{\substack{b_n \mid m \\ N/2\le n\le N}} 1 \\
&< \frac{4}{\pi}\sqrt{2AN\log x}B(N)\cdot \max_{m\le 2x} \sum_{\substack{k\mid m \\ N/2\le k\le NB(N)}} 1 \\
&\le \frac{4}{\pi}\sqrt{2AN\log x}B(N)\cdot \max_{m\le 2x} \sum_{\substack{k\mid m \\ N/2\le k\le NB(N)}} \left(\frac{NB(N)}{k}\right)^w \\
&= \frac{4}{\pi}N^{\frac{1}{2}+w}B(N)^{1+w}\sqrt{2A\log x}\cdot \max_{m\le 2x} \sum_{\substack{k\mid m \\ N/2\le k\le NB(N)}} k^{-w} \\
&\le \frac{4}{\pi}N^{\frac{1}{2}+w}B(N)^{1+w}\sqrt{2A\log x}\cdot \max_{m\le 2x} \sigma_{-w}(m) \\
&< \frac{4}{\pi}N^{\frac{1}{2}+w}B(N)^{1+w}\sqrt{2A\log x}\cdot \exp\left(C_1\frac{(\log 2x)^{1-w}}{\log\log 2x}\right),
\end{aligned}$$

where C_1 is the constant of Ramanujan's theorem, as it is stated in Lemma 3.4 of [2].

Let $w = w(x) := 1/2 - \varphi(x)$, where φ is a positive decreasing function that we will choose later. Then we have

$$B(N)^{1+w} = \exp\left(\left(\frac{3}{2} - \varphi(x)\right)\beta(N)\log N\right),$$

and so we would be done if we showed that

$$N^{1-\varphi(x)+(3/2-\varphi(x))\beta(N)}\sqrt{\log x}\cdot\exp\left(C_1\frac{(\log 2x)^{1/2+\varphi(x)}}{\log\log 2x}\right) = o(N),$$

that is

$$\sqrt{\log x}\cdot\exp\left(C_1\frac{(\log 2x)^{1/2+\varphi(x)}}{\log\log 2x}\right) = o(N^{\varphi(x)+(\varphi(x)-3/2)\beta(N)}).$$

Hence we must have

$$\varphi(x) + (\varphi(x) - 3/2)\beta(N) > 0,$$

that is

$$\beta(N) < \frac{\varphi(x)}{3/2 - \varphi(x)} \approx \frac{2}{3}\varphi(x).$$

Since φ is decreasing and we want to maintain the same range for x as in [2], that is $x \in \left[N, \exp\left(C'(\log N)^2\right)\right]$, we need to have

$$\beta(N) \lesssim \frac{2}{3}\varphi\left(\exp\left(C'(\log N)^2\right)\right).$$

Let us take $\varphi(x) = (\log\log 2x)^{-1}$ and $\beta(N)$ such that for $x \in \left[N, \exp\left(C'(\log N)^2\right)\right]$ it holds

$$\beta(N) \le \frac{2}{3J}\varphi(x) = \frac{2}{3J}\frac{1}{\log\log 2x}, \tag{6}$$

where $J \in \mathbb{R}$, $J > 1$. Then we would achieve our goal if we showed that

$$\sqrt{\log x}\cdot\exp\left(C_1 \mathrm{e}\frac{(\log 2x)^{1/2}}{\log\log 2x}\right) = o\left(\exp\left(\left(1-\frac{1}{J}+o(1)\right)\frac{\log N}{\log\log 2x}\right)\right),$$

that is

$$\exp\left(C_1 \mathrm{e}\frac{(\log 2x)^{1/2}}{\log\log 2x} - \left(1-\frac{1}{J}+o(1)\right)\frac{\log N}{\log\log 2x} + \frac{1}{2}\log\log x\right) = o(1).$$

This condition is equivalent to

$$C_1 \mathrm{e}\frac{(\log 2x)^{1/2}}{\log\log 2x} - \left(1-\frac{1}{J}+o(1)\right)\frac{\log N}{\log\log 2x} + \frac{1}{2}\log\log x \to -\infty.$$

Taking into account the ranges for x, we see that it is sufficient to have

$$\frac{1}{\log\log N}\left[C_1\sqrt{C'}\,\mathrm{e}\log N(1+o(1)) - \left(1-\frac{1}{J}\right)\log N + O\left((\log\log N)^2\right)\right] \to -\infty.$$

We recall that, by our choice of x and N, we have $\log\log x \asymp \log\log N$. Hence, we just need to take C' sufficiently small, in a way that

$$C' < \left(\frac{J-1}{C_1 \mathrm{e} J}\right)^2, \tag{7}$$

to guarantee that $D(N, y(\overline{\delta}), x) < N/4$ for large N. For the sake of simplicity, we take $J = 2$ and the proposition is proved as stated. □

Remark 21. We remark here that condition (4) on β, which we assumed to prove the proposition, was necessary to ensure the existence of the function φ satisfying all the properties we needed, and in particular (6).

Corollary 22. *Let A be a fixed positive constant and β satisfy (4). Then $|\rho(x)| < x^{-A}$ for all sufficiently large $x \in \mathbb{R}$.*

Proof. It holds

$$|\rho(x)| = \left|\rho_{\lfloor x\rfloor+1}(x)\prod_{n>\lfloor x\rfloor+1}\cos(\pi x a_n)\right| < x^{-A}.\square$$

Theorem 23. *Let $C' > 0$ satisfy (7) and β satisfy (4). Then for all intervals $I \subseteq \mathbb{R}$ of length $|I| > \exp(-C'(\log N)^2)$ one has*

$$\mathbb{P}[X_N \in I] = \int_I g(x)\,\mathrm{d}x + o(|I|),$$

as $N \to \infty$, where

$$g(x) := 2\int_0^\infty \cos(2\pi u x) \prod_{n=1}^{\infty} \cos\left(\frac{2\pi u}{b_n}\right) \mathrm{d}u = 2\int_0^\infty \cos(2\pi u x)\rho(2u)\,\mathrm{d}u.$$

The proof follows along the same lines as Theorem 2.1 in [2] and we omit the details for brevity.

