ZIGZAG PERSISTENCE

GUNNAR CARLSSON AND VIN DE SILVA

Abstract. We describe a new methodology for studying persistence of topological features across a family of spaces or point-cloud data sets, called zigzag persistence. Building on classical results about quiver representations, zigzag persistence generalises the highly successful theory of persistent homology and addresses several situations which are not covered by that theory. In this paper we develop theoretical and algorithmic foundations with a view towards applications in topological statistics.

1. Introduction

1.1. **Overview.** In this paper, we describe a new methodology for studying persistence of topological features across a family of spaces or point-cloud data sets. This theory of *zigzag persistence* generalises the successful and widely used theory of persistence and persistent homology. Moreover, zigzag persistence can handle several important situations that are not currently addressed by standard persistence.

The zigzag persistence framework is activated whenever one constructs a *zigzag diagram* of topological spaces or vector spaces: a sequence of spaces $S_1, \ldots, S_n$ where each adjacent pair is connected by a map $S_i \to S_{i+1}$ or $S_i \leftarrow S_{i+1}$. The novelty of our approach is that the direction of each linking map is arbitrary, in contrast to the usual theory of persistence where all maps point in the same direction.

This paper has three principal objectives:

- To describe several scenarios in applied topology where it is natural to consider zigzag diagrams (Section 1).
- To develop a mathematical theory of persistence for zigzag diagrams (Sections 2 and 3).
- To develop algorithms for computing zigzag persistence (Section 4).

There is one subsidiary objective:

- To introduce the *Diamond Principle*, a calculational tool analogous in power and effect to the Mayer–Vietoris theorem in classical algebraic topology (Section 5).

This is a theoretical paper rather than an experimental paper, and we devote most of our effort to covering the mathematical foundations adequately. The technical basis for zigzag persistence comes from the theory of graph representations, also known as quiver theory. We are deeply indebted to the practitioners of that theory; what is new here is the emphasis on algorithmics and on applications to topology (particularly Sections 1, 4 and 5).

Date: November 30, 2008.

This work has been supported by DARPA grants HR0011-05-1-0007 and HR0011-07-1-0002.

1.2. **Persistence.** One of the principal challenges when attempting to apply algebraic topology to statistical data is the fact that traditional invariants — such as the Betti numbers or the fundamental group — are extremely non-robust when it comes to discontinuous changes in the space under consideration. Persistent homology [8, 13] is the single most powerful existing tool for addressing this problem.

A typical workflow runs as follows [6]. The input is a *point cloud*, that is, a finite subset of some Euclidean space or more generally a finite metric space. After an initial filtering step (to remove undesirable points or to focus on high-density regions of the data, say), a set of vertices is selected from the data, and a simplicial complex S is built on that vertex set, according to some prearranged rule. In practice, the simplicial complex depends on a coarseness parameter ϵ, and what we have is a nested family $\{S_\epsilon\}_{\epsilon\in[0,\infty]}$, which typically ranges from a discrete set of vertices at S_0 to a complete simplex at S_∞.

Persistent homology takes the entire nested family $\{S_\epsilon\}$ and produces a *barcode* or *persistence diagram* as output. A barcode is a collection of half-open subintervals $[b_j, d_j) \subseteq [0, \infty)$, which describes the homology of the family as it varies over ϵ. An interval $[b_j, d_j)$ represents a homological feature which is born at time b_j and dies at time d_j. This construction has several excellent properties:

- There is no need to select a particular value of ϵ.
- Features can be evaluated by interval length. Long intervals are expected to indicate essential features of the data, whereas short intervals are likely to be artefacts of noise.
- There exists a fast algorithm to compute the barcode [13].
- The barcode is a complete invariant of the homology of the family of complexes [13].
- The barcode is provably stable with respect to changes in the input [4]. In contrast, any individual homology group $H_k(S_\epsilon)$ is highly unstable.

The major limitation of persistence is that it depends crucially on the family $\{S_\epsilon\}$ being nested, in the sense that $S_\epsilon \subseteq S_{\epsilon'}$ whenever $\epsilon \leq \epsilon'$. This applies to the current theoretical understanding as well as the algorithms. Zigzag persistence addresses this limitation.

