Università degli Studi di Padova – Université de Strasbourg

Dipartimento di Matematica – Institute de Recherche Mathématique Avancée

Scuola di Dottorato in Scienze Matematiche – École Doctorale Mathématiques, Sciences de l'information et de l'ingénieur

PhD Thesis

Periodic Sequences and Persistent Homology

Theoretical foundations and new results

Candidate:
Riccardo C. Gilblas

Advisors:
Prof. Luisa Fiorot
Prof. Moreno Andreatta

XXXVI cycle

Acknowledgements

First I want to thank my supervisors. Luisa has always been kind and supportive during these years. I could not have hoped for a better guide to be introduced to mathematical research. Moreno gave me the warmest welcome in Strasbourg and effectively introduced me to the numerous topics in math/music research, allowing me to deepen several aspects and to meet many experts from different domains.

I thank Alberto, who was always available to give answers to my questions and doubts, and who made the research activity more interesting and fun. I thank all the experts with whom I had the privilege of collaborating: Victoria, Florence, Louis and in particular Andy, who was a great host at Marcs and gave me new and fertile perspectives on my research.

I thank Momo, Karim, Giacomo and the other colleagues in Padua, for the time spent together, lightening up the long working hours. I thank Ibrahim and Isabelle, who made me feel at home in Strasbourg. I thank Milena, Valentine, Patrick, Gil and the other colleagues at Marcs, great companions for discovering a new world.

I thank Giulia, who was always there for me during all these years. Finally, I want to thank my parents, who believed in me and supported my choices.

Introduction

Periodic sequences

Anatol Vieru's serialism. In the context of 1960's musical serialism, the Romanian composer Anatol Vieru plays a prominent role. Heir of Aram Khachaturian, his opus consists of seven symphonies and three operas, plus several chamber music composition and concertos. The compositional process of numerous of his works is based on sequences of numbers and the manipulation of such sequences by algorithmic processes. The composer started from a sequence and collected several others obtained from the first by applying a sum (resp. difference) operator. Then he assigned a musical meaning to each sequence (the melody, the rhythm, the harmony, etc.) for each instrument or orchestral section. Various of his compositions, like Symphony n. 2 (1973) and *Zone d'Oubli*, originated from this approach. Later, in his *Book of Modes* ([18]), Anatol Vieru explains in detail his compositional choices and the operations he performs on the sequences of modular integers.

Mathematical studies and open questions. From the repeated application of the sum operator to certain sequences, some mathematical properties and an interesting behaviour in their length (which we will call *period*) arise. Some observations were made by Anatol Vieru himself in [18]: he noticed that the period of the sequences generally increases when the sum operator is applied and for certain sequences it is always a power of 2. More, for particular sequences there are some values that proliferate. The first mathematical formalisation was done by D.T. Vuza in 1982 ([19]) and in [2] were presented the main results regarding the decomposition into nilpotent and idempotent sequences (called reducible and reproducible respectively). In [4] the problem of the proliferation of certain values in the primitives of the sequence $[2, 1, 2, 8, 4, 1, 4, 8] \in \mathbf{P}_{12}$ was studied in a computational way. More recently, in [13] cellular automata formalism was deployed to study the dual of the operator Δ.

Here, we give a complete overview of the theory of modular periodic sequences developed until now and we present the results we obtained on the period of primitives and the proliferation of values. The main new results (Lemma 1.3.5, Proposition 1.5.10, Theorems 1.5.6, 1.5.9 and 2.2.1, the recursive lemmas of Section 2.3.2 and the results of Section 2.4) are part of a joint work with Luisa Fiorot and Alberto Tonolo (see [8, 9]).

Persistent Homology and Harmonic Analysis

Topological Data Analysis. Persistent Homology is one of the main tools in Topological Data Analysis, a domain that is quickly acquiring importance in data analysis thanks to the deep geometrical information it provides. Its popularity among data analysts has been growing since the early 2000's, and it has been employed to face classification problems and feature extraction in several scientific domains. Persistent homology is at the core of TDA and it allows to reconstruct the topological features of a dataset. More specifically, given a filtration of simplicial complexes, which usually approximates the underlying geometrical structure of a set of points, persistent homology represents the evolution of simplicial homology (precisely, of Betti numbers in the form of persistent diagrams or barcodes) in the filtration. This allows one to gather topological information from the dataset and features from the barcodes that can be further used in Machine Learning algorithms for clustering and classification purposes.

The main reference for persistent homology and its current implementation is [42], which presents the mathematical setting and an efficient algorithm for computing persistent homology, while a wide introduction to its use in topological data analysis can be found in [28, 36]. An interesting review of Machine Learning techniques applied to persistent barcodes is [35]. Persistent homology has been successfully used in classification problems, from fingerprint classification ([31]) to general image recognition ([21]).

Persistent homology and music. More recently, persistent homology has been used also in math-music research, with the aim of linking the topological features of barcodes to the musical properties of datasets. In particular, in [23] persistent homology is used in conjunction with the Tonnetz for automatic style classification. In [24], Vietoris-Rips filtration is used to compute persistent homology of a musical score and to visualise thematic features of musical pieces.

In this work, we focus on harmony and, more precisely, on harmonic complexity. Similarly to [23], the aim is recognise the harmonic properties (the *harmonic density*) of sequences of chords associated to musical pieces. To do so, we make use of two main musical databases ([33, 26]) which contain the harmonic analysis of respectively The Beatles corpus and several classical music corpora. With respect to [23], we abandon the Tonnetz in favour of a more flexible object: a graph of chords. This allows us to freely adjust chord relations and also to consider non-symmetric distances between chords. The latter plays, in our opinion, a central role in harmonic progressions, as it well describes the temporal asymmetry of human perception of harmony. To construct simplicial complexes over directed graphs, so without a symmetric distance, we make use of the Dowker filtration, whose benefits have been shown in [22, 25].

Introduzione

Sequenze periodiche

Il serialismo di Anatol Vieru. Nel contesto del serialismo musicale degli anni '60, il compositore rumeno Anatol Vieru ha un ruolo di primaria importanza. Pupillo di Aram Khachaturian, la sua opera conta sette sinfonie e tre opere, oltre a varie composizioni per musica da camera e concerti. Il processo compositivo di numerose sue composizioni si basa su sequenze di interi e sulla loro manipolazione tramite algoritmi aritmetici. Il compositore parte da una sequenza e da questa ne ottiene di nuove applicando un operatore di somma finita o differenze finite. Successivamente, egli assegna a ciascuna sequenza un significato musicale (melodia, ritmo, armonia) per ciascuno strumento musicale o sezione orchestrale. Svariate sue composizioni, come la Sinfonia n.2 del 1973 e *Zone d'Oubli* sono state composte con questa tecnica. Successivamente, nel suo *Book of Modes* ([18]), Anatol Vieru stesso spiega nel dettaglio le sue scelte compositive e le trasformazioni applicate alle sequenze di interi modulari.

Studi matematici e problemi aperti. Dall'applicazione ripetuta dell'operatore somma ad alcune sequenze, riguardo la loro lunghezza (detta *periodo* di sequito) si osservano comportamenti interessanti e proprietà matematiche ricorrenti. Alcune osservazioni furono fatte da Anatol Vieru stesso in [18]: egli osservò che il periodo delle sequenze tende generalmente ad aumentare quando si applica l'operatore di somma, e per alcune sequenze questo rimane sempre una potenza di 2. Inoltre, per specifiche sequenze ci sono valori che tendono a proliferare. La prima formalizzazione matematica fu fatta da D.T. Vuza nel 1982 ([19]) e in [2] sono stati presentati i risultati fondamentali sulla decomposizione in sequenze idempotenti e nilpotenti (chiamate riducibili e riproducibili rispettivamente). In [4] è stato studiato computazionalmente il problema della proliferazione di alcuni valori nelle primitive della sequenza $[2, 1, 2, 8, 4, 1, 4, 8] \in \mathbf{P}_{12}$ was studied in a computational way. Più recentemente, in [13] il formalismo degli automi (*cellular automata*) è stato impiegato per studiare le proprietà del duale dell'operatore Δ.

Riassumiamo qui la teoria delle sequenze periodiche a valori modulari sviluppata finora e presentiamo i risultati ottenuti sul periodo delle primitive e sulla proliferazione dei valori. I risultati principali (Lemma 1.3.5, Proposition 1.5.10, Theorems 1.5.6, 1.5.9 and 2.2.1, Section 2.3.2, Section 2.4) sono parte di un lavoro congiunto con Luisa Fiorot e Alberto Tonolo (vedi [8, 9]).

Omologia persistente ed analisi armonica

Analisi Topologica dei Dati L'omologia persistente è uno dei principali strumenti nell'Analisi Topologica dei Dati, un dominio che sta rapidamente acquisendo importanza nell'analisi dei dati grazie alle profonde informazioni geometriche che fornisce. La sua popolarità tra gli analisti è in crescita dai primi anni 2000, ed è stata impiegata per affrontare problemi di classificazione ed estrazione di features in diversi settori scientifici. L'omologia persistente è al centro di TDA e permette di ricostruire le funzionalità topologiche di un insieme di dati. Data una filtrazione del complesso simpliciale, che tipicamente approssima la struttura geometrica sottostante di un insieme di punti, l'omologia persistente rappresenta l'evoluzione dell'omologia simpliciale (più precisamente, dei numeri di Betti sotto forma di diagrammi di persistenza o di codici a barre) nella filtrazione. Questo permette di raccogliere informazioni topologiche dal set di dati e features dai codici a barre che possono essere ulteriormente utilizzate negli algoritmi di Machine Learning per scopi di clustering e classificazione.

Il principale riferimento per l'omologia persistente e la sua attuale implementazione è [42], che presenta la formalizzazione matematica e un algoritmo efficiente per il calcolo dell'omologia persistente, mentre un'ampia introduzione al suo uso nell'analisi dei dati può essere trovata in [28, 36]. Un'interessante rassegna delle tecniche di Machine Learning applicate ai diagrammi di persistenza è [35]. L'omologia persistente è stata utilizzata con successo in problemi di classificazione, dalla classificazione delle impronte digitali ([31]) al riconoscimento generale delle immagini ([21]).

Omologia persistente e musica. Più recentemente, l'omologia persistente è stata utilizzata anche nella ricerca matematico-musicale, con l'obiettivo di collegare le caratteristiche topologiche dei codici a barre alla proprietà musicali dei dataset. In particolare, in [23] l'omologia persistente è usata insieme al Tonnetz per la classificazione automatica degli stili. In [24], la filtrazione Vietoris-Rips viene utilizzata per calcolare l'omologia persistente di una partitura musicale e per visualizzare le caratteristiche tematiche dei brani musicali.

In questo lavoro ci concentriamo sull'armonia e, più precisamente, sulla complessità armonica. Analogamente a [23], l'obiettivo è riconoscere le proprietà armoniche (*densità armonica*) di sequenze di accordi associati a brani musicali. Per farlo, utilizziamo due principali database musicali ([33, 26]) che contengono l'analisi armonica rispettivamente del corpus dei Beatles e di diversi corpora di musica classica. Rispetto [23], abbandoniamo il Tonnetz in favore di un oggetto più flessibile: un grafo di accordi. Questo ci permette di modificare liberamente le relazioni tra accordi e anche di considerare distanze non simmetriche tra di essi. Quest'ultimo aspetto svolge, a nostro avviso, un ruolo centrale nelle progressioni armoniche, in quanto descrive bene l'asimmetria temporale della percezione umana dell'armonia. Per costruire complessi simpliciali su grafi orientati, quindi senza una distanza simmetrica, sfruttiamo della filtrazione Dowker, i cui benefici sono stati mostrati in [22, 25].

Resumé

Séquences périodiques

Le sérialisme de Anatol Vieru. Dans le contexte du sérialisme musical des années 1960, le compositeur roumain Anatol Vieru joue un rôle de premier plan. Élève d'Aram Khachaturian, son œuvre compte sept symphonies et trois opéras, ainsi que diverses compositions pour musique de chambre et concerts. Le processus de composition de plusieurs de ses compositions est basé sur des séquences d'entiers et leur manipulation par des algorithmes arithmétiques. Le compositeur part d'une séquence et en obtient de nouvelles en appliquant un opérateur de somme finie ou des différences finies. Ensuite, il attribue à chaque séquence une signification musicale (mélodie, rythme, harmonie) pour chaque instrument de musique ou section orchestrale. Plusieurs de ses compositions, comme la Symphonie n.2 de 1973 et *Zone d'Oubli* ont été composées avec cette technique. Ensuite, dans son *Book of Modes* ([18]), Anatol Vieru lui-même explique en détail ses choix de composition et les transformations appliquées aux séquences d'entiers modulaires.

Mathematical studies and open questions. À partir de l'application répétée de l'opérateur somme à certaines séquences, en ce qui concerne leur longueur (appelée *période*), on observe des comportements intéressants et des propriétés mathématiques récurrentes. Certaines observations ont été faites par Anatol Vieru lui-même dans [18] : il a observé que la période des séquences tend généralement à augmenter quand on applique l'opérateur de somme, et pour certaines séquences cela reste toujours une puissance de 2. En outre, pour des séquences spécifiques, il y a des valeurs qui ont tendance à proliférer. La première formalisation mathématique a été faite par D.T. Vuza en 1982 ([19]) et en [2] ont été présentés les résultats fondamentaux sur la décomposition en séquences idempotenti et nilpotente (appelées réductibles et reproductibles respectivement). Dans [4] le problème de la prolifération de certaines valeurs dans les primitives de la séquence $[2, 1, 2, 8, 4, 4, 8] \in \mathbf{P}_{12}$ was studied in a computational way a été étudié. Plus récemment, in [13] le formalisme des automates (*cellular automata*) a été utilisé pour étudier les propriétés duales de l'opérateur Δ.

Nous résumons ici la théorie des séquences périodiques à valeurs modulaires développée jusqu'à présent et présentons les résultats obtenus sur la période des primitives et sur la prolifération des valeurs. Les résultats principaux (Lemma 1.3.5, Proposition 1.5.10, Theorems 1.5.6, 1.5.9 and 2.2.1, Section 2.3.2, Section 2.4) font partie d'un travail conjoint avec Luisa Fiorot et Alberto Tonolo (voir [8, 9]).

Homologie Persistent et Analyse Harmonique

Analyse topologique des données L'homologie persistante est l'un des principaux outils de l'Analyse Topologique des Données, un domaine qui a aquis rapidement de l'importance dans l'analyse des données grâce aux informations géométriques profondes qu'il fournit. Sa popularité parmi les analystes a augmenté depuis le début des années 2000, et a été utilisé pour traiter des problèmes de classification et d'extraction de caractéristiques dans différents domaines scientifiques. L'homologie persistante est au cœur de la TDA et permet de reconstruire les fonctionnalités topologiques d'un ensemble de données. Etant donné une filtration des complexes simpliciaux, qui se rapproche typiquement à la structure géométrique sous-jacente d'un ensemble de points, l'homologie persistante représente l'évolution de l'homologie simpliciale (plus précisément, des nombres de Betti sous forme de diagrammes de persistance ou de codes-barres) dans la filtration. Cela permet de reconstituer des informations topologiques de l'ensemble de données et des caractéristiques, à partir des codes-barres, qui peuvent être utilisées dans les algorithmes de Machine Learning à des fins de clustering et de classification.

La principale référence pour l'homologie persistante et sa implémentation actuelle est [42], qui présente la formalisation mathématique et un algorithme efficace pour le calcul de l'homologie persistante, alors qu'une large introduction à son utilisation dans l'analyse des données peut être trouvée dans [28, 36]. Une analyse intéressante des techniques d'apprentissage automatique appliquées aux diagrammes de persistance est [35]. L'homologie persistante a été utilisée avec succès dans les problèmes de classification, de la classification des empreintes digitales ([31]) à la reconnaissance générale des images ([21]).

Homologie persistante et musique. Plus récemment, l'homologie persistante a également été utilisée dans la recherche mathématique et musicale, dans le but de relier les caractéristiques topologiques des codes-barres aux propriétés musicales des ensembles de données. En particulier, dans [23] l'homologie persistante est utilisée avec le Tonnetz pour le classement automatique des styles. Dans [24], la filtration Vietoris-Rips est utilisée pour calculer l'homologie persistante d'une partition de musique et pour afficher les caractéristiques thématiques des morceaux de musique.

Nous nous concentrons sur l'harmonie et, plus précisément, sur la complexité harmonique. Comme en [23], l'objectif est de reconstituer les propriétés harmoniques (*densité harmonique*) des séquences d'accords associées à des morceaux de musique. Nous utilisons deux bases de données musicales ([33, 26]) qui contiennent l'analyse harmonique du corpus des Beatles et de plusieurs corpus de musique classique respectivement. Par rapport à [23], nous abandonnons le Tonnetz au profit d'un objet plus souple : un graphe d'accords. Cela nous permet de modifier librement les relations entre les accords et aussi de considérer des distances non symétriques entre eux. Ce dernier aspect joue, à notre avis, un rôle central dans les progressions harmoniques, car il décrit bien l'asymétrie temporelle de la perception humaine de l'harmonie. Pour construire des complexes simpliciaux sur des graphes orientés, donc sans avoir une distance symétrique, nous utilisons la filtration Dowker, dont les avantages ont été présentés dans [22, 25].

Contents

Part I

On Vieru's periodic sequences

Chapter 1

Periodic sequences

In this chapter we study periodic sequences with values over modular integers. We will mainly focus on the period and its behaviour with respect to the main operators on such sequences: the differential operator Δ and the primitive operator Σ. In Section 1.1 we introduce the periodic sequences over $\mathbb{Z}_m$ and the decomposition in p-parts. In Section 1.2 we introduce the main operators and their first properties. In Section 1.3 we study the properties of nilpotent and idempotent sequences and their link with the base prime. In Section 1.4 we introduce the notion of generating vector of a sequence and provide some basic results about it. In Section 1.5 we provide a formula for the period of a definitive primitive (Theorem 1.5.6) of a generic sequence, through the study of the period of primitives of constant sequences. This completely answers the questions on the evolution of the period of a periodic sequence when the operator Σ is repeatedly applied.

1.1 First definitions and properties

1.1.1 The shifting operator on sequences

Let us fix a positive integer m and denote by $\mathbb{N}$ the non-negative integers. Denote $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$ the ring of integers modulo m and consider the $\mathbb{Z}_m$-algebra $\mathbf{S}_m = \mathbb{Z}_m^{\mathbb{N}}$ of all the functions from $\mathbb{N}$ to $\mathbb{Z}_m$, i.e. $\mathbf{S}_m$ is a $\mathbb{Z}_m$-module together with the component-wise multiplication: if $f, g \in \mathbf{S}_n$ and $a \in \mathbb{Z}_m$, define

$$\begin{aligned}(a \cdot f)(n) :=& af(n)\\ (f+g)(n) :=& f(n)+g(n)\\ (fg)(n) :=& f(n)g(n).\end{aligned}$$

The diagonal embedding of $\mathbb{Z}_m$ into $\mathbf{S}_m$ is a ring morphism and its image is given by constant sequences. Notice that an element f of $\mathbf{S}_m$ can be identified with the sequence of its values:

$$(f(n))_{n\geq 0}.$$

We will usually call *sequences* the elements of $\mathbf{S}_m$ and we will call $f(n)$ the n-th coefficient of f.

We denote by $\mathrm{id} \in \mathrm{End}(\mathbf{S}_m)$ the identity operator and by $\theta \in \mathrm{End}(\mathbf{S}_m)$ the shifting operator, defined as:

$$\theta(f)(n) := f(n+1)$$

for every $f \in \mathbf{S}_m$.

Example. Consider $f \in \mathbf{S}_m$ defined by $f(n) = n \mod m$ for every $n \in \mathbb{N}$. Then $f = (0, 1, 2, 3, 4, 5, \dots) \in \mathbf{S}_m$ and $\theta f = (1, 2, 3, 4, 5, 6, \dots)$. Here we wrote the numbers $1, 2, 3$ meaning their coset modulo m, as $f(n) \in \mathbb{Z}_m$ for every n. Notice that $f(n + hm) = n + hm = n = f(n)$ for every $h \in \mathbb{N}$, thus f is periodic. For example if $m = 3$, the sequence f is $(0, 1, 2, 0, 1, 2, ...)$. In this case, notice that if we apply 3 times θ, we obtain again the sequence f:

$$\theta^3 f = \theta^2(1, 2, 0, 1, 2, 0, \dots) = \theta(2, 0, 1, 2, , 0, 1, \dots) = (0, 1, 2, 0, 1, 2, \dots).$$

These kinds of sequences will be the ones of our interest and we formalize the periodicity property in the following definition.

Definition. Given a sequence $f \in \mathbf{S}_m$, we say that it is *periodic* if there exists $j \geq 1$ such that $\theta^j(f) = f$, i.e. $f \in \ker(\theta^j - \mathrm{id})$.

Let us denote $\mathbf{P}_m^j := \ker(\theta^j - \mathrm{id})$ and consider

$$\mathbf{P}_m := \bigcup_{j\geq 1} \mathbf{P}_m^j \subset \mathbf{S}_m$$

the set of all periodic sequences in $\mathbf{S}_m$. It is a $\mathbb{Z}_m$-subalgebra of $\mathbf{S}_m$, since if $f \in \mathbf{P}_m^j$ and $g \in \mathbf{P}_m^i$, then $f + g, fg \in \mathbf{P}_m^\ell$ where $\ell = \mathrm{lcm}(j, i)$.

Definition. Given a periodic sequence $f \in \mathbf{P}_m$, we say that it has period τ if $f \in \mathbf{P}_m^\tau \setminus \mathbf{P}_m^d$ for every proper divisor d of τ, i.e. if τ is the minimum integer such that $\theta^\tau f = f$.

This defines the function:

$$\begin{aligned} \mathfrak{p} : \mathbf{P}_m &\longrightarrow \mathbb{N}_{>0} \\ f &\longmapsto \mathfrak{p}(f). \end{aligned}$$

Example. In our previous example $f = (0, 1, 2, 0, 1, 2, \dots) \in \mathbf{S}_3$, the period of f is 3. More generally, the sequence $(0, 1, 2, 3, 4, 5, 6, \dots)$ in $\mathbf{S}_m$ has period m.

Remark. One could extend the definition of the period on all $\mathbf{S}_m$ considering the function

$$\mathfrak{p} : \mathbf{S}_m \longrightarrow \mathbb{N}_{>0} \cup \{\, \infty \,\}$$

defining

$$\mathfrak{p}(f) = \infty \quad \forall f \in \mathbf{S}_m \setminus \mathbf{P}_m.$$

We can extend the sum and the order of $\mathbb{N}_{>0}$ to $\mathbb{N}_{>0} \cup \{\, \infty \,\}$ with the usual conventions regarding ∞.

Notation: a periodic sequence $f \in \mathbf{P}_m$ is clearly determined by its coefficients $\{f(0), \dots, f(\mathfrak{p}(f) - 1)\}$, so we denote the sequence f as:

$$f =: [f(0), \dots, f(\tau - 1)].$$

So [1] is the constant sequence with all coefficients equal to 1.

Sequences indexed in $\mathbb{Z}$. One could be interested in working with sequences indexed in $\mathbb{Z}$ instead of $\mathbb{N}$, i.e. functions $\mathbb{Z} \to \mathbb{Z}_m$. In this case, one could easily adapt what has previously been said and just consider sequences with unlimited indices on both sides: on the left going to $-\infty$ and on the right going to $+\infty$. However, once one restricts to the study of periodic sequences, this difference does not matter, as all the information of a function is carried by just a finite number of indices. More formally, the truncating morphism

$$\begin{aligned} T : \mathbb{Z}_m^{\mathbb{Z}} &\longrightarrow \mathbb{Z}_m^{\mathbb{N}} = \mathbf{S}_m \\ (f(n))_{n \in \mathbb{Z}} &\longmapsto (f(n))_{n \geq 0} \end{aligned}$$

restricts to an isomorphism between the periodic sequences in $\mathbb{Z}_m^{\mathbb{Z}}$ and the periodic sequences in $\mathbb{Z}_m^{\mathbb{N}}$. Indeed the inverse of T restricted to periodic sequences can be constructed by extending on the left periodic sequences by periodicity. More precisely, if $f \in \mathbf{P}_m$ has period τ, one defines the extension $\tilde{f} \in \mathbb{Z}_m^{\mathbb{Z}}$ of f setting:

$$\tilde{f}(n) = \begin{cases} f(n) & \text{if } n \geq 0 \\ f(r_n) & \text{if } n < 0 \end{cases}$$

where r_i is the reminder of the euclidean division of $-i$ by τ.

A good point of using sequences with indices in $\mathbb{Z}$ is that the shifting operator θ is invertible as an endomorphism of $\mathbb{Z}_m^{\mathbb{Z}}$, while it is just right-invertible on $\mathbf{S}_m$. Indeed when applied to a generic sequence $f \in \mathbf{S}_m$, we loose the information about the coefficient $f(0)$. Anyway, as far as we are treating periodic sequences, this difference will not be relevant, since θ gives an isomorphism on $\mathbf{P}_m$.

The shifting operator and the period. An easy result relates the period with the shifting operator θ:

Lemma 1.1.1. *Given $f \in \mathbf{P}_m$, if $j \geq 1$ is such that $\theta^j f = f$, then $\mathfrak{p}(f) \mid j$.*

Proof. We proceed by contradiction; suppose that $\tau := \mathfrak{p}(f)$ does not divide j. We can perform the euclidean division $j = \tau s + r$ with $0 < r < \tau$. But now we have:

$$f = \theta^j(f) = \theta^{r+\tau s}(f) = \theta^r(\theta^{\tau s}(f)) = \theta^r(f),$$

hence $f \in \ker(\theta^r - \mathrm{id})$. This contradicts the minimality of τ. □

Using that θ is a ring morphism one easily verifies

$$\mathfrak{p}(f+g), \mathfrak{p}(fg) \mid \mathrm{lcm}(\mathfrak{p}(f), \mathfrak{p}(g)).$$

One could hope that $\mathfrak{p}(f+g) = \mathrm{lcm}(\mathfrak{p}(f), \mathfrak{p}(g)) = \mathfrak{p}(fg)$; unfortunately this is not always the case, as the next example shows.

Example. Consider the sequences $[2,1]$ and $[1,2]$ in $\mathbf{P}_3$: they have both period 2, however their sum $[2+1, 1+2] = [0]$ is the constant null sequence and has period 1. The constant null sequence $[0] \in \mathbf{P}_6$ is also obtained multiplying $[2,3]$ by $[3,2]$ in $\mathbf{P}_6$.
More in general, for every m the sequences $[1,0], [0,1] \in \mathbf{P}_m$ have period 2 but $[1,0] \cdot [0,1] = [0]$ has period 1.

Let us see firstly a result about the sum, as it is the operation we will mostly be interested in.

Lemma 1.1.2. *Let $f, g \in \mathbf{P}_m$ be periodic sequences such that:*

$$\mathfrak{p}(f) = \prod_{i=1}^{s} p_i^{d_i} \qquad \mathfrak{p}(g) = \prod_{i=1}^{s} p_i^{e_i}$$

with p_i distinct primes, d_i, e_i non negative integers for every i. For every j such that $d_j \neq e_j$, we have

$$p_j^{\max\{d_j, e_j\}} \mid \mathfrak{p}(f+g).$$

Proof. Set $\tau := \mathfrak{p}(f+g)$ for brevity; we know that

$$\tau \mid \mathrm{lcm}(\mathfrak{p}(f), \mathfrak{p}(g)) = \prod_{i=1}^{s} p_i^{\max(d_i, e_i)}.$$

Fix an index j such that $d_j \neq e_j$; without loss of generality, we may suppose $d_j > e_j$ thus $\max\{d_j, e_j\} = d_j$. Now suppose by contradiction that $p_j^{d_j} \nmid \tau$; thus the maximal power of p_j that divides τ is $p_j^{d_j-1}$. Hence:

$$\tau \mid p_j^{d_j-1} \prod_{i \neq j} p_i^{\max(d_i,e_i)} =: \xi.$$

Now $\mathfrak{p}(g) \mid \xi$ so $\theta^\xi g = g$; we obtain:

$$f + g = \theta^\xi(f+g) = \theta^\xi(f) + \theta^\xi(g) = \theta^\xi(f) + g$$

which gives $f = \theta^\xi(f)$, forcing $\mathfrak{p}(f) \mid \xi$. This is a contradiction since $p_j^{d_j} \mid \mathfrak{p}(f)$ but $p_j^{d_j} \nmid \xi$. □

Remark. By previous lemma, every prime in the factorization of $\mathfrak{p}(f)$ which does not divide $\mathfrak{p}(g)$, divides $\mathfrak{p}(f+g)$ with its power in the factorization of $\mathfrak{p}(f)$. Hence if f and g have coprime periods, the sum $f+g$ has period precisely the product of the periods.

We state a weaker version of the previous lemma for the product:

Lemma 1.1.3. *Let $f, g \in \mathbf{P}_m$ be periodic sequences such that for each $j \geq 0$, $g(j)$ is not a zero divisor in $\mathbb{Z}_m$. Let*

$$\mathfrak{p}(f) = p^h \prod_{i=0}^{s} \alpha_i^{d_i} \qquad \mathfrak{p}(g) = \prod_{i=0}^{s} \alpha_i^{e_i}$$

with α_i, p distinct primes, h, d_i, e_i integers for every $i \leq s$ and $h \neq 0$. Then $p^h \mid \mathfrak{p}(fg)$.

Proof. The proof of the previous lemma works for this lemma by replacing every sum of sequences with the multiplication. Notice that we need $g(j)$ not to be a zero divisor in $\mathbb{Z}_m$ (i.e. invertible) for every j in order to have the implication

$$fg = \theta^\xi(f) g \quad \Longrightarrow \quad f = \theta^\xi(f)$$

in the last part of the proof. □

1.1.2 Decomposition in p-parts

If $m \in \mathbb{N}_{>0}$ is a positive integer, the factorization in primes

$$m = \prod_{i=1}^{s} p_i^{\ell_i}$$

gives a decomposition of the abelian group $\mathbb{Z}/m\mathbb{Z}$ into the direct sum of the abelian groups $\mathbb{Z}/p_i^{\ell_i}\mathbb{Z}$; formally, there is an isomorphism of groups:

$$\begin{aligned}\phi : \mathbb{Z}/m\mathbb{Z} &\longrightarrow \bigoplus_{i=1}^{s} \frac{\mathbb{Z}}{p_i^{\ell_i}\mathbb{Z}}\\ x \bmod m &\longmapsto (x \bmod p_i^{\ell_i})_{1\leq i\leq s}\end{aligned}$$

This is the well known Chinese Remainder Theorem. This result can be extended to the sequences in $\mathbf{P}_m$ just taking the component-wise decomposition:

$$\begin{aligned}\Phi : \mathbf{P}_m &\longrightarrow \bigoplus_{i=1}^{s} \mathbf{P}_{p_i^{\ell_i}}\\ (f(j))_{0\leq j<\mathfrak{p}(f)} &\longmapsto ((f(j) \bmod p_i^{\ell_i})_{0\leq j<\mathfrak{p}(f)})_{1\leq i\leq s}.\end{aligned}$$

This is an isomorphism of abelian groups, so whenever we have a sequence $f \in \mathbf{P}_m$ and we want to study it, we can study instead its projections in $\mathbf{P}_{p_j^{\ell j}}$. We denote by

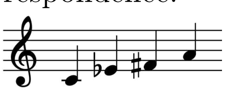
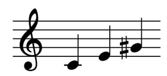
$$\pi_i : \mathbf{P}_m \longrightarrow \mathbf{P}_{p_i^{\ell_i}}$$

the projection and we call $f_{p_i} := \pi_i(f)$ the p_i-part of f . One has $f = \sum \tilde{f_{p_i}}$ where $f_{p_i} = \Phi^{-1}(f_{p_i})$. To keep a clean notation, we can omit the power index ℓ_i in the definition of π_i since ℓ_i is completely determined by m.

Example. Let us have a closer look at $\mathbf{P}_{12}$. Indeed $\mathbb{Z}_{12}$ is the most natural choice in a musical context, as it represents the pitch classes of the notes in an octave. With this correspondence, the decomposition in p-parts of $\mathbb{Z}_{12}$ (thus of $\mathbf{P}_{12}$) can be translated in the decomposition of an octave as the three possible translations of a diminished chord. Indeed if

$$\mathbb{Z}_4 \oplus \mathbb{Z}_3 \simeq \{0, 3, 6, 9\} \oplus \{0, 4, 8\},$$

we obtain the correspondence:

$\{0, 3, 6, 9\} \leftrightarrow$  and $\{0, 4, 8\} \leftrightarrow$ 

It is clear that every pitch class in $\mathbb{Z}_{12}$ can be uniquely described as a sum of a note from the diminished 7th chord $\{0, 3, 6, 9\}$ and a note from the augmented chord $\{0, 4, 8\}$.

Notice that in general the p_i-parts f_{p_i} of $f \in \mathbf{P}_m$ may not have the same period of f, as the following example shows.

Example. Consider the following sequence of period 16 in $\mathbf{P}_{12}$

$$f = [8, 10, 11, 1, 5, 1, 2, 10, 2, 4, 5, 7, 11, 7, 8, 4]$$

As $12 = 2^2 \cdot 3$, we know that

$$\frac{\mathbb{Z}}{12\mathbb{Z}} \simeq \frac{\mathbb{Z}}{2^2\mathbb{Z}} \oplus \frac{\mathbb{Z}}{3\mathbb{Z}}$$

and the p-parts of f are respectively:

$$\begin{aligned} f_2 =& [0, 2, 3, 1, 1, 1, 2, 2, 2, 0, 1, 3, 3, 3, 0, 0] \in \mathbb{Z}/4\mathbb{Z} \\ f_3 =& [2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1] \in \mathbb{Z}/3\mathbb{Z}. \end{aligned}$$

Notice that in fact $f_3 = [2, 1]$ is a sequence of period 2; so f projects to a sequence of period 16 in $\mathbf{P}_4$ and to a sequence of period 2 in $\mathbf{P}_3$. The Chinese Remainder Theorem gives:

$$f = \tilde{f}_2 + \tilde{f}_3 = [0, 6, 3, 9, 9, 9, 6, 6, 6, 0, 9, 3, 3, 3, 0, 0] + [8, 4].$$

We prove in the following lemma that the period of a sequence coincides with the least common multiple of the periods of its p-parts.

Lemma 1.1.4. *Consider $m = \prod_{i=1}^{s} p_i^{\ell_i}$ and take $f \in \mathbf{P}_m$. Then $\mathfrak{p}(f) = \operatorname{lcm}\{\,\mathfrak{p}(f_{p_i})\,\}_{1 \le i \le s}$.*

Proof. We denote shortly $h := \operatorname{lcm}\{\,\mathfrak{p}(f_{p_i})\,\}$. Since $f = \sum_{i=1}^{s} \tilde{f}_{p_i}$ and $\mathfrak{p}(\tilde{f}_{p_i}) = \mathfrak{p}(f_{p_i})$, by Lemma 1.1.2 one obtains $\mathfrak{p}(f) \mid h$.

For the converse: it is clear that the operator θ commutes with the projections π_i. So the following diagram is commutative:

$$\begin{tikzcd} \mathbf{P}_m \arrow[r, "\pi_i"] \arrow[d, "\theta^{\mathfrak{p}(f)}"] & \mathbf{P}_{p_i^{\ell_i}} \arrow[d, "\theta^{\mathfrak{p}(f)}"] \\ \mathbf{P}_m \arrow[r, "\pi_i"] & \mathbf{P}_{p_i^{\ell_i}} \end{tikzcd}$$

thus $\theta^{\mathfrak{p}(f)} f_{p_i} = f_{p_i}$ and by Lemma 1.1.1 we obtain $\mathfrak{p}(f_{p_i}) \mid \mathfrak{p}(f)$. Since this holds for every $i = 0, \ldots, s$, we obtain $h \mid \mathfrak{p}(f)$ as wanted.

□

This lemma, together with the isomorphism $\mathbf{P}_m \simeq \bigoplus \mathbf{P}_{p_i^{\ell_i}}$, allows us to reduce to sequences in $\mathbf{P}_{p_i^{\ell_i}}$ in several cases.

1.2 Differential and primitive operators

Similarly to the derivation and the integration of a real valued function, one may consider the analogous operations for function defined on $\mathbb{N}$. One finds respectively the operator Δ and the operator Σ. In this section we introduce these operators and their first properties.

1.2.1 Differential operator

Definition. Given a positive integer m, we define the *differential* operator

$$\Delta := \theta - \mathrm{id} \in \mathrm{End}(\mathbf{S}_m).$$

It is immediate to verify that Δ commutes with θ, i.e. $\Delta \circ \theta = \theta \circ \Delta$. Hence Δ restricts to an endomorphism $\Delta \in \mathrm{End}(\mathbf{P}_m)$. From now on, we will always consider Δ as an operator on $\mathbf{P}_m$.

Example. Consider the sequence $f = [0, 1, 2, 3] \in \mathbf{P}_4$; we have:

$$\Delta f = \theta f - f = [1, 2, 3, 0] - [0, 1, 2, 3] = [1].$$

Clearly Δ does not preserve the period.