Corollary 24. *Let β satisfy (4). For all $\tau \in \mathbb{R}$ and $C' > 0$ satisfying (7), we have*

$$\#\left\{ (s_1,\dots,s_N) \in \{-1,+1\}^N : \left|\tau - \sum_{n=1}^{N} \frac{s_n}{b_n}\right| < \delta \right\} \sim 2^{N+1} g(\tau)\delta(1 + o_{C',\tau}(1))$$

as $N \to \infty$ and $\delta \to 0$, uniformly in $\delta \geq \exp(-C'(\log N)^2)$. In particular, for large enough N, one has $\mathfrak{m}_N(\tau) < \exp(-C'(\log N)^2)$.

Remark 25. We have imposed condition (4) for β to keep the same range of validity for x as in [2]. We remark that the hypotheses on β could be relaxed at the price of restricting this range: for example, we could take

$$\beta(N) = \frac{\log\log\log N}{\log\log N},$$

and obtain the result of Proposition 20 for $x \in [N, \exp(\log^a N)]$, where $a \in (1,2)$ is a suitable constant. In fact, this would weaken directly the estimates that we have just found in Theorem 23 and Corollary 24, where $\exp(-C'(\log N)^2)$ would be replaced by $\exp(-\log^a N)$.

3. Products of k primes

We now leave the general case and concentrate on primes and products of k distinct primes. Hence, we define

$$\mathcal{P}_k := \{\, n \in \mathbb{N} \mid n \text{ is the product of } k \text{ distinct primes} \,\};$$

we will denote by $b_n^{(k)}$ the n-th element of the ordered set $\mathcal{P}_k$. Let us recall the definition of $\mathcal{S}(N,\delta,x)$ in the case $a_n = 1/b_n^{(k)}$:

$$\mathcal{S}(N,\delta,x) := \{\, n \in \{1,\dots,N\} : \|x/b_n^{(k)}\| \geq \delta \,\}.$$

We remark that, since we left the general case, we can now take $B(n) = b_n^{(k)}/n$, and denote it by $B_k(n)$. In 1900, Landau [8] proved that

$$\pi_k(t) := |\mathcal{P}_k \cap \{ n \in \mathbb{N} \mid n \le t \}| = \frac{t}{\log t} \frac{(\log\log t)^{k-1}}{(k-1)!} + O\left(\frac{t(\log\log t)^{k-2}}{\log t} \right),$$

which implies that

$$B_k(n) \sim \log n \frac{(k-1)!}{(\log\log n)^{k-1}}. \tag{8}$$

We can now start with a refinement of Proposition 20, where we extend the interval of validity for x in the case $b_n = b_n^{(k)}$.

Proposition 26. *Let A be a fixed positive constant, $k \in \mathbb{N}$ be fixed and $a_n = 1/b_n^{(k)}$, where $b_n^{(k)}$ is the n-th element of the ordered set $\mathcal{P}_k$. Then $|\rho_N(x)| < x^{-A}$ for all sufficiently large positive integers N and for all $x \in [U, \exp(f(N))]$, where $\log N = o(f(N))$ and*

$$f(N) = o\left(\left(\frac{N}{B_k^2(N)} \right)^{1/(2k+1)} \right),$$

and $U > 1$ is a constant depending on f.

Proof. Let $x \in [N, \exp(f(N))]$. As in the proof of Proposition 20, we need to show that $D(N, y(\bar\delta), x) < N/4$, where $\bar\delta$ is chosen in the same way and $y(\bar\delta) = \bar\delta N B_k(N)$. Since now we are considering $x \ge N$, it is easy to see that for sufficiently large N we have $y(\bar\delta) \le x$. We recall here that the prime omega function $\omega(n)$ is defined as the number of different prime factors of n, and that

$$\omega(n) \ll \frac{\log n}{\log\log n},$$

as a consequence of the prime number theorem. In this case, we have

$$\begin{aligned} D(N, y(\bar\delta), x) &:= \sum_{x - y(\bar\delta) < m < x + y(\bar\delta)} \sum_{\substack{b_n^{(k)} \mid m \\ N/2 \le n \le N}} 1 \le \sum_{x - y(\bar\delta) < m < x + y(\bar\delta)} \sum_{\substack{p_1 \cdots p_k \mid m \\ p_i \text{ distinct primes}}} 1 \\ &\le \sum_{x - y(\bar\delta) < m < x + y(\bar\delta)} \omega(m)^k \le (2y(\bar\delta) + 1) \max_{m < x + y(\bar\delta)} \omega(m)^k \\ &\ll (N \log x)^{1/2} B_k(N) \left(\frac{\log 2x}{\log\log 2x} \right)^k \ll N^{1/2} B_k(N) (\log x)^{k+1/2}, \end{aligned}$$

where we used the trivial bound for the prime omega function. If we show that this quantity is $o(N)$, we are done. So we need

$$\log x = o\left(\left(\frac{N}{B_k^2(N)} \right)^{1/(2k+1)} \right).$$

Hence we can take any f that satisfies

$$f(N) = o\left(\left(\frac{N}{B_k^2(N)}\right)^{1/(2k+1)}\right),$$

where we recall that B_k satisfies (8). The theorem is then proved for $x \in [N, \exp(f(N))]$. If $x < N$, it holds

$$|\rho_N(x)| \leq |\rho_{\lfloor x \rfloor}(x)|,$$

hence the result we have just proved holds also whenever $x \leq \exp\left(f(\lfloor x \rfloor)\right)$. But there must exist $U > 0$ such that this holds for any $x > U$, since $\log x = o(f(x))$. □

We are now ready to prove a more general version of Theorem 2.1 of [2] for the sequence $\left(b_n^{(k)}\right)_{n\in\mathbb{N}}$.