If we discretise the variable ϵ to a finite set of values, the family of simplicial complexes can be thought of as a diagram of spaces

$$S_1 \to S_2 \to \cdots \to S_n$$

where the arrows represent the inclusion maps. If we apply the k-dimensional homology functor $H_k(,\mathbb{F})$ with coefficients in a field $\mathbb{F}$, this becomes a diagram of vector spaces

$$V_1 \to V_2 \to \cdots \to V_n$$

and linear maps, where $V_i = H_k(S_i; \mathbb{F})$. Such a diagram is called a *persistence module*. What makes persistence work is that there is a simple algebraic classification of persistence modules up to isomorphism; each possible barcode corresponds to an isomorphism type.

Our goal is to achieve a similar classification for diagrams in which the arrows may point in either direction. This is zigzag persistence, in a nutshell.

1.3. **Zigzag diagrams in applied topology.** We consider some problems which arise quite naturally in the computational topology of data.

Example 1.1. Some of the most interesting properties of a point cloud are contained in the estimates of the probability density from which the data are sampled. Deep structure is

sometimes revealed after thresholding according to a density estimate (see [3] for an example drawn from visual image analysis). However, the construction of a density estimation function ρ invariably depends on choosing a smoothing parameter: for instance $\rho(x)$ might be defined to be the number of data points within distance r of x; here r is the smoothing parameter.

It happens that different choices of smoothing parameter may well reveal different structures in the data; a particularly striking example of this occurs in [3]. Statisticians have invented useful criteria for determining what the 'appropriate' value of such a parameter might be for a particular data set; but another point of view would be to analyse all values of the parameter simultaneously, and to study how the topology changes as the parameter varies.

The problem with doing this is that there is no natural relationship between, say, the 25% densest points as measured using two different parameter values. This means that one cannot build an increasing family of spaces using the change in parameters, and so one cannot use persistence to analyze the evolution of the topology. On the other hand, there are natural zigzag sequences which can be used to study this problem. Select a sequence of parameter values $r_1 < r_2 < \cdots < r_n$ and a percentage p, and let X_r^p denote the densest $p\%$ of the point cloud when measured according to parameter value r. We can then consider the union sequence

$$X_{r_1}^p \to X_{r_1}^p \cup X_{r_2}^p \leftarrow X_{r_2}^p \to X_{r_2}^p \cup X_{r_3}^p \leftarrow X_{r_3}^p \to \cdots \leftarrow X_{r_n}^p$$

or the intersection sequence

$$X_{r_1}^p \leftarrow X_{r_1}^p \cap X_{r_2}^p \to X_{r_2}^p \leftarrow X_{r_2}^p \cap X_{r_3}^p \to X_{r_3}^p \leftarrow \cdots \to X_{r_n}^p.$$

As we see in Section 5.3, there is essentially no difference between the zigzag persistent homology of the union and intersection sequences of a sequence of spaces. Here that assertion needs to be filtered through the process of representing the data subsets X_r^p as simplicial complexes.

Example 1.2 (Topological bootstrapping)**.** Suppose we are given a very large point cloud X. If it is too large to process directly, we may take a sequence of small samples $X_1, \ldots, X_n$ and estimate their topology individually, perhaps obtaining a persistence barcode for each one. How does this reflect the topology of the original sample X? On one hand, if most of the barcodes have similar appearance, then one might suppose that X itself will have the same barcode. On the other hand, one needs to be able to distinguish between a single feature detected repeatedly, and multiple features detected randomly but one at a time. If we detect n features in X_i on average, are we detecting n featuresof X with detection probability 1, or kn features with detection probability $1/k$?

Once again, there is a need to correlate features across different instances of the construction. The union sequence comes to the rescue:

$$X_1 \to X_1 \cup X_2 \leftarrow X_2 \to \cdots \leftarrow X_n$$

In this case, the intersection sequence is not useful at the level of samples, because two sparse samples are unlikely to intersect very much.

The approach in this example is analogous to bootstrapping in statistics, where measurements on a large data set are estimated by making repeated measurements on a set of samples.

Example 1.3. In computational topology, there exist several techniques for modelling a point cloud data set X by a simplicial complex S: the Cech complex, the Vietoris–Rips complex, the alpha complex [9], the witness complex [6], and so on. The witness complex $W(X; L)$, in particular, depends on the choice of a small subset of 'landmark' points $L \subset X$ which will serve as the vertex set of S. Roughly speaking, a simplex σ with vertices in L is included in $W(X; L)$ if there is some $x \in X$ which witnesses it, by being close to all the vertices.