Example 1.2.1. Consider the following sequences:

$$\begin{aligned} j =& [2, 11, 9, 7, 2, 2, 11, 9, 7, 4, 4, 0, 11, 9, 6, 2, 2, 0, 9, 11] \in \mathbf{P}_{12} \\ r =& [1, 1, 1, 1, 4] \in \mathbf{P}_5. \end{aligned}$$

Consider the correspondence:

$$\begin{aligned} \mathbb{Z}_{12} \longleftrightarrow& \{C, C\sharp, D, D\sharp, E, F, F\sharp, G, G\sharp, A, A\sharp, B\} \\ \mathbb{Z}_5 \longleftrightarrow& \{0, \text{♪}, \text{♩}, \text{♩.}, \text{𝅗𝅥}\} \end{aligned}$$

where 0 represents the null duration. Together, the sequences j and r provide the first notes of Jingle Bells theme, as shown in Figure 1.1

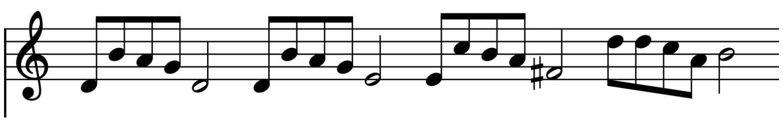


Figure 1.1: Jingle Bells theme

Here j describes the pitch classes of the theme and r the duration of each note. In musical terms, the operator Δ gives the sequence of intervals of the input sequence. So Δj is the sequence of the intervals between the pitch classes of j, while Δr is the sequence of differences of the durations of r. Hence, from a musical point of view, the operator Δ is quite natural, especially if one considers it acting on pitch classes: a fixed theme can be described by the pitch classes of its notes, or by the datum of a starting pitch class and the sequence of intervals. From a compositional point of view, a natural idea is to consider the sequence of intervals of a starting theme as a sequence of pitch classes that describes a new theme. So for example, keeping the durations of r and using the sequence

$$\Delta j = [9, 10, 10, 7, 0, 9, 10, 10, 9, 0, 8, 11, 10, 9, 8, 0, 10, 9, 2, 3]$$

for the pitch classes, one obtains the theme shown in Figure 1.2

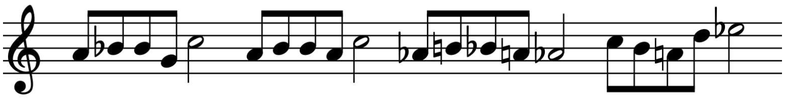


Figure 1.2: Theme obtained by deriving the pitch classes of Jingle Bells

Returning to the mathematical properties, a first observation is that the period of $\Delta(f)$ divides the period of f. Indeed if f has period τ, we have:

$$\Delta f(n) = f(n+1) - f(n) = f(n+1+\tau) - f(n+\tau) = \Delta f(n+\tau)$$

and we conclude with Lemma 1.1.1.

Notice that Δ is linear and satisfies a sort of Leibniz rule, as we expect from a derivation:

$$\begin{aligned} \Delta(f+g) =& \Delta(f) + \Delta(g) \\ \Delta(fg) =& \Delta(f)\theta(g) + f\Delta(g). \end{aligned}$$

Even if the product formula seems asymmetric with respect to f and g, a quick computation ensures that $\Delta(fg) = \Delta(gf)$, as the interested reader can verify.

We introduce some definitions:

Definition. Fix a positive integer m and consider a periodic sequence $f \in \mathbf{P}_m$. We say that:

- f is Δ-nilpotent if there exists $\eta \in \mathbb{N}_{>0}$ such that $\Delta^\eta(f) = 0$. We call the minimum η satisfying this property the *nilpotency index* of f.
- f is Δ-idempotent if there exists $\eta \in \mathbb{N}_{>0}$ such that $\Delta^\eta(f) = f$. We call the minimum η satisfying this property the *idempotency index* of f.

We will often write just *nilpotent* and *idempotent*, omitting Δ, as there is no possible confusion.

Example. Any constant sequence is nilpotent of index 1: $\Delta[c] = \theta[c] - [c] = 0$.

A less trivial case is given by (2 1) $\in \mathbf{P}_4$:

$$\begin{aligned}\Delta[2,1] =&[1,2] - [2,1] = [3,1]\\ \Delta[3,1] =&[1,3] - [3,1] = [2]\\ \Delta[2] =& \,0\end{aligned}$$

i.e. $\Delta^3[2,1] = 0$, so $[2,1]$ is nilpotent of index 3.

An example of an idempotent sequence is $[2,1]$ itself but considered in $\mathbf{P}_3$:

$$\Delta[2,1] = [1,2] - [2,1] = [2,1].$$

In this case $[2,1]$ is idempotent of index 1.

Remark. From a musical point of view, the meaning of nilpotent and idempotent sequences is quite natural. Indeed if we consider the melodic interpretation of periodic sequences, where the coefficients represent the pitch classes of a theme, a theme is idempotent if it can be obtained as the interval sequence of one of its derivatives. A theme is nilpotent if one obtains the constant sequence [0] after a suitable repetition of the interval operator Δ. In the context of Vieru's compositional approach of collecting sequences from an initial one by repeatedly applying Δ, these concepts clearly play a relevant role, as idempotent and nilpotent sequences both pose a limit in the number of sequences collectible with this procedure. Unsurprisingly, the notation used in previous works ([1, 2, 3, 4, 19]) is more evocative from the musical point of view: idempotent (resp. nilpotent) sequences have been named *reproducible* (resp. *reducible*). Here we chose to adopt *idempotent* and *nilpotent*, which are more common in mathematics.

Definition. Fix a positive integer $m = \prod_{i=1}^{s} p_i^{\ell_i}$. We define:

$$\begin{aligned}\mathbf{N}_m &:= \bigcup_{j\geq 1} \ker(\Delta^j) \subset \mathbf{P}_m\\ \mathbf{I}_m &:= \bigcup_{j\geq 1} \ker(\Delta^j - \mathrm{id}) \subset \mathbf{P}_m.\end{aligned}$$

$\mathbf{N}_m$ is the set of the nilpotent sequences and $\mathbf{I}_m$ is the set of idempotent sequences. In fact they are submodules of $\mathbf{P}_m$, as the sum of two nilpotent (resp. idempotent) sequences is again nilpotent (resp. idempotent), since the nilpotency (resp. idempotency) index of a sum is less or equal than (resp. divides) the maximum (resp. least common multiple) of the nilpotency (resp. idempotency) indices.

Lemma 1.2.2. *([3, Prop. 6, Prop. 13]) With the notation above, let us consider a sequence $f \in \mathbf{P}_m$ with p_i-parts f_{p_i}. Then f is nilpotent (resp. idempotent) if and only if its p_i-part f_{p_i} is nilpotent (resp. idempotent) for every i. The nilpotency (resp. idempotency) index η coincides with the maximum (resp. least common multiple) of the nilpotency (resp. idempotency) indices η_i of f_{p_i}.*

Proof. From the decomposition in p-parts we have $\mathbf{P}_m \simeq \bigoplus \mathbf{P}_{p_i^{\ell_i}}$ and the projections $\pi_i : \mathbf{P}_m \to \mathbf{P}_{p_i^{\ell_i}}$ commute with θ. Thus they commute also with Δ so:

$$\begin{aligned} \Delta^j f = 0 &\Longleftrightarrow \Delta^j f_{p_i} = 0 \quad \forall i \text{ and} \\ \Delta^j f = f &\Longleftrightarrow \Delta^j f_{p_i} = f_{p_i} \quad \forall i. \end{aligned}$$

The previous equation permits to easily prove the statement.

□

The Fitting Lemma. We look now at the $\mathbb{Z}_m$-module

$$\mathbf{P}_m^j = \ker(\theta^j - \mathrm{id}) \subset \mathbf{S}_m.$$

It is a finite module, hence in particular it is both artinian and Noetherian. Furthermore, it is Δ-stable so we can consider the restriction of Δ as an endomorphism of $\mathbf{P}_m^j$. In these hypotheses, Fitting Lemma holds:

Lemma 1.2.3. *If f is an endomorphism of a left R-module M that is both artinian and Noetherian, we have the decomposition:*

$$M = f^\infty M \oplus f^{-\infty} 0$$

where $f^\infty M = \bigcap_{i=0}^\infty f^i M$ and $f^{-\infty}0 = \bigcup_{i=1}^\infty \ker f^i$. Furthermore, the restriction of f to $f^\infty M$ is an automorphism and the restriction of f to $f^{-\infty}0$ is nilpotent.

Proof. See Jacobson, Basic Algebra 2, pg. 114. □

In our situation, we get the decomposition:

$$\mathbf{P}_m^j = \Delta^\infty \mathbf{P}_m^j \oplus \Delta^{-\infty} 0$$

where in fact

$$\Delta^{-\infty}0 = \mathbf{N}_m \cap \mathbf{P}_m^j.$$

Furthermore, since $\mathbf{P}_m^j$ is finite, every non zero element in $\Delta^\infty \mathbf{P}_m^j$ is idempotent and clearly every idempotent element in $\mathbf{P}_m^j$ is in $\Delta^\infty \mathbf{P}_m^j$, thus we get the equality

$$\Delta^\infty \mathbf{P}_m^j = \mathbf{I}_m \cap \mathbf{P}_m^j.$$

Moreover, the decomposition of the Fitting Lemma can be extended to the whole $\mathbf{P}_m$: indeed for every $f \in \mathbf{P}_m$, there exists j such that $f \in \mathbf{P}_m^j$ and for this j the Fitting Lemma holds. We have in fact:

$$\mathbf{P}_m = \mathbf{I}_m \oplus \mathbf{N}_m.$$

So every periodic sequence can be decomposed in a unique way as the sum of a idempotent sequence and a nilpotent sequence.

This result can be proven also in a more direct and explicit way. Take $f \in \mathbf{P}_m$ of period τ and consider the set $A = \{\Delta^j f \mid j \in \mathbb{N}\}$. A is a subset of the set of sequences having the period dividing τ, hence A is finite. So there must exist $i \neq j$ such that $\Delta^i f = \Delta^j f$. So take the minimal $M \in \mathbb{N}$ such that there exists $u < M$ satisfying

$$\Delta^M f = \Delta^u f.$$

If $t := M - u$ then $\Delta^{t+u} f = \Delta^u f$. Define $\bar{k}$ to be the minimal $k \in \mathbb{N}$ such that $kt \geq u$. Denote:

$$f_I := \Delta^{\bar{k}t} f \qquad f_N := f - f_I. \tag{1.1}$$

Lemma 1.2.4. *With the above notation, $f = f_I + f_N$ is the unique decomposition of f as a sum of an idempotent and a nilpotent sequence. The sequence f_N (resp. f_I) has nilpotency (resp. idempotency) index u (resp. t).*

Proof. The sequence f_I is idempotent since

$$\Delta^t f_I = \Delta^t(\Delta^{\bar{k}t} f) = \Delta^{\bar{k}t-u}(\Delta^{u+t} f) = \Delta^{\bar{k}t-u}(\Delta^u f) = \Delta^{\bar{k}t} f = f_I.$$

The minimality of t comes from the fact that $\{\Delta^{x+i} f \mid 0 \leq i \leq t-1\}$ has cardinality t. The sequence f_N is nilpotent since

$$\Delta^u f_N = \Delta^u(f - f_I) = \Delta^u f - \Delta^{\bar{k}t+u} f = 0.$$

The minimality of u follows from the minimality of $\bar{k}$.
This decomposition is unique: by contradiction take $f = f_I' + f_N'$. One has that $f_I - f_I' = f_N' - f_N$ is both nilpotent and idempotent thus it is equal to 0. □

The primes decomposition, Lemma 1.2.2 and Lemma 1.2.4 imply the following isomorphisms:

$$\mathbf{P}_m = \bigoplus_{i=1}^{t} \mathbf{I}_{p_i^{\ell_i}} \oplus \mathbf{N}_{p_i^{\ell_i}} \qquad \mathbf{I}_m = \bigoplus_{i=1}^{t} \mathbf{I}_{p_i^{\ell_i}} \qquad \mathbf{N}_m = \bigoplus_{i=1}^{t} \mathbf{N}_{p_i^{\ell_i}}.$$

Let us see what happens to the period with respect to this decomposition:

Lemma 1.2.5. *Let $f \in \mathbf{P}_n$ be a period sequence of period τ. If*

$$f = f_I + f_N$$

is the decomposition of f where $f_I \in \mathbf{I}_n$ and $f_N \in \mathbf{N}_n$, then both f_I and f_N have period dividing τ.

Proof. By Lemma 1.1.1, it is enough to show that $\theta^\tau(f_I) = f_I$ and $\theta^\tau(f_N) = f_N$. To do so, we apply θ^τ to the equality:

$$f = f_I + f_N;$$

we get:

$$f = \theta^\tau(f) = \theta^\tau(f_I + f_N) = \theta^\tau(f_I) + \theta^\tau(f_N).$$

Since θ commutes with Δ, $\theta^\tau(f_N)$ is nilpotent and $\theta^\tau(f_I)$ is idempotent.This provides another decomposition of f into nilpotent and idempotent part and the uniqueness of this decomposition forces:

$$\theta^\tau(f_I) = f_I \qquad \theta^\tau(f_N) = f_N.$$

□

1.2.2 Primitive operator

Definition. Given $f \in \mathbf{P}_m$, we say that F is a discrete primitive of f if $\Delta(F) = f$.

Now we fix periodic sequence $f \in \mathbf{P}_m$ of period τ and a constant $c \in \mathbb{Z}_m$. We define the primitive of f with constant c:

$$\Sigma_c f(n) = \begin{cases} c \text{ if } n = 0 \\ f(n-1) + \Sigma_c f(n-1) \text{ if } n > 0. \end{cases}$$

Lemma 1.2.6. *Given $f \in \mathbf{P}_n$, we have that*

$$\Sigma_c f(i+1) = c + \sum_{j=0}^{i} f(j) \quad \forall i \in \mathbb{N}.$$

Proof. We proceed by induction on $i \in \mathbb{N}$:

- $i = 0$: we have by definition $\Sigma_c f(1) = f(0) + \Sigma_c f(0) = f(0) + c$;

- suppose that the statement is true for i, we prove it for $i+1$:

$$\Sigma_c f(i+1) = f(i) + \Sigma_c f(i) = f(i) + \sum_{j=0}^{i-1} f(j) + c = c + \sum_{j=0}^{i} f(j).$$

□

Remark. From the previous lemma we easily see that $\Sigma := \Sigma_0$ is linear and that the following identities hold:

$$\begin{aligned}\Sigma_c f =& c + \Sigma f\\ \Sigma_c(f+g) =& \Sigma f + \Sigma_c g = \Sigma_c f + \Sigma g.\end{aligned}$$

We will call Σf the *primitive* of f. To avoid heavy notation, for every $s \in \mathbb{N}_{\geq 1}$ we will denote by $f^s = \Sigma^s f$ the s-th primitive of f.

We would like to understand what is the behaviour of the period when we take the primitive of a sequence: unfortunately it does not remain fixed, but it evolves as shown in the following lemma, which requires a preliminary definition.

Definition. If $f \in \mathbf{P}_m$ has period τ, the *trace* of f is defined as $\mathrm{tr} f := \sum_{i=0}^{\tau-1} f(i)$.

Lemma 1.2.7. *Given $f \in \mathbf{P}_m$ with period τ. The period of Σf is $h\tau$, where h is the additive order of $\mathrm{tr} f$ in $\mathbb{Z}_m$.*

Proof. First we verify that $\Sigma f(n+h\tau) = \Sigma f(n)$ for every $n \in \mathbb{N}$. By Lemma 1.2.6 we have:

$$\Sigma f(n) = \sum_{i=0}^{n-1} f(i) \qquad \Sigma f(n+h\tau) = \sum_{i=0}^{n+h\tau-1} f(i).$$

Hence it suffices to show that for every $n \in \mathbb{N}$

$$\sum_{i=n}^{n+h\tau-1} f(i) = 0.$$

Since f has period τ, the following equalities hold:

$$\sum_{i=n}^{n+h\tau-1} f(i) = \sum_{i=1}^{h\tau} f(i) = h \sum_{i=1}^{\tau} f(i) = h\mathrm{tr} f = 0.$$

Now we show that $h\tau$ is precisely the period of Σf. Let ℓ be the period of Σf. Clearly $\ell \mid h\tau$. Furthermore, as $\Delta(\Sigma f) = f$, one has:

$$\mathfrak{p}(\Delta(\Sigma f)) = \mathfrak{p}(f) = \tau \mid \mathfrak{p}(\Sigma f) = \ell$$

thus we can write $\ell = h'\tau$ for some positive integer h'. Putting together the two relations, one gets:

$$h'\tau = \ell \mid h\tau$$

hence $h' \mid h$. To conclude it suffices to show that $h \mid h'$. By definition of the period ℓ, we have for every $n \in \mathbb{N}$:

$$0 = \Sigma f(n+\ell) - \Sigma f(n) = \sum_{i=n}^{n+\ell-1} f(i) = \sum_{i=n}^{n+h'\tau-1} f(i) = \sum_{i=1}^{h'\tau} f(i) = h'\mathrm{tr}f.$$

Since h is the order of $\mathrm{tr}f$, one has $h \mid h'$ thus $h = h'$ and $\ell = h\tau$, as wanted. □

Example. As an example of computation of primitives, we consider the sequences

$$\begin{aligned} j =&[2, 11, 9, 7, 2, 2, 11, 9, 7, 4, 4, 0, 11, 9, 6, 2, 2, 0, 9, 11] \in \mathbf{P}_{12} \\ r =&[1, 1, 1, 1, 4] \in \mathbf{P}_5. \end{aligned}$$

from Example 1.2.1. One has:

$$\begin{aligned} \Sigma j =&[0, 2, 1, 10, 5, 7, 9, 8, 5, 0, 4, 8, 8, 7, \dots] \in \mathbf{P}_{12} \\ \Sigma r =&[0, 1, 2, 3, 4, 3, 4, 0, 1, 2, 1, 2, 3, 4, 0, 4, 0, 1, 2, 3, 2, 3, 4, 0, 1] \in \mathbf{P}_5. \end{aligned}$$

Since $\mathrm{tr}(j) = 10$ has order 6 in $\mathbb{Z}_{12}$, by Lemma 1.2.7 the sequence Σj has period 120. The sequence r has trace equal to $3 \in \mathbb{Z}_5$ which has order 5, thus Σr has period 25.

From the musical point of view, the operator Σ can be interpreted as the inverse of Δ: given the sequence j, we consider it as a sequence of intervals and Σj provides the corresponding sequence of pitch classes with starting note 0, i.e. C. For example, in Figure 1.3 the theme is obtained using Σj as pitch classes and Σr as durations, where the value 0 corresponds to a 8th note rest.

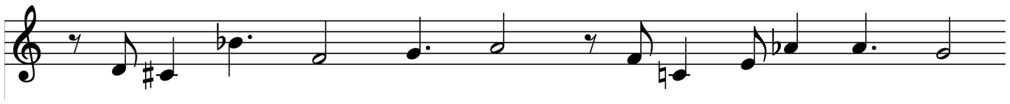


Figure 1.3: Pitches given by Σj and durations by Σr.

1.3 On nilpotent and idempotent sequences.

The decomposition of Lemma 1.2.4 turns out to be a powerful tool for studying periods of sequences. In this section, we present some fundamental links between the period of nilpotent (resp. idempotent) sequences in $\mathbf{P}_{p^\ell}$ and the base prime p. We will make great use of these result in the following.

1.3.1 Nilpotent sequences.

We now study in more detail the properties of nilpotent sequences. We will often consider sequences on $\mathbf{P}_{p^\ell}$: nilpotency turns out to be strongly related to the base prime, as the following theorem shows.

Theorem 1.3.1. *Let $f \in \mathbf{P}_{p^\ell}$ be a periodic sequence. Then:*

1. *$f \in \mathbf{N}_{p^\ell}$ if and only if $\mathfrak{p}(f) = p^t$ for $t \in \mathbb{N}$;*

2. *if $f \in \mathbf{N}_{p^\ell}$ with period p^t and nilpotency index η, then $\eta \leq \ell p^t$.*

Proof. We proceed by induction on ℓ; we denote resp. by (1_k) and (2_k) the statements 1. and 2. at the step k.

- $(k = 1)$ We first show (1_1). If $f \in \mathbf{N}_p$ then there exists $h \geq 1$ such that $\Delta^h(f) = 0$ and if t is such that $p^t \geq h$ we have $\Delta^{p^t} f = 0$. But then:

  $$0 = \Delta^{p^t} f = (\theta - \mathrm{id})^{p^t} f \overset{(*)}{=} (\theta^{p^t} - \mathrm{id}) f = \theta^{p^t} f - f$$

  where the equality $(*)$ holds since in $\mathbb{Z}_p$ one has $(a-b)^{p^s} = a^{p^s} - b^{p^s}$. Now we get $\theta^{p^t} f = f$, thus the period of f divides p^t and so it is a power of p.

  Now suppose $\mathfrak{p}(f) = p^t$. Similarly to the previous part, we have:

  $$0 = (\theta^{p^t} - \mathrm{id}) f = (\theta - \mathrm{id})^{p^t} f = \Delta^{p^t} f$$

  thus $f \in \mathbf{N}_{p^k}$.

  For (2_1): looking at the previous equivalence, we immediately get $\eta \leq p^t$ as η is the minimum integer satisfying $\Delta^\eta f = 0$. Notice that in the case $k = 1$ we have also $p^{t-1} < \eta$, since $p^{t-1} \geq \eta$ would imply $0 = \Delta^{p^{t-1}} f = (\theta^{p^{t-1}} - \mathrm{id}) f$ and then $\mathfrak{p}(f) = p^t \mid p^{t-1}$, a contradiction.

- We suppose that (1_k) and (2_k) are true and we prove (1_{k+1}) and (2_{k+1}). We show (1_{k+1}):

  $(\Longleftarrow)$ Suppose $f \in \mathbf{P}_{p^{k+1}}$ of period $\mathfrak{p}(f) = p^t$. We show that f is Δ-nilpotent and if η is the nilpotency order, one has $\eta \leq (k+1)p^t$. First we consider the usual projection $\rho_1 : \mathbb{Z}_{p^{k+1}} \longrightarrow \mathbb{Z}_p$. This morphism gives a map:

  $$\begin{aligned} \bar{\rho}_1 : \mathbf{P}_{p^{k+1}} &\longrightarrow \mathbf{P}_p \\ f &\longmapsto \bar{\rho}_1(f) := \rho_1 \circ f. \end{aligned}$$

  If we denote by θ_p (resp. $\theta_{p^{k+1}}$) the shifting operator in $\mathbf{P}_p$ (resp. $\mathbf{P}_{p^{k+1}}$), one notices immediately that

  $$\theta_p(\rho_1 \circ f) = \rho_1(\theta_{p^{k+1}}(f)) \tag{1.2}$$

so $\mathfrak{p}(\rho_1 \circ f) \mid p^t$. So by the inductive hypothesis (1_1) we have $\rho_1 \circ f \in \mathbf{N}_p$ and by (2_1) we have $\eta_{\rho_1 \circ f} \leq \mathfrak{p}(\rho_1 \circ f) \leq p^t$. Then

$$\Delta^{p^t}(\rho_1 \circ f) = 0.$$

Since ρ_1 and Δ commute by (1.2), we obtain that $(\rho_1 \circ \Delta^{p^t})f = 0$ so

$$\Delta^{p^t} f \in \ker \bar{\rho}_1 \lesssim \mathbf{P}_{p^k},$$

i.e. $\ker \bar{\rho}_1$ is isomorphic to a submodule of $\mathbf{P}_{p^k}$.

Hence we have $\Delta^{p^t} f \in \mathbf{P}_{p^k}$ and $\mathfrak{p}(\Delta^{p^t} f) \mid \mathfrak{p}(f) = p^t$. Thus by inductive hypothesis (1_k), $\Delta^{p^t} f \in \mathbf{N}_{p^k}$ holds and by (2_k) holds also $\eta_{\Delta^{p^t} f} \leq kp^t$. Then

$$0 = \Delta^{kp^t}(\Delta^{p^t} f) = \Delta^{(k+1)p^t} f$$

and we get that $f \in \mathbf{N}_{p^{k+1}}$. We obtain also $\eta \leq (k+1)p^t$ so we proved (2_{k+1}).

$(\Longrightarrow)$ Let $f \in \mathbf{N}_{p^{k+1}}$ be Δ-nilpotent. We show that $\mathfrak{p}(f) = p^t$ for a suitable $t \in \mathbb{Z}$. Take the projection $\rho_k : \mathbb{Z}_{p^{k+1}} \longrightarrow \mathbb{Z}_{p^k}$ and consider as above the induced map:

$$\bar{\rho}_k : \mathbf{P}_{p^{k+1}} \longrightarrow \mathbf{P}_{p^k}.$$

Similarly to the previous case, $\bar{\rho}_k \circ \theta_{p^{k+1}} = \theta_{p^k} \circ \bar{\rho}_k$ so we have that

$$0 = \bar{\rho}_k(\Delta^\eta f) = \Delta^\eta \bar{\rho}_k f.$$

Thus $\bar{\rho}_k f \in \mathbf{N}_{p^k}$ and by inductive hypothesis (1_k) one has $\mathfrak{p}(\bar{\rho}_k f) = p^t$. Now there exists $n \in \mathbb{N}$ such that both the following are true:

$$\theta^{p^n}(\bar{\rho}_k f) = \bar{\rho}_k f$$
$$\Delta^{p^n} f = 0.$$

Indeed it suffices to take n a multiple of m such that $p^n \geq \eta$. We have that

$$g := (\theta^{p^n} - \mathrm{id})f \in \ker(\bar{\rho}_k) \lesssim \mathbf{P}_p.$$

Now $g \in \mathbf{N}_p$ since

$$\Delta^{p^n} g = \Delta^{p^n}\theta^{p^n} f - \Delta^{p^n} f = \theta^{p^n}\Delta^{p^n} f - \Delta^{p^n} f = 0.$$

Using the hypotheses (1_1) and (2_1) one obtains $\mathfrak{p}(g) = p^h$ and $p^{h-1} < \eta_g \leq p^h$ [1]. Since $\Delta^{p^n} g = 0$, one has $\eta_g \leq p^n$ thus $p^{h-1} < \eta_g \leq p^n$ holds and so $p^h \mid p^n$. In particular, $\theta^{p^n}(g) = g$. Then we can use the identities:

$$\begin{cases} \theta^{p^n} g = g \\ \theta^{p^n} f = f + g \end{cases}$$

[1] In the case $k = 1$ we proved also $p^{h-1} < \eta_g$

to develop the following:

$$\begin{aligned}\theta^{p^{n+1}} f = \theta^{pp^n} f =&\theta^{(p-1)p^n} \circ \theta^{p^n}(f) = \theta^{(p-1)p^n}(f+g)\\ =&\theta^{(p-2)p^n} \circ \theta^{p^n}(f+g) = \theta^{(p-2)p^n}(\theta^{p^n} f + \theta^{p^n} g)\\ =&\theta^{(p-2)p^n}(f+g+g) = \ldots = f + pg = f\end{aligned}$$

using $pg = 0$ as $g \in \mathbf{P}_p$. Hence we conclude that $\mathfrak{p}(f) \mid p^{n+1}$, as wanted.

□

The previous theorem and Lemma 1.2.2 allow to easily prove the following corollary:

Corollary 1.3.2. *Consider $f \in \mathbf{N}_m$ with $m = \prod p_i^{\ell_i}$ and denote $\tau_i = \mathfrak{p}(f_{p_i})$. Then the nilpotency index η of f satisfies:*

$$\eta \leq \max_i \ell_i p_i^{\tau_i}.$$

The first statement of Theorem 1.3.1 has been already proven in [2] and it provides a powerful condition to describe and recognise nilpotent sequences over a base prime p. The second statement of the theorem and the previous corollary give an upper-bound for the nilpotent index: we will use it in Section 1.3.3.

The p-periodised sequence. Fix $\tau \in \mathbb{N}_{>0}$ and let p^r, $r \geq 0$, be the maximum power of p dividing τ, so $\tau = p^r q$ with $p \nmid q$. We consider the following operator on $\mathbf{P}_m$:

$$\mathrm{per}_{\tau,p} := \sum_{j=0}^{q-1} \theta^{jp^r} : \mathbf{P}_m \to \mathbf{P}_m.$$

Lemma 1.3.3. *The operator $\mathrm{per}_{\tau,p}$ sends a sequence of period dividing τ to a sequence whose period divides p^r.*

Proof. Take f of period dividing τ; then $\theta^\tau f = f$. Now:

$$\begin{aligned}\theta^{p^r}\,\mathrm{per}_{\tau,p}(f) =&\theta^{p^r} \sum_{j=0}^{\tau/p^r-1} \theta^{jp^r} f = \sum_{j=0}^{\tau/p^r-1} \theta^{p^r}\theta^{jp^r} f = \sum_{j=0}^{\tau/p^r-1} \theta^{(j+1)p^r} f\\ =&\sum_{j=1}^{\tau/p^r} \theta^{jp^r} f \overset{(*)}{=} \sum_{j=0}^{\tau/p^r-1} \theta^{jp^r} f = \mathrm{per}_{\tau,p}(f)\end{aligned}$$

where the equality $(*)$ holds since for $j = \tau/p^r$ we have

$$\theta^{\frac{\tau}{p^r}p^r} f = \theta^\tau f = f = \theta^0 f.$$

Thus the period of $\mathrm{per}_{\tau,p}(f)$ divides p^r. □

Definition. If $f \in \mathbf{P}_m$ has period τ, $\operatorname{per}_{\tau,p}(f)$ is called the p-periodised sequence of f.

It is possible to recover explicitly the nilpotent part of a sequence in $\mathbf{P}_{p^k}$ using the p-periodised sequence.

Lemma 1.3.4. *Let $f \in P(p^\ell)$ be a sequence of period τ and let p^r be the maximum power of p dividing τ, i.e. $\tau = p^r q$ with $p \nmid q$. Then the p^r-periodised sequence of f coincides with its nilpotent component multiplied by the coefficient $q \mod p^\ell$.*

Proof. We consider the decomposition

$$f = f_I + f_N$$

with $f_I \in \mathbf{N}_{p^\ell}$ and $f_N \in \mathbf{I}_{p^\ell}$; notice that Lemma 1.3.1 forces $\mathfrak{p}(f_N) = p^s$ with $s \in \mathbb{N}$ and Lemma 1.2.5 gives $\mathfrak{p}(f_N) \mid \mathfrak{p}(f) = \tau = p^r q$ with $p \nmid q$, so $s \leq r$ and $\theta^{p^r}(f_N) = f_N$. We write just $\operatorname{per}_p(f)$ for the p-periodised sequence of f as there is no possible confusion. We get:

$$\operatorname{per}_p(f) = \operatorname{per}_p(f_I) + \operatorname{per}_p(f_N) = \sum_{j=0}^{q-1} \theta^{jp^r} f_I + \sum_{j=0}^{q-1} \theta^{jp^r} f_N.$$

Since θ commutes with Δ, we have that $\operatorname{per}_{\tau,p}(f_I)$ is Δ-idempotent, by Lemma 1.3.3 it has order dividing p^r and then it is also Δ-nilpotent by Lemma 1.3.1. Thus we obtain

$$\operatorname{per}_p(f_I) = 0.$$

On the other hand, since $\theta^{p^r} f_N = f_N$, we get:

$$\operatorname{per}_p(f_N) = \sum_{j=0}^{q-1} \theta^{jp^r} f_N = \sum_{j=0}^{q-1} f_N = q f_N.$$

So we proved $\operatorname{per}_p(f) = q f_N$ as wanted. □

The following lemma gives a very useful description of nilpotent sequences using primitives of constant sequences.

Lemma 1.3.5. *Every sequence $f \in \mathbf{N}_n$ is a finite sum of primitives of constant sequences.*

Proof. We construct such primitives of constant sequences. Let η be the nilpotency order of f; for all $i = 1, \dots, \eta$ define

$$\delta_f^i = \Delta^i f(0).$$

We state that:

$$f = \sum_{i=0}^{\eta-1} \Sigma^i \delta_f^i.$$

To prove this, we proceed by induction on η :

- $\eta = 1$ means $\Delta f = 0$, so $f = c$ constant. In this case, we correctly have
$$f = c = f(0).$$
- Suppose that the statement holds for η, we prove it for $\eta + 1$: if f has nilpotency order $\eta + 1$, then Δf has nilpotency order η and by inductive hypothesis we have:
$$\Delta f = \sum_{i=0}^{\eta-1} \Sigma^i(\delta_{\Delta f}^i) = \sum_{i=0}^{\eta-1} \Sigma^i(\Delta^i(\Delta f)(0)) = \sum_{i=0}^{\eta-1} \Sigma^i(\Delta^{i+1} f(0)) = \sum_{i=0}^{\eta-1} \Sigma^i(\delta_f^{i+1}).$$

But now we know that

$$f = \Sigma_{f(0)} \Delta f = f(0) + \Sigma \Delta f,$$

and substituting Δf with the previous equation we get:

$$\begin{aligned} f =& f(0) + \Sigma(\sum_{i=0}^{\eta-1} \Sigma^i(\delta_f^{i+1}) \\ =& f(0) + \sum_{i=0}^{\eta-1} \Sigma^{i+1}(\delta_f^{i+1}) \\ =& \delta_f^0 + \sum_{i=1}^{\eta} \Sigma^i(\delta_f^i) \\ =& \sum_{i=0}^{\eta} \Sigma^i(\delta_f^i). \end{aligned}$$

□

Example. Consider the sequence $f = [3, 1, 2, 0, 1, 3, 0, 2] \in \mathbf{P}_4$; its derivatives are:

$$\Delta f = [2, 1] \qquad \Delta^2 f = [3, 1] \qquad \Delta^3 f = [2] \qquad \Delta^4 f = [0].$$

Hence with the notation of the previous lemma, one has:

$$\delta_f^0 = [3] \qquad \delta_f^1 = [2] \qquad \delta_f^2 = [3] \qquad \delta_f^3 = [2]$$

and

$$\begin{aligned} f =& \delta_f^0 + \Sigma^1 \delta_f^1 + \Sigma^2 \delta_f^2 + \Sigma^3 \delta_f^3 \\ =& [3] + \Sigma^1[2] + \Sigma^2[3] + \Sigma^3[2] \\ =& [3] + [0, 2] + [0, 0, 3, 1, 2, 2, 1, 3] + [0, 0, 0, 2]. \end{aligned}$$

1.3.2 Idempotent sequences.

We focus now on the idempotent sequences. Unfortunately we don't have a general result like Theorem 1.3.1. One could think that the idempotent sequences in $\mathbf{P}_{p^\ell}$ have period prime to p, but this is not true. Of course one has:

Lemma 1.3.6. *Let $f \in \mathbf{P}_{p^\ell}$ be a periodic sequence of period $\tau = p^r h$ with $p \nmid h$. Then $f \in \mathbf{I}_{p^\ell}$ if and only if the p^r-periodised sequence of f is zero.*

Proof. Easily comes from Lemma 1.3.4. □

Example. The sequence

$$f = [1,1,1,0,0,2,0,0,0,2,2,2,0,0,1,0,0,0] \in \mathbf{P}_3$$

has the 3-periodised sequence equal to zero. Hence it is idempotent and has period 18.

Lemma 1.3.7. *If $f \in \mathbf{I}_{p^\ell}$, then* $\mathrm{tr} f = 0$.

Proof. If η is the idempotency index of f, from $\Delta^\eta f = f$ one gets:

$$\Delta^{\eta-1} f = \Sigma f + \gamma$$

where $\gamma = \Delta^{\eta-1} f(0)$. Now $\Delta^j f$ has period τ for every j, as f is idempotent, and from Lemma 1.2.7 Σf has period $h\tau$ where h is the additive order of $\mathrm{tr} f$ in $\mathbb{Z}_{p^\ell}$. Thus $h = 1$ and $\mathrm{tr} f = 0$. □

Corollary 1.3.8. *If $f \in \mathbf{P}_{p^\ell}$ is a sequence of period τ prime to p, then $f \in \mathbf{I}_{p^\ell}$ if and only if* $\mathrm{tr} f = 0$.

Proof. One implication comes from the previous lemma. For the other implication, suppose $\mathrm{tr} f = 0$ and let f_I and f_N the idempotent and the nilpotent part of f respectively. Clearly $\mathrm{tr} f = \mathrm{tr} f_I + \mathrm{tr} f_N$ and from the previous lemma $\mathrm{tr} f_I = 0$, so $\mathrm{tr} f_N = 0$. Now from Lemma 1.3.4 f_N has the same period of the p-periodised of f, which is a constant sequence since τ is prime to p. Thus f_N is a constant sequence and $\mathrm{tr} f_N = 0$ forces $f_N = (0)$. Hence f is idempotent. □

1.3.3 Ranks and indices

In view of the previous results, one question naturally rises: for a fixed period τ, *how many* nilpotent and idempotent sequences exist in $\mathbf{P}_{p^\ell}$? Let us put this in a more formal way. Let us consider $\mathbf{P}^t_{p^\ell} = \ker(\theta^t - \mathrm{id})$, the $\mathbb{Z}_{p^\ell}$-module of sequences having period dividing t. It is a free module and since $\mathbb{Z}_{p^\ell}$ is a PID, every submodule of $\mathbf{P}^t_{p^\ell}$ is free. In particular the modules $\mathbf{I}^t_{p^\ell} := \mathbf{I}_{p^\ell} \cap \mathbf{P}^t_{p^\ell}$ and

$\mathbf{N}^t_{p^\ell} := \mathbf{I}_{p^\ell} \cap \mathbf{P}^t_{p^\ell}$ of idempotent and nilpotent sequences with period dividing t are free and we are interested in finding their ranks and possibly canonical sets of generators. We divide the study in three cases, based on the divisibility of t by p.