Theorem 27. *Let f and a_n be defined as in Proposition 26. Then for all intervals $I \subseteq \mathbb{R}$ of length $|I| > \exp(-f(N))$ one has*

$$\mathbb{P}[X_N \in I] = \int_I g(x)\,\mathrm{d}x + o(|I|),$$

as $N \to \infty$, where

$$g(x) := 2\int_0^\infty \cos(2\pi u x) \prod_{n=1}^\infty \cos\left(\frac{2\pi u}{b_n^{(k)}}\right) \mathrm{d}u = 2\int_0^\infty \cos(2\pi u x)\rho(2u)\,\mathrm{d}u.$$

Proof. The proof follows the one of Theorem 2.1 of [2]. Let $\varepsilon > 0$ be fixed. We define

$$\begin{aligned}
\xi = \xi_{N,-\varepsilon} &:= \exp(-(1-\varepsilon)f(N)),\\
\xi_+ = \xi_{N,+\varepsilon} &:= \exp(-(1+\varepsilon)f(N)),\\
\xi_0 := \xi_{N,0} &= \exp(-f(N)),
\end{aligned}$$

so that $\xi^{-1} < \xi_0^{-1}$ and Proposition 26 holds for $x \in [N, \xi_0^{-1}]$. For an interval $I = [a,b]$ with $b - a > 2\xi_0$, let us define $I^+ := [a-\xi, b+\xi]$ and $I^- := [a+\xi_+, b-\xi_+]$. Then one can construct two smooth functions $\Phi^\pm_{N,\varepsilon,I}(x) : \mathbb{R} \to [0,1]$ (from now on, we will drop the subscripts when they are clear by the context) such that

$$\begin{cases}
\operatorname{supp}\Phi^+ \subseteq I^+ & \\
\Phi^+(x) = 1 & \text{for } x \in I,\\
\operatorname{supp}\Phi^- \subseteq I & \\
\Phi^-(x) = 1 & \text{for } x \in I^-,\\
(\Phi^\pm)^{(j)}(x) \ll_j \xi^{-j} & \text{for all } j \geq 0.
\end{cases}$$

By the last equation, we know that the Fourier transforms of $\Phi^\pm$ satisfy

$$\widehat{\Phi^\pm}(x) \ll_B (1+|x|\xi)^{-B} \quad \text{for any } B > 0 \text{ and } x \in \mathbb{R}. \tag{9}$$

Since

$$\mathbb{E}[\Phi^-(X_N)] \le \mathbb{P}[X_N \in I] \le \mathbb{E}[\Phi^+(X_N)],$$

we just need to show that

$$\mathbb{E}[\Phi^\pm(X_N)] = \int_{\mathbb{R}} \Phi^\pm(x) g(x)\,\mathrm{d}x + o_\varepsilon(|I|).$$

From now on, Φ will indicate either Φ^+ or Φ^-. By Lemma 16 we have

$$\mathbb{E}[\Phi(X_N)] = \frac{1}{2}\int_{\mathbb{R}} \widehat{\Phi}(x/2)\rho_N(x)\,\mathrm{d}x = I_1 + I_2 + I_3,$$

where I_1, I_2 and I_3 are the integrals supported respectively in $|x| < N^\varepsilon$, $|x| \in [N^\varepsilon, \xi^{-(1+\varepsilon)}]$ and $|x| > \xi^{-(1+\varepsilon)}$. Note that $\xi^{-(1+\varepsilon)} = \exp((1-\varepsilon^2)f(N)) > \exp(\varepsilon \log N) = N^\varepsilon$, that $\xi^{-(1+\varepsilon)} = \xi_0^{-(1-\varepsilon^2)} < \xi_0^{-1}$, and that $\xi^{-(1+\varepsilon)} \cdot \xi = \xi^{-\varepsilon} = \xi_0^{-\varepsilon(1-\varepsilon)} \to +\infty$ as $N \to +\infty$. By Lemma 17 and Corollary 22, we have

$$\begin{aligned} I_1 &= \frac{1}{2}\int_{-N^\varepsilon}^{N^\varepsilon} \widehat{\Phi}(x/2)\rho_N(x)\,\mathrm{d}x = \frac{1}{2}\int_{-N^\varepsilon}^{N^\varepsilon} \widehat{\Phi}(x/2)\rho(x)\,\mathrm{d}x + O\left(\|\widehat{\Phi}\|_\infty N^{-1+3\varepsilon}\right) \\ &= \frac{1}{2}\int_{\mathbb{R}} \widehat{\Phi}(x/2)\rho(x)\,\mathrm{d}x + O_A\left(\|\widehat{\Phi}\|_\infty N^{-(A-1)\varepsilon}\right) + O\left(\|\widehat{\Phi}\|_\infty N^{-1+3\varepsilon}\right) \\ &= \int_{\mathbb{R}} \widehat{\Phi}(x)\rho(2x)\,\mathrm{d}x + O_\varepsilon\left(\|\Phi\|_1 N^{-1+3\varepsilon}\right), \end{aligned}$$

where to conclude we chose $A = A(\varepsilon)$ sufficiently large. For the second integral, we use Proposition 26 and obtain