How does the witness complex $W(X; L)$ depend on the choice of landmark set? There is no direct way to compare $W(X; L)$ with $W(X; M)$ for two different choices of landmark sets L, M. However, it turns out that one can define a witness *bicomplex* $W(X; L, M)$ which maps onto each witness complex. The cells are cartesian products $\sigma \times \tau$, where σ, τ have vertices in L, M respectively. A cell $\sigma \times \tau$ is included provided that there exists $x \in X$ which simultaneously witnesses σ for $W(X; L)$ and τ for $W(X; M)$.

Given a sequence $L_1, \ldots, L_n$ of landmark subsets, one can then construct the biwitness zigzag:

$$W(X; L_1) \leftarrow W(X; L_1, L_2) \rightarrow W(X; L_2) \leftarrow \cdots \rightarrow W(X; L_n)$$

Long intervals in the zigzag barcode will then indicate features that persist across the corresponding range of choices of landmark set.

The fundamental requirement is then for a way of assessing, in a zigzag diagram of vector spaces, the degree to which consistent families of elements exist. The point of this paper is that there is such methodology. We will interpret the isomorphism classes of zig-zag diagrams as a special case of the classification problem for quiver representations (see [7] for background on this theory). There turns out to be a theorem of Gabriel [10] which classifies arbitrary diagrams based on Dynkin diagrams, and which shows in particular that the set of isomorphism classes of zigzag diagrams of a given length is parametrised by barcodes — just as persistence modules are. Long intervals in the classification define large families of consistent elements, hence indicate the presence of features stable across samples, landmark sets, or parameter values for a density estimator.

1.4. **Organisation of the paper.** In Section 2 we describe the theory of decompositions of zigzag modules. These decompositions produce zigzag persistence barcodes analogous to the barcodes of persistent homology. The foundational theorem of Gabriel is stated without proof. In Section 3 we develop the machinery of right-filtrations, which turn out to be the right tool for accessing the decomposition structure of a zigzag module. This is an important section for the reader who wishes to make serious use of zigzag persistence. In Section 4, we present a general-purpose algorithmic framework for calculating zigzag persistence, and we show how this operates in a practical class of examples. The algorithm is based on a proof of Gabriel's theorem for zigzag modules, included for completeness. Section 5 is devoted to a localisation principle which gives another approach to zigzag barcode calculations. We apply this to prove the Diamond Principle. We use this in turn to compare the zigzag barcodes for two natural zigzag diagrams obtained from a sequence of simplicial complexes.

2. Zigzag Diagrams of Vector Spaces

We work over a field $\mathbb{F}$ which remains fixed throughout this paper. There is no significance to the choice of $\mathbb{F}$. All vector spaces are finite-dimensional.

2.1. **Zigzag modules.** Let $\mathbb{V}$ denote a sequence of vector spaces and linear maps, of length n:

$$V_1 \overset{p_1}{\longleftrightarrow} V_2 \overset{p_2}{\longleftrightarrow} \cdots \overset{p_{n-1}}{\longleftrightarrow} V_n$$

Each $\overset{p_i}{\longleftrightarrow}$ represents either a forward map $\overset{f_i}{\longrightarrow}$ or a backward map $\overset{g_i}{\longleftarrow}$. The object $\mathbb{V}$ is called a **zigzag diagram** of vector spaces, or simply a **zigzag module**, over $\mathbb{F}$.

The sequence of symbols f or g is the **type** of $\mathbb{V}$. For instance, a diagram of type $\tau = fgg$ looks like this:

$$V_1 \overset{f_1}{\longrightarrow} V_2 \overset{g_2}{\longleftarrow} V_3 \overset{g_3}{\longleftarrow} V_4$$

The length of a type τ is the length of any diagram of type τ. For example, we say that fgg has length 4. We will usually be considering zigzag modules of a fixed type τ of length n. Such diagrams are called **τ-modules**, and the class of τ-modules is denoted τMod.

Persistence modules (see [8, 13]) are zigzag modules where all the maps have the forward orientation; in other words, where $\tau = ff\ldots f$. As explained in [13], persistence modules can be viewed as graded modules over the polynomial ring $\mathbb{F}[t]$. This observation simplifies the analysis of persistence modules quite considerably.