- If $t = p^j$ is a power of p, then from Theorem 1.3.1 all the sequences with period dividing t are nilpotent. Thus for every j one has $\mathbf{P}^{p^j}_{p^\ell} = \mathbf{N}^{p^j}_{p^\ell}$ and $\mathbf{I}^{p^j}_{p^\ell} = 0$. In particular, $\mathbf{N}^{p^j}_{p^\ell}$ has rank p^j and one can take the canonical set of generators $\{\varepsilon_i\}_{0\le i<p^j}$ where ε_i is the periodic sequence having as coefficients 1 in the position i and 0 otherwise.
- If t is prime to p, then by Corollary 1.3.8 the idempotent sequences are precisely the ones having zero trace. Thus $\mathbf{I}^{p^j}_{p^\ell}$ has rank $t-1$ and one can take the set of generators given by $\{\varepsilon_i - \varepsilon_{t-1}\}_{0\le i<t-1}$.
- If $t = p^j q$ with q prime to p: again by Theorem 1.3.1 the nilpotent sequences are exactly the ones with period dividing p^j. Thus $\mathbf{N}^t_{p^\ell} \simeq \mathbf{N}^{p^j}_{p^\ell}$ has rank p^j and hence $\mathbf{I}^t_{p^\ell}$ has rank $t - p^j = p^j(q-1)$. The generators can be taken as in the first case.

1.3.4 Recurrence relations for idempotent sequences.

Consider $f \in \mathbf{I}_{p^\ell}$ having period τ and idempotency order s, i.e. $\Delta^s f = f$. Expanding the latter condition, one has for every $0 \le n < \tau$:

$$f(n) = \Delta^s f(n) = \sum_{i=0}^{s} \binom{s}{i} (-1)^{s-i} f(n+i).$$

If we denote $x_j := f(j)$, this gives a linear recurrence relation:

$$\begin{aligned} 0 &= \left(\sum_{i=0}^{s} \binom{s}{i} (-1)^{s-i} x_{n+i} \right) - x_n \\ &= c_0 x_n + c_1 x_{n+1} + \cdots + c_{s-1} x_{n+s-1} + c_s x_{n+s} \end{aligned}$$

where $c_0 = \binom{s}{i}(-1)^s - 1$ and $c_i = \binom{s}{i}(-1)^{s-i}$. The characteristic polynomial associated to this recurrence relation is:

$$\chi(x) = \sum_{i=0}^{s} c_i x^i \in \mathbb{Z}_{p^\ell}[x].$$

In [16], linear recurrence relations over $\mathbb{R}$ are studied in detail and it is proven that all the solutions are linearly generated by the roots of the associated characteristic polynomial.

Case $\ell = 1$: polynomials over finite field $\mathbb{F}_p$. Berlekamp's algorithm, eventually going to an extension of $\mathbb{F}_p$. Case $\ell \geq 2$: Hensel's lemma can lift a factorization over $\mathbb{Z}_p$ to a factorization over $\mathbb{Z}_{p^\ell}$ for every $\ell \geq 2$. In order to keep the results from [16], one needs to restrict to non zero-divisors roots of the characteristic polynomial.

1.3.5 Σ-idempotent sequences

Definition. Given $f \in \mathbf{P}_n$, we say that f is Σ-idempotent if there is $k \in \mathbb{N}_{>0}$ such that $\Sigma^k_{f(0)} f = f$. We will denote the subset of Σ-idempotent sequence of $\mathbf{P}_n$ with I_n^Σ; in fact it is a submodule of $\mathbf{P}_n$ (using the linearity of Σ).

Lemma 1.3.9. *Given $f \in \mathbf{P}_n$, one has for every i:*

$$\Sigma^i_{f(0)} f(j) = 2^j f(0) \quad \forall j \leq i.$$

Proof. See Ancellotti Lemma 2.2.6. □

The previous lemma helps us to describe the Σ-idempotent sequences: indeed if f is Σ-idempotent, i.e. there exists n such that

$$\Sigma^n_{f(0)} = f,$$

for all $j \leq n$ we get:

$$f(j) = \Sigma^n_{f(0)} f(j) = 2^j f(0).$$

This identity can be extended to all x, since we can take a multiple of n which is greater than the period of f (and replace n with it in the previous equalities). Thus a necessary condition for a periodic sequence f to be Σ-idempotent is to be of the type:

$$f(j) = 2^j f(0) \quad \forall j.$$

Furthermore, one verifies that a periodic sequence of the type $2^j f(0)$ is Δ-idempotent of order 1:

$$\Delta f = \Delta(2^j f(0)) = 2^{j+1} f(0) - 2^j f(0) = 2^x f(0) = f.$$

But applying $\Sigma_{f(0)}$ to both sides, we get:

$$f = \Sigma_{f(0)} \Delta f = \Sigma_{f(0)} f$$

hence we proved that if f is a Σ-idempotent periodic sequence, then its idempotency order (with respect to both Δ and Σ) must be 1. Thus to check that a periodic sequence f is Σ-idempotent, it is enough to check if $\Sigma_{f(0)} f = f$.

1.4 Generating vector

Consider the following map:

$$\begin{aligned}\mathbf{vect} : \mathbf{S}_{p^\ell} &\longrightarrow \mathbf{S}_{p^\ell}\\ f &\longmapsto \mathbf{vect}(f) := \left(\Delta^i f(0)\right)_{i\geq 0}.\end{aligned}$$

We will call $\mathbf{vect}(f)$ the *generating vector* of f. By the properties of Δ, $\mathbf{vect} \in \mathrm{End}(\mathbf{S}_{p^\ell})$. Moreover, it is an isomorphism, as the values $\Delta^i f(0)$ uniquely determine the sequence f. This means that we can describe a sequence $f \in \mathbf{S}_{p^\ell}$ either by using its coefficients

$$(f(0), f(1), f(2), \dots)$$

or using the coefficients of its generating vector:

$$(f(0), \Delta f(0), \Delta^2 f(0), \dots).$$

Notice that $f \in \mathbf{I}_{p^\ell}$ of index η if and only if $\mathbf{vect}(f)$ is a periodic sequence of period η. Also, $f \in \mathbf{N}_{p^\ell}$ of index η if and only if $\mathbf{vect}(f)$ is null after η, i.e. $\mathbf{vect}_f(i) = 0$ for every $i \geq \eta$. We say that $f \in \mathbf{S}_{p^\ell}$ is a definitively null sequence if $\mathbf{vect}_f(i) = 0$ for every $i \geq M$, for a suitable $M \in \mathbb{N}$. Notice that $\mathbf{vect}$ sends idempotent sequences in periodic sequences and nilpotent sequences in definitively null sequences. From the decomposition of a generic periodic sequence in idempotent and nilpotent part, one obtains that $\mathbf{vect}$ induces an isomorphism $\mathbf{P}_{p^\ell} \simeq \mathbf{DP}_{p^\ell}$ where:

$$\mathbf{DP}_{p^\ell} := \{f \in \mathbf{S}_{p^\ell} \mid \exists\, t, M \in \mathbb{N} \text{ s.t. } \mathbf{vect}_f(M+i) = \mathbf{vect}_f(M+i+t)\ \forall\, i \geq 0\}$$

is the set of definitively periodic sequences.

$$\begin{tikzcd}
\mathbf{S}_{p^\ell} \arrow[r, "\mathbf{vect}"] & \mathbf{S}_{p^\ell} \\
\mathbf{P}_{p^\ell} \arrow[u, hook] \arrow[r, "\sim"] & \mathbf{DP}_{p^\ell} \arrow[u, hook]
\end{tikzcd}$$

If $f \in \mathbf{P}_{p^\ell}$, we can express $\mathbf{vect}_f$ with finitely many coefficients:

$$\mathbf{vect}_f = [d_0, \dots, d_{M-1}, d_M, \dots, d_{M+t-1}]$$

where M is the nilpotency index of f_N and t is the idempotency index of f_I. Indeed the coefficient d_{M+at+i} coincides with d_{M+i} for every $a \in \mathbb{N}$ and $0 \leq i < t$. Notice that the coefficients of $\mathbf{vect}_f$ correspond to the Discrete Taylor formula of f centred in 0 (see [16, Theorem 6.53]).

Definition. Given $f \in \mathbf{P}_{p^\ell}$, we call *leading component* of f the last entry of $\mathbf{vect}_f$ with minimal p-adic valuation.

1.5 Primitives of sequences and constants

In this section we will study the behaviour of the operator Σ acting on periodic sequences. Lemma 1.1.1 gives already a useful result, but it does not allow to compute immediately the period of the generic primitive $\Sigma^s f$ of a sequence $f \in \mathbf{P}_m$. In this section, we will provide a more direct formula for the period. To do so, we first prove a formula for the period of constant sequences, linking this problem to the study of binomial coefficients modulo the power of a prime. Then we will reduce the primitives of nilpotent and idempotent sequences to the sum of primitives of constants and we will give a formula for the period of their *definitive* primitives. Finally, we will put all the results together to have an explicit formula for the definitive primitives of a generic sequence $f \in \mathbf{P}_m$.

1.5.1 Constant sequences.

We proceed here to study the constant sequences in more detail. In particular we give an explicit formula for their primitives, which is the origin of the link with modular binomial coefficients.

Lemma 1.5.1. *Given the constant sequence* $[1] \in \mathbf{P}_n$*, one has:*

$$\Sigma^k[1](x) = \binom{x}{k}.$$

Proof. We proceed by induction on k:

- $(k = 1)$ We have:
  - for $x = 0$: trivially $\Sigma[1](0) = 0$;
  - for $x > 0$:
  $$\Sigma[1](x) = \sum_{i=0}^{x} 1 = x = \binom{x}{1}.$$
- Now we suppose true the statement for k and we prove it for $k+1$. We proceed by induction on x: for $x = 0$, we have:
  $$\Sigma^{k+1}[1](0) = 0 = \binom{0}{k+1}.$$
  Now suppose that $\Sigma^{k+1}[1](x) = \binom{x}{k+1}$; we prove the statement for $x + 1$:
  $$\Sigma^{k+1}[1](x+1) = \Sigma^k[1](x) + \Sigma^{k+1}[1](x) = \binom{x}{k} + \binom{x}{k+1} = \binom{x+1}{k+1}.$$

□

Corollary 1.5.2. *The constant sequence* $(a) \in \mathbf{P}_n$ *is such that:*

$$\Sigma^k(a)(x) = a \cdot \binom{x}{k}.$$

Proof. The result is immediate: as $(a) = a \cdot [1]$ and since the primitive operator Σ is linear, we have:

$$\Sigma^k(a)(x) = \Sigma^k a \cdot [1](x) = a \cdot \Sigma^k[1](x) = a \cdot \binom{x}{k}$$

using the previous lemma. □

Lemma 1.5.3. *If* $c = p^m b$ *with* $p \nmid b$, *then* $\Sigma^\ell c$ *is a sequence in* $p^m\mathbb{Z}/p^n\mathbb{Z}$ *for every* l.

Lemma 1.5.4. *Let* (c) *be a non zero constant sequence in* $\mathbf{P}_{p^n}$ *with* $(c, p) = 1$. *Let* $s \in \mathbb{N}$ *and*

$$[a_k a_{k-1} \cdots a_1 a_0]_p$$

with $a_k \neq 0$ *the representation of* s *in base* p. *If we denote* $f^s = \Sigma^s(c)$ *the* s*-th primitive sequence of* (c), *then the following equalities hold:*

$$\sum_{x=0}^{p^{n+k}-1} f^s(x) = 0 \text{ in } \frac{\mathbb{Z}}{p^n\mathbb{Z}} \qquad \text{if } s \neq \sum_{i=0}^{k}(p-1)p^i$$

$$\sum_{x=0}^{p^{n+k}-1} f^s(x) = p^{n-1}u \text{ in } \frac{\mathbb{Z}}{p^n\mathbb{Z}},\ (p, u) = 1 \qquad \text{if } s = \sum_{i=0}^{k}(p-1)p^i.$$

Proof. We are going to use the Kummer's theorem, which states that given $a \geq b \geq 0$ integers, p a prime number, then the p-adic valuation $v_p(\binom{a}{b})$ (i.e. the largest power of p dividing $\binom{a}{b}$) coincides with the number of remainders obtained when performing the sum between b and $a - b$ in base p.

First, from Corollary 1.5.2 we have:

$$\sum_{x=0}^{p^{n+k_s}-1} f^s(x) = \sum_{x=0}^{p^{n+k_s}-1} \binom{x}{s} = \binom{p^{n+k}}{s+1};$$

for the last equality see [16]. Hence we reduced to study whether $\binom{p^{n+k}}{s+1}$ is divisible by p^n or by p^{n-1}. If $s \neq \sum_{i=0}^{k}(p-1)p^i$, then $s + 1 < p^{k+1}$ so

$$s + 1 = \sum_{i=0}^{k} b_i p^i.$$

This means that in base p we can write $s+1=[b_k\cdots b_0]_p$, while clearly $p^{n+k}=10\cdots 0_p$ with $n+k$ zeros. Suppose that $h\leq k$ is the smallest index such that $b_h\neq 0$. We want to use Kummer's theorem; we write $p^{n+k}-(s+1)$ in base p:

$$\underbrace{(p-1)\cdots(p-1)}_{n}\underbrace{(p-1-b_k)(p-1-b_{k-1})\cdots(p-1-b_{h+1})(p-b_h)}_{k+1-h}\underbrace{0\cdots 0_p}_{h}$$

and it is clear that performing $(p^{n+k}-(s+1))+(s+1)$ in base p gives $n+k-h$ reminders; thus Kummer's theorem says that p^{n+k-h} divides $\binom{p^{n+k}}{s+1}$, hence also

$$p^n \mid \binom{p^{n+k}}{s+1}$$

as $h\leq k$. So we can conclude $\binom{p^{n+k}}{s+1}=0\in\mathbb{Z}/p^n\mathbb{Z}$ as wanted.

Now suppose $s=\sum_{i=0}^{k}(p-1)p^i$; in this case $s+1=p^{k+1}$. Now we write $p^{n+k}-p^{k+1}$ in base p:

$$\underbrace{(p-1)\cdots(p-1)}_{n-1}\underbrace{0\cdots 0_p}_{k+1}$$

with $k+1$ zeros and $n-1$ times $(p-1)$. Now the number of reminder in the computation $(p^{n+k}-p^{k+1})+p^{k+1}$ is $n-1$, thus from Kummer's theorem we get:

$$p^{n-1} \mid \binom{p^{n+k}}{s+1}$$

but $p^n\nmid\binom{p^{n+k}}{s+1}$, as wanted. □

Theorem 1.5.5. *Let (c) be a non zero constant sequence in $\mathbf{P}_{p^n}$ with $c=p^\ell b$, $p\nmid b$. Let $s\in\mathbb{N}$ and $[a_ka_{k-1}\cdots a_1a_0]_p$, $a_k\neq 0$, the representation of s in base p. Then the sequence $\Sigma^s[c]$ has period $p^{n-\ell+k}$.*

Proof. Denote $f=[c]$ so $f^s=\Sigma^s[c]$. It is enough to show the case c prime to p. Indeed take $c=p^\ell b\in p^\ell\mathbb{Z}/p^n\mathbb{Z}$ with $(b,p)=1$ and suppose that the statement is true for b; using the isomorphism

$$\begin{aligned}\epsilon:\frac{p^\ell\mathbb{Z}}{p^n\mathbb{Z}}&\longrightarrow\frac{\mathbb{Z}}{p^{n-\ell}\mathbb{Z}}\\ p^\ell b&\longmapsto b.\end{aligned}$$

Then the period of f^s coincides with the period of $\Sigma^s[b]$ in $\mathbb{Z}/p^{n-\ell}\mathbb{Z}$ which is equal to $p^{n-\ell}p^k$.

So we can suppose that c is prime to p. We want to show that f^s has period p^{n+k} where $s=\sum_{i=0}^{k}a_ip^i$.

We proceed by induction on s; we will write $k=k_s$ the index of the maximal non zero coefficient in the p-adic expansion of s.

- For $s = 1$ the statement is clear: in this case $k_s = 0$ and
$$f^s = c \cdot (0, 1, 2, 3, \dots, p^n - 1)$$
has period p^n as we are supposing c prime to p.

- Suppose that the statement is true for s, we prove that it is true for $s + 1$:
  - If $s \neq \sum_{i=0}^{k_s}(p-1)p^i$, notice that $k_s = k_{s+1}$. Lemma 1.5.4 says
$$\sum_{x=0}^{p^{n+k_s}-1} f^s(x) = 0$$
so by Lemma 1.2.7 one has
$$\mathfrak{p}(f_{s+1}) = \mathfrak{p}(f^s) = p^{n+k_s} = p^{n+k_{s+1}}$$
as wanted.
  - If $s = \sum_{i=0}^{k_s}(p-1)p^i$, then $k_{s+1} = k_s + 1$. Lemma 1.5.4 says
$$\sum_{x=0}^{p^{n+k_s}-1} f^s(x) = p^{n-1}u$$
with u prime to p, hence again by Lemma 1.2.7 we have
$$\mathfrak{p}(f_{s+1}) = p\mathfrak{p}(f^s) = pp^{n+k_s} = p^{n+k_s+1} = p^{n+k_{s+1}}.$$

□

1.5.2 Leading terms

Now that we have a formula for the period of the primitives of constant sequences, we provide a formula for the period of nilpotent and idempotent sequences and finally for generic sequences. Practically, we will need to study what happens to the period when we sum primitives of constants. As already observed, the period does not behave perfectly fine with respect to the sum, i.e. it is not enough to take the least common multiple of the periods of the summands. The next example gives a counterexample with primitives of constants:

Example. The sequences $\Sigma[1] = [0, 1, 2, 3]$ and $\Sigma^2[2] = [0, 0, 2, 2]$ in $\mathbf{P}_4$ have both period 4 but their sum is the sequence $[0, 1]$, having period 2.

Theorem 1.5.6. *If e_γ is the leading component of the generating vector $\boldsymbol{vect}(f)$ of a nilpotent periodic sequence $f \in \mathbf{N}_{p^\ell}$, then*

$$\mathfrak{p}(\Sigma^s f) = \mathfrak{p}(\Sigma^{s+\gamma}[e_\gamma]) \quad \textit{for } s \gg 0.$$

Proof. If $\mathbf{vect}(f) = (e_0, \ldots, e_{\eta-1})$, by Lemma 1.3.5

$$f = \sum_{i=0}^{\eta-1} [e_i]^i = \sum_{i=0}^{\eta-1} \Sigma^i [e_i].$$

If e_γ is the leading component, then $\nu_p(e_\gamma) \leq \nu_p(e_i)$, $0 \leq i < \gamma$, and $\nu_p(e_\gamma) < \nu_p(e_i)$, $\gamma < i < \eta$. Let us prove that $\mathfrak{p}(\Sigma^s f) = \mathfrak{p}(\Sigma^{s+\gamma}[e_\gamma])$ for $s \gg 0$. Let μ be the minimal natural number such that $\eta - \gamma - 1 < p^\mu(p-1)$. Notice that for any $k \geq \mu$ both p^k and $p^k + \eta - \gamma - 1$ are strictly less than p^{k+1} and hence they have the same number of digits in base p. In order to conclude the proof, we show that for any $k \geq \mu$ one has:

$$\mathfrak{p}(f^s) = \mathfrak{p}([e_\gamma]^{s+\gamma}) \qquad \forall\, p^k - \gamma \leq s < p^{k+1} - \gamma$$

hence the statement holds for any $s \geq p^\mu - \gamma$.

For $s = p^k - \gamma$ we have:

$$f^{p^k-\gamma} = [e_0]^{p^k-\gamma} + [e_1]^{1+p^k-\gamma} + \cdots + [e_\gamma]^{\gamma+p^k-\gamma} + \cdots + [e_{\eta-1}]^{\eta-1+p^k-\gamma}.$$

By Theorem 1.5.5, $[e_\gamma]^{p^k}$ has period $p^{\ell-\nu(e_\gamma)+k}$. The other summands have period strictly dividing $p^{\ell-\nu(e_\gamma)+k}$:

- For every $\gamma < i < \eta$, $p^k + i - \gamma < p^{k+1}$ by construction and so $p^k + i - \gamma$ has $k+1$ digits in base p, hence the period of $[e_i]^{p^k+i-\gamma}$ is $p^{\ell-\nu_p(e_i)+k} \mid p^{\ell-\nu_p(e_\gamma)-1+k}$ (since $\nu_p(e_i) > \nu_p(e_\gamma)$).
- For every $0 \leq i < \gamma$, $\nu_p(e_\gamma) \leq \nu_p(e_i)$ and $p^k + i - \gamma < p^k$; hence $p^k + i - \gamma$ has at most k digits in base p. Thus the period of $[e_i]^{p^k+i-\gamma}$ is a divisor of $p^{\ell-\nu_p(e_i)+k-1}$ and so it divides $p^{\ell-\nu_p(e_\gamma)+k-1}$.

Thus the period $\mathfrak{p}(f^{p^k-\gamma})$ is equal to $p^{\ell-\nu_p(e_\gamma)+k}$.

For $p^k - \gamma < s < p^{k+1} - \gamma$, the period of $[e_\gamma]^{s+\gamma}$ is $p^{\ell-\nu_p(e_\gamma)+k}$, and by Lemma 1.2.7 $\mathfrak{p}(f^s) \geq \mathfrak{p}(f^{p^k-\gamma}) = p^{\ell-\nu_p(e_\gamma)+k}$. Furthermore, since $p^{k+1} + \eta - \gamma - 1 < p^{k+2}$ we have

$$p^k - \gamma < s \leq s + \eta - 1 < p^{k+1} - \gamma + \eta - 1 < p^{k+2}.$$

Then $\mathfrak{p}([e_i]^{s+i}) \leq p^{\ell-(\nu_p(e_\gamma)+1)+k+1}$ for $\gamma < i \leq \eta - 1$, and $\mathfrak{p}([e_i]^{s+i}) \leq p^{\ell-\nu_p(e_\gamma)+k}$ for $0 \leq i \leq \gamma - 1$. Thus $\mathfrak{p}(f^s) \leq p^{\ell-\nu_p(e_\gamma)+k}$, and hence $\mathfrak{p}(f^s) = p^{\ell-\nu_p(e_\gamma)+k}$. □

Corollary 1.5.7. *Denoted by e_γ the leading component of the generating vector of $f \in \mathbf{N}_{p^\ell}$, one has that for $t \gg 0$:*

$$\mathfrak{p}(\sum_{i=0}^{t} f^i) = \mathfrak{p}([e_\gamma]^{t+\gamma}).$$

Remark. In the proof of Theorem 1.5.6 we computed explicitly how big s has to be for the statement to hold. Precisely we have $s \geq p^\mu - \gamma$ where μ is minimal with respect to $\eta - \gamma - 1 < p^\mu(p-1)$. In particular if $\gamma = \eta - 1$, the condition becomes $s \geq 0$.

Example. Let us consider the sequence $V_2 = [2,1,2,0,0,1,0,0] \in \mathbf{N}_4$. It has nilpotency index 5 and $\mathbf{vect}(V_2) = (2,3,2,3,2)$. Then

$$V_2 = [2] + \Sigma[3] + \Sigma^2[2] + \Sigma^3[3] + \Sigma^4[2].$$

The leading component is $e_3 = 3$ and by Theorem 1.5.6 one has $\mathfrak{p}(\Sigma^s V_2) = \mathfrak{p}(\Sigma^{s+3}[3])$ for $s \gg 0$. By Section 1.5.2 one easily checks that for any $s \geq 0$, if $s+3 = \lfloor a_k \cdots a_1 a_0 \rfloor_2$, $a_k > 0$, we have $\mathfrak{p}(\Sigma^s V_2) = \mathfrak{p}(\Sigma^{s+3}[3]) = 2^{2+k}$.

Corollary 1.5.8. *With the notation of the theorem above, if $\ell = 1$, i.e. considering sequences on the finite field $\mathbb{Z}_p$, the leading component of a nilpotent sequence is always $e_{\eta-1}$. Hence $\mathfrak{p}(\Sigma^s f) = \mathfrak{p}(\Sigma^{s+\eta-1}[e_{\eta-1}])$ for any $s \geq 0$.*

In the following theorem we show that constant sequences and their primitives also have a key role in studying the primitives of idempotent sequences. For $z \in \mathbb{Z}$ and $0 < \eta \in \mathbb{N}$, denote by $0 \leq \bar{z} < \eta$ the remainder in the division of z by η.

Theorem 1.5.9. *Consider $f \in \mathbf{I}_{p^\ell}$ with idempotency index η and generating vector $\boldsymbol{vect}(f) = (e_0, \ldots, e_{\eta-1})$. For every $s \geq 1$, one has:*

$$\Sigma^s f = \Delta^{\overline{-s}} f - \sum_{j=0}^{s-1} \Sigma^j [e_{\overline{j-s}}].$$

This provides the explicit decomposition in idempotent and nilpotent part. Moreover if e_γ is the leading component of $\boldsymbol{vect}(f)$, one has

$$\mathfrak{p}(\Sigma^s f) = \operatorname{lcm}\big(\mathfrak{p}(f), \mathfrak{p}(\Sigma^{s-\eta+\gamma}[e_\gamma])\big) \quad \forall s \gg 0.$$

Proof. Observe that, by definition, the constants e_i and e_j coincide if $i \equiv_\eta j$. We proceed by induction on s.

- For $s = 1$ one has $\Sigma f = \Sigma(\Delta^\eta f) = \Delta^{\eta-1} f - [e_{\eta-1}] = \Delta^{\overline{-1}} f - [e_{\overline{-1}}]$, hence the thesis.

- Suppose that the statement is true for $1 \leq s = t\eta + \bar{s}$, $t \geq 0$; let us prove it for $s+1 = t'\eta + \overline{s+1}$. Notice that $(t', \overline{s+1}) = (t+1, 0)$ if $\bar{s} = \eta - 1$ and $(t', \overline{s+1}) = (t, \bar{s}+1)$ otherwise. By inductive hypothesis we have:

$$\begin{aligned} f^{s+1} = \Sigma(f^s) =& \Sigma(\Delta^{\overline{-s}} f - \sum_{j=0}^{s-1} [e_{\overline{j-s}}]^j) \\ =& \Delta^{\overline{-s-1}} f - [e_{\overline{-s-1}}] - \sum_{j=1}^{s} [e_{\overline{j-1-s}}]^j \\ =& \Delta^{\overline{-(s+1)}} f - \sum_{j=0}^{s} [e_{\overline{j-(s+1)}}]^j. \end{aligned}$$

This proves the first part of the statement.

For the period, first we have $\tau\left(\Delta^j f\right) = \mathfrak{p}(f)$ for any $j \in \mathbb{N}$ (since $f \in \mathbf{I}_{p^\ell}$). Now let us denote by g the nilpotent sequence $\sum_{j=0}^{\eta-1} [e_j]^j$. The nilpotency index of g is $\eta_g := \max\{j : e_j \neq 0\} + 1$; clearly $\eta_g \leq \eta$, but the leading component of $\mathbf{vect}(g)$ and $\mathbf{vect}(f)$ is the same: $[e_\gamma]^\gamma$. Notice that for any $s \geq \eta$:

$$\sum_{j=0}^{s-1} [e_{\overline{j-s}}]^j = \sum_{0 \leq j < \bar{s}} [e_{\overline{j-s}}]^j + \sum_{j=\bar{s}}^{s-1} [e_{\overline{j-s}}]^j = \sum_{0 \leq j < \bar{s}} [e_{\overline{j-s}}]^j + \sum_{i=0}^{t-1} g^{i\eta+\bar{s}}.$$

By Corollary 1.5.7, for $s \gg 0$ one has:

$$\tau\left(\sum_{j=0}^{s-1} [e_{\overline{j-s}}]^j\right) = \tau\left(\sum_{i=0}^{t-1} g^{i\eta+\bar{s}}\right) = \tau\left([e_\gamma]^{(t-1)\eta+\bar{s}+\gamma}\right) = \tau\left([e_\gamma]^{s-\eta+\gamma}\right).$$

Setting $w := \eta - \gamma$ one obtains the statement:

$$\tau\left(\Sigma^s f\right) = \operatorname{lcm}\left(\mathfrak{p}(f), \tau\left(\Sigma^{s-w}[e_\gamma]\right)\right).$$

□

Remark. In the proof of Theorem 1.5.9, to make the statement relative to the period true, we did two assumptions about s. First s has to be greater or equal than the idempotency index η, second $(t-1)\eta + \bar{s} = s - \eta$ has to be greater or equal than $p^\mu - \gamma$ where μ is minimal with respect to $\eta - \gamma - 1 < p^\mu(p-1)$ to have he possibility to apply Theorem 1.5.6 as observed in Section 1.5.2.

Example. Consider the sequence $f = [1, 3, 0] \in \mathbf{P}_4$. It is idempotent of index $\eta = 6$ and $\mathbf{vect}(f) = (1, 2, 3, 1, 0, 1)$. The leading constant is $e_5 = 1$. By Theorem 1.5.9 for $s = 8$, one has:

$$\Sigma^8 f = \Delta^{\overline{-8}} f - \sum_{j=0}^{7} \Sigma^j [e_{\overline{j-8}}] = \Delta^4 f - \sum_{j=0}^{7} \Sigma^j [e_{\overline{j+4}}].$$

Indeed $\Sigma^8 f = [0,0,0,0,0,0,0,0,0,1,3,0,1,1,2,1,1,1,2,1,1,2,1,1,2,2,0,2,2,2,0,$ $2,2,3,3,2,3,3,2,3,3,3,3,2,3,3,0,1,3,0]$, $\Delta^4 f = [0,1,3]$, and

$$\sum_{j=0}^{7} \Sigma^j[e_{\overline{j+4}}] = [0,1,3,0,1,3,0,1,2,1,1,2,3,3,2,3].$$

Finally, by Section 1.5.2 if $s \geq \eta = 6$, and $s - \eta = s - 6 \geq 2^0 - 5 = -4$, i.e., globally, $s \geq 6$, we have

$$\mathfrak{p}(\Sigma^s f) = \operatorname{lcm}\left(\mathfrak{p}(f), \mathfrak{p}(\Sigma^{s-1}[1])\right) = 3 \cdot 2^{2+k}$$

where $s - 1 = \lfloor 1a_{k-1}\cdots a_0 \rfloor_2$. In particular for $s = 8$ we have

$$\mathfrak{p}(\Sigma^8 f) = \operatorname{lcm}\left(\mathfrak{p}(f), \mathfrak{p}(\Sigma^7[1])\right) = \operatorname{lcm}(3,16)) = 3 \cdot 2^{2+2} = 48.$$

Period of a generic primitive. As a consequence of the results of this sections, we can provide an explicit formula for the period of the primitives of a generic sequence $f \in \mathbf{P}_m$. If $m = \prod p_i^{\ell_i}$, by Lemma 1.2.2 it is possible to reduce to the p_i-parts of f. Thus we can suppose $f \in \mathbf{P}_{p^\ell}$ and we denote by f_I (resp. f_N) its idempotent (resp. nilpotent) part and by η_I (resp. η_N) the idempotency (resp. nilpotency) index. By Lemma 1.3.5 and Theorem 1.5.9 we can write:

$$\Sigma^s f = \Sigma^s f_I + \Sigma^s f_N = \Delta^{\eta_I - s'} f_I + \sum_{i=1}^{s} \Sigma^{s-i}[\varepsilon^{\eta_I - i'}] + \sum_{j=0}^{\eta_N - 1} \Sigma^{s+j}[\delta^j] \tag{1.3}$$

where $\varepsilon^n = \Delta^n f_I(0)$, $\delta^j = \Delta^j f_N(0)$ and $0 \leq i', s' < \eta_I$ are such that $i = i' \mod p^\ell$ and $s = s' \mod p^\ell$. Then the following result holds:

Proposition 1.5.10. *With the notation above, there is one constant $c \in \{\varepsilon^i, \delta^j\}$ such that it definitively leads the period of f, i.e. there exist $M \in \mathbb{N}$ and $u \in \mathbb{Z}$ such that for any $s \geq M$:*

$$\mathfrak{p}(\Sigma^s f) = \operatorname{lcm}\{\mathfrak{p}(f_I), \mathfrak{p}(\Sigma^{s+u}[c])\}.$$

Proof. From Equation (1.3) it is clear that

$$(\Sigma^s f)_I = \Delta^{\eta_I - s'} f_I \qquad (\Sigma^s f)_N = \sum_{i=1}^{s} \Sigma^{s-i}[\varepsilon^{\eta_I - i'}] + \sum_{j=0}^{\eta_N - 1} \Sigma^{s+j}[\delta^j].$$

In particular the idempotent part of $\Sigma^s f$ has always period $\mathfrak{p}(f_I)$. To conclude the proof, by Lemma 1.2.2 it suffices to show that there is a fixed constant c that leads the period of the nilpotent part of $\Sigma^s f$. Notice that we can rewrite the latter as:

$$(\Sigma^s f)_N = \sum_{i=1-\eta_N}^{s} \Sigma^{s-i}[c_i]$$

where

$$c_i = \begin{cases} \delta^{-i} \text{ if } 1-\eta_N \leq i \leq 0 \\ \varepsilon^{\eta_I - i'} \text{ with } i = i' \mod p^\ell. \end{cases}$$

Observe that for any $i \geq 0$ one has $c_i = c_{i+\eta_I}$, thus for any $s \geq \eta_I$ we can suppose $c \in \{c_{1-\eta_N}, c_{2-\eta_N}, \dots, c_{\eta_I - 1}\}$. Indeed the following constants coincides with these first ones but are integrated fewer times, therefore they generally have smaller period (in light of Theorem 1.5.5). Now a reasoning similar to the proof of Theorem 1.5.6 allows to conclude. □

Vieru's primitives with different constants We conclude this chapter with a complete analysis of the period of the primitives of a specific sequence, which we will call Vieru's sequence:

$$V = (2,1,2,4,8,1,8,4) \in \mathbf{P}_{12}.$$

Its idempotent part is $V_3 := [8,4]$ which corresponds to the 3-part of V:

$$v_3 = (2,1) \in \mathbf{P}_3$$

while its nilpotent part is $V_2 := [6,9,6,0,0,9,0,0]$ which corresponds to the 2-part

$$v_2 = (2,1,2,0,0,1,0,0) \in \mathbf{P}_4$$

whose primitives' period has been already studied in Section 1.5.2.

We focus here on the period of the primitives of the sequence V, eventually considering integration constants which are different from zero. This extends the computational study done in [4] and provides a definitive answer to the question of the periods, which dates back to Anatol Vieru himself.

The first case we study is the integration of V with the constant 8. One has:

$$\Sigma_8 V = \Sigma_8(V_2 + V_3) = \Sigma V_2 + \Sigma V_3 + [8].$$

Now it is worth observing that $\Sigma V_3 + [8] = \Sigma_8 V_3 = V_3$, so for any $s \in \mathbb{N}_{\geq 1}$ one has:

$$\Sigma_8^s V = \Sigma^s V_2 + \Sigma_8^s V_3 = \Sigma^s V_2 + V_3.$$

Since V_3 is idempotent and $\Sigma^s V_2$ is nilpotent $\forall\, s \geq 1$, by Lemma 1.2.5 one has:

$$\mathfrak{p}(\Sigma_8^s V) = \operatorname{lcm}(\mathfrak{p}(\Sigma^s V_2), \mathfrak{p}(V_3)) = \mathfrak{p}(\Sigma^s V_2) = \mathfrak{p}(\Sigma^s v_2).$$

For every $c \in \mathbb{Z}_{12}$ such that $c \equiv 2 \mod 3$ one has a result similar to the case $c = 8$. Indeed for example for $c = 5$ one has:

$$\Sigma_5 V = \Sigma_5(V_2 + V_3) = \Sigma V_2 + \Sigma V_3 + [5] = \Sigma V_2 + \Sigma V_3 + [8] + [9].$$

Hence for any $s \in \mathbb{N}_{\geq 1}$ one obtains:

$$\Sigma_5^s V = \Sigma^s V_2 + \sum_{i=0}^{s-1} \Sigma^i [9] + V_3.$$

Now $\sum_{i=0}^{s-1} \Sigma^i[9]$ is a nilpotent sequence that projects to 0 in $\mathbf{P}_3$, hence by Theorem 1.3.1 it has period equal to a power of 2 for every s. Nonetheless, by Theorem 1.5.6 the sequence $\Sigma^s V_2$ continues to lead the period of the nilpotent part, hence again one obtains:

$$\mathfrak{p}(\Sigma_5^s V) = \mathfrak{p}(\Sigma^s V_2) = \mathfrak{p}(\Sigma^s v_2).$$

The computation for the cases $c = 2, 11$ are perfectly similar.