$$\begin{aligned} |I_2| &\le \|\widehat{\Phi}\|_\infty \int_{N^\varepsilon}^{\xi^{-(1+\varepsilon)}} |\rho_N(x)|\,\mathrm{d}x \le \|\Phi\|_1 \int_{N^\varepsilon}^{\xi^{-(1+\varepsilon)}} x^{-A}\,\mathrm{d}x \le \|\Phi\|_1 \int_{N^\varepsilon}^{+\infty} x^{-A}\,\mathrm{d}x \\ &\ll_\varepsilon \|\Phi\|_1 N^{-A\varepsilon+\varepsilon} \ll_\varepsilon \|\Phi\|_1 N^{-1}, \end{aligned}$$

where, as before, to conclude we took $A = A(\varepsilon)$ sufficiently large. For the last integral, we recall that trivially $|\rho_N(x)| \le 1$; using the bound (9), we obtain

$$\begin{aligned} |I_3| &\le \int_{|x|>\xi^{-(1+\varepsilon)}} |\widehat{\Phi}(x/2)|\,\mathrm{d}x \ll_B \int_{\xi^{-(1+\varepsilon)}}^{+\infty} (1+x\xi)^{-B}\,\mathrm{d}x = (B-1)(\xi^{-1}+\xi^{-(1+\varepsilon)})^{1-B} \\ &\ll_B \xi_0^{B-1} = o_\varepsilon(\xi_0) = o_\varepsilon(|I|), \end{aligned}$$

where to conclude we chose $B = B(\varepsilon)$ sufficiently large. We can now put these results together: using Parseval's theorem and the fact that $\|\Phi\|_1 = O_\varepsilon(|I|)$, we get

$$\mathbb{E}[\Phi(X_N)] = \int_{\mathbb{R}} \widehat{\Phi}(x)\rho(2x)\,\mathrm{d}x + O_\varepsilon\left(\|\Phi\|_1 N^{-1+3\varepsilon}\right) + o_\varepsilon(|I|) = \int_{\mathbb{R}} \Phi(x) g(x)\,\mathrm{d}x + o_\varepsilon(|I|)$$

and the theorem is then proved. □

Remark 28. By Corollary 22, for any $n \in \mathbb{N}$ it holds

$$\int_{-\infty}^{+\infty} |t^n \rho(t)|\,\mathrm{d}t < \infty,$$

which implies by standard arguments (see e.g. §5 of [12]) that the density g is a smooth strictly positive function. Besides, by the same corollary, $g(x) \ll_D x^{-D}$ for any $D > 0$.

Corollary 29. *For all $\tau \in \mathbb{R}$, we have*

$$\#\left\{ (s_1,\dots,s_N) \in \{-1,1\}^N : \left| \tau - \sum_{n=1}^{N} \frac{s_n}{b_n^{(k)}} \right| < \delta \right\} \sim 2^{N+1} g(\tau)\delta(1+o_\tau(1))$$

as $N \to \infty$ and $\delta \to 0$, uniformly in $\delta \geq \exp(-f(N))$, where f is defined as in Proposition 26. In particular, for N large enough, one has $\mathfrak{m}_N(\tau) < \exp\big(-f(N)\big)$.

4. Addendum (by J. Benatar and A. Nishry): proof of Corollary 15

Proof. Let c_m denote the m-th non-prime integer, so that $c_1 = 1$, $c_2 = 4$, $c_3 = 6, \dots$ We first approximate τ with a restricted harmonic sum of the form $\sum_{m \leq M} s_m c_m$, where $M = M(N) = N - \pi(N)$. Since $C_m := c_m/m \sim 1$, we may apply Theorem 12 to obtain a sequence of signs $(s_n)_{n \leq M} \in \{-1,1\}^M$ such that

$$-1 \leq \tau' := \sum_{m \leq M} s_m c_m - \tau \leq 1.$$

Moreover, taking $(p_n)_{n\in\mathbb{N}}$ to be the sequence of primes, we have that $B(n) \sim \log n$ and hence we may apply Theorem 13 to get a choice of signs $(\sigma_n)_{n \leq \pi(N)} \in \{-1,1\}^{\pi(N)}$ such that

$$\left| \tau' - \sum_{n \leq \pi(N)} \frac{\sigma_n}{p_n} \right| \ll_{\tau,\varepsilon} \exp\Big(-N^{1/3-\varepsilon}\Big).\square$$

References

[1] S. Bettin, G. Molteni, and C. Sanna, *Greedy approximations by signed harmonic sums and the Thue–Morse sequence*, Adv. Math. 366, no. 3 (2020), 1–42.

[2] S. Bettin, G. Molteni, and C. Sanna, *Small values of signed harmonic sums*, C. R. Math. Acad. Sci. Paris **356** (2018), no. 11-12, 1062–1074.

[3] M. N. Bleicher and P. Erdős, *The number of distinct subsums of* $\sum_1^N 1/i$, Math. Comp. **29** (1975), 29–42.

[4] M. N. Bleicher and P. Erdős, *Denominators of Egyptian fractions. II*, Illinois J. Math. **20** (1976), no. 4, 598–613.

[5] M. Cafferata, A. Gambini, R. Tonon, and A. Zaccagnini, Sequence A332399 in The On-Line Encyclopedia of Integer Sequences (2020), published electronically at https://oeis.org/A332399.

[6] R. E. Crandall, *Theory of ROOF walks*, Unpublished. Available at http://www.reed.edu/physics/faculty/crandall/papers/ROOF11.pdf, 2008.

[7] L. Euler, *Variae observationes circa series infinitas*, Commentarii academiae scientiarum imperialis Petropolitanae **9** (1737), 160–188.

[8] E. Landau, *Sur quelques problèmes relatifs à la distribution des nombres premiers*, Bull. Soc. Math. France **28** (1900), 25–38.