More generally, one can consider **graph representations** of arbitrary oriented graphs. Zigzag modules constitute the special case where the graph is A_n (a path with n vertices and $n-1$ edges) and the orientation is specified by the type τ. In 1972, Gabriel showed that the Dynkin–Coxeter graphs A_n, D_n, E_6, E_7, E_8 (arbitrarily oriented) have an especially well-behaved representation theory [10]. The theory of **quivers** was launched from this starting block; see [7] for a beautiful and transparent introduction. Zigzag persistence is enabled by the good behaviour of A_n graph representations.

Remark. τMod has the structure of an abelian category. Given two τ-modules $\mathbb{V}, \mathbb{W}$, a morphism $\alpha : \mathbb{V} \to \mathbb{W}$ is defined to be a collection of linear maps $\alpha_i : V_i \to W_i$ which satisfy the commutation relations $\alpha_{i+1} f_i = h_i \alpha_i$ or $\alpha_i g_i = k_i \alpha_{i+1}$ for each i. (Here the forward and backward maps for $\mathbb{W}$ are written h, k respectively.) Morphisms can be composed in the obvious way, and have kernels, images, and cokernels: for instance $\mathbb{K} = \mathrm{Ker}(\alpha)$ is the τ-module with spaces $K_i = \mathrm{Ker}(V_i \to W_i)$ and maps $f_i|_{K_i}$ and $g_i|_{K_{i+1}}$ defined by restriction. The set of morphisms $\mathrm{Hom}(\mathbb{V}, \mathbb{W})$ is naturally a vector space over $\mathbb{F}$, and the endomorphism ring $\mathrm{End}(\mathbb{V}) = \mathrm{Hom}(\mathbb{V}, \mathbb{V})$ is a non-commutative $\mathbb{F}$-algebra. We can view $\mathrm{End}(\mathbb{V})$ as the subalgebra of $\mathrm{End}(V_1) \times \cdots \times \mathrm{End}(V_n)$ defined by the commutation relations.

2.2. **Decompositions of zigzag modules.** We wish to understand zigzag modules by decomposing them into simpler parts. Accordingly, a **submodule** $\mathbb{W}$ of a τ-module $\mathbb{V}$ is defined by subspaces $W_i \leq V_i$ such that $f_i(W_i) \subseteq W_{i+1}$ or $g_i(W_{i+1}) \subseteq W_i$ for all i. These conditions guarantee that $\mathbb{W}$ is itself a τ-module, with maps given by the restrictions $f_i|_{W_i}$ or $g_i|_{W_{i+1}}$. We write $\mathbb{W} \leq \mathbb{V}$.

A submodule $\mathbb{W}$ is called a **summand** of $\mathbb{V}$ if there exists a submodule $\mathbb{X} \leq \mathbb{V}$ which is complementary to $\mathbb{W}$, in the sense that $V_i = W_i \oplus X_i$ for all i. In that case, we say that $\mathbb{V}$ is the **direct sum** of $\mathbb{W}, \mathbb{X}$ and write $\mathbb{V} = \mathbb{W} \oplus \mathbb{X}$.

Example 2.1. As a rule, most submodules are not summands. $\mathbb{V} = (\mathbb{F} \overset{1}{\longrightarrow} \mathbb{F})$ has the submodule $\mathbb{W} = (0 \longrightarrow \mathbb{F})$. However, $\mathbb{W}$ is not a summand because the only possible complement is $(\mathbb{F} \longrightarrow 0)$, and that is not a submodule of $\mathbb{V}$.

Remark. The direct sum can also be defined as an 'external' operation: given τ-modules $\mathbb{V}, \mathbb{W}$ their direct sum $\mathbb{V} \oplus \mathbb{W}$ is defined to be the τ-module with spaces $V_i \oplus W_i$ and maps $f_i \oplus h_i$ or $g_i \oplus k_i$. (Here the forward and backward maps for $\mathbb{W}$ are written h, k respectively.)

A τ-module $\mathbb{V}$ is **decomposable** if it can be written as a direct sum of nonzero submodules, and **indecomposable** otherwise. Any τ-module $\mathbb{V}$ has a **Remak decomposition**; in other words we can write $\mathbb{V} = \mathbb{W}_1 \oplus \cdots \oplus \mathbb{W}_N$, where the summands $\mathbb{W}_j$ are indecomposable. The existence of such a decomposition is proved by induction on the total dimension $\sum_i \dim(V_i)$: if $\mathbb{V}$ is decomposable, say $\mathbb{V} = \mathbb{W} \oplus \mathbb{X}$, then we may assume inductively that $\mathbb{W}, \mathbb{X}$ have Remak decompositions, and therefore so does $\mathbb{V}$. (Base case: if $\mathbb{V}$ is indecomposable, then it has a Remak decomposition with one term.)