If one takes $c \in \mathbb{Z}_{12}$ as constant of integration with $c \not\equiv 8 \mod 3$, then the computation is slightly different. Indeed in this case we have:

$$\Sigma_c V = \Sigma V_2 + \Sigma V_3 + [c].$$

Since $\Sigma V_3 = V_3 + [4]$ one has:

$$\Sigma_c^s V = \Sigma^s V_2 + V_3 + \sum_{i=0}^{s-1} \Sigma^i [c + 4].$$

Now we reduce to study separately the 2-part and the 3-part. If c_2 (resp. c_3) is the projection of c in $\mathbb{Z}_4$ (resp. in $\mathbb{Z}_3$), the 2-part and the 3-part coincide respectively with:

$$\Sigma^s v_2 + \sum_{i=0}^{s-1} \Sigma^i [c_2] \in \mathbf{P}_4 \qquad v_3 + \sum_{i=0}^{s-1} \Sigma^i [c_3 + 1] \in \mathbf{P}_3.$$

By Theorem 1.5.6 the term $\Sigma^s v_2$ continues to lead the period of the 2-part. For the 3-part, v_3 has period equal to 2, while the period of $\Sigma^{s-1}[c_3 + 1]$ is equal to 3^t for a suitable $t \in \mathbb{N}$ by Theorem 1.5.5. Hence:

$$\mathfrak{p}(\Sigma^s V) = \mathrm{lcm}(\mathfrak{p}(\Sigma^s v_2), 3^t).$$

Chapter 2

Proliferation of values

In this chapter we face another mathematical question arisen from Vieru's observation: the proliferation of values. In particular, we explain why the values 4 and 8 proliferate in the primitives of the sequence

$$V = [2, 1, 2, 4, 8, 1, 8, 4] \in \mathbf{P}_{12}.$$

More explicitly, one notices that the values 4 and 8 appear more and more frequently among the coefficients of $\Sigma^s V$ for $s \in \mathbb{N}_{\geq 1}$ (see [4, App. A] for more details). We provide here an algebraic explanation for this behaviour.

First, notice that studying the proliferation of such values in V is equivalent to studying the proliferation of zeros in the sequence

$$v_2 = [2, 1, 2, 0, 0, 1, 0, 0] \in \mathbf{P}_4,$$

which corresponds to its 2-part. Indeed for every $s \in \mathbb{N}_{\geq 1}$ one has $\Sigma^s V(n) \in \{4, 8\}$ if and only if $\Sigma^s v_2(n) = 0$. Thus we can reduce to studying the proliferation of zeros in the primitives of

$$v := v_2 = [2] + \Sigma[3] + \Sigma^2[2] + \Sigma^3[3] + \Sigma^4[2].$$

Let us denote $Z(s) := |\{0 \leq n < \mathfrak{p}(v^s) \mid v^s(n) = 0\}|$ the number of zeros among the coefficients of the sequence v^s. In Figure 2.1, the values of $Z(s)$ are depicted.

The values of $Z(s)$ tend to increase, as we can expect from the definition of the operator Σ. But looking closely at the successive peaks, one can observe a recursive pattern. Compare for example the values of $Z(s)$ for $29 \leq s < 61$ with the values for $93 \leq s < 125$, as well as the values for $61 \leq s < 125$ with the values for $125 \leq s < 192$. One notices that there is a periodic structure in the peaks.

Also the quantity $\frac{Z(s)}{\mathfrak{p}(v^s)}$, depicted in Figure 2.2, has some interesting aspects. It tends to increase periodically up to more than 90%, reaching its peak in $2^k - 5$ for $k \in \mathbb{N}_{\geq 0}$, before immediately dropping down to around 10%.

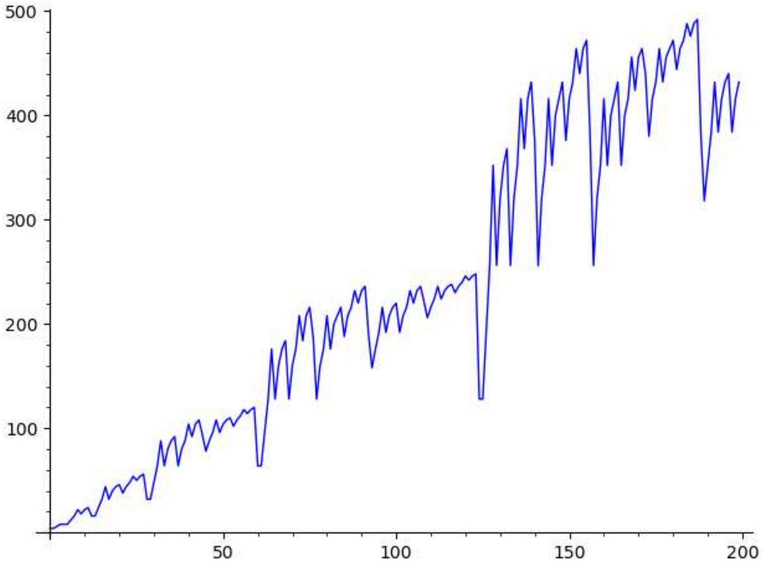


Figure 2.1: The values of $Z(s)$ for $1 \leq s \leq 200$.

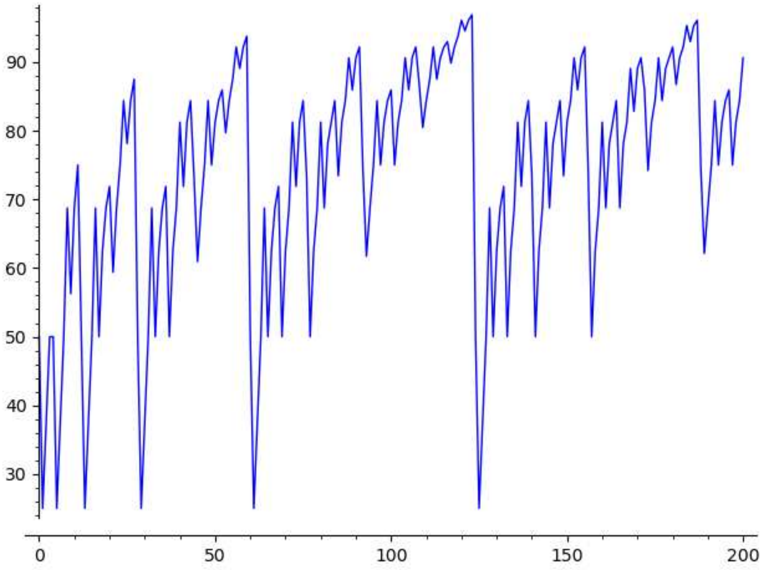


Figure 2.2: The values of $\frac{Z(s)}{\mathfrak{p}(v^s)}$ for $1 \leq s \leq 200$.

This chapter is devoted to the explanation of these observations. We will need to introduce some preliminary tools that will allow us to prove the final recursive formula for the zeros of v.

In Section 2.1, we study the primitives of the sequence $[2] \in \mathbf{P}_4$.

In Section 2.2, the primitives of the sequence v are studied in detail, giving a precise local description of the periodic peaks of $Z(s)$.
In Section 2.3, we will study modular binomial functions and prove three fundamental recursive lemmas about them.
In Section 2.4, we get back to study the sequence v and we prove the main recursive formula.

2.1 The primitives of $[2] \in \mathbf{P}_4$

Let us consider the constant sequence $f = [2]$ in $\mathbf{P}_4$. The behaviour of its period is the same as the one of the constant sequence $[1] \in \mathbf{P}_2$. We want to study f^s, the s-th primitive of f for every integer $s \geq 1$. Let the expression of s in base 2 be:

$$s = \lfloor a_k a_{k-1} \dots a_1 a_0 \rfloor_2$$

with $a_k = 1$. By Theorem 1.5.5,the sequence f^s has period 2^{k+1}. Then the following proposition holds:

Proposition 2.1.1. *Given f and $s = \lfloor a_k a_{k-1} \dots a_1 a_0 \rfloor_2$ as above, we have:*

i. *$f^s(n) = 0$ for every $0 \leq n \leq s$.*

ii. *if $a_k = a_{k-1} = \dots = a_0 = 1$, then $f^s(n) = 0$ for all $0 \leq n \leq 2^{k+1} - 1$ and $f^s(2^{k+1} - 1) = 2$.*

iii$_k$. *if $s = 2^k = \lfloor 100 \dots 00 \rfloor_2$, then $f^s(2^k + n) = 2$ for every $0 \leq n < 2^k$;*

iv$_k$. *$f^s(2^k + n) = f^{s'}(n)$ for every $0 \leq n < 2^k$ with $s' = \lfloor a_{k-1} \dots a_0 \rfloor_2$.*

Proof. *i.* By Corollary 1.5.2, one has:

$$f^s(n) = 2 * \binom{n}{s}$$

hence for every $s < n$ we have $f^s(n) = 0$.

ii. If $a_k = a_{k-1} = \dots = a_0 = 1$, by the previous point we have

$$f^s(n) = 0 \quad \forall\, 0 \leq n < 2^{k+1} - 1.$$

Notice that $\mathfrak{p}(f^s) = 2^{k+1}$ while the sequence f^{s+1} has period 2^{k+2} since $s + 1 = 2^{k+1}$, hence by Lemma 1.2.7 one has

$$0 \neq \operatorname{tr}(f^s) = \sum_{i=0}^{2^{k+1}-1} f^s(n) = f^s(2^{k+1} - 1).$$

Thus $f^s(2^{k+1} - 1) = 2$.

iii$_k$. We proceed by induction on k; the case $k = 0$ can be verified by hand. Now suppose the statement holds for $s = 2^{k-1}$, we prove that it holds for $s = 2^k$; we proceed by induction on n. For $n = 0$ we have by Corollary 1.5.2:

$$f^{2^k}(2^k) = 2 * \binom{2^k}{2^k} = 2.$$

Now suppose that the statement is true for n, we show it holds for $n + 1$:

$$f^{2^k}(2^k + n + 1) = f^{2^k - 1}(2^k + n) + f^{2^k}(2^k + n) = 0 + 2 = 2.$$

iv$_k$. Again by induction on k. Notice that the previous points already prove the statement for $s = 2^k$ for every k. Now we suppose that the statement is true for $s' = \lfloor a_k a_{k-1} \dots a_0 \rfloor_2$ and we show that it is true for $s = \lfloor a_{k+1} a_k \dots a_0 \rfloor_2$. Now

$$s = \sum_{j=0}^{k+1} a_j 2^j = 2^{k+1} + \sum_{j=0}^{k} a_j 2^j =: 2^{k+1} + s'$$

and one has:

$$f^s(2^{k+1} + n) = \Sigma^s f(2^k + n) = \Sigma^{s'} f^{2^k+1}(2^{k+1} + n) = \Sigma^{s'} f(n) = f^{s'}(n).$$

□

For the number of zeros of the primitives of the constant sequence [2] in $\mathbb{Z}/4\mathbb{Z}$, the following result shows an interesting link with the Hamming weight:

Proposition 2.1.2. *Let $s = \lfloor a_{k-1} \dots a_0 \rfloor_2$ be the index of the primitives of $f = [2] \in \mathbb{Z}/4\mathbb{Z}$. Remember that f^s has period 2^k by Theorem 1.5.5. If we denote*

$$wt(s) = |\{ 0 \leq i < k \mid a_i = 1 \}|,$$

the Hamming weight of s, then the sequence f^s has $\sum_{i=1}^{wt(s)} 2^{k-i}$ zeros.

Proof. We proceed by induction on k. For $k = 0, 1$ the statement can be verified by hand. Now suppose that the statement is true for every $j < k - 1$, we prove it is true for $k - 1$. The first 2^{k-1} coefficients of f^s are equal to zero, since $f^s(n) = 2 * \binom{n}{s} = 0$ for every $n < s$. For the coefficients $2^{k-1} + n$ we can use the previous result:

$$f^{\lfloor a_{k-1} \dots a_0 \rfloor_2}(2^{k-1} + n) = f^{\lfloor a_{k-2} \dots a_0 \rfloor_2}(n)$$

and by induction we conclude that in the last 2^{k-1} coefficients there are $\sum_{j=1}^{wt(\lfloor a_{k-2} \dots a_0 \rfloor_2)} 2^{k-1-j}$ zeros. One has

$$wt(\lfloor a_{k-2} \dots a_0 \rfloor_2) = wt(\lfloor 1 a_{k-2} \dots a_0 \lfloor_2) - 1$$

so

$$\sum_{j=1}^{wt(\lfloor a_{k-2}\ldots a_0\rfloor_2)} 2^{k-1-j} = \sum_{h=2}^{wt(\lfloor 1a_{k-2}\ldots a_0\rfloor_2)} 2^{k-h}$$

thus the number of zeros in f^s is:

$$2^{k-1} + \sum_{h=2}^{wt(\lfloor 1a_{k-2}\ldots a_0\rfloor_2)} 2^{k-h} = \sum_{h=1}^{wt(\lfloor a_{k-1}a_{k-2}\ldots a_0\rfloor_2)} 2^{k-h}.$$

□

Lemma 2.1.3. *Let $f = [2] \in \mathbf{P}_4$. For every $s \in \mathbb{N}$, the sequence $f^s + f^{s+2}$ has the same number of zeros of the sequence f^{s+2}.*

Proof. As previously observed, we can consider f to be the constant sequence [1] in $\mathbf{P}_2$. Let

$$s = \lfloor a_k \ldots a_0\rfloor_2 \qquad s+2 = \lfloor a'_k \ldots a'_0\rfloor_2$$

the expressions of the indices s and $s+2$ in base 2, with $a'_k = 1$; notice that we took k to be the same index, as one can eventually consider a_k to be 0. We are going to study separately the cases when $a_1 = 0$ and $a_1 = 1$.

- $a_1 = 0$: in this case one has $a_i = a'_i$ for every $i \neq 1$ and $a'_1 = 1$. Now let b be the index of a generic entry of f^{s+2}; by Theorem 1.5.5, f^{s+2} has period 2^k so we can write $b = \lfloor b_k \ldots b_0\rfloor_2$. We compute:

$$\begin{aligned} f^s + f^{s+2}(b) =& (f^{\lfloor a_k\ldots a_0\rfloor_2} + f^{\lfloor a'_k\ldots a'_0\rfloor_2})(\lfloor b_k \ldots b_0\rfloor_2) \\ =& \prod_{i=2}^{k} \binom{b_i}{a_i} (\binom{b_1}{a_1} + \binom{b_1}{a'_1}) \binom{b_0}{a_0} \end{aligned}$$

using that $a_i = a'_i$ for $i \neq 1$. Now $a_1 = 0$ and $a'_1 = 1$, so:

$$f^s + f^{s+2}(b) = \prod_{i=2}^{k} \binom{b_i}{a_i} (\binom{b_1}{0} + \binom{b_1}{1}) \binom{b_0}{a_0}.$$

Now if f^{s+2} has a zero in the entry b, then some of the coefficients $\binom{b_j}{a'_j}$ must be 0 (remember that we are working in $\mathbb{Z}_2$). If this is the case for some $j \neq 1$, then from the expression above also $f^s + f^{s+2}$ has a zero in b. If all the binomial coefficients (of f^{s+2}) but $\binom{b_1}{1}$ are non-zero, we consider $b' = \lfloor b_k \ldots b_2 1 b_0\rfloor_2$, i.e. $b' = b + 2$: in this entry f^{s+2} does not present a zero, since $\binom{b'_1}{a'_1} = \binom{1}{1} = 1$ and we supposed $\binom{b_j}{a'_j} \neq 0$ for every $j \neq 1$; nevertheless, the sequence $f^s + f^{s+2}$ has a zero in this position: indeed

$$\binom{b'_1}{a_1} + \binom{b'_1}{a'_1} = \binom{1}{0} + \binom{1}{1} = 1 + 1 \equiv_2 0.$$

Hence we constructed a bijection between the set of zero entries of f^{s+2} and the set of zero entries of $f^s + f^{s+2}$, so their cardinality is the same, i.e. f^{s+2} and $f^s + f^{s+2}$ have the same number of zeros.

- $a_1 = 1$. We have:

$$s = \lfloor a_k \dots a_{h+3} 0 \underbrace{1 \dots 1}_{h} 1 a_0 \rfloor_2$$

for a suitable $h \leq k - 3$. When we add 2 in base 2, we have then h reminders and we get:

$$s + 2 = \lfloor a_k \dots a_{h+3} 1 \underbrace{0 \dots 0}_{h} 0 a_0 \rfloor_2.$$

As before, consider the index b of a generic entry of f^{s+2} and its expression in base 2: $b = \lfloor b_k \dots b_0 \rfloor_2$. Again, one has:

$$\begin{aligned} f^s + f^{s+2}(b) =& (f^{\lfloor a_k \dots a_0 \rfloor_2} + f^{\lfloor a'_k \dots a'_0 \rfloor_2})(\lfloor b_k \dots b_0 \rfloor_2) \\ =& \binom{b_0}{a_0} \prod_{i=h+3}^{k} \binom{b_i}{a_i} \Big(\binom{b_{h+2}}{0} \prod_{i=1}^{h+1} \binom{b_i}{1} + \binom{b_{h+2}}{1} \prod_{i=1}^{h+1} \binom{b_i}{0} \Big). \end{aligned}$$

Each zero entry of f^{s+2} given by a factor of

$$\binom{b_0}{a_0} \prod_{i=h+3}^{k} \binom{b_i}{a_i}$$

gives a zero entry of the sum $f^s + f^{s+2}$. Suppose now that the product above is non zero; the only possibility to have a zero entry of f^{s+2} is given by the zeros of the coefficient

$$\binom{b_{h+2}}{1},$$

since for every $i = 1, \dots, h+1$ the coefficient $\binom{b_i}{0}$ is always non zero. Now the coefficient $\binom{b_{h+2}}{1}$ is zero when $b_{h+2} = 0$. Let us consider $b' = \lfloor b'_k \dots b'_0]$ with $b'_i = b_i$ for $i \in \{0\} \cup \{h+3, \dots, k\}$ and $b'_i = 1$ for $i = 1, \dots, h+2$. Notice now that by construction f^{s+2} does not have a zero in the entry b', but the sequence $f^s + f^{s+2}$ has a zero in b' since:

$$\begin{aligned} f^s + f^{s+2}(b') =& \binom{b'_0}{a_0} \prod_{i=h+3}^{k} \binom{b'_i}{a_i} (\binom{b'_{h+2}}{0} \prod_{i=1}^{h+1} \binom{b'_i}{1} + \binom{b'_{h+2}}{1} \prod_{i=1}^{h+1} \binom{b'_i}{0}) \\ =& 1 * (\binom{1}{0} + \binom{1}{1}) = 2 \equiv_2 0. \end{aligned}$$

So again we constructed a bijection between the zero entries of f^{s+2} and the zero entries of $f^s + f^{s+2}$.

□

2.2 The proliferation of values in Vieru's sequence

2.2.1 Lucas's Theorem generalisation

For what follows, we will need a powerful generalisation of Lucas's Theorem ([6, 7]), due to Davis and Webb ([5]), which is the main tool to study residue classes of binomial coefficients modulo a prime power. Let us recall it. Let $0 \leq B \leq A$ be integers and p be a prime. Denote by $A = \lfloor a_s ... a_1 a_0 \rfloor_p$ and $B = \lfloor b_s ... b_1 b_0 \rfloor_p$, $0 \leq a_i, b_i < p$, and $a_s \neq 0$, the representations of A and B in base p.

- Lucas's Theorem claims:

$$\binom{A}{B} \equiv \binom{a_s}{b_s} \binom{a_{s-1}}{b_{s-1}} \cdots \binom{a_1}{b_1} \binom{a_0}{b_0} \quad \text{mod } p.$$

  In particular, $p \mid \binom{A}{B}$ if and only if $a_i < b_i$ for some $0 \leq i \leq s$.

- Davis's and Webb's Theorem generalises this result for residue classes of a power of a prime. Consider the numbers

$$A_{i,j} = [a_i a_{i-1} \dots a_{j+1} a_j]_p, \quad B_{i,j} = [b_i b_{i-1} \dots b_{j+1} b_j]_p \quad \forall 0 \leq j \leq i \leq s.$$

  Define the following modified binomial coefficient:

  - if $B_{i,j} \leq A_{i,j}$: $\left\langle {A_{i,j} \atop B_{i,j}} \right\rangle := \binom{[A_{i,j}]_p}{[B_{i,j}]_p}$.
  - if $B_{i,i} > A_{i,i}$: $\left\langle {A_{i,i} \atop B_{i,i}} \right\rangle := p$.
  - if $B_{i,j} > A_{i,j}$: $\left\langle {A_{i,j} \atop B_{i,j}} \right\rangle := p \left\langle {A_{i-1,j} \atop B_{i-1,j}} \right\rangle$ for all $j \leq i-1$.

Let $2 \leq r \leq s+1$. If $\left\langle {A_{i,j} \atop B_{i,j}} \right\rangle = p^t\alpha$ with $p \nmid \alpha$, we define

$$\left\langle {A_{i,j} \atop B_{i,j}} \right\rangle^{-1} = p^{-t}\alpha^{-1}$$

where α^{-1} is such that $\alpha\alpha^{-1} \equiv 1 \mod p^r$. Then for each $2 \leq r \leq s+1$:

$$\binom{A}{B} \equiv \left\langle {A_{r-1,0} \atop B_{r-1,0}} \right\rangle \prod_{j=1}^{s-r+1} \left\langle {A_{j+r-1,j} \atop B_{j+r-1,j}} \right\rangle \left\langle {A_{j+r-2,j} \atop B_{j+r-2,j}} \right\rangle^{-1} \mod p^r.$$

In the sequel, we denote by $\left\langle\left\langle {A_{i,j} \atop B_{i,j}} \right\rangle\right\rangle := \left\langle {A_{i,j} \atop B_{i,j}} \right\rangle \left\langle {A_{i-1,j} \atop B_{i-1,j}} \right\rangle^{-1}$.

Example. Consider $38 = [1102]_3$ and $12 = [0110]_3$; let us compute the residue class of $\binom{32}{12}$ modulo 9. We have $p = 3$, $s = 3$, $r = 2$. One gets

$$\begin{aligned}
\binom{38}{12} = \binom{[1102]_3}{[0110]_3} &\equiv \left\langle {02 \atop 10} \right\rangle \left\langle {10 \atop 11} \right\rangle \left\langle {0 \atop 1} \right\rangle^{-1} \left\langle {11 \atop 01} \right\rangle \left\langle {1 \atop 1} \right\rangle^{-1} \mod 9 \\
&\equiv 3 \times \binom{2}{0} \times 3 \times 3 \times 3^{-1} \times \binom{4}{1} \times 1^{-1} \mod 9 \\
&\equiv 36 \equiv 0 \mod 9.
\end{aligned}$$

Now we can proceed to study the zeros of the sequence $v \in \mathbf{P}_4$. Given its decomposition in constants:

$$v = [2] + \Sigma[3] + \Sigma^2[2] + \Sigma^3[3] + \Sigma^4[2]$$

by Theorem 1.5.6 the leading term is $\Sigma^3[3]$. If $s + 3 = [a_k \cdots a_1 a_0]_2$ is the representation of $s+3$ in base 2, we have $\mathfrak{p}(v^s) = 2^{k*2}$. $\Sigma^3[3]$ being the leading term suggests the following change of indexing of the primitives of v:

Definition. For every $s \geq 3$, define $g^s := \Sigma^{s-3}V_2$ and denote by $z(s)$ the number of zeros among the coefficients in a period of g^s.

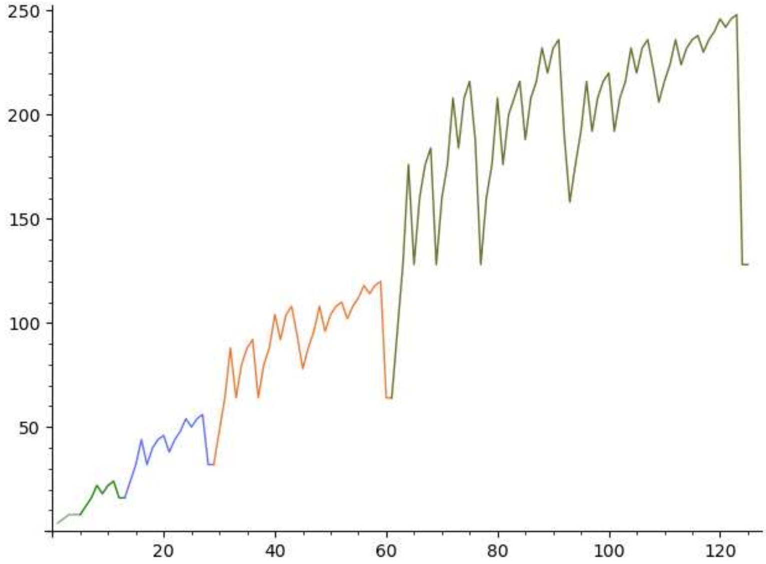


Figure 2.3: $z(s)$ for $1 \le s \le 127$.

In Figure 2.3, the function $z(s)$ is represented using different colours to highlight the change of period of g^s.

Remark. The ratio $z(s)/2^{k+2}$ represents the percentage of zeros in a period of $\Sigma^{s-3}v$, hence the percentage of 4 and 8 in a period of $(s-3)$-primitive $\Sigma^{s-3}V$ of Vieru's sequence for $2^k \le s < 2^{k+1}$.

As we are interested in the sequence $g_s \in \mathbf{P}_4$, it is convenient to compute preventively the quantities $\langle {ab \atop cd} \rangle$ and $\langle\langle {ab \atop cd} \rangle\rangle$ for $a, b, c, d \in \{0, 1\}$. One has:

$$\langle\langle {00 \atop 00} \rangle\rangle = 1;\ \langle\langle {00 \atop 01} \rangle\rangle = 2;\ \langle\langle {00 \atop 10} \rangle\rangle = 2;\ \langle\langle {00 \atop 11} \rangle\rangle = 2$$

$$\langle\langle {01 \atop 00} \rangle\rangle = 1;\ \langle\langle {01 \atop 01} \rangle\rangle = 1;\ \langle\langle {01 \atop 10} \rangle\rangle = 2;\ \langle\langle {01 \atop 11} \rangle\rangle = 2$$

$$\langle\langle {10 \atop 00} \rangle\rangle = 1;\ \langle\langle {10 \atop 01} \rangle\rangle = 1;\ \langle\langle {10 \atop 10} \rangle\rangle = 1;\ \langle\langle {10 \atop 11} \rangle\rangle = 2$$

$$\langle\langle {11 \atop 00} \rangle\rangle = 1;\ \langle\langle {11 \atop 01} \rangle\rangle = 3;\ \langle\langle {11 \atop 10} \rangle\rangle = 3;\ \langle\langle {11 \atop 11} \rangle\rangle = 1$$

and

$$\langle {00 \atop 00} \rangle = 1;\ \langle {00 \atop 01} \rangle = 0;\ \langle {00 \atop 10} \rangle = 2;\ \langle {00 \atop 11} \rangle = 0$$

$$\langle {01 \atop 00} \rangle = 1;\ \langle {01 \atop 01} \rangle = 1;\ \langle {01 \atop 10} \rangle = 2;\ \langle {01 \atop 11} \rangle = 2$$

$$\langle {10 \atop 00} \rangle = 1;\ \langle {10 \atop 01} \rangle = 2;\ \langle {10 \atop 10} \rangle = 1;\ \langle {10 \atop 11} \rangle = 0$$

$$\langle {11 \atop 00} \rangle = 1;\ \langle {11 \atop 01} \rangle = 3;\ \langle {11 \atop 10} \rangle = 3;\ \langle {11 \atop 11} \rangle = 1$$

2.2.2 Local and global peaks of $z(s)$

The following result shows that $z(s)$ is not strictly increasing and if one restricts to the values where g^s has fixed period, i.e. when $2^k \leq s < 2^{k+1}$, it has maximum value $2^{k+2} - 8$ in the point $s = 2^{k+1} - 2$.

Theorem 2.2.1. *For every $k \geq 3$ we have:*

$$2^{k+1} + 1 = z(2^{k+1} - 1) < z(2^{k+1} - 2) = 2^{k+2} - 8.$$

More precisely, one has:

$$g^{2^{k+1}-2} = \Sigma^{2^{k+1}-5} V_2 = [\underbrace{0, \ldots, 0}_{2^{k+1}-5}, 2, 3, 1, 0, 0, \underbrace{0, \ldots, 0}_{2^k-4}, 2, 2, 0, 0, \underbrace{0, \ldots, 0}_{2^k-5}, 2, 1, 3, 0, 0].$$

Proof. Fix

$$s = 2^{k+1} - 2 = \lfloor \underbrace{11 \cdots 1}_{k} 0 \rfloor_2$$

and look at the sequence

$$g^s = \Sigma^{\lfloor 1 \cdots 1011 \rfloor_2}[2] + \Sigma^{\lfloor 1 \cdots 100 \rfloor_2}[3] + \Sigma^{\lfloor 1 \cdots 101 \rfloor_2}[2] + \Sigma^{\lfloor 1 \cdots 110 \rfloor_2}[3] + \Sigma^{\lfloor 1 \cdots 111 \rfloor_2}[2].$$

- $\Sigma^{\lfloor 1 \cdots 1011 \rfloor_2}[2]$ has period 2^{k+1} and the coefficients are

  $$2 * \binom{n}{\lfloor 1 \cdots 1011 \rfloor_2} \quad \text{for } 0 \leq n < \lfloor \underbrace{1 \cdots 1}_{k+1} \rfloor_2.$$

  Clearly the binomial is 0 for $n < \lfloor 1 \cdots 1011 \rfloor_2$. One can use Lucas's Theorem to compute the remaining coefficients since the primitives of the sequence $(2) \in \mathbf{P}_4$ correspond to the primitives of $(1) \in \mathbf{P}_2$. We get:

  $$\Sigma^{\lfloor 1 \cdots 1011 \rfloor_2}[2] = [\underbrace{0, \ldots, 0}_{2^{k+1}-5}, 2, 0, 0, 0, 2].$$

- $\Sigma^{\lfloor 1 \cdots 100 \rfloor_2}[3]$ has period 2^{k+2} and the coefficients are

  $$3 * \binom{n}{\lfloor 1 \cdots 100 \rfloor_2} \quad \text{for } 0 \leq n < \lfloor \underbrace{1 \cdots 1}_{k+2} \rfloor_2.$$

  Again if $n < \lfloor 1 \cdots 1011 \rfloor_2$ the binomial is zero, hence the first half of the sequence is:

  $$[\underbrace{0, \ldots, 0}_{2^{k+1}-4}, 1, *, *, *].$$

For the second half of the sequence, we have $n = \lfloor 1a_k \cdots a_0 \rfloor_2$ and we use Kummer's Theorem: the binomial coefficient

$$\binom{\lfloor 1a_k \cdots a_0 \rfloor_2}{\lfloor 01 \cdots 100 \rfloor_2}$$

is zero modulo 4 if there are at least two borrows in performing the difference

$$\lfloor 1a_k \ldots a_0 \rfloor_2 - \lfloor 01 \cdots 100 \rfloor_2. \tag{2.1}$$

Notice that for $2 \leq j \leq k-1$, the equality $a_j = 0$ implies that there are two borrows in performing the difference in Equation (2.1), thus the coefficient is zero modulo 4. Hence the sequence $\Sigma^{\lfloor 1 \cdots 100 \rfloor_2}(3)$ is:

$$[\underbrace{0, \ldots, 0}_{2^{k+1}-4}, 3, *, *, *, \underbrace{0, \ldots, 0}_{2^k-4}, *, *, *, *, \underbrace{0, \ldots, 0}_{2^k-4}, *, *, *, *].$$

The remaining coefficients can be computed using the generalisation of Lucas's Theorem. For $2^{k+1} - 3 \leq n \leq 2^{k+1} - 1$, one has:

$$3 * \binom{\lfloor 1 \cdots 101 \rfloor_2}{\lfloor 1 \cdots 100 \rfloor_2} = 3 * \big(\prod_{i=1}^{k-2} \langle\langle \genfrac{}{}{0pt}{}{11}{11} \rangle\rangle \big) \langle\langle \genfrac{}{}{0pt}{}{10}{10} \rangle\rangle \langle \genfrac{}{}{0pt}{}{01}{00} \rangle = 3.$$

$$3 * \binom{\lfloor 1 \cdots 110 \rfloor_2}{\lfloor 1 \cdots 100 \rfloor_2} = 3 * \langle\langle \genfrac{}{}{0pt}{}{11}{10} \rangle\rangle \langle \genfrac{}{}{0pt}{}{10}{00} \rangle = 1.$$

$$3 * \binom{\lfloor 1 \cdots 111 \rfloor_2}{\lfloor 1 \cdots 100 \rfloor_2} = 3 * \langle\langle \genfrac{}{}{0pt}{}{11}{10} \rangle\rangle \langle \genfrac{}{}{0pt}{}{11}{00} \rangle = 1.$$

For $2^{k+2} - 2^k - 4 \leq n \leq 2^{k+2} - 2^k - 1$, one has:

$$3 * \binom{\lfloor 101 \cdots 100 \rfloor_2}{\lfloor 011 \cdots 100 \rfloor_2} =3 * \langle\langle \genfrac{}{}{0pt}{}{10}{01} \rangle\rangle \langle\langle \genfrac{}{}{0pt}{}{01}{11} \rangle\rangle = 3 * 1 * 2 = 2.$$

$$3 * \binom{\lfloor 101 \cdots 101 \rfloor_2}{\lfloor 011 \cdots 100 \rfloor_2} =3 * \langle\langle \genfrac{}{}{0pt}{}{10}{01} \rangle\rangle \langle\langle \genfrac{}{}{0pt}{}{01}{11} \rangle\rangle \langle \genfrac{}{}{0pt}{}{01}{00} \rangle = 2.$$

$$3 * \binom{\lfloor 101 \cdots 110 \rfloor_2}{\lfloor 011 \cdots 100 \rfloor_2} =3 * \langle\langle \genfrac{}{}{0pt}{}{10}{01} \rangle\rangle \langle\langle \genfrac{}{}{0pt}{}{01}{11} \rangle\rangle \langle\langle \genfrac{}{}{0pt}{}{11}{10} \rangle\rangle = 2.$$

$$3 * \binom{\lfloor 101 \cdots 111 \rfloor_2}{\lfloor 011 \cdots 100 \rfloor_2} =3 * \langle\langle \genfrac{}{}{0pt}{}{10}{01} \rangle\rangle \langle\langle \genfrac{}{}{0pt}{}{01}{11} \rangle\rangle \langle\langle \genfrac{}{}{0pt}{}{11}{10} \rangle\rangle \langle \genfrac{}{}{0pt}{}{11}{00} \rangle = 2.$$

For $2^{k+2}-4 \leq n \leq 2^{k+2}-1$, one has:

$$\begin{aligned}
3 * \binom{\lfloor 111\cdots 100\rfloor_2}{\lfloor 011\cdots 100\rfloor_2} &= 3 * \langle\langle {11 \atop 01} \rangle\rangle = 3 * 3 = 1.\\
3 * \binom{\lfloor 111\cdots 101\rfloor_2}{\lfloor 011\cdots 100\rfloor_2} &= 3 * \langle\langle {11 \atop 01} \rangle\rangle \, \langle {01 \atop 00} \rangle = 1.\\
3 * \binom{\lfloor 111\cdots 110\rfloor_2}{\lfloor 011\cdots 100\rfloor_2} &= 3 * \langle\langle {11 \atop 01} \rangle\rangle \, \langle\langle {11 \atop 10} \rangle\rangle = 3.\\
3 * \binom{\lfloor 111\cdots 111\rfloor_2}{\lfloor 011\cdots 100\rfloor_2} &= 3 * \langle\langle {11 \atop 01} \rangle\rangle \, \langle\langle {11 \atop 10} \rangle\rangle \, \langle {11 \atop 00} \rangle = 3.
\end{aligned}$$

Hence the sequence $\Sigma^{\lfloor 1\cdots 100\rfloor_2}[3]$ is:

$$[\underbrace{0,\ldots,0}_{2^{k+1}-4},3,3,1,1,\underbrace{0,\ldots,0}_{2^k-4},2,2,2,2,\underbrace{0,\ldots,0}_{2^k-4},1,1,3,3].$$

- The sequences $\Sigma^{\lfloor 1\cdots 101\rfloor_2}[2]$ and $\Sigma^{\lfloor 1\cdots 111\rfloor_2}[3]$ can be treated as the first point: their period is 2^{k+1} and they have respectively the following shape:

$$\begin{aligned}
\Sigma^{\lfloor 1\cdots 101\rfloor_2}[2] &= [\underbrace{0,\ldots,0}_{2^{k+1}-3},2,0,2]\\
\Sigma^{\lfloor 1\cdots 111\rfloor_2}[2] &= [\underbrace{0,\ldots,0}_{2^{k+1}-3},0,0,2].
\end{aligned}$$

- We study the sequence $\Sigma^{\lfloor 1\cdots 110\rfloor_2}[3]$ as we did for $\Sigma^{\lfloor 1\cdots 100\rfloor_2}[3]$ previously. It has period 2^{k+2}. Using Kummer's Theorem and the generalisation of Lucas's Theorem, one finds:

$$\Sigma^{\lfloor 1\cdots 110\rfloor_2}[3] = [\underbrace{0,\ldots,0}_{2^{k+1}-2},3,1,\underbrace{0,\ldots,0}_{2^k-2},2,2,\underbrace{0,\ldots,0}_{2^k-2},1,3].$$

Now we are ready to study g^s by looking at the sum of the single components. Of course the primitives of [2], having half of the period of the primitives of [3], have to be repeated twice to perform the sum. Thus we find:

$$\begin{array}{rl}
\Sigma^{\lfloor 1\cdots 1011\rfloor_2}[2]: & [\underbrace{0,\ldots,0}_{2^{k+1}-5},2,0,0,0,2\,\underbrace{0,\ldots\ldots\ldots\ldots\ldots\ldots\ldots.0}_{2^{k+1}-5},\ 2,0,0,0,2]+\\
\Sigma^{\lfloor 1\cdots 100\rfloor_2}[3]: & [\underbrace{0,\ldots,0}_{2^{k+1}-5},0,3,3,1,1,\underbrace{0,\ldots,0}_{2^{k}-4},2,2,2,2,\underbrace{0,\ldots,0}_{2^{k}-5},0,1,1,3,3]+\\
\Sigma^{\lfloor 1\cdots 101\rfloor_2}[2]: & [\underbrace{0,\ldots,0}_{2^{k+1}-5},0,0,2,0,2,\underbrace{0,\ldots\ldots\ldots\ldots\ldots\ldots\ldots.0}_{2^{k+1}-5},0,0,2,0,2]+\\
\Sigma^{\lfloor 1\cdots 110\rfloor_2}[3]: & [\underbrace{0,\ldots,0}_{2^{k+1}-5},0,0,0,3,1,\underbrace{0,\ldots,0}_{2^{k}-4},0,0,2,2,\underbrace{0,\ldots,0}_{2^{k}-5},0,0,0,1,3]+\\
\Sigma^{\lfloor 1\cdots 111\rfloor_2}[2]: & [\underbrace{0,\ldots,0}_{2^{k+1}-5},0,0,0,0,2,\underbrace{0,\ldots\ldots\ldots\ldots\ldots\ldots\ldots.0}_{2^{k+1}-5},\ 0,0,0,0,2]=\\
\hline
g^{s}: & [\underbrace{0,\ldots,0}_{2^{k+1}-5},2,3,1,0,0,\underbrace{0,\ldots,0}_{2^{k}-4},2,2,0,0,\underbrace{0,\ldots,0}_{2^{k}-5},\ 2,1,3,0,0]
\end{array}$$

So g^{2^k+1-2} has all but up to 8 entries equal to 0, hence we can conclude $\mathbb{Z}(s) = 2^{k+2} - 8$.