[9] K. E. Morrison, *Cosine products, Fourier transforms, and random sums*, Amer. Math. Monthly **102** (1995), no. 8, 716–724.

[10] K. E. Morrison, *Random walks with decreasing steps*, Unpublished manuscript, California Polytechnic State University, 1998.

[11] S. Ramanujan, *Highly composite numbers*, Proc. London Math. Soc. **14** (1915), 347–409.

[12] B. Schmuland, *Random harmonic series*, Amer. Math. Monthly **110** (2003), no. 5, 407–416.

[13] R. T. Worley, *Signed sums of reciprocals. I*, J. Austral. Math. Soc. Ser. A **21** (1976), no. 4, 410–413.

[14] R. T. Worley, *Signed sums of reciprocals. II*, J. Austral. Math. Soc. Ser. A **21** (1976), no. 4, 414–417.

Alessandro Gambini
Dipartimento di Matematica Guido Castelnuovo
Sapienza Università di Roma
Piazzale Aldo Moro, 5
00185 Roma, Italia
email (AG): alessandro.gambini@uniroma1.it

Remis Tonon, Alessandro Zaccagnini
Dipartimento di Scienze, Matematiche, Fisiche e Informatiche
Università di Parma
Parco Area delle Scienze, 53/a
43124 Parma, Italia
email (RT): remis.tonon@unimore.it
email (AZ): alessandro.zaccagnini@unipr.it

Lavoro pervenuto in redazione il 02.10.2019.

1. Numerical data

N	$\mathfrak{m}_N(0)\cdot p_1\cdots p_N$
1	1
2	1
3	1
4	23
5	43
6	251
7	263
8	21013
9	1407079
10	4919311
11	818778281
12	2402234557
13	379757743297
14	3325743954311
15	54237719914087
16	903944329576111
17	46919460458733911
18	367421942920402841
19	17148430651130576323
20	1236225057834436760243
21	4190310920096832376289
22	535482916756698482410061
23	29119155169912957197310753
24	44328424890849151628867 1253
25	284387814834969303966896382 31
26	10196503226925713726754541885481
27	13751219812531776626796813776 5087
28	55728212024753056062119855537860 81
29	77833992457426020006787481021085581
30	24244850423688161715955346535954790877
31	20303493347784199953241194396599940861 31
32	76860130392109667765387079377871685276909
33	51919706244457608828445331682701847213186 37
34	32964320927134843189509655079215913228392 0307
35	19171590315567357340242017182966253037383120953
36	581923784909774304868513653323528747823328 7403
37	837477642920747839191618216897250374978659503996169
38	13066546626103391941444189280002540864243236444837 2023
39	75415501694072326086891495259849678983989478052962 16009
40	23868339955752715692132986729285170427530832996153507207

Table 1: The values, multiplied by $p_1\cdots p_N$, of the smallest signed harmonic sums with the first N primes, with N up to 40. See also OEIS A332399.

N	$\mathfrak{m}_N(0)\cdot p_1\cdots p_N$
41	3343165792500492306892396976512891068137770193474133826457
42	472332689319626425103031695114936015175668001545378672380 57
43	939153294398682057461561638052904417551519861279479163756267 93
44	503134391484163245811276101556411501279873182605691723310335931 81
45	203570378824611321145575301458424678266473772064479301689195508719 7
46	193768861589178044091624877468627581772116464350368833881209864412247
47	466412854952040265053303054101346780628864888074165457806800584527117 7
48	252294099680710988063673862003152188841680135741161924018446904086039541
49	164152705533632496799540344537262942048356425519773153500697538193607343 3
50	254364245054513324419283194746564713368741676550473667747021878822748940640 63
51	178002407776132876331812856270329912040466608132314917858262023648082741528925 9
52	115533643751466097619699345183033980786661230484621892531131629910924364040946261
53	346445205731766592295370811989346241267385291503362454494739411253204971046538171 09
54	736966896305166158296639261731963300962552237561129405178436540109047122094638759278 9
55	199963258224846876335793874247507216756651341869412816388166951273778698828707537479531 7
56	151351981933638637742621357138936533979590998748883750430193460129876391573603481014628429
57	153027249026981884500276849749805539399879910740134024375798662329813718469262266844584069 69
58	626908543267515513547777358925056214956392632737317661747337955522213761579292221419596422528 1
59	429918790837116674905123858093668694474961832761345115366942177591943696826657060080682245858603
60	115809464188499233574522294110279752895686365776568444548440426304978721966632473743873345620708313

Table 2: The values, multiplied by $p_1\cdots p_N$, of the smallest signed harmonic sums with the first N primes, with N between 41 and 60. See also OEIS A332399.

N	$\Delta_N \cdot p_1 \cdots p_N$
1	1
2	1
3	1
4	2
5	22
6	35
7	263
8	4675
9	24871
10	104006
11	2356081
12	6221080
13	141769355
14	6096082265
15	6928889495
16	367231143235
17	1283811918935
18	78312527055035
19	5246939312687345
20	372532691200801495
21	8815359347599933286
22	223849990729887044174
23	6148176498383067879445
24	17984783728793716081796 3
25	6630243946027524253731 30

Table 3: The values, multiplied by $p_1 \cdots p_N$, of the shortest distances Δ_N between different signed harmonic sums with the first N primes, with N up to 25.