Remak decompositions themselves are not unique. However, the Krull–Schmidt principle from commutative algebra tells us that the summands in a Remak decomposition are unique up to reordering:

Proposition 2.2. *(Krull–Remak–Schmidt.) Suppose a τ-module $\mathbb{V}$ has Remak decompositions*

$$\mathbb{V} = \mathbb{W}_1 \oplus \cdots \oplus \mathbb{W}_M \quad \textit{and} \quad \mathbb{V} = \mathbb{X}_1 \oplus \cdots \oplus \mathbb{X}_N.$$

Then $M = N$ and there is some permutation σ of $\{1, \ldots, N\}$ such that $\mathbb{W}_j \cong \mathbb{X}_{\sigma(j)}$ for all j.

Proof. The proof of Theorem 7.5 of Lang [12], which is stated for modules in the ordinary sense, can be applied verbatim to our present context; all the required algebraic operations can be carried out within $\mathrm{End}(\mathbb{V})$. Since our τ-modules have finite total dimension, the ascending and descending chain conditions (ACC and DCC) are automatic. □

For further context, we refer the reader to an elegant article by Atiyah [1]; the Krull–Schmidt principle applies in any exact abelian category to objects which satisfy ACC and DCC, or a weaker 'bi-chain condition' defined by Atiyah. Our category, τMod, is included by this formulation.

Thus we can use the multiset $\{\mathbb{W}_j\}$ as an isomorphism invariant of $\mathbb{V}$. For this to be useful, we need to identify the set of indecomposable τ-modules. We now describe a natural collection of indecomposables. For each subinterval $[b, d]$ of the integer sequence $\{1, \ldots, n\}$ there is an associated τ-module.

Definition 2.3. Let τ be a type of length n and consider integers $1 \le b \le d \le n$. The **interval** τ-module with birth time b and death time d is written $\mathbb{I}_\tau(b, d)$ and defined with spaces

$$I_i = \begin{cases} \mathbb{F} & \text{if } b \le i \le d, \\ 0 & \text{otherwise;} \end{cases}$$

and with identity maps between adjacent copies of $\mathbb{F}$, and zero maps otherwise. When τ is implicit, we will usually suppress it and simply write $\mathbb{I}(b, d)$.

Example. If $\tau = fgg$ then $\mathbb{I}(2, 3)$ is the zigzag module

$$0 \xrightarrow{0} \mathbb{F} \xleftarrow{1} \mathbb{F} \xleftarrow{0} 0.$$

Proposition 2.4. *Interval τ-modules are indecomposable.*

Proof. Suppose $\mathbb{I}(b,d) = \mathbb{V} \oplus \mathbb{W}$ and consider two adjacent terms $\mathbb{F}$ connected by an identity map. Since $\mathbb{V}, \mathbb{W}$ are submodules, the dimensions of $\mathbb{V}$ and $\mathbb{W}$ cannot decrease in the direction of the map; nor, since they are complements, can they increase. Thus $\dim(V_i)$ and $\dim(W_i)$ are constant over $b \leq i \leq d$, and in particular one of $\mathbb{V}, \mathbb{W}$ must be zero. □

Here is the foundation stone for the theory of zigzag persistence.

Theorem 2.5 (Gabriel)**.** *The indecomposable τ-modules are precisely the intervals $\mathbb{I}(b,d)$, where $1 \leq b \leq d \leq n = \text{length}(\tau)$. Equivalently, every τ-module can be written as a direct sum of intervals.*

Proof. This is the simplest special case of Gabriel's theorem, for the graphs A_n. The original reference (in German) is [10]. See [7] for an accessible overview. □

Thus, any τ-module can be described completely up to isomorphism as an unordered list of intervals $[b,d]$, which correspond to its indecomposable summands. This is in exact accordance with the special case of ordinary persistence, where the result is comparatively easy to prove: it is simply the classification of finitely-generated graded modules over the polynomial ring $\mathbb{F}[t]$ (see [13]).

The philosophical point is that the decomposition theory of graph representations is somewhat independent of the orientation of the graph edges (see Kac [11]). Even in our case this is surprising, because there is no obvious congruence between persistence modules and zigzag modules of an arbitrary type τ. However, if we accept this principle, then the generalisation from ordinary persistence to zigzag persistence is not surprising: interval decomposition for persistence modules implies interval decomposition for zigzag modules.