To complete the proof, let us study the sequence g^{s+1}:

$$g^{s+1} = \Sigma^{\lfloor 1\cdots 100\rfloor_2}[2] + \Sigma^{\lfloor 1\cdots 101\rfloor_2}[3] + \Sigma^{\lfloor 1\cdots 110\rfloor_2}[2] + \Sigma^{\lfloor 1\cdots 111\rfloor_2}[3] + \Sigma^{\lfloor 10\cdots 000\rfloor_2}[2].$$

Notice that the index $\lfloor 10\cdots 000\rfloor_2$ of the last primitive in this sum has 2^{k+1} digits in base 2, hence $\Sigma^{\lfloor 10\cdots 000\rfloor_2}[2]$ has period 2^{k+2}. As above, one can compute the coefficients of $\Sigma^{\lfloor 10\cdots 000\lfloor_2}[2]$ with Lucas's Theorem and get:

$$[\underbrace{0\cdots 0}_{2^{k+1}},\underbrace{2\cdots 2}_{2^{k+1}}].$$

Now with the same considerations applied to the study of g^s, we arrive to write the sum defining g^{s+1}:

$$\begin{array}{rl}
\Sigma^{\lfloor 1\cdots 100\rfloor_2}[2]: & [\underbrace{0,\dots,0}_{2^{k+1}-4},2,2,2,2\,\underbrace{0,\dots\dots\dots\dots\dots\dots,0}_{2^{k+1}-4},\ 2,2,2,2]+\\
\Sigma^{\lfloor 1\cdots 101\rfloor_2}[3]: & [\underbrace{0,\dots,0}_{2^{k+1}-4},0,3,2,3,\underbrace{0,\dots,0}_{2^{k}-3},2,0,2,\underbrace{0,\dots,0}_{2^{k}-4},0,1,2,1]+\\
\Sigma^{\lfloor 1\cdots 110\rfloor_2}[2]: & [\underbrace{0,\dots,0}_{2^{k+1}-4},0,0,2,2,\underbrace{0,\dots\dots\dots\dots\dots\dots,0}_{2^{k+1}-4},\ 0,0,2,2]+\\
\Sigma^{\lfloor 1\cdots 111\rfloor_2}[3]: & [\underbrace{0,\dots,0}_{2^{k+1}-4},0,0,0,3,\underbrace{0,\dots,0}_{2^{k}-3},0,0,2,\underbrace{0,\dots,0}_{2^{k}-4},\ 0,0,0,1]+\\
\Sigma^{\lfloor 10\cdots 000\rfloor_2}[2]: & [\underbrace{0,\dots,0}_{2^{k+1}-4},0,0,0,0,\underbrace{2,\dots\dots\dots\dots\dots\dots,2}_{2^{k+1}-3},\ 2,2,2,2]=\\
\hline
g^{s+1}: & [\underbrace{0,\dots,0}_{2^{k+1}-4},2,1,2,2,\underbrace{2,\dots,2}_{2^{k}-3},2,0,0,\underbrace{2,\dots,2}_{2^{k}-4},\ 0,1,0,0].
\end{array}$$

Thus in the sequence g^{s+1} we find $2^{k+1}-4+2+3=2^{k+1}+1$ zeros. This completes the proof. □

Remark. It is interesting to study what happens when there is a change of period in the primitives of g. By Theorem 1.5.5 and Theorem 1.5.6, this happens when $s=2^k$ with $k\in\mathbb{N}_{\geq 1}$. We can then study the coefficients of g^{2^k} as done in Theorem 2.2.1, using the generalisation of Lucas's Theorem. One obtains:

$$\begin{array}{rl}
\Sigma^{\lfloor 1\cdots 101\rfloor_2}[2]: & [\underbrace{0,\dots,0}_{2^{k+1}-3},2,0,2]+\\
\Sigma^{\lfloor 1\cdots 110\rfloor_2}[3]: & [\underbrace{0,\dots,0}_{2^{k+1}-2},3,1,\underbrace{0,\dots,0}_{2^{k}-2},2,2,\underbrace{0,\dots,0}_{2^{k}-2},1,3]+\\
\Sigma^{\lfloor 1\cdots 111\rfloor_2}[2]: & [\underbrace{0,\dots,0}_{2^{k+1}-1},2]+\\
\Sigma^{\lfloor 10\cdots 000\rfloor_2}[3]: & [\underbrace{0,\dots,0}_{2^{k+1}},\underbrace{3,\dots,3}_{2^{k-1}},\underbrace{1,\dots,1}_{2^{k-1}},\underbrace{2,\dots,2}_{2^{k}},\underbrace{1,\dots,1}_{2^{k-1}},\underbrace{3,\dots,3}_{2^{k-1}}]+\\
\Sigma^{\lfloor 10\cdots 001\rfloor_2}[2]: & [\underbrace{0,\dots,0}_{2^{k+1}},\underbrace{0,2,0,2,\dots,0,2}_{2^{k+1}}]=\\
\hline
g^{2^k}: & [\underbrace{0,\dots,0}_{2^{k+1}-3},2,3,1,\underbrace{3,1,\dots,1}_{2^{k-1}-1},1,\underbrace{2,\dots,2}_{2^{k}-1},0,\underbrace{1,3,\dots,3}_{2^{k-1}},\underbrace{3,1,\dots,3}_{2^{k-1}-1},3,0,0].
\end{array}$$

This observation, together with Theorem 2.2.1, proves the existence of a local maximum for $z(s)$ in $s=2^k-2$, which is in fact a global maximum if one restricts to the indices $2^k\leq s<2^{k+1}$.

Let us look at the percentage $z(s)/\mathfrak{p}(g^s)$ of zeros among the coefficients of g^s. The previous results shows that for a fixed period $2^{k+2} = \mathfrak{p}(g^s)$, i.e. for $2^k \leq s < 2^{k+1}$, the maximum of the percentage is in $s = 2^{k-2}$ with value $1-8/2^{k+2}$. Thus if one considers the subsequence $\big(z(2^k-1)\big)_{k\geq 2}$, the corresponding percentages are:

$$\big(1 - 8/2^{k+2}\big)_{k\geq 2} \overset{k\to\infty}{\longrightarrow} 1.$$

Remark. The previous results give a complete answer to the observations made in [4, App. A]. In particular, one can compare the formula of Theorem 2.2.1 for $k = 3, 4, 6$ with the explicit computation of the corresponding levels $5, 13, 61$ done in [4].

2.3 Modular binomial functions

2.3.1 First properties

In this section, we focus on the s-th binomial function:

$$\begin{aligned} \mathbf{b}_s : \mathbb{N} &\longrightarrow \mathbb{Z}_{p^\ell} \\ n &\longmapsto \binom{n}{s}. \end{aligned}$$

As shown in previous section, this function coincides with the s-th primitives $\Sigma^s[1]$ of the constant sequence $[1] \in \mathbf{P}_{p^\ell}$. If the expression of s in base p is one of the following:

$$\lfloor b_k \cdots b_{k-m} \underbrace{(p-1) \cdots (p-1)}_{\ell} b_{k-m-\ell-1} \cdots b_0 \rfloor_p$$
$$\lfloor b_k \cdots b_{k-m} \underbrace{0 \quad \cdots \quad 0}_{\ell} b_{k-m-\ell-1} \cdots b_0 \rfloor_p$$
$$\lfloor b_k \cdots b_{k-m} \underbrace{(p-1)\ 0 \ \cdots \ 0}_{\ell} \ b_{k-m-\ell-1} \cdots b_0 \rfloor_p$$

where $k > \ell$ and $0 \leq m \leq k - \ell - 1$, we prove that it is possible to link the s-th binomial function $\mathbf{b}_s$ to $\mathbf{b}_{s'}$ where s' is obtained from s by removing one of the explicit coefficients in its p-base expression. Of course, in the general case this formula can be combined with the usual ones. Notice that when $p = 2$ and $\ell = 2$, this provides a complete recursive formula for the binomial function $\mathbf{b}_s$.

First we need some definitions.

Definition. Given a sequence $f \in \mathbf{S}_m := \mathbb{Z}_m^{\mathbb{N}}$, a prime q and an integer $t \geq 1$, we call the j-th q^t-subsequence of f the element $h_j \in \mathbb{Z}_m^{q^t}$ defined as

$$h_j = (f(jq^t), f(jq^t + 1), \ldots, f((j+1)q^t - 1)) \qquad j \in \mathbb{N}.$$

We denote by $\mathrm{R}(f,q^t) \in \mathbf{S}_m$ the sequence obtained repeating q times the ordered q^t-subsequences of f:

$$\mathrm{R}(f,q^t) = (\underbrace{h_0,\dots,h_0}_{q},\underbrace{h_1,\dots,h_1}_{q},\dots).$$

We denote by $\mathrm{A}(f,q^t) \in \mathbf{S}_m$ the sequence obtained alternating $(q-1)q^t$ zeros and the ordered q^t-subsequences of f:

$$\mathrm{A}(f,q^t) = (\underbrace{0,\dots,0}_{(q-1)q^t},h_0,\underbrace{0,\dots,0}_{(q-1)q^t},h_1,\dots).$$

Proposition 2.3.1. *For any $f \in \mathbf{S}_m$, $t \geq 1$, and $n' = \lfloor a_r \dots a_t a_{t-1} \dots a_0 \rfloor_q$ one has*

$$\mathrm{R}(f,q^t)(n) = f(n') \qquad \text{if } n = \lfloor a_r \dots a_t\,\alpha\, a_{t-1} \dots a_0 \rfloor_q,\ \forall\, 0 \leq \alpha < q.$$
$$\mathrm{A}(f,q^t)(n) = \begin{cases} f(n') & \text{if } n = \lfloor a_r \dots a_t (q-1) a_{t-1} \dots a_0 \rfloor_q \\ 0 & \text{otherwise.} \end{cases}$$

Proof. By Definition 2.3.1, given $\xi \in \mathbb{N}$, $0 \leq \alpha < q$, $0 \leq i < q^t$, one has

$$\mathrm{R}(f,q^t)(\xi q^{t+1} + \alpha q^t + i) = f(\xi q^t + i)$$
$$\mathrm{A}(f,q^t)(\xi q^{t+1} + \alpha q^t + i) = \begin{cases} f(\xi q^t + i) & \text{if } \alpha = q-1, \\ 0 & \text{otherwise.} \end{cases}$$

Translating in the q-adic representation, we get the claim. □

Example.

- The set of 2-subsequences of $f = [0,1,2,3,4,5] \in \mathbf{P}_7$ is

$$\{[0,1],[2,3],[4,5]\}.$$

- If $h = [1,2,3,4,5,6,7,8] \in \mathbf{P}_{11}$, then:

$$\mathrm{R}(h,2^2) = [1,2,3,4,1,2,3,4,5,6,7,8,5,6,7,8]$$
$$\mathrm{A}(h,2^2) = [0,0,0,0,1,2,3,4,0,0,0,0,5,6,7,8].$$

Moreover

$$\mathrm{R}(h,2^2)(2^3+2^2+3) = 8 = h(2^2+3) \qquad \mathrm{R}(h,2^2)(2^3+2) = 7 = h(2^2+2)$$
$$\mathrm{A}(h,2^2)(2^3+2^2+3) = 8 = h(2^2+3) \qquad \mathrm{A}(h,2^2)(2^3+2) = 0.$$

Remark. Observe the following facts:

- For any q^t both R and A are linear operators: for any $c_1, c_2 \in \mathbb{Z}_m$ and $f_1, f_2 \in \mathbf{S}_m$, it is

$$\begin{aligned} \mathrm{R}(c_1 f_1 + c_2 f_2, q^t) =& c_1 \mathrm{R}(f_1, q^t) + c_2 \mathrm{R}(f_2, q^t) \\ \mathrm{A}(c_1 f_1 + c_2 f_2, q^t) =& c_1 \mathrm{A}(f_1, q^t) + c_2 \mathrm{A}(f_2, q^t). \end{aligned}$$

- If $f \in \mathbf{P}_m$ has period τ and $q^t \mid \tau$, then both $\mathrm{R}(f, q^t)$ and $\mathrm{A}(f, q^t)$ have period $q\tau$.

Definition. If $f, g \in \mathbf{P}_{p^\ell}$, we write:

- $f \equiv_\nu g$ if for any $n \geq 0$, $f(n) = 0$ if and only if $g(n) = 0$, otherwise $\nu_p(f(n)) = \nu_p(g(n)) \in \{0, \cdots, \ell - 1\}$.
- $\Pi_i(f) := \#\{f(x) \mid 0 \leq x < \mathfrak{p}(f),\ \nu_p(f(x)) = i\}$ the number of coefficients with p-adic valuation i, for every $0 \leq i < \ell$.
- $Z(f) := \#\{f(x) \mid 0 \leq x < \mathfrak{p}(f),\ f(x) = 0\}$ the number of zeros.

2.3.2 Recursive lemmas

Let us consider now the primitives $\Sigma^s[1] = \mathbf{b}_s$ of the constant sequence $[1]$ in $\mathbf{P}_{p^\ell}$. Suppose that $p^k \leq s < p^{k+1}$. The next results allow to link in certain cases the quantities $\Pi_i(\mathbf{b}_s), Z(\mathbf{b}_s)$ to the quantities $\Pi_i(\mathbf{b}_{s'}), Z(\mathbf{b}_{s'})$ for some s' with $p^{k-1} \leq s' < p^k$.

Lemma 2.3.2. *With the notation above, suppose that $k > \ell$, $0 \leq m \leq k - \ell - 1$, and that the expression of s in base p is:*

$$s = \lfloor b_k \cdots b_{k-m} \underbrace{(p-1) \cdots (p-1)}_{\ell} b_{k-m-\ell-1} \cdots b_0 \rfloor_p.$$

Denote by

$$\begin{aligned} s' :&= s - \big(b_k p^k + (b_{k-1} - b_k) p^{k-1} + \cdots + (p - 1 - b_{k-m}) p^{k-m-1}\big) \\ &= \lfloor b_k \cdots b_{k-m} \underbrace{(p-1) \cdots (p-1)}_{\ell-1} b_{k-m-\ell-1} \cdots b_0 \rfloor_p. \end{aligned}$$

Then $\mathbf{b}_s \equiv_\nu \mathrm{A}(\mathbf{b}_{s'}, p^{k-m-\ell})$. In particular $\Pi_i(\mathbf{b}_s) = \Pi_i(\mathbf{b}_{s'})$ and $Z(\mathbf{b}_s) = Z(\mathbf{b}_{s'}) + (p-1)p^{k+\ell-1}$.

Proof. The sequence $\mathbf{b}_s$ has period $p^{\ell+k}$ by Theorem 1.5.5. For any $0 \leq n < p^{\ell+k}$, let $n = \lfloor a_{k+\ell-1} \ldots a_0 \rfloor_p$ be its expression in base p. The n-th coefficient of $\mathbf{b}_s$ is:

$$\begin{pmatrix} a_{k+\ell-1} & \cdots & a_{k+1} & a_k & \cdots & a_{k-m} & a_{k-m-1} \cdots a_{k-m-\ell} & a_{k-m-\ell-1} & \cdots & a_0 \\ & & & b_k & \cdots & b_{k-m} & \underbrace{(p-1) \cdots (p-1)}_{\ell} & b_{k-m-\ell-1} & \cdots & b_0 \end{pmatrix}.$$

Let n' be obtained from n by removing the coefficient $a_{k-m-\ell}$. The n'-th coefficient of $\mathbf{b}_{s'}$ is:

$$\begin{pmatrix} a_{k+\ell-1} & \cdots & a_{k+1} & a_k & \cdots & a_{k-m} & a_{k-m-1} \cdots a_{k-m-\ell+1} & a_{k-m-\ell-1} & \cdots & a_0 \\ & & & b_k & \cdots & b_{k-m} & \underbrace{(p-1) \cdots (p-1)}_{\ell-1} & b_{k-m-\ell-1} & \cdots & b_0 \end{pmatrix}.$$

By Proposition 2.3.1, to conclude that $\mathbf{b}_s \equiv_\nu \mathrm{A}(\mathbf{b}_{s'}, p^{k-m-\ell})$, it is enough to show that $\nu_p(\mathbf{b}_s(n)) = \nu_p(\mathbf{b}_{s'}(n'))$ if $a_{k-m-\ell} = p-1$ and $\mathbf{b}_s(n) = 0$ otherwise. To prove this, we use Kummer's Theorem studying the number of borrows in the subtractions $n - s$ and $n' - s'$ in base p:

- If $a_{k-m-\ell} = p-1$:
  - If $a_{k-m-\ell}$ lends, the number of borrows in s is one more than the number of borrows in s'. However in both binomials there are at least ℓ borrows (given by the remaining $(\ell-1)$ coefficients equal to $p-1$), hence both binomials are zero modulo p^ℓ.
  - If $a_{k-m-\ell}$ does not lend, the number of borrows is the same for s and s'.
- If $a_{k-m-\ell} < p-1$: the binomial $\mathbf{b}_s(n) = 0$ since again there are at least ℓ borrows.

From the considerations above, we conclude that $\mathbf{b}_s \equiv_\nu \mathrm{A}(\mathbf{b}_{s'}, p^{k-m-\ell})$. Then immediately follows

$$\Pi_i(\mathbf{b}_s) = \Pi_i(\mathbf{b}_{s'}) \qquad Z(\mathbf{b}_s) = Z(\mathbf{b}_{s'}) + (p-1)\mathfrak{p}(\mathbf{b}_{s'}) = Z(\mathbf{b}_{s'}) + (p-1)p^{k+\ell-1}.$$

□

Remark. With the notation above, it is possible to verify that the proof of the previous lemma holds also for the case $s = \lfloor \underbrace{(p-1) \cdots (p-1)}_{\ell} b_{k-\ell} \cdots b_0 \rfloor_p$ (which corresponds to $m = -1$).

Lemma 2.3.3. *With the notation above, suppose that $k > \ell$, $0 \leq m \leq k - \ell - 1$ and that the expression of s in base p is:*

$$s = \lfloor b_k \cdots b_{k-m} \underbrace{0 \cdots 0}_{\ell} b_{k-m-\ell-1} \cdots b_0 \rfloor_p.$$

Denote by

$$\begin{aligned} s' &:= s - \big(b_k p^k + (b_{k-1} - b_k)p^{k-1} + \cdots + (b_{k-m} - b_{k-m+1})p^{k-m} - b_{k-m}p^{k-m-1}\big) \\ &= \lfloor b_k \cdots b_{k-m} \underbrace{0 \cdots 0}_{\ell-1} b_{k-m-\ell-1} \cdots b_0 \rfloor_p. \end{aligned}$$

Then $\mathbf{b}_s \equiv_\nu \mathrm{R}(\mathbf{b}_{s'}, p^{k-m-1})$. In particular, $\Pi_i(\mathbf{b}_s) = p \cdot \Pi_i(\mathbf{b}_{s'})$ and $Z(\mathbf{b}_s) = p \cdot Z(\mathbf{b}_{s'})$.

Proof. The sequence $\mathbf{b}_s$ has period $p^{\ell+k}$ by Theorem 1.5.5. Similarly to the previous lemma, for $0 \leq n < p^{\ell+k}$ with $n = \lfloor a_{k+\ell-1} \ldots a_0 \rfloor_p$, the coefficient $\mathbf{b}_s(n) = \binom{n}{s}$ is:

$$\begin{pmatrix} a_{k+\ell-1} & \cdots & a_{k+1} & a_k & \cdots & a_{k-m} & a_{k-m-1} \cdots a_{k-m-\ell} & a_{k-m-\ell-1} & \cdots & a_0 \\ & & & b_k & \cdots & b_{k-m} & \underbrace{0 \quad \cdots \quad 0}_{\ell} & b_{k-m-\ell-1} & \cdots & b_0 \end{pmatrix}.$$

Let n' be obtained from n by removing the coefficient a_{k-m-1}, hence the n'-th coefficient of $\mathbf{b}_{s'}$ is:

$$\begin{pmatrix} a_{k+\ell-1} & \cdots & a_{k+1} & a_k & \cdots & a_{k-m} & a_{k-m-2} \cdots a_{k-m-\ell} & a_{k-m-\ell-1} & \cdots & a_0 \\ & & & b_k & \cdots & b_{k-m} & \underbrace{0 \quad \cdots \quad 0}_{\ell-1} & b_{k-m-\ell-1} & \cdots & b_0 \end{pmatrix}.$$

By Proposition 2.3.1, to conclude that $\mathbf{b}_s \equiv_\nu \mathrm{R}(\mathbf{b}_{s'}, p^{k-m-1})$, it is enough to show that, for any value of a_{k-m-1}, $\mathbf{b}_{s'}(n') = 0$ whenever $\mathbf{b}_s(n) = 0$, otherwise $\nu_p(\mathbf{b}_s(n)) = \nu_p(\mathbf{b}_{s'}(n'))$. To prove this, we use Kummer's Theorem studying the number of borrows in the subtractions $n - s$ and $n' - s'$ in base p:

- If a_{k-m-1} lends, then $a_{k-m-2} = \cdots = a_{k-m-\ell} = 0$ and they all lend. So in this case in both s and s' there are at least ℓ borrows (notice that a_{k-m} lends in s'); so the binomials are both equal to zero.

- If a_{k-m-1} does not lend, then the number of borrows remains the same in both the binomials.

Henceforth we can conclude that $\mathbf{b}_s \equiv_\nu \mathrm{R}(\mathbf{b}_{s'}, p^{k-m-1})$, thus:

$$\Pi_i(\mathbf{b}_s) = p \cdot \Pi_i(\mathbf{b}_{s'}) \qquad Z(\mathbf{b}_s) = p \cdot Z(\mathbf{b}_{s'}).$$

□

Remark. Observe that if $\ell = 1$, i.e. the base ring $\mathbb{Z}/p\mathbb{Z}$ is a field, Lemma 2.3.3 (resp. Lemma 2.3.2) reduces to removing a coefficient equal to 0 (resp. equal to $p-1$) in the expression of s in base p. This is just a consequence of Lucas's theorem on binomial coefficients modulo p.

In order to present the last result of this section, we need some preliminary definitions.

Definition. Given $s = \lfloor b_k \cdots b_{k-m} \underbrace{(p-1)\ 0 \cdots 0}_{\ell}\ b_{k-m-\ell-1} \cdots b_0 \rfloor_p \in \mathbb{N}$ with $k > \ell$ and $0 \leq m \leq k-\ell-1$, we denote by E_s the following subset of $\{0, \ldots, p^{k+\ell}-1\}$:

$$E_s := \Big\{ n \in \mathbb{N} : 0 \leq n < p^{k+\ell},\ n = \lfloor a_{k+\ell-1} \ldots a_0 \rfloor_p \text{ such that:}$$
$$a_{k-m-1} = p-1 \qquad a_{k-m-2} \neq 0$$
$$a_{k-m-i} = 0 \quad \forall\, 3 \leq i \leq \ell \qquad a_{k-m-\ell-1} < b_{k-m-\ell-1}$$
$$a_j \geq b_j \quad \forall\, j \in \{0, \ldots, k-m-\ell-2\} \cup \{k-m, \ldots, k\} \Big\}.$$

We denote by $\chi_{E_s} \in \mathbf{P}_{p^\ell}$ the sequence:

$$\chi_{E_s} = [e_0, \ldots, e_{p^{k+\ell}-1}] \quad \text{where } e_i = \begin{cases} 1 \text{ if } i \in E_s \\ 0 \text{ otherwise.} \end{cases}$$

The definition above makes sense also for $m = -1$: in such a way we include also the case $s = \lfloor \underbrace{(p-1)\ 0 \cdots 0}_{\ell}\ b_{k-\ell} \cdots b_0 \rfloor_p$. It is easy to check that

$$|E_s| = p^{\ell-1} \left(\prod_{j=k-m}^{k} (p - b_j) \right) (p-1)\, b_{k-m-\ell-1} \left(\prod_{i=0}^{k-\ell-m-2} (p - b_i) \right)$$

and hence $E_s = \emptyset$ if $b_{k-m-\ell-1} = 0$.

Lemma 2.3.4. *With the notation above, suppose that $k > \ell, 0 \leq m \leq k-\ell-1$ and that the expression of s in base p is:*

$$s = \lfloor b_k \cdots b_{k-m} \underbrace{(p-1)\ 0 \cdots 0}_{\ell}\ b_{k-m-\ell-1} \cdots b_0 \rfloor_p .$$

Denote by

$$\begin{aligned} s' &:= s - \big(b_k p^k + (b_{k-1} - b_k) p^{k-1} + \cdots + (p-1-b_{k-m}) p^{k-m-1} - (p-1) p^{k-m-2} \big) \\ &= \lfloor b_k \cdots b_{k-m} \underbrace{(p-1)\ 0 \cdots 0}_{\ell-1}\ b_{k-m-\ell-1} \cdots b_0 \rfloor_p . \end{aligned}$$

Then $\mathbf{b}_s \equiv_\nu \mathrm{R}(\mathbf{b}_{s'}, p^{k-m-2}) + p^{\ell-1}\chi_{E_s}$ *and thus*

$$\begin{aligned} \Pi_i(\mathbf{b}_s) &= p \cdot \Pi_i(\mathbf{b}_{s'}) \qquad 0 \le i \le \ell - 2 \\ \Pi_{\ell-1}(\mathbf{b}_s) &= p \cdot \Pi_{\ell-1}(\mathbf{b}_{s'}) + |E_s| \\ Z(\mathbf{b}_s) &= p \cdot Z(\mathbf{b}_{s'}) - |E_s|. \end{aligned}$$

Proof. The sequence $\mathbf{b}_s$ has period $p^{\ell+k}$ by Theorem 1.5.5. Similarly to the previous lemmas, for $0 \le n < p^{\ell+k}$ with $n = \lfloor a_{k+\ell-1} \ldots a_0 \rfloor_p$, the coefficient $\mathbf{b}_s(n) = \binom{n}{s}$ is:

$$\begin{pmatrix} a_{k+\ell-1} \cdots a_{k+1} & a_k \cdots a_{k-m} & a_{k-m-1} & a_{k-m-2} \cdots a_{k-m-\ell} & a_{k-m-\ell-1} \cdots a_0 \\ & b_k \cdots b_{k-m} & p-1 & \underbrace{0 \ \ \cdots \ \ 0}_{\ell-1} & b_{k-m-\ell-1} \cdots b_0 \end{pmatrix}.$$

Let n' be obtained from n by removing the coefficient a_{k-m-2}, hence the n'-th coefficient of $\mathbf{b}_{s'}$ is:

$$\begin{pmatrix} a_{k+\ell-1} \cdots a_{k+1} & a_k \cdots a_{k-m} & a_{k-m-1} & a_{k-m-3} \cdots a_{k-m-\ell} & a_{k-m-\ell-1} \cdots a_0 \\ & b_k \cdots b_{k-m} & p-1 & \underbrace{0 \ \ \cdots \ \ 0}_{\ell-2} & b_{k-m-\ell-1} \cdots b_0 \end{pmatrix}.$$

Let us use Kummer's Theorem to study the number of borrows in the subtractions $n - s$ and $n' - s'$ in base p:

- if $a_{k-m-\ell-1}$ does not lend, the two binomials have the same number of borrows.
- if $a_{k-m-\ell-1}$ lends, we have the following cases:
  - if $a_{k-m-2} = a_{k-m-3} = \cdots = a_{k-m-\ell} = 0$, then both binomials have at least ℓ borrows and hence they are zero.
  - If $a_{k-m-3} = \cdots = a_{k-m-\ell} = 0$ but $a_{k-m-2} \neq 0$, there are at least ℓ borrows in s'. In this situation there are at least $\ell - 1$ borrows in s and they are precisely $\ell - 1$ when $n \in E_s$.
  - In the remaining cases, there exists an index $k - m - \ell \le i \le k - m - 3$ such that $a_i \neq 0$, thus a_{k-m-2} does not lend, so the borrows in s and s' are the same.

This proves the statement. □

Remark. With the notation above, it is possible to verify that the proof of the previous lemma holds also for the case $s = \lfloor \underbrace{(p-1)\, 0 \cdots 0}_{\ell}\, b_{k-\ell} \cdots b_0 \rfloor_p$ (which corresponds to $m = -1$). Moreover, observe that Lemma 2.3.4 generalises Lemma 2.3.3 if $p = 2$: indeed the hypotheses of Lemma 2.3.3 imply $b_{k-m-\ell-1} = 0$ in Lemma 2.3.4 and hence $E_s = \emptyset$.

Remark. Let $s = \lfloor b_k \cdots b_0 \rfloor$. The construction of s' in Lemmas 2.3.2 to 2.3.4 does not depend on the $(k-m-\ell)$-tail $b_{k-m-\ell-1} ... b_0$. Therefore if s and $s+i$ differ only on their $(k-m-\ell)$-tails, then $(s+i)' = s' + i$.

2.4 The case of $\mathbb{Z}_4$ and Vieru's sequence

Let us focus on $\mathbb{Z}_4$: with the notation of the previous section, we are considering $p = 2$ and $\ell = 2$. Notice that in this case Lemmas 2.3.2 and 2.3.4 allow us to reduce each binomial coefficient to a smaller one, permitting one to link any primitive $[1]^s$ with $2^k \leq s < 2^{k+1}$ to a primitive $[1]^{s'}$ with $2^{k-1} \leq s' < 2^k$.

As an example we provide a recursive formula for the zeros $Z(s) := Z(v^s)$ of the primitives $v^s := \Sigma^s v$ of Vieru's sequence

$$v = [2, 1, 2, 0, 0, 1, 0, 0] \in \mathbf{P}_4,$$

when $2^k \leq s < 2^{k+1}$ for $k \geq 5$. The zeros of specific primitives of the sequence v were already studied in Section 2.2

The sequence $Z(s)$ is clearly a sequence of natural numbers.

2.4.1 Preliminary results

To state our formula, we need some technical results. First observe that since $2 \cdot 2 = 0$ in $\mathbb{Z}_4$, if $2^k \leq s, t < 2^{k+1}$, then

$$2\chi_{E_s \triangle E_t} := 2(\chi_{E_s} + \chi_{E_t})(n) = \begin{cases} 2 & \text{if } n \in E_s \triangle E_t, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore if $s = \lfloor 10 b_{k-2} \dots b_0 \rfloor_2$, the quantity $|E_s|$ is linked with the number $\mathfrak{z}(s)$ of 0's in the binary expansion of s in the following way:

$$\begin{aligned} |E_s| &= 2 \cdot b_{k-2} \cdot 2^{\mathfrak{z}(\lfloor b_{k-3} \cdots b_0 \rfloor_2)} \\ &= b_{k-2} \cdot 2^{\mathfrak{z}(s)} = \begin{cases} 2^{\mathfrak{z}(s)} & \text{if } b_{k-2} = 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The coefficients $\Pi_0(f)$, $\Pi_1(f)$, $Z(f)$ introduced in Section 2.3.1 represent the number of 1 or 3, the number of 2, and the number of 0 in f, respectively.

If $s = \lfloor 1 b_{k-1} \cdots b_0 \rfloor_2$ and $t = \lfloor 1 b'_{k-1} \cdots b'_0 \rfloor_2$, denote by $(s \mid t)$ the bitwise OR of s and t, i.e., the number whose 2-adic representation has 1 in each bit position for which the corresponding bit of either s or t is 1.

Lemma 2.4.1. *Let $k \geq 5$ and $2^k + 2^{k-2} \leq s < 2^k + 2^{k-1} - 4$. Set $\mathfrak{d}_k$ equal to the $(2^{k-2} - 4)$-sequence $\mathfrak{d}_k(s) := \Pi_1(2(\chi_{E_{s+1} \triangle E_{s+3}}))$. Then*

$$\mathfrak{d}_k(s) = 2^{\mathfrak{z}(s+1)} + 2^{\mathfrak{z}(s+3)} - 2 \times 2^{\mathfrak{z}((s+1|s+3))}$$

and

$$\mathfrak{d}_5 = (4, 8, 4, 4) \quad \textit{and} \quad \mathfrak{d}_{k+1} = (2 \times \mathfrak{d}_k, 4, 2^{k-1}, 2^{k-2}, 2^{k-2}, \mathfrak{d}_k)\ \forall k \geq 5.$$

Proof. Observe that, by Remark 2.3.2

$$\mathfrak{d}_k(s) = |E_{s+1}| + |E_{s+3}| - 2 \times |E_{s+1} \cap E_{s+3}| = 2^{\mathfrak{z}(s+1)} + 2^{\mathfrak{z}(s+3)} - 2 \times 2^{\mathfrak{z}((s+1|s+3))}$$

since $|E_{s+1} \cap E_{s+3}| = 2^{\mathfrak{z}((s+1|s+3))}$ (see definition of E_s in the proof of Lemma 2.3.4 which, in our case, reduces $a_j \geq b_j$).

If $k = 5$, then $s \in \{40, 41, 42, 43\}$. It is easy to verify that

$$\mathfrak{d}_5 = (2^{\mathfrak{z}(41)} + 2^{\mathfrak{z}(43)} - 2^{\mathfrak{z}(41|43)+1}, ..., 2^{\mathfrak{z}(44)} + 2^{\mathfrak{z}(46)} - 2^{\mathfrak{z}(44|46)+1}) = (4, 8, 4, 4).$$

Fixed k, the binary representation of the numbers s between $2^k + 2^{k-2}$ and $2^k + 2^{k-1} - 4$ are of the following three types

- I_k: $2^k + 2^{k-2} \leq s < 2^k + 2^{k-2} + 2^{k-3} - 4$,
- II_k: $2^k + 2^{k-2} + 2^{k-3} - 4 \leq s < 2^k + 2^{k-2} + 2^{k-3}$,
- III_k: $2^k + 2^{k-2} + 2^{k-3} \leq s < 2^k + 2^{k-1} - 4$.

Given $s' \in III_{k+1}$, hence $s' = 2^{k+1} + 2^{k-1} + 2^{k-2} + t$ with $0 \leq t < 2^{k-2} - 4$. Set $s = 2^k + 2^{k-2} + t$, we get

$$(\mathfrak{z}(s'+1),\ \mathfrak{z}(s'+3),\ \mathfrak{z}(s'+1 \mid s'+3)) = (\mathfrak{z}(s+1),\ \mathfrak{z}(s+3),\ \mathfrak{z}(s+1 \mid s+3)).$$

Given $s' \in I_{k+1}$, hence $s' = 2^{k+1} + 2^{k-1} + t$ with $0 \leq t < 2^{k-2} - 4$. Set $s = 2^k + 2^{k-2} + t$, we get

$$(\mathfrak{z}(s'+1),\ \mathfrak{z}(s'+3),\ \mathfrak{z}(s'+1 \mid s'+3)) = (1+\mathfrak{z}(s+1),\ 1+\mathfrak{z}(s+3),\ 1+\mathfrak{z}(s+1 \mid s+3)).$$

Finally the binary representation of s' in the group II_{k+1} is the following:

$$s': \quad \lfloor 10101_{k+1-5}00 \rfloor_2,\ \lfloor 10101_{k+1-5}01 \rfloor_2,\ \lfloor 10101_{k+1-5}10 \rfloor_2,\ \lfloor 10101_{k+1-5}11 \rfloor_2.$$

Therefore $(\mathfrak{z}(s'+1), \mathfrak{z}(s'+3), \mathfrak{z}(s'+1 \mid s'+3))$ are

1. $(3, 2, 2)$ for $s' = \lfloor 10101_{k+1-5}00 \rfloor_2$,

2. $(3, k-1, 2)$ for $s' = \lfloor 10101_{k+1-5}01 \rfloor_2$,

3. $(2, k-2, 1)$ for $s' = \lfloor 10101_{k+1-5}10 \rfloor_2$,

4. $(k-1, k-2, k-2)$ for $s' = \lfloor 10101_{k+1-5}11 \rfloor_2$.