Rendiconti Sem. Mat. Univ. Pol. Torino
Vol. 78, 1 (2020), 143 – 147

G. Zaghloul

ZEROS OF GENERALIZED HURWITZ ZETA FUNCTIONS

Abstract. In [6], Davenport and Heilbronn proved that the classical Hurwitz zeta function $\zeta(s,\alpha)$ has infinitely many zeros in the half-plane $\sigma > 1$, provided that $\alpha \notin \{1, \frac{1}{2}\}$ is rational or transcendental. The algebraic irrational case was later settled by Cassels [3]. This note is a survey of the main results about zeros of Dirichlet series in the region of absolute convergence. In particular, we focus on the results in [12], where a recent contribution by Chatterjee and Gun [4] is improved. Given a function $f(n)$ periodic of period $q \geq 1$ and a real number $0 < \alpha \leq 1$, in [12] it is shown that the series $F(s,f,\alpha) = \sum_{n=0}^{\infty} \frac{f(n)}{(n+\alpha)^s}$ has infinitely many zeros for $\sigma > 1$.

1. Introduction

In the literature, several results show that non-trivial linear combinations of L-functions may have infinitely many zeros in the region of absolutely convergence, so in particular they do not satisfy the Riemann hypothesis. For instance, in 1935 Potter and Titchmarsh showed the Epstein zeta function has infinitely many zeros on the critical line $\sigma = \frac{1}{2}$, but they also gave an example of an Epstein zeta function, without a Euler product, which has a zero in the critical strip not lying on the critical line (cf. [9]).
In 1936, Davenport and Heilbronn [6] completed the analysis of the Epstein zeta function and they also studied the Hurwitz zeta function, defined by

$$\zeta(s,\alpha) = \sum_{n=0}^{\infty} \frac{1}{(n+\alpha)^s}, \tag{1}$$

for $s = \sigma + it \in \mathbb{C}$ with $\sigma > 1$ and $\alpha \in (0,1]$. They proved that if $\alpha \notin \{1, \frac{1}{2}\}$ is either rational or transcendental, $\zeta(s,\alpha)$ has infinitely many zeros in $\sigma > 1$.
In 1961, the remaining, more difficult, case of α algebraic irrational was settled by Cassels [3], by means of a lemma of algebraic number theory. The following theorem summarizes the results for the zeros of the Hurwitz zeta function.

THEOREM 1 (Davenport-Heilbronn, Cassels). *Let $\alpha \in (0,1]$. If $\alpha \notin \{1, \frac{1}{2}\}$, then $\zeta(s,\alpha)$ has infinitely many zeros in $\sigma > 1$.*

REMARK 1. We observe that if $\alpha = 1$, $\zeta(s,1) = \zeta(s)$, while for $\alpha = \frac{1}{2}$ we get $\zeta(s,\frac{1}{2}) = (2^s - 1)\zeta(s)$. So, for this values the Hurwitz zeta function does not vanish in the half-plane $\sigma > 1$.

The proof of Theorem 1 is based on Bohr's theory of equivalent Dirichlet series. Let $f(s)$ and $g(s)$ be two general Dirichlet series

$$f(s) = \sum_{n=1}^{\infty} a(n) e^{-\lambda(n)s} \quad \text{and} \quad g(s) = \sum_{n=1}^{\infty} b(n) e^{-\lambda(n)s}$$

and let B be a *basis* for the sequence of exponents $\Lambda = \{\lambda(n)\}$ (cf. [1] for more details). Then, $f(s)$ and $g(s)$ are *equivalent* with respect to the basis B if there exists a sequence of real numbers $Y = \{y(n)\}$ such that

$$b(n) = a(n)e^{i(R_B Y)_n} \quad \text{for all} \quad n \geq 1$$

where R_B is a matrix such that $\Lambda = R_B B$.
An important property of Dirichlet series used in the proof is *almost periodicity*.

DEFINITION 1. *A holomorphic function $f(s) = f(\sigma + it)$ defined in some vertical strip $-\infty \leq \sigma_1 < \sigma < \sigma_2 \leq +\infty$ is* Bohr almost periodic *in (σ_1, σ_2) if, for any $\varepsilon > 0$, the set*

$$\{\tau \in \mathbb{R} || f(s + i\tau) - f(s)| < \varepsilon \text{ for all } \sigma_1 < \sigma < \sigma_2, t \in \mathbb{R}\}$$

is relatively dense, *i.e. there exists $\ell = \ell(f, \varepsilon, \sigma_1, \sigma_2) > 0$ such that any interval of length ℓ contains at least an element of the above set.*

Another key ingredient is Rouché's theorem.

THEOREM 2 (Rouché). *Let two functions $f(s)$ and $g(s)$ be analytic inside and on a closed simple curve $\mathcal{C}$. Assume that*

$$|f(s)| > |g(s)| \quad \text{on} \quad \mathcal{C}.$$

Then, $f(s)$ and $f(s) + g(s)$ have the same number of zeros inside $\mathcal{C}$.

The idea is to find a Dirichlet series equivalent to $\zeta(s, \alpha)$ whit a zero $s_0 = \sigma_0 + it_0$ in $\sigma > 1$. Then, by almost periodicity and Rouché's theorem, one concludes that, for any $\varepsilon > 0$ and for T sufficiently large,

$$|\{s = \sigma + it | \zeta(s, \alpha) = 0, \sigma_0 - \varepsilon < \sigma < \sigma_0 + \varepsilon, a < t < T + a\}| \gg T.$$

The same argument has been used by Conrey and Ghosh in [5] to prove the existence of infinitely many zeros with $\sigma > 1$ for the Dirichlet series associated to the square of the Ramanujan's τ function.
In 2009, Saias and Weingartner [11] proved that Dirichlet series with periodic coefficients can be written as linear combinations of the form

$$F(s) = \sum_{j=1}^{N} P_j(s) L(s, \chi_j), \tag{2}$$

where for $j = 1, \ldots, N$, $P_j(s)$ is a Dirichlet polynomial and $L(s, \chi_j)$ is the Dirichlet L-function associated to the primitive character χ_j. They also show that, if (2) does not reduce to a single term, then it has infinitely many zeros in $\sigma > 1$. They were inspired by Kaczorowski and Kulas [8], who proved that (2) has infinitely many zeros for $\frac{1}{2} < \sigma < 1$ if $N \geq 2$, using a strong universality property. However, since the universality property introduced in [8] does not hold in strips in the half-plane $\sigma > 1$, Saias and

Weingartner proved a sort of *weak joint universality property* of Dirichlet L-functions. In 2014, Booker and Thorne [2] extended the result to combinations of L-functions coming from automorphic representations, under Ramanujan conjecture. In 2016, Righetti [10] modified the proof obtaining the analogous result in a more general setting, (combinations of Dirichlet series with an Euler product, bounded coefficients and satisfying orthogonality relations).