We will devote much of this paper to constructing a stand-alone proof of Theorem 2.5. This provides technical support towards our two main goals: to provide algorithms for computing the interval summands of a given τ-module; and to make rigorous statements about the output of those algorithms.

2.3. **Zigzag persistence.** We now define zigzag persistence and develop some of its elementary properties.

Definition 2.6. Let $\mathbb{V}$ be a zigzag module (of arbitrary type). The **zigzag persistence** of $\mathbb{V}$ is defined to be the multiset

$$\text{Pers}(\mathbb{V}) = \{[b_j, d_j] \subseteq \{1, \ldots, n\} \mid j = 1, \ldots, N\}$$

of integer intervals derived from a decomposition $\mathbb{V} \cong \mathbb{I}(b_1, d_1) \oplus \cdots \oplus \mathbb{I}(b_N, d_N)$. The Krull–Schmidt principle asserts that this definition is independent of the decomposition.

Graphically, $\text{Pers}(\mathbb{V})$ can be represented as a set of lines measured against a single axis with labels $\{1, \ldots, n\}$ (the **barcode**), or as a multiset of points in $\mathbb{R}^2$ lying on or above the diagonal in the positive quadrant (the **persistence diagram**). See Figure 1 for an example presented in each style.

Remark. In the special case of persistence modules, this agrees with the standard treatment (see [8, 13]) except in the following particular: the closed integer intervals $[b_j, d_j] \subseteq \{1, \ldots, n\}$ are replaced by half-open real intervals $[b_j, d_j + 1) \subset \mathbb{R}$ in the standard treatment. This is particularly natural when the indexing parameter is continuous: an interval $[b,d)$ indicates a feature born at time b that survives right up to, but vanishes at, time d. Our convention is

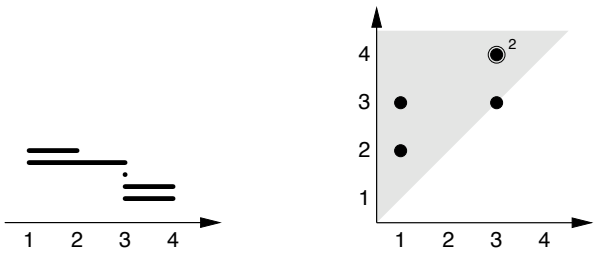


FIGURE 1. Barcode (left) and persistence diagram (right) representations of the persistence $\{[1,2],[1,3],[3,3],[3,4],[3,4]\}$ of a zigzag module of length 4.

motivated by the desire to maintain symmetry between the forward and backward directions. We advise the reader to take particular care in handling the different conventions.

The transition from a zigzag module to its interval decomposition presents certain hazards which are not present in the case of persistence modules. We now draw attention to these hazards.

Definition 2.7. Let $\mathbb{V}$ be a zigzag module and let $\mathbb{V}[p,q]$ denote the restriction of $\mathbb{V}$ to the index set $p \leq i \leq q$. A **feature** of $\mathbb{V}$ over the time interval $[p,q]$ is a summand of $\mathbb{V}[p,q]$ isomorphic to $\mathbb{I}(p,q)$.

With persistence modules, there are several equivalent ways to recognise the existence of a feature. Here is a sample result.

Proposition 2.8. *Let $\mathbb{V}$ be a persistence module of length n, and let $1 \leq p \leq q \leq n$. The following are equivalent:*

1. *The composite map $V_p \to V_q$ is nonzero.*
2. *There exist nonzero elements $x_i \in V_i$ for $p \leq i \leq q$, such that $x_{i+1} = f_i(x_i)$ for $p \leq i < q$.*
3. *There exists a submodule of $\mathbb{V}[p,q]$ isomorphic to $\mathbb{I}(p,q)$.*
4. *There exists a summand of $\mathbb{V}[p,q]$ isomorphic to $\mathbb{I}(p,q)$, i.e. a feature over $[p,q]$.*

Proof. It is easy to verify that (1), (2), (3) are equivalent. For (1) $\Rightarrow$ (2), begin by choosing $x_p \in V_p$ that maps to a nonzero element in V_q, and let x_i be the image of x_p in V_i. For (2) $\Rightarrow$ (3), define $\mathbb{I}$ by $I_i = \mathrm{Span}(x_i)$. For (3) $\Rightarrow$ (1), note that the restriction $I_p \to I_q$ is nonzero.