Thus we get the wanted claim. □

Remark. Let us link the sequence $\mathfrak{d}_k$ to two well known integer sequences. Fixed $k \geq 5$ the sequence $\mathfrak{d}_k$ coincides with

$$(2^{k-a(4)}, 2^{k-a(5)}..., 2^{k-a(2^{k-2}-1)})$$

where $a(2^t) = t + 1$ and $a(2^t + i) = 1 + a(i)$ for $t \geq 0$ and $0 < i < 2^t$ (see A063787 in the OEIS, the online encyclopedia of integer sequences). Noticed that $a(2^{t_1} + \cdots + 2^{t_h}) = h + t_h$ for $t_1 > \cdots > t_h \geq 0$, one can directly prove that $\mathfrak{d}_k(2^k + 2^{k-2} + 2^{t_1} + \cdots 2^{t_h} - 4) = 2^{k-h-t_h}$.

Denoted by $wt(n)$ the Hamming weight of n, i.e. the number of 1's in the binary expansion of n, we have, for $2^k + 2^{k-2} \leq s < 2^k + 2^{k-1} - 4$

$$\mathfrak{d}_k(s) = 2^{wt(2^k+2^{k-1}-4-s)+1}.$$

The recurrence relation for $\mathfrak{d}_k(s)$ permits to compute a recurrence relation for the Hamming weight. Denoted by w_h the Hamming weight of the numbers $\lfloor 1 \rfloor_2$, ..., $\lfloor 2^{h+1} - 4 \rfloor_2$, we have

$$w_2 = (1, 1, 2, 1), \quad w_{h+1} = (w_h, h, h, h+1, 1, w_h + 1) \quad \forall h \geq 2$$

where $w_h + 1$ is the sequence obtained by w_h increasing by one each entrance.

2.4.2 Main Recursive Formula

Now we are ready to state the theorem on the zeros of the primitives of Vieru's sequence:

Theorem 2.4.2. *For $k \geq 5$ and $2^k \leq s < 2^{k+1}$, denote:*

$$\begin{aligned}
(c_1, c_2, c_3, c_4) &:= 2^{k-5}(48, 32, 40, 44) \\
(c_1', c_2', c_3', c_4') &:= 2^{k-5}(48, 40, 44, 48) \\
(c_1'', c_2'', c_3'', c_4'') &:= 2^{k-5}(32, 32, 48, 64) \\
\mathcal{Z}_k &:= \big(Z(s)\big)_{2^k \leq s < 2^{k+1}}.
\end{aligned}$$

The initial condition is

$$\mathscr{Z}_5 = (32, 48, 64, 88, 64, 80, 88, 92, 64, 80, 88, 104, 92, 104, 108, 94,\\ 78, 88, 96, 108, 96, 104, 108, 110, 102, 108, 112, 118, 114, 118, 120, 64).$$

For $k \geq 6$, the 2^k-tuple $\mathscr{Z}_k$ coincides with

$$Z(s) = \begin{cases} 2Z(s-2^{k-1}) & \textit{if } 2^k \leq s \leq 2^k + 2^{k-2} - 5 \quad (\mathbf{A}) \\ Z(s-2^{k-1}-2^{k-3}) + c_i & \textit{if } s = 2^k + 2^{k-2} - 5 + i,\ i = 1,2,3,4 \quad (\mathbf{B}) \\ 2Z(s-2^{k-1}) - \mathfrak{d}_k(s) & \textit{if } 2^k + 2^{k-2} \leq s \leq 2^k + 2^{k-1} - 5 \quad (\mathbf{C}) \\ Z(s-2^{k-1}-2^{k-2}) + c'_i & \textit{if } s = 2^k + 2^{k-1} - 5 + i,\ i = 1,2,3,4 \quad (\mathbf{D}) \\ Z(s-2^k) + 2^{k+1} & \textit{if } 2^k + 2^{k-1} \leq s \leq 2^{k+1} - 5 \quad (\mathbf{E}) \\ Z(s-2^k) + c''_i & \textit{if } s = 2^{k+1} - 5 + i,\ i = 1,2,3,4 \quad (\mathbf{F}). \end{cases}$$

Proof. The s-primitive of the sequence $v = [2, 1, 2, 0, 0, 1, 0, 0]$ is equal to

$$v^s = 2\,\mathbf{b}_{s+4} + 3\,\mathbf{b}_{s+3} + 2\,\mathbf{b}_{s+2} + 3\,\mathbf{b}_{s+1} + 2\,\mathbf{b}_s \quad \forall s \geq 0. \tag{2.2}$$

In base 2 we have

$$2^k = \lfloor 10_k \rfloor_2 := \lfloor 1\underbrace{0\cdots0}_{k \text{ times}} \rfloor_2,$$

therefore $\lfloor 10_k \rfloor_2 \leq [s]_2 \leq \lfloor 11_k \rfloor_2$. Set $h = k - 5$, we will consider in order the following cases:

$$\begin{aligned} \mathbf{A}: &\quad \lfloor 1000_h000 \rfloor_2 \leq s \leq \lfloor 1001_h011 \rfloor_2; \\ \mathbf{C}: &\quad \lfloor 1010_h000 \rfloor_2 \leq s \leq \lfloor 1011_h011 \rfloor_2; \\ \mathbf{E}: &\quad \lfloor 1100_h000 \rfloor_2 \leq s \leq \lfloor 1111_h011 \rfloor_2; \\ \mathbf{B}: &\quad \lfloor 1001_h100 \rfloor_2 \leq s \leq \lfloor 1001_h111 \rfloor_2; \\ \mathbf{D}: &\quad \lfloor 1011_h100 \rfloor_2 \leq s \leq \lfloor 1011_h111 \rfloor_2; \\ \mathbf{F}: &\quad \lfloor 1111_h100 \rfloor_2 \leq s \leq \lfloor 1111_h111 \rfloor_2. \end{aligned}$$

In the cases **A**, **C** and **E**, the primitive indices of all the summands in Equation (2.2) have the same prefix: 10 in the first two cases, and 11 in the last. This allows to apply in parallel the recursive lemmas of Section 2.3. The remaining twelve cases require *ad hoc* analysis.

Using a generic computer algebra system one can easily compute the sequence $\mathscr{Z}_5$, the initial condition for the recursive formula.

Cases A and C. In both the cases s, $s+1$, $s+2$, $s+3$, and $s+4$ have a binary representation $\lfloor 10b_{k-2}...b_0\rfloor$ with the two most representative figures equal to 10. If $f \in \mathbf{P}_4$, we denote shortly

$$\mathrm{R}\,f^s := \mathrm{R}(f^s, 2^{k-1}), \quad \mathrm{A}\,f^s := \mathrm{A}(f^s, 2^{k-1}).$$

Using Lemma 2.3.4, we reduce the study of $\mathbf{b}_s$, ..., $\mathbf{b}_{s+4}$ to the study of $\mathbf{b}_{s'}$, ..., $\mathbf{b}_{s'+4}$ where $s' = s - (2^k - 2^{k-1}) = s - 2^{k-1}$. It is

$$\begin{aligned} v^s &= 2\,\mathbf{b}_{s+4} + 3\,\mathbf{b}_{s+3} + 2\,\mathbf{b}_{s+2} + 3\,\mathbf{b}_{s+1} + 2\,\mathbf{b}_s \\ &\equiv_\nu 2\left(\mathrm{R}\,\mathbf{b}_{s'+4} + 2\chi_{E_{s+4}}\right) + 3\left(\mathrm{R}\,\mathbf{b}_{s'+3} + 2\chi_{E_{s+3}}\right) + 2\left(\mathrm{R}\,\mathbf{b}_{s'+2} + 2\chi_{E_{s+2}}\right) + \\ &+ 3\left(\mathrm{R}\,\mathbf{b}_{s'+1} + 2\chi_{E_{s+1}}\right) + 2\left(\mathrm{R}\,\mathbf{b}_{s'} + 2\chi_{E_s}\right) \\ &\equiv_\nu \mathrm{R}\,v^{s'} + 3\cdot 2\chi_{E_{s+1}} + 3\cdot 2\chi_{E_{s+3}} \\ &\equiv_\nu \mathrm{R}\,v^{s'} + 2\chi_{E_{s+1}\triangle E_{s+3}}. \end{aligned}$$

In case **A** it is $E_{s+1} = \emptyset = E_{s+3}$ and hence $v^s \equiv_\nu \mathrm{R}\,v^{s'}$. Therefore

$$Z(v^s) = Z\left(\mathrm{R}\,v^{s'}\right) = 2 \times Z(v^{s'}).$$

In case **C**, if $n \in E_{s+1}\triangle E_{s+3}$, then $\mathrm{R}\,v^{s'}(n)$ is equal to zero. Indeed it is easy to check that $n = \lfloor a_{k+1}\dots a_0\rfloor_2 \in E_{s+1}\triangle E_{s+3}$ implies $a_k = 1$, $a_{k-1} = 1$ and $a_{k-2} = 0$. Since the binary representation of $t \in \{s, s+1, s+2, s+3, s+4\}$ is $\lfloor 101b_{k-3}\dots b_0\rfloor_2$, using Kummer's Theorem one has for $t' = t - 2^{k-1}$:

$$\mathrm{R}\,\mathbf{b}_{t'}(n) = \mathbf{b}_{t'}(n') = \binom{\lfloor a_{k+1}10a_{k-3}\dots a_0\rfloor_2}{\lfloor 11b_{k-3}\dots b_0\rfloor_2} = 0,$$

hence $\mathrm{R}\,v^{s'}(n) = 0$. Therefore, we have

$$Z(v^s) = Z\left(\mathrm{R}\,v^{s'}\right) - \Pi_1(2\chi_{E_{s+1}\triangle E_{s+3}}) = 2 \times Z(v^{s'}) - \mathfrak{d}_k(s).$$

Case E. The numbers s, $s+1$, $s+2$, $s+3$, and $s+4$ have a binary representation $\lfloor 11b_{k-2}...b_0\rfloor$ with the two most representative figures equal to 11. If $f \in \mathbf{P}_4$, we denote shortly

$$\mathrm{R}\,f^s := \mathrm{R}(f^s, 2^{k-1}), \quad \mathrm{A}\,f^s := \mathrm{A}(f^s, 2^{k-1}).$$

Using Lemma 2.3.2, we reduce the study of $\mathbf{b}_s$, ..., $\mathbf{b}_{s+4}$ to the study of $\mathbf{b}_{s'}$, ..., $\mathbf{b}_{s'+4}$ where $s' = s - 2^k$. Thanks to the linearity of A we have:

$$\begin{aligned} v^s &= 2\,\mathbf{b}_{s+4} + 3\,\mathbf{b}_{s+3} + 2\,\mathbf{b}_{s+2} + 3\,\mathbf{b}_{s+1} + 2\,\mathbf{b}_s \\ &\equiv_\nu 2\,\mathrm{A}\,\mathbf{b}_{s'+4} + 3\,\mathrm{A}\,\mathbf{b}_{s'+3} + 2\,\mathrm{A}\,\mathbf{b}_{s'+2} + 3\,\mathrm{A}\,\mathbf{b}_{s'+1} + 2\,\mathrm{A}\,\mathbf{b}_{s'} \\ &\equiv_\nu \mathrm{A}\left(2\,\mathbf{b}_{s'+4} + 3\,\mathbf{b}_{s'+3} + 2\,\mathbf{b}_{s'+2} + 3\,\mathbf{b}_{s'+1} + 2\,\mathbf{b}_{s'}\right) \\ &\equiv_\nu \mathrm{A}\,v^{s'}. \end{aligned}$$

Therefore

$$Z(v^s) = Z\left(\mathrm{A}\, v^{s'}\right) = Z(v^{s'}) + 2^{k+1}.$$

Case B and D. The number s has a binary representation $\lfloor 10b_{k-2}1_h1b_1b_0\rfloor$ with $b_0, b_1, b_{k-2} \in \{0,1\}$. If $f \in \mathbf{P}_4$, we denote shortly

$$\mathrm{R}\, f^s := \mathrm{R}(f^s, 2^{k-4}), \quad \mathrm{A}\, f^s := \mathrm{A}(f^s, 2^{k-4}).$$

B. Using Lemma 2.3.2 with $m = 2$ and Lemma 2.3.3 with $m = 3$, we lead back the study of $\mathbf{b}_s$, ..., $\mathbf{b}_{s+4}$ to the study of $\mathbf{b}_{s'}$, ..., $\mathbf{b}_{s'+4}$ where $s' = s - 2^{k-1} - 2^{k-3}$ in case **B**, and $s' = s - 2^{k-1} - 2^{k-2}$ in case **D**.

- If $(b_1b_0) = (00)$, then we have

$$s+1 = \lfloor 10b_{k-2}1_h101\rfloor_2,\ s+2 = \lfloor 10b_{k-2}1_h110\rfloor_2,\ s+3 = \lfloor 10b_{k-2}1_h111\rfloor_2,$$

and $s+4 = \lfloor 1b'_{k-1}b'_{k-2}0_h000\rfloor_2$ with $b'_{k-1}b'_{k-2} = 01$ in case **B** and $b'_{k-1}b'_{k-2} = 10$ in case **D**. By Lemma 2.3.2 with $m = 2$ for $s+i$, $i = 0,1,2,3$ and Lemma 2.3.3 with $m = 3$ for $s+4$ we have

$$\mathbf{b}_{s+i} \equiv_\nu \mathrm{A}\, \mathbf{b}_{s'+i},\ i = 0,1,2,3,\ \text{ and }\ \mathbf{b}_{s+4} \equiv_\nu \mathrm{R}\, \mathbf{b}_{s'+4}\,.$$

Then

$$v^s \equiv_\nu 2\,\mathrm{R}\,\mathbf{b}_{s'+4} + 3\,\mathrm{A}\,\mathbf{b}_{s'+3} + 2\,\mathrm{A}\,\mathbf{b}_{s'+2} + 3\,\mathrm{A}\,\mathbf{b}_{s'+1} + 2\,\mathrm{A}\,\mathbf{b}_{s'}\,.$$

Analysing the previous equation in blocks of length 2^{k-4}, one obtains:

$$Z(v^s) = Z(v^{s'}) + Z\left(2\,\mathbf{b}_{s'+4}\right).$$

Since $s'+4 = \lfloor 1b'_{k-1}b'_{k-2}0_{h-1}000\rfloor_2$, applying h-times Lemma 2.3.3, we get

$$\begin{aligned} Z\left(2\,\mathbf{b}_{s'+4}\right) &= Z\left(\mathbf{b}_{s'+4}\right) + \Pi_1\left(\mathbf{b}_{s'+4}\right) \\ &= \begin{cases} 2^h\Big(Z\left(\mathbf{b}_{20}\right) + \Pi_1\left(\mathbf{b}_{20}\right)\Big) = 48 \cdot 2^{k-5} & \text{in case } \mathbf{B}\ , \\ 2^{h-1}\Big(Z\left(\mathbf{b}_{24}\right) + \Pi_1\left(\mathbf{b}_{24}\right)\Big) = 48 \cdot 2^{k-5} & \text{in case } \mathbf{D} \end{cases} \end{aligned}$$

Therefore in both the cases **B** and **D** we have $Z(v^s) = Z(v^{s'}) + 2^{k-5} \times 48$.

- If $(b_1b_0) = (01)$, then we have $Z(v^s) = \begin{cases} Z(v^{s'}) + 2^{k-5} \times 32 & \text{in case } \mathbf{B}, \\ Z(v^{s'}) + 2^{k-5} \times 40 & \text{in case } \mathbf{D}. \end{cases}$

- If $(b_1b_0) = (10)$, then we have $Z(v^s) = \begin{cases} Z(v^{s'}) + 2^{k-5} \times 40 & \text{in case } \mathbf{B}, \\ Z(v^{s'}) + 2^{k-5} \times 44 & \text{in case } \mathbf{D}. \end{cases}$

- If $(b_1b_0) = (11)$, then we have $Z(v^s) = \begin{cases} Z(v^{s'}) + 2^{k-5} \times 44 & \text{in case } \mathbf{B}, \\ Z(v^{s'}) + 2^{k-5} \times 48 & \text{in case } \mathbf{D}. \end{cases}$

Case F. The number s has a binary representation $\lfloor 1111_h 1 b_1 b_0 \rfloor$ with $b_0, b_1 \in \{0,1\}$. If $f \in \mathbf{P}_4$, and $2^k \leq t < 2^{k+1}$we denote shortly

$$\mathrm{R}\, f^t := \mathrm{R}(f^t, 2^{k-2}), \quad \mathrm{A}\, f^t := \mathrm{A}(f^t, 2^{k-1}).$$

We lead back the study of $\mathbf{b}_s$, ..., $\mathbf{b}_{s+4}$ to the study of $\mathbf{b}_{s'}$, ..., $\mathbf{b}_{s'+4}$ where $s' = s - 2^k$

- If $(b_1 b_0) = (00)$, then we have

$s{+}1 = \lfloor 1111_h 101 \rfloor_2$, $s{+}2 = \lfloor 1111_h 110 \rfloor_2$, $s{+}3 = \lfloor 1111_h 111 \rfloor_2$, $s{+}4 = \lfloor 10000_h 000 \rfloor_2$. For $0 \leq i \leq 3$ the sequence $\mathbf{b}_{s+i}$ has period 2^{k+2}, while $\mathbf{b}_{s+4}$ has period 2^{k+3}. Nevertheless, the period of

$$v^s = 2\,\mathbf{b}_{s+4} + 3\,\mathbf{b}_{s+3} + 2\,\mathbf{b}_{s+2} + 3\,\mathbf{b}_{s+1} + 2\,\mathbf{b}_s$$

is 2^{k+2}: indeed the sequence $2\,\mathbf{b}_{s+4}$ has period 2^{k+2} by Theorem 1.5.5. By Lemma 2.3.2 with $m = -1$, Lemma 2.3.3 with $m = 1$, and Section 2.3.2 we have

$$v^s = 2\,\mathrm{R}\,\mathbf{b}_{s'+4} + 3\,\mathrm{A}\,\mathbf{b}_{s'+3} + 2\,\mathrm{A}\,\mathbf{b}_{s'+2} + 3\,\mathrm{A}\,\mathbf{b}_{s'+1} + 2\,\mathrm{A}\,\mathbf{b}_{s'}$$

where $s' = s - 2^k$. Then one gets

$$Z(v^s) = Z(v^{s'}) + Z(2\,\mathbf{b}_{s'+4}).$$

Notice that $Z(2\,\mathbf{b}_{s'+4}) = \frac{1}{2}\big(Z\,(\mathbf{b}_{s'+4}) + \Pi_1\,(\mathbf{b}_{s'+4})\,\big)$. Indeed $2\,\mathbf{b}_{s'+4}$ has period equal to one half of the period of $\mathbf{b}_{s'+4}$ and the $0s$ of $2\,\mathbf{b}_{s'+4}$ correspond to the $0s$ and $2s$ of $\mathbf{b}_{s'+4}$. Applying h-times Lemma 2.3.3 with $m = 1$, we get

$$Z\,(\mathbf{b}_{s'+4}) + \Pi_1\,(\mathbf{b}_{s'+4}) = 2^h\big(Z(\mathbf{b}_{32}) + \Pi_1(\mathbf{b}_{32})\big) = 2^{k-5} \cdot 64.$$

Hence $Z(v^s) = Z(v^{s'}) + 2^{k-5} \cdot 32$.

- If $(b_1 b_0) = (01)$, we have

$s{+}1 = \lfloor 1111_h 110 \rfloor_2$, $s{+}2 \lfloor 1111_h 111 \rfloor_2$, $s{+}3 = \lfloor 10000_h 000 \rfloor_2$, $s{+}4 = \lfloor 10000_h 001 \rfloor_2$. By Lemma 2.3.2 with $m = -1$, Lemma 2.3.3 with $m = 1$, and Section 2.3.2 we have

$$v^s = 2\,\mathrm{R}\,\mathbf{b}_{s'+4} + 3\,\mathrm{R}\,\mathbf{b}_{s'+3} + 2\,\mathrm{A}\,\mathbf{b}_{s'+2} + 3\,\mathrm{A}\,\mathbf{b}_{s'+1} + 2\,\mathrm{A}\,\mathbf{b}_{s'}\,.$$

Observe that $3\,\mathrm{R}\,\mathbf{b}_{s'+3}$ has period 2^{k+3}, while $2\,\mathrm{R}\,\mathbf{b}_{s'+4}$, $\mathrm{A}\,\mathbf{b}_{s'+i}$, $i = 0, 1, 2$, have period 2^{k+2}. We have that

$$Z(v^s) = Z(v^{s'}) + Z(2\,\mathbf{b}_{s'+4} + 3\,\mathbf{b}_{s'+3}).$$

Applying h times Lemma 2.3.3 with $m = 1$, we get

$$Z(2\,\mathbf{b}_{s'+4} + 3\,\mathbf{b}_{s'+3}) = Z(2R^h\,\mathbf{b}_{33} + 3R^h\,\mathbf{b}_{32}) = 2^h Z(2\,\mathbf{b}_{33} + 3\,\mathbf{b}_{32}) = 2^{k-5} \cdot 32.$$

Therefore $Z(v^s) = Z(v^{s'}) + 2^{k-5} \cdot 32$.

- If $(b_1 b_0) = (10)$, then we have $Z(v^s) = Z(v^{s'}) + 2^{k-5} \cdot 48$.
- If $(b_1 b_0) = (11)$, then we have $Z(v^s) = Z(v^{s'}) + 2^{k-5} \cdot 64$.

□

In conclusion, we can write in the following compact way the recursive result for $Z(v^s)$ when $2^k \leq s < 2^{k+1}$. Denote:

$$\begin{aligned} u_i &:= 2^{r-i} \qquad i = 1, 2, 3 \\ s' &:= s - u_1 \\ (c_1, c_2, c_3, c_4) &:= 2^{k-3}(12, 8, 10, 11) \\ (c'_1, c'_2, c'_3, c'_4) &:= 2^{k-3}(12, 10, 11, 12). \end{aligned}$$

The initial condition for the recursive formula is the 2^5-tuple $(Z(s))_s$ for $k = 5$, i.e., $2^5 \leq s < 2^6$:

$$\begin{aligned} (32, 48, 64,& 88, 64, 80, 88, 92, 64, 80, 88, 104, 92, 104, 108, 94, \\ & 78, 88, 96, 108, 96, 104, 108, 110, 102, 108, 112, 118, 114, 118, 120, 64). \end{aligned}$$

For $k \geq 6$, the 2^k-tuple $(Z(s))_{2^k \leq s < 2^{k+1}}$ coincides with:

$$\begin{aligned} (\, & \underbrace{2Z(s'), \dots, 2Z(s')}_{2^{k-2}-1}, \underbrace{Z(s'-u_3)+c_1, \dots, Z(s'-u_3)+c_4}_{4}, \underbrace{2Z(s')-\mathfrak{d}_r(s), \dots, 2Z(s')-\mathfrak{d}_r(s)}_{2^{k-2}-4}, \\ & \underbrace{Z(s'-u_2)+c'_1, \dots, Z(s'-u_2)+c'_4}_{4}, \underbrace{Z(s'-u_1)+2^{k+1}, \dots, Z(s'-u_1)+2^{k+1}}_{2^{k-1}-4}, \underbrace{2Z(s'-u_1)}_{1}). \end{aligned}$$

Recall that $s' = s - 2^{k-1}$ and so in the tuple above the first coefficient is computed using $s = 2^k$, the second one using $s = 2^k + 1$, the last one using $s = 2^{k+1} - 1$.

2.4.3 Interpretation of the formula.

In order to visualise the previous result, one can have a look at the following graphs. In Figure 2.4, it is depicted the sequence $Z(s)$ for $2^6 \leq s < 2^7$ while in Figure 2.5, the same sequence is represented for $2^7 \leq s < 2^8$. In the latter we have used different colours to highlight the six cases of Theorem 2.4.2, where the recursive formula has different definitions. As the recursive formula states, in the case **A** we recognise the first half of the graph of Figure 2.4 with doubled values, while in the case **E** we recognise the entire graph of Figure 2.4 with values augmented by $2^8 = 256$. In the case **C**, one can recognise again the first half of the graph of Figure 2.4, where we notice that the translation given by $\mathfrak{d}_7$ does not drastically modify the behaviour of the sequence $Z(s)$.

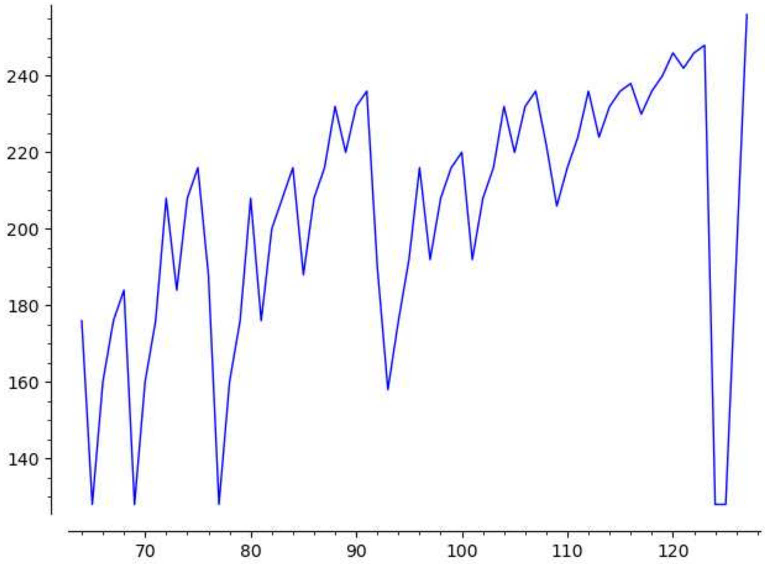


Figure 2.4: The values $Z(s)$ for $2^6 \leq s < 2^7$.

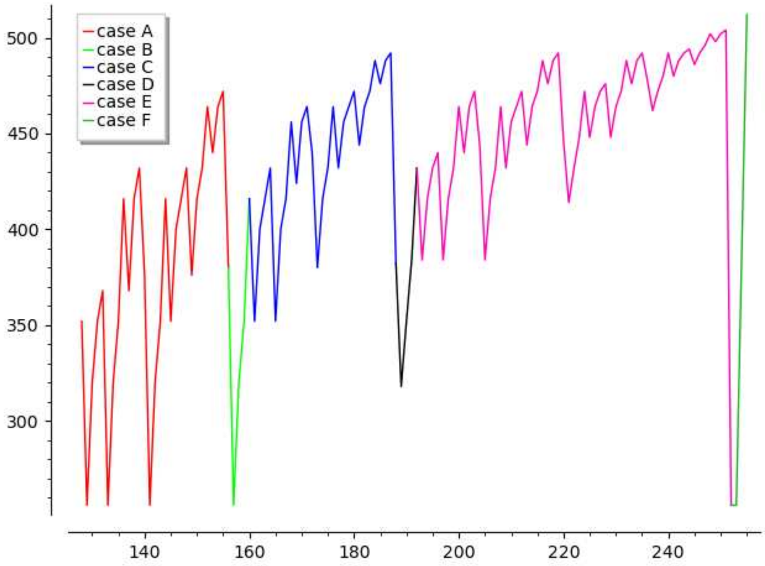


Figure 2.5: The values $Z(s)$ for $2^6 \leq s < 2^7$.

Part II

Persistent homology and harmonic analysis

Chapter 3

Introduction to persistent homology and harmonic analysis

In this chapter we briefly introduce persistent homology and we attempt to present some fundamental concepts of musical harmonic analysis in a formal language. In Section 3.1 we introduce simplicial complexes and the persistent homology of a filtration of complexes. In Section 3.2 we present two kind of constructions that can be used to build a filtration of simplicial complexes from a dataset: the Vietoris-Rips filtration and the Dowker filtration. Finally, in Section 3.3 we formalise some basic notions of music theory and we define, at least intuitively, harmonic complexity and tonality.

3.1 Simplicial complexes

Where not differently specified, we consider m to be an integer bigger than every integer index we will introduce in the rest of the section.

3.1.1 First definitions

Definition. Consider the category SCpx of finite simplicial complexes defined as follows:

- an object of SCpx is a finite set K together with a collection of non-empty subsets $\Delta \subset \mathcal{P}(K)$ that contains all the singletons and is closed under the operation of taking subsets, i.e. if $\sigma \in \Delta$ and $\tau \subset \sigma$, then $\tau \in \Delta$.

- a morphism from (K, Δ) to (T, Γ) is a set-theoretic map $f : K \to T$ such that for any $\sigma \in \Delta$, $f(\sigma) \in \Gamma$.

A set $\sigma \in \Delta$ with cardinality $\#\sigma = j+1$ is called a *simplex* of dimension j, or a j-simplex for brevity. The dimension of the simplicial complex (K, Δ) is defined as $\max\{\dim \sigma \mid \sigma \in \Delta\}$. We will denote a simplicial complex just by K if the datum of Δ can be omitted.

Definition. Given a simplicial complex (K, Δ), we say that (K', Δ') is a *sub-complex* of (K, Δ) if it is a simplicial complex, $K' \subset K$ and $\Delta' \subset \Delta$. Minimal sub-complexes (i.e. singletons) are called *vertices* and maximal sub-complexes are called *faces*, having dimension respectively 0 and $\dim(\Delta) - 1$.

On simplices we can define an *orientation*: given a simplex $\sigma = \{p_1, \dots, p_s\} \in \Delta$, an orientation is a class of equivalence of σ^s where $(p_1, \dots, p_s) \sim (p_{\alpha(1)}, \dots, p_{\alpha(s)})$ if α is a permutation of the indices with signature 1. A simplex (resp. simplicial complex) with an orientation is called *oriented simplices* (resp. *oriented simplicial complex*).

Example. Consider a finite set of points $K = \{p_1, \dots, p_k\} \subset \mathbb{R}^n$. Given an integer $j \leq k$ and a set $J \in \mathcal{P}(K)$ of cardinality j, we can uniquely associate to J the j-simplex generated by its points in $\mathbb{R}^n$, i.e. the j-dimensional polytope obtained by the convex hull of the points $p_i \in J$, which we denote by $[p_i]_i$. For example, $[p_1, p_2]$ denotes the line in $\mathbb{R}^n$ connecting p_1 and p_2, while $[p_1, p_2, p_3]$ denotes the triangle having those points as vertices. If $k = 3$ and we denote $\Delta = \mathcal{P}(K)$, one has that (K, Δ) is a simplicial complex and the associate set

$$D = \{[p_1, p_2, p_3], [p_1, p_2], [p_1, p_3], [p_2, p_3], [p_1], [p_2], [p_3]\}$$

describes the triangle $[p_1, p_2, p_3]$ in $\mathbb{R}^n$ together with all its faces and vertices.

We call the *standard j-simplex* the convex hull of the basis vector $e_0, \dots, e_j$ in $\mathbb{R}^{j+1}$. To each simplicial complex (K, Δ) one can associate a topological space called its *geometric realisation* and denoted by $|K|$. This is done by associating to each abstract j-simplex $\sigma \in \Delta$ a copy of a standard j-simplex in $\mathbb{R}^n$, and then gluing together according to the structure of Δ. More formally, the following result holds:

Theorem 3.1.1. *[34, sec. 4.3] Every (eventually oriented) simplicial complex has a geometrical realisation.*

We will not use the notion of geometrical realisation in detail, but it is sometimes convenient to visualise an abstract simplicial complex through its realisation, i.e. an object in a suitable $\mathbb{R}^n$ obtained by gluing together standard j-simplices.

Let us fix an oriented simplicial complex K of dimension m. For every $0 \leq k \leq m$, let us denote by $\mathcal{C}_k$ the free abelian group generated by the oriented k-simplices of K. Consider the boundary morphism $\partial_k : \mathcal{C}_k \longrightarrow \mathcal{C}_{k-1}$ obtained by linear extension of the map defined on a k-simplex $\sigma = (p_0, \ldots, p_k)$ by:

$$\partial_k(\sigma) = \sum (-1)^i (p_0, \ldots, \hat{p_i}, p_k)$$

where $(p_0, \ldots, \hat{p_i}, p_k)$ is the $(k-1)$-simplex obtained from σ by removing p_i. Observe that these maps satisfy the null-composition property: $\partial_k \circ \partial_{k+1} = 0$.

Definition. Given a simplicial complex K, the groups $\mathcal{C}_k$ together with the maps ∂_k define the *associated chain complex* $\mathcal{C}_\bullet$:

$$\cdots \longrightarrow \mathcal{C}_{k+1} \xrightarrow{\partial_{k+1}} \mathcal{C}_k \xrightarrow{\partial_k} \mathcal{C}_{k-1} \longrightarrow \cdots$$

With a little abuse, we will sometimes call simplicial complex also the chain complex $\mathcal{C}_\bullet$ associated to a complex K. Following the usual notation in algebraic topology, we define the *cycle group* Z_k and the *boundary group* B_k as follows:

$$Z_k := \ker \partial_k \qquad B_k := \operatorname{im} \partial_{k+1}$$

The property $\partial_k \circ \partial_{k+1} = 0$ ensures that for any k one has $B_k \subset Z_k$. Thus for any k we can consider the k-th homology group $H_k := Z_k / B_k$.

More generally, given a simplicial complex K and a commutative ring R with identity, we can define $\mathcal{C}_k$ as the free R-module generated by the k-simplices of K. In this context, Z_k, B_k, H_k are R-modules and if D is a PID, H_k decomposes as a direct sum of cyclic R-modules (see [42, Theorem 2.1]). Hence for $\beta \in \mathbb{Z}$ and suitable $d_i \in R$ one has:

$$H_k \simeq R^\beta \oplus \left(\bigoplus_{i=1}^{t} R/d_i R \right).$$

The rank β of the torsion-free part is called the k-th Betti number of the complex $\mathcal{C}_\bullet$. In the abelian group setting, i.e. when $R = \mathbb{Z}$, the previous decomposition becomes:

$$H_k = \mathbb{Z}^\beta \oplus \left(\bigoplus_{i=0}^{t} (\mathbb{Z}/p_i)^{m_i} \right)$$

where p_i are primes. If R is a field, the torsion part disappears and one gets $H_k = R^\beta$. In the applications, R will be almost always $\mathbb{Z}, \mathbb{R}$ or the finite field $\mathbb{F}_2$.

3.1.2 Filtration of complexes

Definition. A *persistence complex* is a family of chain complexes $\{\mathcal{C}_\bullet^i\}$ over R together with chain maps $f^i : \mathcal{C}_\bullet^i \to \mathcal{C}_\bullet^{i+1}$. If the maps f_i are inclusions, we call $(\mathcal{C}_\bullet^i, f^i)_i$ a *filtration of complexes.*

By the definition of chain maps, the maps f^i and the maps ∂_k are compatible for any i, k: if we denote $f^i_k : \mathcal{C}^i_k \to \mathcal{C}^{i+1}_k$ the k-th component of f^i for any k, one has:

$$f^i_k \circ \partial_{k+1} = \partial_{k+1} \circ f^i_{k+1}.$$

Thus the chain map f^i sends cycles to cycles and boundaries to boundaries and induces maps on the homology groups:

$$f^*_k : H_k(\mathcal{C}^i_\bullet) \longrightarrow H_k(\mathcal{C}^{i+1}_\bullet).$$

Definition. Given a filtration of simplicial complexes $(\mathcal{C}^i_\bullet, f^i)_{i\geq 0}$, we define the *$p$-persistent k-th homology group* of $\mathcal{C}^i_\bullet$ to be:

$$H^{i,p}_k := \frac{Z^i_k}{B^{i+p}_k \cap Z^i_k}$$

and its rank $\beta^{i,k}_k$ is called the p-persistent k-th Betti number of $\mathcal{C}^i_\bullet$.

In order to compute the persistence in the filtration, one needs to find a compatible basis for all the persistent k-th homology groups. A fundamental result is the Correspondence Theorem, that provides an elegant and abstract equivalence of categories which turns out to be very useful in the applications as well. Here we briefly summarise the construction and state this result, which is fundamental in order to understand the meaning of persistence barcodes and persistence diagrams that will be extensively used.

Definition. A *persistence module* M is a family of R-modules M^i together with homomorphisms $\phi^i : M^i \to M^{i+1}$. A persistence module is of *finite type* if for every i the module M^i is finitely generated and the maps ϕ^i are isomorphisms for $i \geq N$ for some integer N.

In the case of our interest, the modules M^i are the chain complexes $\mathcal{C}^i_\bullet$ and the maps ϕ^i will be the inclusions of the filtration. This gives a persistence module of finite type.