2. The generalized Hurwitz zeta function

Let now $f(n)$ be a periodic function of period $q \geq 1$ and let $\alpha \in (0,1]$. The *generalized Hurwitz zeta function* is defined for $\sigma > 1$ by

$$F(s,f,\alpha) = \sum_{n=0}^{\infty} \frac{f(n)}{(n+\alpha)^s}.$$

Since the coefficients are periodic, it can be easily observed that

$$F(s,f,\alpha) = \frac{1}{q^s} \sum_{b=0}^{q-1} f(b) \zeta\left(s, \frac{b+\alpha}{q}\right).$$

Then, it follows by the well-known properties of the classical Hurwitz zeta function (cf. e.g. [7]) that $F(s,f,\alpha)$ admits a meromorphic continuation to the whole complex plane with a possible simple pole at $s=1$ with residue

$$\mathrm{Res}_{s=1} F(s,f,\alpha) = q^{-1} \sum_{b=0}^{q-1} f(b).$$

In 2014, Chatterjee and Gun [4] proved that $F(s,f,\alpha)$ has infinitely many zeros in $\sigma > 1$ if α is irrational under some restrictive conditions.

THEOREM 3 (Chatterjee−Gun). *Let α be a positive transcendental number and let f be a real valued periodic function with period $q \geq 1$. If $F(s,f,\alpha)$ has a pole at $s=1$, then $F(s,f,\alpha)$ has infinitely many zeros for $\sigma > 1$.*

THEOREM 4 (Chatterjee−Gun). *Let α be a positive algebraic irrational number and let f be a positive valued periodic function with period $q \geq 1$. Moreover, let $c := \frac{\max_n f(n)}{\min_n f(n)} < 1.15$. If $F(s,f,\alpha)$ has a pole at $s=1$, then $F(s,f,\alpha)$ has infinitely many zeros for $\sigma > 1$.*

In [12], we showed that these assumptions can be removed, proving the result in full generality, including the case of α rational.

THEOREM 5. *Let $f(n)$ be a non identically zero periodic function with period $q \geq 1$ and let $0 < \alpha \leq 1$ be a real number. If $\alpha \notin \{1, \frac{1}{2}\}$, or if $\alpha \in \{1, \frac{1}{2}\}$ and $F(s,f,\alpha)$ is not of the form $P(s)L(s,\chi)$, where $P(s)$ is a Dirichlet polynomial and $L(s,\chi)$ is the*

L-function associated to a Dirichlet character χ, then $F(s,f,\alpha)$ has infinitely many zeros with $\sigma > 1$.

We now briefly sketch the idea of the proof of Theorem 5.

Case α rational. It can be easily verified that $F(s,f,\alpha)$ can be written as a linear combination of Dirichlet L-functions, i.e.

$$F(s,f,\alpha) = \sum_{\chi \in C} P_\chi(s) L(s,\chi), \tag{3}$$

where $\mathcal{C}$ is a set of primitive Dirichlet characters and $P_\chi(s)$ is a Dirichlet polynomial. Then, by the result of Saias and Weingartner, the sum (3) does not vanish in the half-plane $\sigma > 1$ if and only if it reduces to a single term. It can be shown that (3) can be of the form $P(s)L(s,\chi)$ only if $\alpha = 1, \frac{1}{2}$.

Case α transcendental. The argument of Davenport and Heilbronn for the Hurwitz zeta function applies also to $F(s,f,\alpha)$, since

$$\sum_{n=0}^{\infty} \frac{|f(n)|}{(n+\alpha)^\sigma} \to +\infty \quad \text{as} \quad \sigma \to 1^+,$$

so the assumption of the existence of the pole at $s = 1$ can be avoided.

Case α algebraic irrational. The proof is based on a modification of Cassels' original lemma. Let $K = \mathbb{Q}(\alpha)$, let O_K be its ring of integers and let $\mathfrak{a} = \{r \in O_K \mid r \cdot (\alpha) \subseteq O_K\}$ be the denominator ideal of α.

LEMMA 1. *Given an integer $q \geq 1$, fix $b \in \{0,\dots,q-1\}$. There exists an integer $N_0 > 10^6 q$, depending on α and q, satisfying the following property: for any integer $N > N_0$ put $M = \lfloor 10^{-6} N \rfloor$, then at least $0.54\frac{M}{q}$ of the integers $n \equiv b \pmod q$, $N < n \leq N+M$ are such that $(n+\alpha)\mathfrak{a}$ is divisible by a prime ideal $\mathfrak{p}_n$ for which*

$$\mathfrak{p}_n \nmid \prod_{\substack{m \leq N+M \\ m \neq n}} (m+\alpha)\mathfrak{a}.$$

The proof of the lemma is based on facts from algebraic number theory and it follows Cassels' argument, with some small modifications.