Clearly (4) $\Rightarrow$ (3). We now show that (1) $\Rightarrow$ (4). Consider an interval decomposition $\mathbb{V}[p,q] = \mathbb{I}(b_1,d_1) \oplus \cdots \oplus \mathbb{I}(b_N,d_N)$. On each summand, the map $I_p(b_j,d_j) \to I_q(b_j,d_j)$ is zero unless $b_j = p$ and $d_j = q$. Thus at least one of the summands is isomorphic to $\mathbb{I}(p,q)$. $\square$

The intuitions supported by Proposition 2.8 break down in the general case.

Caution 2.9. Let $\mathbb{V}$ be a zigzag module of arbitrary type. Statement (1) has no clear interpretation at this stage (something can be said in terms of the right-filtration functor of Section 3). Consider the following statements:

2. There exist nonzero elements $x_i \in V_i$ for $p \leq i \leq q$, such that $x_{i+1} = f_i(x_i)$ or $x_i = g_i(x_{i+1})$ (whichever is applicable) for $p \leq i < q$.
3. There exists a submodule of $\mathbb{V}[p,q]$ isomorphic to $\mathbb{I}(p,q)$.
4. There exists a summand of $\mathbb{V}[p,q]$ isomorphic to $\mathbb{I}(p,q)$, i.e. a feature over $[p,q]$.

It is easy to verify that (2) $\Leftrightarrow$ (3) and that (4) implies (2), (3). However, the next two examples demonstrate that (2), (3) do not in general imply (4).

Example 2.10. Let $\tau = gf$ and consider the τ-module $\mathbb{V}$ defined as follows:

$$\begin{array}{ccccc} \mathbb{F} & \xleftarrow{g_1} & \mathbb{F}^2 & \xrightarrow{f_2} & \mathbb{F} \\ x & \longleftarrow\!\shortmid & (x,y) & \longmapsto & y \end{array}$$

The interval decomposition is $\mathbb{V} = \mathbb{I}(1,2) \oplus \mathbb{I}(2,3)$, where the summands are

$$\begin{array}{ccccc} \mathbb{F} & \longleftarrow & \mathbb{F} \oplus 0 & \longrightarrow & 0 \\ x & \longleftarrow\!\shortmid & (x,0) & & \end{array}$$

and

$$\begin{array}{ccccc} 0 & \longleftarrow & 0 \oplus \mathbb{F} & \longrightarrow & \mathbb{F} \\ & & (0,y) & \longmapsto & y \end{array}$$

respectively. If this example appeared in a statistical topology setting, the feature corresponding to the generator of the $\mathbb{F}$ at V_1 would be regarded as unrelated to the feature corresponding to the generator of the $\mathbb{F}$ at V_3.

On the other hand, $\mathbb{V}$ does have a submodule (in fact, many submodules) isomorphic to $\mathbb{I}(1,3)$. Indeed, let $\Delta = \{(x,x) \mid x \in \mathbb{F}\}$ denote the diagonal subspace of $\mathbb{F}^2$. Then

$$\begin{array}{ccccc} \mathbb{F} & \longleftarrow & \Delta & \longrightarrow & \mathbb{F} \\ x & \longleftarrow\!\shortmid & (x,x) & \longmapsto & x \end{array}$$

is a submodule $\mathbb{W} \leq \mathbb{V}$ isomorphic to $\mathbb{I}(1,3)$. The quotient τ-module $\mathbb{V}/\mathbb{W}$ is isomorphic to $\mathbb{I}(2,2)$ but $\mathbb{W}$ has no complementary τ-module in $\mathbb{V}$. Indeed, that would contradict the Krull–Schmidt theorem. More concretely, any complement of $\mathbb{W}$ must be isomorphic to $(0 \longleftarrow \mathbb{F} \longrightarrow 0)$, but that would require a 1-dimensional subspace of $\mathrm{Ker}(g_1) \cap \mathrm{Ker}(f_2) = 0$.