Now consider the polynomial ring $R[t]$ with the standard grading and define a graded module over $R[t]$ as:

$$\alpha(M) := \bigoplus_{i=0}^{\infty} M^i$$

where the action of the variable t is the translation of the components:

$$t \cdot (m^0, m^1, m^2, m^3, \dots) = (0, \phi^0 m^0, \phi^1 m^1, \phi^2 m^2, \dots).$$

Theorem 3.1.2 (Correspondence)**.** *The map α above defines an equivalence of categories between persistence modules of finite type over R and finitely generated graded modules over $R[t]$.*

If R is a field, the graded ring $R[t]$ is a PID and its ideals are generated by t^j, hence a finitely generated graded $R[t]$-module M can be decomposed as:

$$M \simeq \left(\bigoplus_{\ell=1}^{n} \theta^{\beta_\ell} R[t] \right) \oplus \left(\bigoplus_{j=1}^{m} \theta^{\gamma_j} \frac{R[t]}{(t^{\epsilon_j})} \right) \tag{3.1}$$

where θ^α denotes the upward shift of the grading of M and $\beta_i, \gamma_j, \epsilon_j$ are positive integers.

Let us spend a moment to understand the relationship between this decomposition and the evolution of Betti numbers in a filtration of simplicial complexes. Deeper analysis and proofs can be found in [42]. Let us consider $\mathbb{F}$ as base field, $(\mathcal{C}^i_\bullet, \iota^i)_i$ a filtration of simplicial complexes and $\big(H_k(\mathcal{C}^i_\bullet), (\iota^i_k)^*\big)$ the induced chain of morphisms on the k-th homology groups (which in fact are $\mathbb{F}$-vector spaces), for any k. By construction, $\big(H_k(\mathcal{C}^i_\bullet), (\iota^i_k)^*\big)$ is a persistence module of finite type, hence thanks to the correspondence theorem we can uniquely associate to it a finitely generated graded module over $\mathbb{F}[t]$, which has a decomposition as in Equation (3.1).

In this setting, the integer β_ℓ represents a generator of the homology group $H_k^{\beta_\ell}$ that is not in the image of $(\iota_k^{\beta_\ell - 1})^*$. For brevity, we say that this generator is *born* in the filtration at the index β_ℓ. β_ℓ coming from the torsion-free part of the decomposition, this generator is mapped to a generator of the groups H_k^a for every $a \geq \beta_\ell$ via the (composition of the) maps (ι_k^a). We also say that this generator *dies* at index ∞. In particular, it is a generator of the p-persistent homology groups $H_k^{\beta_\ell, p}$ for any p.

Similarly, the integers γ_j and ϵ_j represent a generator that is born in the group $H_k^{\gamma_j}$ and that lives in the filtration up to the index ϵ_j, where it dies as it becomes trivial in $H_k^{\epsilon_j}$.

Henceforth, the points of the type $(\beta_\ell, \infty) \in \mathbb{R}^2 \cup \{\infty\}$ completely describe the torsion-free part of the filtration, while the points $(\gamma_j, \epsilon_j) \in \mathbb{R}$ describe the torsion part. Ultimately, the set of these points provides all the information about persistence in the filtration and it is the object we are going to analyse in the rest of the chapter.

3.1.3 Representations and distances

We now consider the set of points in the extended real plane which have on the x-axis (resp. y-axis) the index of birth (resp. death) of a generator of the persistent homology groups associated to a filtration of complexes. Such a subset of $\mathbb{R}^2 \cup \{\infty\}$

is called the *persistence diagram* associated to the filtration. In the applications, it is common to consider the persistence diagram in order to proceed to further analysis.

Another way of representing the (birth, death) points associated to a filtration is a *persistence barcode*. This is defined as the multi-set of real intervals $[b_g, d_g)$ where (b_g, d_g) is the point of the persistence diagram associated to the generator g of the homology groups.

One can define a distance between persistence diagrams:

Definition. Fix $p \in [1, \infty)$ and consider two persistence diagrams $X, Y \subset \mathbb{R}^2$ and a distance d on $\mathbb{R}^2$. The p-th Wasserstein distance between X and Y on $(\mathbb{R}^2, d)$ is defined as:

$$W_p^d(X,Y) := \inf_{\phi: X \to Y} \big(\sum_{x \in X} d(x, \phi(x))^p \big)^{1/p}$$

where ϕ ranges over the matchings between X and Y. For $p = \infty$, we define:

$$W_\infty^d(X,Y) := \inf_{\phi: X \to Y} \sup_{x \in X} d(x, \phi(x)).$$

In particular, for $d = L_\infty$, one has that $W_\infty^{L_\infty}$ is the so-called *bottleneck distance.*

Definition. Given a persistence diagram $X = \{(x_i, y_i) \mid x_i < y_i,\ x_i, y_i \in \mathbb{R}\}$, if we denote $S = \sum(y_i - x_i)$, the *entropy* of X is defined as:

$$\varepsilon_X = \sum_i -\frac{y_i - x_i}{S} \log \big(\frac{y_i - x_i}{S} \big).$$

3.2 Construction of filtrations

Now that we introduced the basic concepts of persistent homology, we are ready to focus on its use for Topological Data Analysis. The starting point in TDA is a set S that represents the data we want to study. Depending on the data, this set may be endowed with extra structures that can be used in the analysis. For example, one may typically be interested in defining a distance on S that well describes the features of the data in that specific context. Another common setting is S being a graph, eventually directed and weighted. In this section, we present two constructions that allow to obtain a simplicial complex from such structures (metric space and directed graph). This procedure allow us to use persistent homology to reconstruct the topological properties of the dataset and extract useful geometric features.

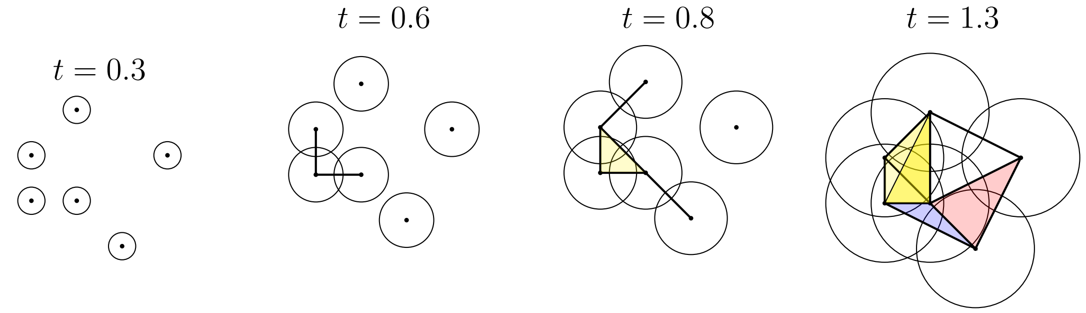
3.2.1 Vietoris-Rips filtration

The Vietoris-Rips filtration is named after Leopold Vietoris, who introduced the construction in [40] and Eliyahu Rips, who applied it to the study of hyperbolic groups. It allows us to build a filtration of simplicial complexes associated to a metric space.

Let us consider (S, d) a metric space and fix an order on the indices of the points $x_1, \dots, x_n \in S$. The associated Vietoris-Rips complex of parameter $t \geq 0$ is the oriented simplicial complex $\mathcal{R}_t$ whose oriented k-simplices are of the type $\sigma = (x_0, \dots, x_k)$ such that $d(x_i, x_j) \leq t$ for every $0 \leq i < j \leq k$. It is easy to verify that this actually gives a simplicial complex and for $t_1 < t_2$ one has $\mathcal{R}_{t_1} \hookrightarrow \mathcal{R}_{t_2}$, hence the chain of inclusions provides a filtration of complexes $(\mathcal{R}_t)_{t \geq 0}$.

The following is a visual example of the construction, with

$$S = \{(1,2), (1,3), (2,2), (2,4), (3,1), (4,3)\}.$$



Example. A simple example that shows the effectiveness of Vietoris-Rips construction on recovering the topological properties of the data set is the case of a point cloud on a surface. Consider a torus $\mathbb{T}$ immersed in $\mathbb{R}^3$ and fix a set of points $S \subset \mathbb{T}$ on the torus. Now let us pretend to forget that the set S is a subset of the torus: we would like to recover this information directly from S itself. This is a simplification of the common problem of having a point cloud and wanting to reconstruct its topological features. Persistent homology of Vietoris-Rips complex associated to S (with respect to the euclidean distance induced by $\mathbb{R}^3$) addresses precisely this problem.

Indeed, the filtration at parameter $t = 0$ coincides with the set of points S and as t increases we start to add geometric structure and to recover the *shape* of the torus. When t becomes big enough, we lose accuracy as the simplices become all homologically equivalent. The information that we are interested in retaining is the persistence of the topological features in the filtration: the longer lasting properties are the most valuable and representative ones.

As the previous example suggests, the Vietoris-Rips complex is in fact an approximation of the Čech complex, which is widely used in algebraic geometry.

Theorem 3.2.1. *[29, pag. 74] Given S a subset of an Euclidean space, denote $\mathcal{R}_t$ (resp. $\mathcal{C}_t$) the Vietoris-Rips complex (resp. the Čech complex) associated to S with parameter $t \geq 0$. Then one has:*

$$\mathcal{C}_t \subseteq \mathcal{R}_t \subseteq \mathcal{C}_{\sqrt{2}t}.$$

The Vietoris-Rips construction is preferred because of computational aspects: while being a good approximation of Čech complex, it is significantly simpler and quicker to compute, making it much more efficient to use in practical problems.

Non-directed graphs. Consider now the case where the dataset is a finite non-directed graph G, eventually with weighted edges. If G is a non-directed graph, one can use the Vietoris-Rips construction to associate a simplicial complex to G. Indeed given two vertices $u, v \in G$, one can collect the set $\Gamma_{u,v}$ of all the paths γ in G connecting u and v. Then, if $\Sigma\gamma$ denotes the total cost of the path γ, i.e. the sum of the weights of its edges, one looks for the path $\bar{\gamma} \in \Gamma_{u,v}$ that minimises Σ and defines:

$$d(u,v) := \Sigma\bar{\gamma}. \tag{3.2}$$

The map d defines a distance on set of vertices of G, hence we can use it to build the Vietoris-Rips complex as shown above for metric spaces. We will use this construction in Section 4.2.

3.2.2 Dowker filtration

What if we want to study a directed graph? This could be the case in many real life problems, like asymmetric networks, road maps, etc. In Section 4.1.3 , we will model the musical chords as a directed graph and we will need to associate a filtration of simplicial complexes to it. Notice that in the case of digraphs, the minimal path distance is not well defined, as it lacks symmetry. There is the possibility to *symmetrize* the definition in Equation (3.2) in order to obtain a well defined distance, but this causes the loss of the information regarding the asymmetry of the digraph.

Another option is to abandon Vietoris-Rips construction and use another filtration. Here we focus on Dowker filtration. This filtration manages to retain the information associated to the asymmetry of the digraph, hence it suits better the study of cases where asymmetry plays an important role for the dataset. In [25, Section 5.2], a family of digraphs is provided as an example of the efficacy of the Dowker filtration and its capability to reconstruct significant features of the graph, in comparison with Vietoris-Rips filtration that remains blind to asymmetry.

First, let us consider a directed weighted graph (G, ω_G), where $\omega_G : G \times G \to \mathbb{R}$ denotes the weights function. We define the function:

$$\begin{aligned}\bar{\omega}_G : G \times G \longrightarrow & \mathbb{R} \\ (v, v') \longmapsto & \max\{\omega_G(v, v'), \omega_G(v, v), \omega_G(v', v')\}.\end{aligned}$$

Given a parameter $t \geq 0$, we define:

$$R_{t,G} := \{(v, v') \in G \times G \mid \bar{\omega}_G(v, v') \leq t\}$$

Clearly if $t_1 \leq t_2$, one has $R_{t_1,G} \subset R_{t_2,G}$. We define the *Dowker t-sink simplicial complex* $\mathfrak{D}^{\mathrm{si}}_{t,G}$ to be:

$$\mathfrak{D}^{\mathrm{si}}_{t,G} = \{\sigma = (v_0, \ldots, v_k) \mid \exists\, v' \in G \text{ s.t. } (v_i, v') \in R_{t,G} \text{ for all } v_i\}.$$

If v' satisfies the condition above for σ, we say that it is a *t-sink* for σ. The filtration $(R_{t,G})_{t\geq 0}$ gives rise to the filtration

$$(\mathfrak{D}^{\mathrm{si}}_{t,G})_{t\geq 0}$$

of which we can study the persistent homology.

The dual construction is the Dowker source filtration, defined as follows:

$$\mathfrak{D}^{\mathrm{so}}_{t,G} = \{\sigma = (v_0, \ldots, v_k) \mid \exists\, v' \in G \text{ s.t. } (v', v_i) \in R_{t,G} \text{ for all } v_i\}.$$

Here v' is called a *t-source* of σ.

In what follows, we will simply say *Dowker filtration* without specifying if we are using the sink or the source construction. In fact, this does not lead to confusion, as the homologies of the sink and the source filtrations have been proven to be equivalent:

Theorem 3.2.2. *[25, Theorem 17] Let us denote by $Dgm^{si}_k(G)$ (resp. $Dgm^{so}_k(G)$) the persistence diagram associated to the k-th homology groups of the sink (resp. source) Dowker filtration of the weighted directed graph G. Then one has:*

$$Dgm^{si}_k(G) \simeq Dgm^{so}_k(G).$$

Another important result is the stability of Dowker filtration with respect to perturbation on data. We need some definitions.

Definition. Given two weighted digraphs (G, ω_G) and (H, ω_H), consider a *correspondence* R between them, i.e. a relation $R \subset G \times H$ such that $\pi_G(R) = G$ and $\pi_H(R) = H$ where $\pi_G : G \times H \to G$ and $\pi_H : G \times H \to H$ denote the projections. We define the *distortion* of R to be:

$$\mathrm{dis}(R) := \max_{(x,y),(x',y') \in R} |\omega_G(x, x') - \omega_H(y, y')|.$$

As it is clear from the definition, the distortion measures how much the correspondence R modifies the path-distances of the graphs.

One can define a distance between digraphs as:

$$d_{\mathcal{N}}(G,H) := \frac{1}{2}\min_{R\in\mathfrak{R}} \operatorname{dis}(R)$$

where $\mathfrak{R}$ denotes the set of all correspondences between G and H. Then the bottleneck distance d_B between the persistence diagrams associated to G and H is controlled by the graph distance $d_{\mathcal{N}}$, as the following result states:

Theorem 3.2.3 (Stability)**.** *[25, Prop. 4] With the notation above, for any degree k one has:*

$$d_B(Dgm_k^{si}(G), Dgm_k^{si}(H)) \leq 2d_{\mathcal{N}}(G,H).$$

3.3 Harmonic analysis programme

In what follows, we will focus on *tonal music*. But what is tonal music? Tonal music is the music composed around a central *chord*, which defines the *tonality* of a musical piece. In this section, we will introduce the basic concepts of chord and tonality and roughly present the main ideas of the functional relation of chords in a tonality. This will lead us to define, at least intuitively, harmony and harmonic complexity, which will be the case of study in the next chapter. We will try to be as rigorous as possible, yet trying to avoid some of the thorny questions regarding tonality and functional relations. The reader with a background in music theory can safely skip this section.

We call *pitch* a key of a standard 88-keys keyboard. We will use the *scientific pitch notation*:

$$A^0, A\sharp^0, B^0, C^1, \ldots$$

with the indices as superscripts instead of subscripts, in order to avoid confusion with the usual notation of mathematical indices. As it is well known, the set $\mathcal{P}$ of pitches can be divided in octaves:

$$\begin{aligned}
\mathcal{O}^0 =&\{A^0, A\sharp^0, B^0\}\\
\mathcal{O}^1 =&\{C^1, C\sharp^1, D^1, D\sharp^1, E^1, F^1, F\sharp^1, G^1, G\sharp^1, A^1, A\sharp^1, B^1\}\\
\mathcal{O}^2 =&\{C^2, C\sharp^2, D^2, \ldots, B^2\}\\
&\vdots\\
\mathcal{O}^7 =&\{C^7, C\sharp^7, D^7, \ldots, B^7\}\\
\mathcal{O}^8 =&\{C^8\}
\end{aligned}$$

We define the function $f : \mathcal{P} \to \mathbb{R}$ that associates to every pitch its audio frequency, so for example $f(A^4) = 440$. Notice that for every pitch $P^i \in \mathcal{P}$ with $i \geq 1$, one has $f(P^i) = 2f(P^{i-1})$.

Definition. We define the set $\bar{\mathcal{P}}$ of *pitch classes* to be the set of pitches modulo octave:

$$\bar{\mathcal{P}} = \{C, C\sharp, D, D\sharp, \dots, B\}.$$

where, for example, C denotes the class $\{C_i\}_{1 \leq i \leq 8}$.

The bijection $\iota : \bar{\mathcal{P}} \to \mathbb{Z}_{12}$ defined as:

$$C \longmapsto 0 \qquad C\sharp \longmapsto 1 \qquad \dots \qquad B \longmapsto 11$$

allows to endow the set $\bar{\mathcal{P}}$ with a sum operation. Given $\bar{P}_1, \bar{P}_2 \in \bar{\mathcal{P}}$, the *interval* between $\bar{P}_1$ and $\bar{P}_2$ is defined as $\operatorname{int}(\bar{P}_1, \bar{P}_2) := \bar{P}_2 - \bar{P}_1$.

Remark. Given $\bar{P} \in \bar{\mathcal{P}}$ and $n \in \mathbb{Z}_{12}$, we will often use the notation $\bar{P} + n$ meaning $\iota^{-1}(\iota(\bar{P}) + n)$. We are sure that this will cause no confusion.

Definition. A *chord* $\mathcal{C}$ is a subset of $\mathcal{P}$ of cardinality $|\mathcal{C}| \geq 3$. We will say that $\mathcal{C} = \{P_1, P_2, P_3, \dots\}$ is *non-degenerate* if

$$\bar{\mathcal{C}} := \{\bar{P}_1, \bar{P}_2, \bar{P}_3, \dots\} \subset \bar{\mathcal{P}}$$

has cardinality ≥ 3, where $\bar{P}_i \in \bar{\mathcal{P}}$ denotes the pitch class of the pitch P_i. We will say that $\bar{\mathcal{C}}$ is the *abstract chord* associated to $\mathcal{C}$ and $\mathcal{C}$ is a *realisation* of $\bar{\mathcal{C}}$.

Example. The abstract chord $\{C, E, G\} \subset \bar{\mathcal{P}}$ represents what in music theory one usually refers to as *C Major chord*, without considering its disposition on the keyboard. The chord $\{C_4, E_3, G_3\} \subset \mathcal{P}$ is one realisation of it, providing what a musicologist would call *C Major in first inversion*, as the lowest pitch on the keyboard is E_3 and not the fundamental C_4.

In what follows we always suppose every chord considered to be non-degenerate. This is a common assumption also in traditional harmony theory. In any case, this is in fact a non restrictive hypothesis. For brevity, we will sometimes say just *chord* instead of *abstract chord*. This is a little abuse of notation with what we introduced above, but we are sure that this will not lead to any misunderstanding.

Definition. Given an abstract chord $\bar{\mathcal{C}} = \{\bar{P}_i\}_i$ and $n \in \mathbb{Z}_{12}$, we denote by $\bar{\mathcal{C}}+n := \{\bar{P}_i + n\}$ the abstract chord obtained by transposing all the pitch classes of $\bar{\mathcal{C}}$ by n. We say that $\bar{\mathcal{C}} + n$ is the *transposition* of $\bar{\mathcal{C}}$ by n.

3.3.1 Main types of chords

We proceed by introducing the main types of chords we will focus on. In music tradition, and especially in tonal music, there are some classes of chords that are recurrent and very commonly used in compositions, like major and minor chords, dominant chords, augmented chords, etc. We introduce them using the previous notation.

Major and minor chords The first type of chords we consider are major and minor chords. These can be consider the foundation of tonal music and maybe even of western music in general.

The set of *major chords* is defined as:

$$\mathfrak{M} = \{\{C, E, G\} + n \mid n \in \mathbb{Z}_{12}\}.$$

So for example $\{C + 4, E + 4, G + 4\} = \{E, G\sharp, B\}$ is a major chord, as well as $\{G, B, D\}$. The pitch class $C + n$ is called *fundamental pitch* of the major chord and we denote the major chord as $(C+n)M$. So for example the chord $\{E, G\sharp, B\}$ is denoted as EM, standing for E major. The pitch class $E + n$ (resp. $G + n$) is called the *third* (resp. *fifth*) of the chord.

Similarly, we define the set of *minor chords* as:

$$\mathfrak{m} = \{\{C, D\sharp, G\} + n \mid n \in \mathbb{Z}_{12}\}.$$

The definitions of fundamental, third and fifth is the same as for major chords and if $\bar{P}$ is the fundamental of a minor chord, we denote the chord by $\bar{P}m$. So for example Em is the chord $\{E, G, B\}$.

Seventh chords Seventh chords are chords consisting of 4 pitches, often obtained from major or minor chords with by adding a pitch class called the *seventh* of the chord. Here we present the ones we will consider:

- *Major seventh chords*:

  $$\mathfrak{M}_7 := \{\{C, E, G, B\} + n \mid n \in \mathbb{Z}_{12}\}$$

  are obtained by adding to major chords a *major seventh*, i.e. the pitch class obtained by the transposition of the fundamental by 11. If $\bar{P}$ is the fundamental, we denote them by $\bar{P}M7$.

- *Dominant seventh chords*:

  $$\mathfrak{do} := \{\{C, E, G, A\sharp\} + n \mid n \in \mathbb{Z}_{12}\}$$

  are obtained by adding to minor chords a *minor seventh*, i.e. the pitch class obtained by the transposition of the fundamental by 10. If $\bar{P}$ is the fundamental, we denote them by $\bar{P}Mm7$.

- *Minor seventh chords*:

  $$\mathfrak{m}_7 := \{\{C, D\sharp, G, A\sharp\} + n \mid n \in \mathbb{Z}_{12}\}$$

  are obtained by adding to minor chords a minor seventh. If $\bar{P}$ is the fundamental, we denote them by $\bar{P}mm7$.

- *Major-seventh minor chords*:

  $$\mathfrak{m}_{\hat{7}} := \{\{C, D\sharp, G, B\} + n \mid n \in \mathbb{Z}_{12}\}$$

  are obtained by adding to minor chords a major seventh. If $\bar{P}$ is the fundamental, we denote them by $\bar{P}mM7$.

- *Diminished seventh chords*:

  $$\mathfrak{di} := \{\{C, D\sharp, F\sharp, A\} + n \mid n \in \mathbb{Z}_{12}\}.$$

  If $\bar{P}$ is the fundamental, we denote them by $\bar{P}\circ$. Notice that there are only 3 types of diminished chords, as

  $$\bar{P}\circ = (\bar{P}+3)\circ = (\bar{P}+6)\circ = (\bar{P}+9)\circ\,.$$

- *Half-diminished chords*:

  $$\mathfrak{di}_{\mathfrak{h}} := \{\{C, D\sharp, F\sharp, A\sharp\} + n \mid n \in \mathbb{Z}_{12}\}.$$

  If $\bar{P}$ is the fundamental, we denote them by $\bar{P}\%$.

Other chords. For the purpose of analysing classical tonal music and contemporary music, several other categories of chords are to be considered. The following list does not pretend to be exhaustive and sum-up all the possible chords that one can find in western music: for the sake of brevity, we will write here only the most common ones.

- *Augmented chords*:
  $$\{\{C, E, G\sharp\} + n \mid n \in \mathbb{Z}_{12}\}.$$
  If $\bar{P}$ is the fundamental, we denote them by $\bar{P}^{+}$. Notice that there are only 4 types of diminished chords, as

  $$\bar{P}^{+} = (\bar{P}+4)^{+} = (\bar{P}+8)^{+}.$$

- *4-suspended chords*:
  $$\{\{C, F, G\} + n \mid n \in \mathbb{Z}_{12}\}.$$
  If $\bar{P}$ is the fundamental, we denote them by $\bar{P}sus4$.

- *2-suspended chords*:

  $$\{\{C, D, G\sharp\} + n \mid n \in \mathbb{Z}_{12}\}.$$

  If $\bar{P}$ is the fundamental, we denote them by $\bar{P}sus2$.

- *augmented sixths*: there are several types of augmented sixths. We introduce the so-called *German*, *Italian* and *French*.

$$\begin{aligned} Ger :=& \{\{C, E, G, A\sharp\} + n \mid n \in \mathbb{Z}_{12}\}. \\ It :=& \{\{C, E, A\sharp\} + n \mid n \in \mathbb{Z}_{12}\}. \\ Fr :=& \{\{C, E, F\sharp, A\sharp\} + n \mid n \in \mathbb{Z}_{12}\}. \end{aligned}$$

The careful reader will notice that the definition of German augmented sixths coincides with the dominant sevenths one. While having the same chords in two different classes results strange from the formal point of view, it turns out to be very useful in the musical context. Indeed in classical music theory dominant chords and German augmented sixths chords appear in different chord progressions and in particular augmented sixths have a very distinctive behaviour. Indeed the chord $C\,Ger$ is followed in most cases by BM, while the chord $CMm7$ is (almost) never followed by BM. This difference essentially defines German augmented sixth chords and allows one to distinguish them from the otherwise identical dominant seventh chords. Here, the choice of having two different classes of chords for the same pitch sets aims to reflect the different use of such chords made by composers. From the mathematical point of view, this choice does not create any problem for the purpose of the present work.

3.3.2 Other musical considerations

Tonality. Music tradition defines relations between chords: some chords are considered strictly related to some of the others. These relations depend from several factors, ranging from historically stable ones (as acoustic reasons) to rather mutable ones (as style of the periods and preferences of the composers). In an effort to maintain a useful flexibility of the concept, yet avoiding excessive relativism, we will assume the following simplification: two chords can be considered more strictly related the more frequently they appear close to each other in a given music corpus.

So for example, given almost any modern music corpus, the chords CM (i.e. $\{C, E, G\}$) and GM ($\{G, B, D\}$) will appear very often close to each other inside the chord progressions of the pieces, hence we will consider them closely related. By converse, the chords CM and $D\sharp m$ (i.e. $\{D\sharp, F\sharp, A\sharp\}$) appear very rarely close to each other, thus we will consider them to be very distantly related. In what follows, we will fix every time a corpus of reference, often called *testing corpus*, which all the relations between chords refer to. When we do not explicitly refer to a testing corpus, we intend to consider the corpus of all tonal music, or more precisely a theoretic corpus that respects all the traditional conventions about tonal music.

Fixed a testing corpus, the choice of a base chord induces a hierarchy on the other chords, based on how strong their relations are with the base chord with respect to the testing corpus. This hierarchy will be modelled as a graph, in the following chapters.

We say that a musical piece is *tonal* if there is a chord, typically either a major chord or a minor chord, that plays the role of *fundamental chord* of the piece. The fundamental chord is often both the first and the last chord of the composition and the hierarchy it defines reflects the frequency of chords in the piece: chords that are in closer relation to the fundamental chord tend to be more present in the chord progression of the piece.

The fundamental chord of a tonal piece is called the *tonality* of the piece. So for example if we consider a Sonata in the tonality of DM (D major), one would expect the chord of DM to be the first and the last chord of the piece, and the chords more strictly related to DM (for example GM, AM, etc.) to be more common than the less related ones. Notice that almost all classical music in the range XV–XIX century and a great part of contemporary music (jazz, pop, rock) can be considered tonal music: that's why *harmonic analysis*, which is the study of chord progressions and of relations between chords, plays a prominent role in music analysis and in music theory more generally.

The role of time and harmonic density. Rather than overall harmonic complexity, we will focus on local harmonic complexity, i.e. *harmonic density.* Indeed the concept of harmonic complexity in music theory is strictly related (even if often implicitly) to a local consideration of the evolution of chords, such as the *amount or variety of chords in a given time.* This reflects the human perception of harmony, which tends to highlight relations between chords that are close in the timeline. For simplicity, we will consider the chords of a piece as a sequence $(\mathbf{C}_i)_{0\leq i\leq k}$, thus ignoring the duration of each chord and possible overlaps of two chords. This allows us to consider *time-chords* as the couple $(\mathbf{C}_i, i)$ and define a distance function that takes the time i into consideration.

Chapter 4

Barcodes of chord sequences

4.1 Graph of chords

In this chapter, we present two possibilities for the construction of a graph that represents the chords of a musical piece (or corpus, in general) and that retains the tonal relations between the chords. Both these constructions provide in fact a sort of generalisation of the Tonnetz, which is one of the main tools for chords analysis in math-music research.

In Section 4.1.1, we briefly discuss the limits of the Tonnetz and the reasons that motivated the present investigation. In Section 4.1.2, we present the construction of a non-directed graph with manually defined basic distances between chords.

In Section 4.1.3, we provide the construction of a directed graph whose distances are extracted from a given musical corpus.

4.1.1 The Tonnetz(e)

One of the main tools for chords representation in math-music theory is the Tonnetz(e). For a recent review, see [41]. Given the aim of characterising tonal music, so considering chords that are largely based on major and minor chords, the Tonnetz (3,4,5) seems the most reasonable one to consider and it is the one we focus on (see Figure 4.1).

In particular, we are focusing on its dual, where minor and major chords are represented in a lattice that originates a torus when the equivalence modulo octave is considered. This has been extensively and successfully used to represent major and minor chords. It has been also extended (see [39]) to represent other classes of chords, like diminished seventh chords.

If we want to consider a distance between chords, we can use the Tonnetz and take the minimal path distance (as in Equation (3.2)). Then, one can use the Vietoris-Rips construction to associate a simplicial complex to a musical piece.

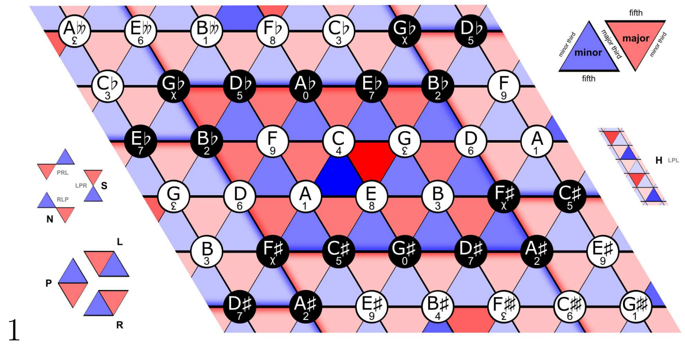


Figure 4.1: T. Piesk, July 2014, Neo-Riemannian Tonnetz with generators (3,4,5), from en.wikipedia.org/wiki/Tonnetz/.

However, the rich but rigid geometric structure of the Tonnetz limits the possibility of deforming it arbitrarily. For example, it is not easy to get two chords closer to each other or, for example, pushing one chord infinitely far away from the others. Yet, these operations could sometimes appear quite natural from the musical point of view, for example if we want to adapt our representation to various musical genres or styles. For example, it could be desirable to be able to consider two different chords to be very similar from the harmonic point of view, which translates to squeeze them close in the Tonnetz, or to consider one chord to be very bizarre and harmonically distant from the others, which would correspond to send it to infinity in the Tonnetz. In performing these deformations, the Tonnetz turns out to be neither flexible nor easy to work with.

Hence we substitute for the Tonnetz a simpler object, that manages to represent the chords and their relations without the limitations of the Tonnetz's rigidity and duality. This leads us to consider a graph of chords, whose weighted edges provide a flexible tool for adapting chords' relations to various contexts.

4.1.2 Non-directed graph and basic relations

The first construction we used is a non-directed graph with manually defined basic relations between chords.

As a first step, we consider only major, minor and dominant chords. So let G be a weighted graph whose vertices are the points $\mathfrak{M} \cup \mathfrak{m} \cup \mathfrak{do}$ (with the notation of Section 3.3.1) and whose edges are $D \cup F \cup H \cup R$ where:

$$\begin{aligned}
D =& \big\{\{PM, (P+7)M7\} \mid PM \in \mathfrak{M}\} \cup \{(Pm, (P+7)M7) \mid Pm \in \mathfrak{m}\big\}\\
F =& \big\{\{PM, (P+7)M\} \mid PM \in \mathfrak{M}\} \cup \{(Pm, (P+7)m) \mid Pm \in \mathfrak{m}\big\}\\
H =& \big\{\{PM, PM7\} \mid PM \in \mathfrak{M}\big\}\\
R =& \big\{\{PM, (P+9)m\} \mid PM \in \mathfrak{M}\big\}.
\end{aligned}$$

Musically, D provides the edges connecting each major and minor chord to its dominant, F connects each major and minor chord to its fifth,H connects each major chord to itself with added minor seventh and R connects each major chord to its relative minor chord.

Given $d, f, h, r \in \mathbb{R}$ and denoted by ω_G the weights function associate to G, we define the following weights:

$$\begin{aligned}
\omega_G(v) &= d \quad \forall v \in D, \qquad & \omega_G(v) &= f \quad \forall v \in F,\\
\omega_G(v) &= h \quad \forall v \in H, \qquad & \omega_G(v) &= r \quad \forall v \in R.
\end{aligned}$$

We refer to the edges in D, F, H, R and the respective choices of weights d, f, h, r as *basic relations*.

As one can notice, this is a very similar construction to the Tonnetz, where instead of the generating translations (3,4,5) we choose a set of generating relations in the graph. This permits to easily tweak the graph in order to match certain expected musical behaviour. For example, by modifying the parameters d, f, h, r or by adding new basic relations, it is possible to adapt the graph to different music styles.

Clearly, one can add other families of chords to the graph G as vertices. Since we need G to be connected to be able to define the minimal path distance, one also needs to define a corresponding basic relation that connects the added vertices to the others.

4.1.3 Directed graph and database extraction

Another construction is using a directed graph. The interest for directed edges arises from the desire to model also the idea of directionality of a chord progression. Indeed, the traditional theory of functional harmonic analysis suggests that in a chord progression each chord plays a role that strongly depends on the progression itself, which involves in particular the temporal order of the chords. For example, in the tonality of C major the progression $GM \to CM$ (a descending fifth) is considered to be very *stable*, often conclusive of the harmonic progression of a phrase or theme or section. In this case, it is usually called *perfect cadence*. Instead, the cadence $CM \to GM$ is considered less stable and it is called *imperfect* or *suspended cadence.* This asymmetry, which is well known and recognised in music theory, suggests to consider separately the progressions $GM \to CM$ and $CM \to GM$ and possibly with different weights. This leads to consider directed graphs, where we can consider distinct edges (GM, CM) and (GM, CM) with different weights.

Another aspect is how to define the weights of the edges of the graph. Indeed, while defining them by hand (as done in the non-directed case) provides the maximal flexibility, it requires the choice of at least two parameters for every class of chords, in order to have G connected. This results in a considerable complexity, where the use of several independent parameters causes a rapid increase in the difficulty of fine-tweaking the model.

A natural choice is to extract these values directly from a given musical corpus, in order to inherit the relations of that corpus and somehow encode the chord progressions style of it in the graph.

Definition. A musical corpus $\mathcal{M}$ is a set of chord sequences of the type $S = (\mathbf{C}_i)_{0 \leq i \leq \tau_S}$ where $\tau \in \mathbb{N}$ and for every i, $\mathbf{C}_i$ is a chord as described in Section 3.3.1. In this case we write $\mathbf{C}_i \in \mathcal{M}$. τ_S is the number of chords in the sequence S and the total number of chords in $\mathcal{M}$ is denoted as $\tau_{\mathcal{M}} := \sum_{S \in \mathcal{M}} \tau_S$.

For every pair of chords $\mathbf{C}_1, \mathbf{C}_2 \in G$, we write $(\mathbf{C}_1, \mathbf{C}_2) \in \mathcal{M}$ if there exists $S \in \mathcal{M}$ that contains the chord progression $(\mathbf{C}_1, \mathbf{C}_2)$. If $(\mathbf{C}_1, \mathbf{C}_2) \in M$, we define the *frequency* $\nu(\mathbf{C}_1, \mathbf{C}_2)$ as the total number of instances of the chord progression in $\mathcal{M}$, divided by $\tau_{\mathcal{M}}$.

Definition. Given a musical corpus $\mathcal{M}$, we define the *associate directed weighted graph* $G_{\mathcal{M}}$ as follows:

- as vertices, we consider the set of all the chords contained in $\mathcal{M}$.
- As directed edges, we consider the set of all the couples $(\mathbf{C}_1, \mathbf{C}_2) \in M$.
- As weights, we take $\omega_G(\mathbf{C}_1, \mathbf{C}_2) = \frac{1}{\nu(\mathbf{C}_1, \mathbf{C}_2)}$.

We denote by $\bar{G}_{\mathcal{M}}$ the *extended graph* associated to $G_{\mathcal{M}}$, where we consider vertices as in $G_{\mathcal{M}}$ and edges:

$$(\mathbf{C}_i, \mathbf{C}_j) \text{ if } (\mathbf{C}_i + n, \mathbf{C}_j + n) \text{ is edge of } G_{\mathcal{M}}, \quad \exists\, n \in \mathbb{Z}_{12}.$$

Remark. Notice that the graph $G_{\mathcal{M}}$ retains the precise information about the chords used in the corpus $\mathcal{M}$, hence including the choice of tonalities. The extended graph $\bar{G}_{\mathcal{M}}$ flattens down the data about the tonalities and which chords are used, and preserves only the informations about the progressions of chords as mutual degree (see [38] for details). This turns out to be very useful in practice, as it allows to analyse a certain musical piece using a corpus $\mathcal{M}$ even if the piece and the corpus don't possess chords in common.