We now briefly describe the proof of Theorem 5 when α is algebraic irrational, referring to [12] for the complete argument. The idea is to rearrange Cassels' argument, applying it to each residue class modulo q.
By Bohr's theory, we know that it is sufficient to find a series equivalent to $F(s,f,\alpha)$ with a zero in $\sigma > 1$. In this case, this means finding a $\sigma \in (1, 1+\delta)$ and a function φ of absolute value 1 multiplicative on the group of ideals of O_K, such that

$$\sum_{n=0}^{\infty} \frac{f(n)\varphi((n+\alpha)\mathfrak{a})}{(n+\alpha)^\sigma} = 0.$$

It can be noticed that it is enough to define $\varphi(\mathfrak{p})$, with $|\varphi(\mathfrak{p})| = 1$, on the prime ideals $\mathfrak{p}$ of O_K dividing $(n+\alpha)\mathfrak{a}$. To this end, the idea is to apply Lemma 1 to each residue class and to use Bohr's results on addition of convex curves, proceeding as in Cassels. The conclusion follows summing over the residue classes modulo q and, as usual, by almost periodicity and Rouché's theorem.

References

[1] H. Bohr, *Zur Theorie der allgemeinen Dirichletschen Reihen*, Math. Ann. 79 (1918), 136156.

[2] A. R. Booker, F. Thorne, *Zeros of L-functions outside the critical strip*, Algebra Number Theory 8 (2014), no. 9, 20272042.

[3] J.W.S. Cassels, *Footnote to a note of Davenport and Heilbronn*, J. London Math. Soc. 36 (1961), 177-184.

[4] T. Chatterjee and S. Gun, *On the zeros of generalized Hurwitz zeta functions*, J. Number Theory, 145 (2014), 352-361.

[5] J. B. Conrey, A. Ghosh, *Turán inequalities and zeros of Dirichlet series associated with certain cusp forms*, Trans. Amer. Math. Soc. 342 (1994), no. 1, 407419.

[6] H. Davenport and H. Heilbronn, *On the zeros of certain Dirichlet series I, II*, Journal London Math. Soc. 11 (1936), 181-185, 3017-312.

[7] R. Garunkstis, A. Laurincikas, *The Lerch Zeta Function*, Springer (2002).

[8] J. Kaczorowski, M. Kulas, *On the non-trivial zeros off the critical line for L-functions from the extended Selberg class*, Monatsh. Math. 150 (2007), no. 3, 217232.

[9] H. S. A. Potter, E. C. Titchmarsh, *The zeros of Epsteins zeta functions*, Proc. London Math. Soc. S2-39 (1935), no. 1, 372-384.

[10] M. Righetti, *Zeros of combination of Euler product for* $\sigma > 1$, Monatsh Math 180 (2016), 337-356.

[11] E. Saias and A. Weingartner, *Zeros of Dirichlet series with periodic coefficients*, Acta Arithmetica 140 (2009), 335-344.

[12] G. Zaghloul, *A note on the zeros of generalized Hurwitz zeta functions*, J. Number Theory (2019), https://doi.org/10.1016/j.jnt.2018.09.016.

AMS Subject Classification: 11M35

Giamila ZAGHLOUL,
Dipartimento di Matematica, Università degli Studi di Genova
via Dodecaneso 35, 16146 Genova, ITALIA
e-mail: giamizaghi@gmail.com

Lavoro pervenuto in redazione il 10.07.2019.

Special Issues and Proceedings published in the Rendiconti

Differential Geometry	(1992)
Numerical Methods in Astrophysics and Cosmology	(1993)
Partial Differential Equations, I-II	(1993–1994)
Problems in Algebraic Learning, I-II	(1994)
Number Theory, I-II	(1995)
Geometrical Structures for Physical Theories, I-II	(1996)
Jacobian Conjecture and Dynamical Systems	(1997)
Control Theory and its Applications	(1998)
Geometry, Continua and Microstructures, I-II	(2000)
Partial Differential Operators	(2000)
Liaison and Related Topics	(2001)
Turin Fortnight Lectures on Nonlinear Analysis	(2002)
Microlocal Analysis and Related Topics	(2003)
Splines, Radial Basis Functions and Applications	(2003)
Polynomial Interpolation and Projective Embeddings - Lecture Notes of the School	(2004)
Polynomial Interpolation and Projective Embeddings - Proceedings of the Workshop of the School	(2005)
Control Theory and Stabilization, I-II	(2005–2006)
Syzygy 2005	(2006)
Subalpine Rhapsody in Dynamics	(2007)
ISASUT Intensive Seminar on Non Linear Waves, Generalized Continua and Complex Structures	(2007)
Lezioni Lagrangiane 2007–2008	(2008)
Second Conference on Pseudo-Differential Operators and Related Topics: Invited Lectures	(2008)
Second Conference on Pseudo-Differential Operators and Related Topics	(2009)
In Memoriam Aristide Sanini	(2009)
Workshop on Hodge Theory and Algebraic Geometry	(2010)
School on Hodge Theory	(2011)
Generalized Functions, Linear and Nonlinear Problems. Proceedings of the International Conference GF 2011	(2011)
Forty years of Analysis in Turin. A conference in honour of Angelo Negro	(2012)
Proceedings of the School (and Workshop) on Invariant Theory and Projective Geometry	(2013)
Stochastic Analysis at the 8th Congress of Isaac	(2013)
Special issue dedicated to Alberto Conte on the occasion of his 70th birthday	(2013)
Rate-independent evolutions and hysteresis modelling	(2014)

RENDICONTI DEL SEMINARIO MATEMATICO-UNIVERSITÀ
E POLITECNICO DI TORINO

Volume 78, N. 1 (2020)

CONTENTS

ISSN 0373-1243

UNIVERSITÀ DEGLI STUDI DI TORINO
DIPARTIMENTO DI MATEMATICA