4.2 The Beatles with non-directed graph

In this section we describe the first attempt to associate to a musical piece a barcode representing its harmonic content. The aim is to manage to find a correspondence between the mathematical properties of the barcodes and the musical properties of the musical piece form the harmonic point of view, as the harmonic complexity of the piece.

To do this, we used as database the annotations of The Beatles's discography and a graph as anticipated in Section 4.1.2.

4.2.1 The database

In order to consider a graph of chords, we need a database containing musical pieces as sequences of chords. The problem of (automatically) extracting the sequence of chords from a musical piece, usually called *harmonic analysis* of the piece in music theory, is a non-trivial one in math-music literature (see [27] for a comprehensive discussion). Indeed there are several delicate aspects in designing an algorithm that produces sequences of chords (or labels) associated to audio files or musical scores. Among the algorithmic ones, there are several musical techniques like anticipation and suspension, which contribute to the complexity of correctly analysing a musical piece. From the music theory point of view, there is also the problem regarding the not unique choices in associating a label to a chord, especially in more harmonically rich settings. This is well known to musicologists, who sometimes debate on which label is the more correct one, given a chord as set of pitches.

The database is the online database on The Beatles's discography of the Centre for Digital Music at Queen Mary, University of London (see [33]). It is built using the algorithm for automatic harmonic labels presented in [32]. In order to be used in the model, some pre-processing functions has been applied to the CSV files associated. In particular, since we used only major, minor and dominant chords, we simplified to these three categories the various chords of the original database. In particular, following the notation of Section 3.3.1, we converted:

- diminished and half-diminished chords into dominants;
- minor seventh and major-seventh minor chords into minors;
- augmented chords and major seventh into majors;
- all the added notes (like 9ths, 11ths etc.) were ignored.

Finally, we set the time duration of each chord to 1, hence we model each song as a sequence $(\mathbf{C}_i, i)_{0\leq i\leq k}$.

4.2.2 Barcodes and analysis

We deepen here the algorithm we used to associate a set of barcodes to each musical piece of The Beatles. First, we defined the graph G of all the chords considered: major, minor and dominant chords. We used the construction for non-directed graphs presented in Section 4.1.2, with the following choices of parameters:

$$d = 10 \qquad f = 11 \qquad h = 7 \qquad r = 12.$$

With these values, the graph G has diameter 51 and we denote by d the minimal-path distance associated. The graph G is plotted in Figure 4.2.

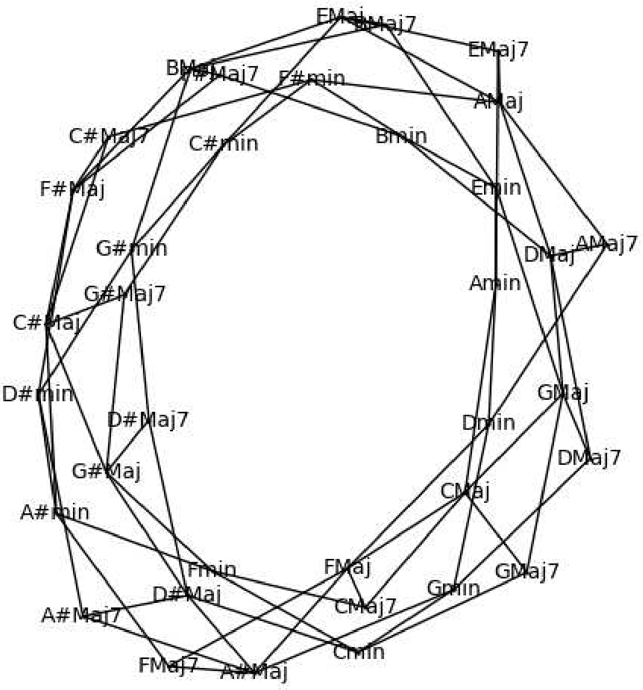
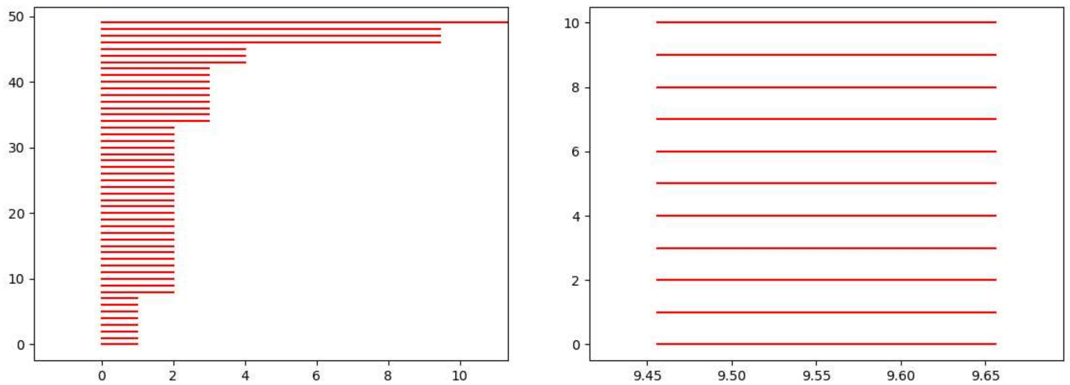


Figure 4.2: The graph of chords G.

Given a Beatles song as the sequence of chords $S = (\mathbf{C}_i)_{1 \le i \le \tau_S}$, we denote the sequence of time-chords as $\bar{S} = (\mathbf{C}_i, i)_{1 \le i \le \tau_S}$. Given two time-chords $(\mathbf{C}_i, i), (\mathbf{C}_j, j)$ in $\bar{S}$, we define the following function:

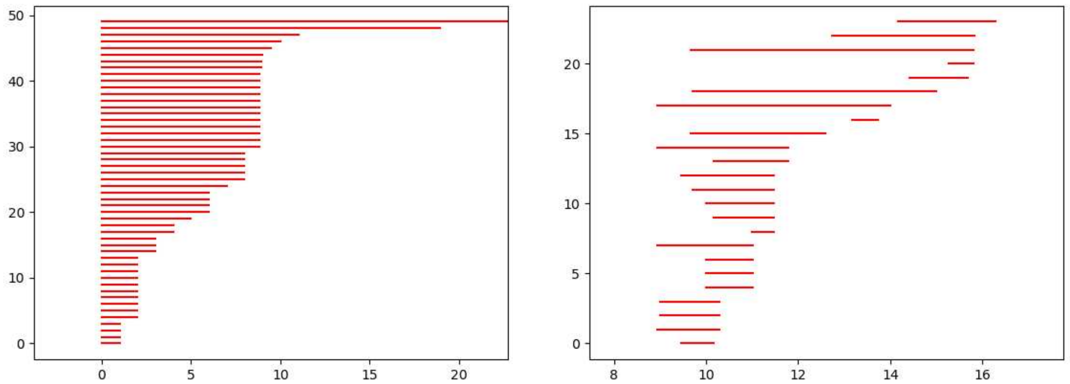
$$\bar{d}\big((\mathbf{C}_i, i), (\mathbf{C}_j, j)\big) = |j - i| + \frac{d(\mathbf{C}_i, \mathbf{C}_j)}{\log(e + |j - i|)}.$$

Indeed, as anticipated in Section 3.3, we are mainly interested in harmonic density and the function $\bar{d}$ serves the purpose of balancing the distance between chords obtain from G with the time distance, on a logarithmic scale. We built the Vietoris-Rips filtration on $\bar{S}$ with respect to $\bar{d}$ using the package PERSIL on SAGEMATH, over the field $\mathbb{F}_2$. The same package allows to compute the homology of the filtration and get the persistent barcodes associated.



(a) Degree 0. (b) Degree 1.

Figure 4.3: Barcodes of *Love Me Do*, from *Please Please Me*.



(a) Degree 0. (b) Degree 1.

Figure 4.4: Barcodes of *She's Leaving Home*, from *Sgt. Pepper's*.

In Figure 4.3 and Figure 4.4 we plot the barcodes of two songs as an example. In order to gather some features from the barcodes, we selected a group of particular values, suspected to be able to describe the concept of harmonic complexity that we are trying to model. In particular, considering only the bars with finite length in the barcodes associated to degree 0 and 1, we focused on the following:

- the maximal length;
- the average length;
- the variance of lengths.

As an example, we depict these values extracted from the album *Please Please Me* in Table 4.1 and from the album *Sgt. Pepper's Lonely Hearts Club Band* in Table 4.2.

Song Name	**0-M**	**0-A**	**0-V**	**1-M**	**1-A**	**1-V**
01 I Saw Her Standing There	16.3	4.0	9.4	0.9	0.3	0.1
02 Misery	10.1	4.2	6.6	0.7	0.4	0.1
03 Anna (Go To Him)	10.3	3.9	10.5	5.6	1.4	4.6
04 Chains	12.6	4.4	6.3	2.0	0.7	0.3
05 Boys	9.5	3.3	2.2	0.2	0.2	0.0
06 Ask Me Why	13.3	5.7	12.9	6.2	2.4	5.1
07 Please Please Me	16.3	4.7	12.7	6.8	1.2	4.2
08 Love Me Do	9.5	2.6	3.6	0.2	0.2	0.0
09 P. S. I Love You	16.9	5.6	22.3	4.9	1.6	2.7
10 Baby It's You	10.1	4.5	6.9	6.6	1.3	4.2
11 Do You Want To Know A Secret	16.9	7.0	11.7	7.4	1.6	4.7
12 A Taste Of Honey	11.0	4.4	12.1	5.5	1.4	2.7
13 There's A Place	12.5	5.2	8.5	6.6	1.6	4.5
14 Twist And Shout	9.5	3.3	1.6	0.2	0.2	0.0

Table 4.1: Maximal length, Average and Variance in 0 and 1 degree barcodes of the album *Please Please Me.*

Song Name	**0-M**	**0-A**	**0-V**	**1-M**	**1-A**	**1-V**
01 Sgt. Pepper's L. H. C. B.	13.1	5.0	7.9	3.0	0.8	0.9
02 With a Little Help	16.9	4.9	15.4	5.1	0.9	1.3
03 Lucy in the Sky with Diamonds	12.5	5.1	7.4	6.0	1.1	2.5
04 Getting Better	10.2	3.5	7.2	1.3	0.7	0.2
05 Fixing a Hole	13.1	3.7	14.4	3.4	1.4	1.8
06 She's Leaving Home	19.0	6.1	13.5	6.2	2.0	2.3
07 Being for the Benefit of Mr. Kite!	15.3	6.0	13.9	4.8	1.5	1.6
08 Within You Without You	18.9	3.2	21.0	0.3	0.3	0.0
09 When I'm Sixty-Four	16.9	6.2	19.5	6.2	2.0	2.7
10 Lovely Rita	9.5	5.4	6.4	5.7	1.4	2.4
11 Good Morning Good Morning	10.3	4.5	8.3	3.1	1.2	0.9
12 Sgt. Pepper's L. H. C. B. (Rep)	15.3	5.0	11.0	1.0	0.4	0.1
13 A Day in the Life	9.5	6.0	5.1	6.5	2.3	4.4

Table 4.2: Maximal length, Average and Variance in 0 and 1 degree barcodes of the album *Sgt Pepper's Lonely Hearts Club Band.*

These two albums are particularly suited to distinguish the relations between these coefficients and the intuitive harmonic complexity. Indeed, *Please Please Me* (which we will shortly denote by *PPM*) is the first of the group, and it is often considered its simplest one, for the '50s rock-and-roll influence being still very strong in the album. On the other hand, *Sgt. Pepper's Lonely Hearts Club Band* (which we will denote by *SGP*), is usually considered the most experimental album and among the richest ones from the harmonic and structural point of view. While the clear musical distinction is not evidently reflected by the barcodes coefficients, it is worth noticing that lower scores in maximal length and average length, both in degree 0 and 1, are associated to songs which can be considered very simple from the harmonic point of view. Indeed observe that the songs with low scores in 0-maximal lengths, as 05,08,14 of *PPM* and 10,13 of *SGP*, have very basic harmonic progression. In particular for *PPM*, the songs have the typical rock-and-roll static harmonic structure and notice that their score also in 0-average length is very low.

On the other hand, songs with more complex harmonic structure, such as 07,11 in *PPM* and several ones in *SGP*, are associated to higher scores for maximal length and average, both in degree 0 and degree 1.

These naive correspondences are confirmed also in the other album analysed, with higher scores for maximal lengths and average generally associated to richer harmonic content.

It is interesting also observing the case of *I saw her standing there*, which has a very basic harmonic structure and a rock-and-roll style, but has a good score in 0-maximal lengths, perhaps influenced by the non-standard cadence (with respect to the chord distance used) at the end of the chorus.

Of course, there are several aspects that influence the barcodes and that can be considered as disturbing elements. First, the chord sequences associated to the various songs have different lengths, a lack of homogeneity that affects the consistency of the scores, considering that chords distant in time are *pushed apart* in the simplicial complex. Moreover, the high approximation of the chords, reduced to only three types (major, minor, dominants), often artificially alters the result with respect to the perceived harmonic complexity of the songs. Indeed the latter not rarely is strongly related to added notes (7ths, 9ths, etc.) that are present in the original score but not taken into consideration here. Finally, the arbitrary choice of the parameters of the chords graph, even if quite reasonable a priori, can have unexpected results in practice.

4.3 Classical Music with directed graph

In this section we describe the second attempt to associate to a musical piece a barcode representing its harmonic content. This time, we focused on classical music using the database of the Digital and Cognitive Musicology Lab at École Polytechnique Fédérale de Lausanne. Considered the richer harmonic structure of this corpus, we included all the types of chords listed in Section 3.3.1 and we built the graph of chords as in Section 4.1.3.

4.3.1 The database

The composers. The database we used has been created by the Digital and Cognitive Musicology Lab at École Polytechnique Fédérale de Lausanne (see [26]). This database contains several music corpora, with a complete analysis of the measures, chords and harmonies. We focused specifically on the following:

- Annotated Beethoven Corpus, containing all Beethoven's string quartets.
- Beethoven's Piano Sonatas.
- Chopin's Mazurkas.
- Corelli's Trio Sonatas.
- Debussy's *Suite Bergamasque.*
- Liszt's *Pelegrinage.*
- Mozart's Piano Sonatas.

- Schumann's *Kinderszenen* op. 15.
- Tchaikovsky's *Seasons* op. 37a.

In particular, we used the analysis from the *harmonies* section of the database. It is worth stressing out that this contains the harmonic annotations of the musical pieces instead of the sequence of chords. This is a non-trivial difference from the music theory point of view, but we preferred to ignore these aspects and consider the analysis conducted from Lausanne Lab as sequences of chords. We are sure that even the most careful reader agrees that harmonic annotations are at least an acceptable approximation of chords analysis. More in detail, the notation used in the database, based on global/local tonalities and scale degrees, was converted in the simpler notation as in Section 3.3.1, with a number $n \in \mathbb{Z}_{12}$ expressing the fundamental and a short string (*"M", "m", "Mm7"*) to denote the type of chord. We denote by $\mathcal{M}$ the musical corpus of the chord sequences obtained by the union of the musical compositions listed above. We used $\mathcal{M}$ both to construct the graph of chords $G_{\mathcal{M}}$ as shown in Section 4.1.3 and as a testing set of chord sequences.

Frequencies extraction and composers' balance. When extracting the frequencies related to chord progression, we took into account the remarkable difference of dimension of the various composers' corpora. For example, the corpus of Beethoven's piano sonatas (resp. string quartets) has a total of 18928 (resp. 24303) chords, while the corpus of Debussy's *Suite Bergamasque* only contributes with 920 chords. We decided to balance the contribution of each composer by multiplying the frequencies deriving from each composer by a suitable correction factor. We preferred to apply this correction instead of just restricting the total number of chords of each composer to a fixed (forcedly small) upper-bound in order to retain the maximum amount of information from the corpora.

The possibility of adjusting the balance of the various contributions to $\mathcal{M}$ highlights the flexibility of the construction: it is very simple to modify how much a certain composer *influences* the corpus by changing the correspondent correction factor.

4.3.2 Barcodes and analysis

The construction of Section 4.1.3 provided the group $G_{\mathcal{M}}$ of chords and relations. We denote by d the minimal-path weight function that associate to every couple of chords $(\mathcal{C}_1, \mathcal{C}_2)$ the weight of the minimal directed path in G from $\mathbf{C}_1$ to $\mathbf{C}_2$. Remember that here d is not a distance, as it lacks symmetry. In Table 4.3 we report some of the significant values of d, for comparison with the definitions used in Section 4.2.2.

Chord progression	**Weight**
$CM \to GM$	25.72
$CM \to GMm7$	21.47
$CM \to FM$	17.57
$CMm7 \to FM$	13.11
$CM \to Am$	118.39

Table 4.3: Some significant weights of the graph of chords $G_{\mathcal{M}}$

As one can notice, the values are very close for very common chord progressions, but quickly increase for rarer ones. For example, the weight of the progression $CM7 \to BmM7$ is 10737.29, as major-seventh minor chords are quite scarce in the corpus. It is interesting to observe also some unexpected results: the progression $CM \to Am$, which one could expect to be very common, has already weight 118.39, bigger than the tritone $CM \to G\sharp M$, which stops at 105.47.

In order to build a testing set, we divided each musical piece in the corpus in several parts having the same number of chords, in order to have comparable results. To do this, we used thresholds τ equal to 5, 10 and 20 chords and obtained chord sequences $S = (\mathbf{C}_i)_{1\leq i\leq \tau}$.

Similarly to The Beatles case, we included time into consideration, so to have a local description of the harmonic content rather than the overall harmonic complexity to The Beatles case, we included time into consideration, so to have a local description of the harmonic content rather than the overall harmonic complexity. In this case, we extended the function d on time-chords $(\mathbf{C}_i, i)$ as follows:

$$\bar{d}\big((\mathbf{C}_i, i), (\mathbf{C}_j, j)\big) = \begin{cases} 0 & \text{if } i = j \\ \log(|j-i|^{\alpha}) & \text{if } \mathbf{C}_i = \mathbf{C}_j \\ \max\{1, \log(d(\mathbf{C}_i, \mathbf{C}_j)|j-i|^{\alpha})\} & \text{otherwise.} \end{cases}$$

Here the parameter $\alpha \in \mathbb{R}$ allows to tweak the importance of time in the computation: by increasing its value one gets a more local harmonic description, while $\alpha = 0$ gives the overall harmonic complexity. In what follows, we consider $\alpha = 1$.

Then, for every music piece S, together with the weight function $\bar{d}$, we built a simplicial complex using the Dowker filtration and we computed the persistent homology. To do so, we used the package SIMPLICIAL, available in JULIA, which is one of the few implementations of the Dowker filtration. Thus, for every chord sequence S we get the barcodes of degree 0 and 1.

As done in Section 4.2.2, we associate to each barcode a set of values (*features*) that allows to conduct further analysis on the barcodes. In particular, restricting to finite length bars of the barcode, we define: maximal length, second maximal length, average length and variance length. One can consider barcode entropy as well as a feature.

In Table 4.4 we depict some of these values of barcodes obtained from musical pieces having chord threshold $\tau = 20$. More precisely, one can read the features **0-ML**, **0-A**, **0-V**, **1-ML**, **1-A** and **1-V**, corresponding to maximal length, average length and variance length for degrees 0 and 1. For maximal length and average, we wrote both the mean and the maximum (with respect to the corpus), while for variance we wrote only the mean.

Corpus	**0-ML**	**0-A**	**0-V**	**1-ML**	**1-A**	**1-V**
Beeth. Quart.	39.6, 125	9.6, 24	152	8, 51	4.8, 31.5	12.5
Beeth. Son.	41.5, 112	9.4, 23.9	159	7.8, 69	4.6, 26.3	10.9
Chopin	34.3, 106	8.5, 25.4	125	7, 41	4.3, 21	8.9
Corelli	43.8, 138	10.6, 27.9	187	9.7, 49	5.1, 26	20.6
Debussy	46.8, 106	11.0, 22.1	200	11.3, 34	5.9, 18	23.4
Liszt	38.2, 99	10.0, 22.7	156	7.7, 40	4.3, 21.5	12.7
Mozart	42.7, 138	9.9, 25.1	169	7.5, 48	4.5, 28	10.6
Schumann	40.6, 100	10.3, 15.6	142	8.6, 17	4.7, 8.5	8.7
Tchaikovsky	43.6, 106	10.2, 20.3	176	10.3, 40	6, 29	19.6

Table 4.4: Values associated to the barcodes from each corpus.

As expected, to higher values (**0-ML**, **0-A**, **1-ML**, **1-A**) correspond more harmonically complex corpora. In the case of Debussy (which has the highest values) and Tchaikovsky, the high complexity probably comes from the wide range of chords used in the corpus, and in particular for Debussy also from the frequent use of non-standard chord progressions from the tonal point of view. Notice that Debussy has high mean values, both in degree 0 and 1, while keeping relatively low maximal values, somehow showing a high average harmonic complexity throughout the corpus without many harmonic density peaks.

Corelli's corpus also has interestingly high values, this time probably caused by the considerable density of harmony and the quick changes of tonality (so called *modulations*), typical of Baroque period and polyphonic music, rather than tonally unusual chord progressions. This interpretation is supported by Corelli's high scores in maximal values, which would suggest the presence of very dense harmonic sections in the corpus.

On the other side of the spectrum, Chopin's corpus are characterised by very low scores, which well describe the limited overall harmonic complexity and the low density of harmonic changes of Mazurkas.

Non surprisingly, Beethoven's corpora have very similar scores and both occupy the middle of the spectrum with respect to almost every value. Also Schumann can be placed more or less in the middle. Liszt's corpus seems more complex to analyse: it has quite low scores in all the indicators except **0-A**. In any case, it is worth noticing that *Pelegrinage* is a very heterogeneous collection of pieces, with remarkably different artistic choices and musical structures.

4.4 Further Analysis

4.4.1 Machine Learning Techniques

The analysis on the barcodes presented until now is limited to some general considerations, trying to link the values obtained from the barcodes with the intuition of the musical properties of the corpora. In order to have stronger and more objective results, we employed some machine learning techniques to further analyse the barcodes and, more specifically, the features extracted from them. We focused on two main tools, whose use in machine learning is well established: support vector machine (SVM) and multi-linear regression. We present these approaches in the following sections.

Support Vector Machine. Support Vector Machine is one of the most common techniques for classification in machine learning. We briefly present here the basic definitions. Consider a training set of parameters $\{\mathbf{x}_i\}_i \subset \mathbb{R}^n$ with a corresponding set of values $\{y_i\}_i \in \mathbb{R}^n$. Here typically $y_i \in \{-1, 1\}$ and represents the property function that one wants to predict. For example, in our case $\S_i$ is the vector of features associated to a barcode of a musical piece and y_i has value 1 if the musical piece has a fixed property, for example being composed by Beethoven, and -1 otherwise. We say that $\mathbf{x}_i$ is a *positive* (resp. *negative*) training point if the corresponding $y_i = 1$ (resp. $y_i = -1$. The aim of the algorithm is to find an hyperplane $\mathbb{H}$ of $\mathbb{R}^n$ (or an algebraic hyper-surface, more generally) that divides the set of positive training points from the one of negative training points. Then one can use $\mathbb{H}$ to estimate the value $y_j \in \{-1, 1\}$ of a testing point $\mathbf{x}'_j \in \mathbb{R}^n$. More formally, the goal is to find $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ solving the following problem:

$$\begin{aligned} \min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{s} \zeta_i \\ y_i(w^T \phi(\mathbf{x_i}) + b) \geq 1 - \zeta_i \\ \zeta_i \geq 0 \text{ for } i = 1, \dots, s \end{aligned}$$

where s is the cardinality of the training set, ϕ is a linear transformation, ζ_i is the error we admit on the training point i and C is a penalty term that forces to stay next to the margin boundary. Indeed with a perfect prediction, i.e. if the two subsets of positive and negative training points are actually perfectly separable by an hyperplane, one has $y_i(w^T\phi(\mathbf{x_i})+b) \geq 1$. In applications, this situation is very rare, and one usually accepts to miss-classify the point $\mathbf{x}_i$ by introducing the parameter ζ_i, which expresses the distance of the point from the boundary of the hyperplane $\mathbb{H}$.

We used SVM to attack the problem of automatic composer classification. We divided the barcodes obtained in Section 4.3.2 in two subsets: Beethoven's chord sequences and non-Beethoven ones. Using the previous notation, we took as parameters $\mathbf{x}_i$ the vectors

$$(\textbf{0-E}, \textbf{0-ML}, \textbf{0-A}, \textbf{0-V}, \textbf{1-E}, \textbf{1-ML}, \textbf{1-A}, \textbf{1-V}) \in \mathbb{R}^8$$

where **0-E** (resp. **1-E**) denotes the entropy of the barcode of degree 0 (resp. degree 1). As anticipated, we set $y_i = 1$ if $\mathbf{x}_i$ is a positive point (i.e. corresponds to a Beethoven's chord sequence) and $y_i = 0$ otherwise.

To perform the computations, we used the package SVM provided by SciKit (see [37]) in Python. In particular, we used the function StandardScaler to properly scale the coefficients and the built-in SVC algorithm. However, we did not obtain particularly interesting results: we never managed to obtain a score higher than 0.56, where 0.5 denotes the score of a random classification function. This remained the case even applying some tweaks to the input: we modified the chord threshold of the chord sequences used to extract the barcodes ($\tau = 5, 10, 20$), and the best result (a score of 0.558) was obtained with $\tau = 20$. We also tried to use an enlarged database composed of the barcodes associated to chord sequences having all three possible thresholds, but this did not significantly improve the accuracy. There are still many tweaks that can be explored, as adjusting the parameter α of time in the computation of time-chord distances, or manually modify the graph of chords to suit it to an expected outcome.

Multi-linear regression. We obtained slightly better results using another machine learning tool: linear regression. The mathematical setting is similar to SVM: given training set $(\mathbf{x}_i, y_i) \in \mathbb{R}^{n+1}$, one wants to determine $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^n$ such that

$$y_i = \beta_0 + \beta^T\mathbf{x}_i + \epsilon_i \qquad \forall\, i$$

where ϵ_i denotes an error variable. This allows to use (β_0, β) to predict the value y_j of a test point $\mathbf{x}_j$.

Again, we used SciKit (see [37]) in Python with its Linear Regression models. Similarly to SVM, we define the features vector as:

$$\mathbf{x}_i = (\textbf{0-E}, \textbf{0-ML}, \textbf{0-A}, \textbf{0-V}, \textbf{1-E}, \textbf{1-ML}, \textbf{1-A}, \textbf{1-V}) \in \mathbb{R}^8$$

and the corresponding value $y_i = 1$ if the chord sequence is composed by Beethoven, $y_i = 0$ otherwise. We considered several possibilities for the chord sequences constituting the database: we tried with chord threshold $\tau = 5, 10, 20$ and even all the three choices combined. When considering a positive database given by all Beethoven's compositions and a negative database composed by all non-Beethoven compositions, we did not get particularly interesting results: the best score was obtained with $\tau = 20$ and it was equal to 0.010, where 0 is the random prediction and 1 is the perfect prediction. Better results were obtained by restricting the dataset. For example, we reduced the positive database to Beethoven's Sonatas only and the negative database to Debussy's, Schumann's and Tchaikovsky's compositions. We made a random selection of Beethoven's barcodes in order to have an even cardinality of positive and negative part. With these modifications, the score was consistently around 0.2 with a peak of 0.27. Of course these results are still too low to get accurate predictions, but they suggest that it is possible to fine tweak the model to increase the accuracy of the regression.

4.4.2 Experiment on perception

In collaboration with Dr. Andrew Milne from MARCS Institute for Brain, Behaviour and Development at Western Sydney University, we conducted an experiment on human perception of harmonic complexity. The aim of the experiment was to gather some information regarding how harmony is perceived by people and to verify a possible correlation with the mathematical characterisation of the chord progressions based on persistent homology analysis. Moreover, considering the big flexibility of the model we used, real world data on perception can be used to tweak the various parameters involved and adjust the algorithmic predictions to better suit the empirical results. The first part of experiment involved psychology students from Western Sydney University, which were considered *naive* from the musical point of view. A second part is currently underway and it involves students from Italian music conservatory, which can be considered *experts*.

We briefly describe the design of the experiment. A database of musical composition was selected, evenly balanced between classical music (Bach, Beethoven, Debussy, Scriabin, Rachmaninov, Prokofiev) and pop-rock music (The Beatles and several other well-known bands), for a total of 48 musical pieces. From each musical piece the sequence of chords was manually extracted and reduced to a fixed length range (between 15 and 23 chords), in order to have sufficiently even examples for persistent homology analysis, yet maintaining the length of the original theme/phrase of the piece. Then the sequence was analysed using Dowker filtration with respect to the directed graph as in Section 4.3.2, obtaining barcode features a previously described. From each chord in the sequence, a 4-notes realisation (as in Section 3.3) was chosen with the aim of respecting the disposition of the original musical piece. The sequence of realised chords was then converted in an MP3 audio file, via the software MuseScore.

In order to have a direct comparison between couple of samples, we presented to the participants two examples, randomly chosen, we asked them to rate their similarity and to choose the most interesting, unusual and enjoyable. Also, to control in-test learning, after every sample we asked if the participant had already heard that example. The question regarding interestingness and unusualness were designed to express indirectly the concept of harmonic complexity, which can be too complicated to present to naive audience. The rate of similarity was intended to analyse a correlation with the barcode distance. In short, the participants were tested with the following scheme (repeated 35 times):

- Listening to the first examples + already-heard question.
- Second examples + already-heard question.
- Rate similarity of the two examples.
- Which one is more unusual/interesting/enjoyable.

We are still analysing the results, again using linear regression to highlight correlated indicators. These are the observations arising from the first tests:

- The participants tended to prefer the second example (unusual, interesting, enjoyable).
- Little inverse correlation between the entropy of barcode in degree 0 and unusualness preference: examples with higher entropy tended to be less chosen as *more unusual* (correlation factor of 0.14).
- Chord sequence from classical music tended to be considered more unusual than chord sequences with simple structure from pop/rock music.

With further analysis, we will attempt to tweak the predictions arising from persistent homology analysis in order to reflect human perception. In doing this, an important step will be involving music experts in this sort of test, which will hopefully provide more accurate responses on which modelling persistent homology parameters.

Bibliography

[1] N. Ancelotti, *On some algebraic aspects of Anatol Vieru Periodic Sequences*, Tesi di Laurea Triennale in Matematica, Universitá degli Sudi di Padova, Relatore L. Fiorot.

[2] M. Andreatta, D.T. Vuza, *On some prperties of periodic sequences in Anatol Vieru's modal theory*, Tatra Mountains Mathematical Publications, **23**, (2001) pp. 1–15.

[3] M. Andreatta, C. Agon and D.T. Vuza, *Analyse et implementation de certaines techniques compositionnelles chez Anatol Vieru*, Actes des Journées d'Informatique Musicale, Marseill (2002), pp.167-176.

[4] M. Andreatta, D.T. Vusa, C. Agon, *On some theoretical and computational aspects of Anatol Vieru's periodic sequences*, Soft Computing **8**, no. 9, (2004), pp. 588–56.

[5] K.S. Davis, W.A. Webb, *Lucas' theorem or prime powers*, Europ. J. Combinatorics **11** (1990), 229-233.

[6] L.E. Dickson, *History of the theory of numbers*, Vol. 1, Chelsea, New York, (1952).

[7] N.J. Fine, *Binomial cofficients modulo a prime*, Am. Math. Monthly, **54** (1947), 589-592.

[8] L. Fiorot, R. Gilblas, A. Tonolo, *The Mystery of Anatol Vieru's Periodic Sequences Unveiled*, Mathematics and Computation in Music, MCM 2022.

[9] L. Fiorot, R. Gilblas, A. Tonolo, *Modular binomials with an application to periodic sequences*, arXiv:2307.02366.

[10] A. Granville, *Arithmetic properties of binomial coefficients. I. Binomial coefficients modulo prime powers*, Organic mathematics (Burnaby, BC, 1995), 23–276.

[11] D.E. Knuth, H. S. Wilf, *The power ofa prime that divides a generalized binomial coefficient*, J. Reine Angew. Math. 396, 212–219, (1989)

[12] E. Kummer, *Über die Ergänzungssätze zu den allgemeinen Reciprociätsgesetzen*, Journal für die reine und angewandte Mathematik, **44** (1852), 93-146.

[13] P. Lanthier, C. Guichaoua and M. Andreata *Reinterpreting and Extending Anatol Vieru's Perioic Sequences Through the Cellular Automata Formalisms*, Proceedings MCM, (2019), Springer, pp. 261–272.

[14] T. Lengyel, *On divisibility properties of some differences o the central binomial coefficients and catalan numbers*, Integers, **13** (2013), A10.

[15] C.J. Lu, S.C. Tsai, *The Periodic Property of Binomial Coefficients Modulo m and Its Applications*, 10th SIAM Conference on Discrete Mathematics, Minneapolis,Minnesota, USA, 2000.

[16] C. Mariconda, A. Tonolo, *Discrete Calculus - Methods for Counting*, Springer UNITEXT 103, (2016), pp. xxi+659, DOI 10.1007/978-3-319-03038-8.

bibitemR C. Riddlesden, *Generalised Fibonacci sequences under modular arithmetic*, Rose-Hulman Undergrad. Math. J. 21, (2020)

[17] M.P. Saikia, J. Vogrinc, *Binomial smbols and prime moduli*, J. Indian Math. Soc. (N.S.) 78 (2011), no. 1-4, 137–143.

[18] A. Vieru, *The Book of Modes*, Editura Muzicala, Bucharest, (1993)

[19] D.T. Vuza, *Aspects mathématiques dan la théorie modale d'Anatol Vieru*, Editura Academiei Republicii Socialiste România (1982)

[20] Ś. Ząbek, *Sur la périodicité modulo m des suites de nombres* $\binom{n}{k}$, Ann. Univ. Mariae Curie-Skłodowska Sect. A **10** (1956), **3–47** (1958)

[21] A.T. Asaad, *Persistent Homology Tools for Image Analysis*, PhD Thesis, University of Buckingham, (2020)

[22] M.E. Atkas, E. Akbas, A. El Atmaoui *Persistent homology of networks: methods and applicatons*, Applied Network Science **4**, 61 (2019)

[23] M.G. Bergomi, A. Baratè, *Homological persistence in time series: an application to music classification*, Journal of Mathematics and Music, (2020), 204–221.

[24] V. Callet, *Persistent Homology on Musical Bars*, Mathematics and Computation in Music. MCM 2022.

[25] S. Chodhury, F. Mémoli, *Persistent Homology of Asymmetric Networks: an Approah based on Dowker Filtrations*, arXiv:1608.05432

[26] Digital and Cognitive Musicology Lab, *https://github.com/DCMLab*, École Polytechnique Fédérale de Lausanne (EPFL)

[27] W. Bas De Haas, et al., *Automatic Functional Harmonic Analysis* Computer Music Journal, **37**, no. 4, (2013) pp. 37–53

[28] H. Edelsbrunner, J. Harer, *Persistet Homology – a Survey*, Discrete & Computational Geometry, **453**, (2009)

[29] H. Edelsbrunner, J. Harer, *Computational topology: an introduction.*, Applied mathematics, Am. Math. Soc., (2010)

[30] R. Ghrist, *Computing Persistent Homology*, Bull. Amer. Math. Soc **45**, 61–75 (2009)

[31] N. Giansiracusa, R. Giansiracusa, C. Moon *Persistent homology machine learning for fingerprint classification*, 10.1109/ICMLA.2019.00201, 1219–1226 (2019)

[32] C. Harte, *Towards Automatic Extraction of Harmony Information from Music Signals*, PhD thesis, Queen Mary, University of London (2010)

[33] Centre for Digital Music, *Annotations: The Beatles*, isophonics.net/content/reference-annotations-beatles

[34] J. Matousek, *Using the Borsuk-Ulam Theorem: Lectures on Topological Methods in Combinatoris and Geometry*, Springer-Verlag, (2007)

[35] C.S. Pun, S.X. Lee, K. Xia, *Persistent-homology-based machine learning: a survey nd comparative study*, Art. Intell. Rev. **55**, 5169–5213 (2022)

[36] N. Oter, M.A. Porter, U. Tillmann, et a. *A roadmap for the computation of persistent homology*, EPJ Data Sci **6**, 17 (2005)

[37] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*, JMLR 12, (2011) 2825–2830.

[38] A. Schönberg, *Structural functions of harmony*, Ernest Benn Limited, (1969)

[39] D. Tymoczko, *The Generalized Tonnetz* Journal of Music Theory, **56**, (2012)

[40] L. Vietoris, *Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen*, Mathematische Annalen, **97**, (1927), 454–472,

[41] J. Yust, *Generalized Tonnetze and Zeitnetze, and the topology of music concepts*, Journal of Mathematics and Music, **14**, (2020), 170–203

[42] A. Zomorodian, G. Carlsson, *Computing Persistent Homology*, Discrete Comput Geom **33**, (2005), 249–274