Example 2.11. We can extend the previous example to arbitrary length. Consider the type $\tau = gf \ldots gf = (gf)^n$, of length $2n+1$. Let $\mathbb{V}$ be the τ-module

$$\mathbb{F} \xleftarrow{\pi_1} \mathbb{F}^2 \xrightarrow{\pi_2} \mathbb{F} \xleftarrow{\pi_1} \cdots \xrightarrow{\pi_2} \mathbb{F} \xleftarrow{\pi_1} \mathbb{F}^2 \xrightarrow{\pi_2} \mathbb{F},$$

where $\pi_1(x,y) = x$, and $\pi_2(x,y) = y$. Then $\mathbb{V}$ is isomorphic to a sum of short intervals

$$\mathbb{I}(1,2) \oplus \{\mathbb{I}(2,4) \oplus \cdots \oplus \mathbb{I}(2n-2,2n)\} \oplus \mathbb{I}(2n,2n+1)$$

but it has a submodule

$$\mathbb{F} \longleftarrow \Delta \longrightarrow \mathbb{F} \longleftarrow \cdots \longrightarrow \mathbb{F} \longleftarrow \Delta \longrightarrow \mathbb{F}$$

isomorphic to the long interval $\mathbb{I}(1,2n+1)$.

Moral. *In zigzag persistence it is necessary to respect the distinction between submodules and summands. Features are defined in terms of summands; never submodules.*

We have defined features in terms of a chosen subinterval $[p,q]$. Features behave as expected when zooming to a larger or smaller window of observation. The following proposition illustrates what we mean.

Proposition 2.12. *Let $\mathbb{V}$ be a zigzag module of length n and let $1 \leq p \leq q \leq n$. The following statements are equivalent.*

(1) *There exists a summand of* $\mathbb{V}[p,q]$ *isomorphic to* $\mathbb{I}(p,q)$*, i.e. a feature over* $[p,q]$*.*
(2) *There exists a summand of* $\mathbb{V}$ *isomorphic to* $\mathbb{I}(p',q')$*, for some* $[p',q'] \supseteq [p,q]$*.*

Indeed, there is a bijection between intervals $[p,q]$ *in* $\mathrm{Pers}(\mathbb{V}[p,q])$ *and intervals* $[p',q'] \supseteq [p,q]$ *in* $\mathrm{Pers}(\mathbb{V})$*.*

Proof. Consider an interval decomposition $\mathbb{V} = \mathbb{I}(b_1,d_1) \oplus \cdots \oplus \mathbb{I}(b_N,d_N)$. By restriction, this induces an interval decomposition of $\mathbb{V}[p,q]$ into intervals $\mathbb{I}(b_j,d_j)[p,q]$. This induces the claimed bijection, because $[b_j,d_j]$ restricts to $[p,q]$ if and only if $[b_j,d_j] \supseteq [p,q]$. □

Operating invisibly in this proof is the Krull–Schmidt principle, which allows us to select the interval decompositions most convenient to us when calculating $\mathrm{Pers}(\mathbb{V})$ and $\mathrm{Pers}(\mathbb{V}[p,q])$.

Remark. Sometimes it is useful to reduce the resolution of $\mathrm{Pers}(V)$. Let $K \subset \{1,\dots,n\}$ be any subset. We define the **restriction** of $\mathrm{Pers}(\mathbb{V})$ to K to be the multiset

$$\mathrm{Pers}(\mathbb{V})|_K = \{I \cap K \mid I \in \mathrm{Pers}(\mathbb{V}),\, I \cap K \neq \emptyset\}\,.$$

For instance, Proposition 2.12 amounts to the observation that $\mathrm{Pers}(\mathbb{V}[p,q]) = \mathrm{Pers}(\mathbb{V})|_{[p,q]}$.

3. From Zigzag Modules to Filtrations

3.1. **The right-filtration operator.** Our strategy is to understand (and construct) decompositions of a τ-module $\mathbb{V}$ by an iterative process, moving from left to right and retaining the necessary information at each stage. The bulk of this information is encoded as a filtration on the rightmost vector space V_n.

Definition 3.1. The **right-filtration** $\mathrm{R}(\mathbb{V})$ of a τ-module $\mathbb{V}$ of length n takes the form

$$\mathrm{R}(\mathbb{V}) = (R_0, R_1, \dots, R_n),$$

where the R_i are subspaces of V_n satisfying the inclusion relations

$$0 = R_0 \leq R_1 \leq \cdots \leq R_n = V_n.$$

$\mathrm{R}(\mathbb{V})$ is defined recursively as follows.

Base case:

- If $\mathbb{V}$ has length 1, then $\mathrm{R}(\mathbb{V}) = (0, V_1)$.

Recursive step. Suppose we have already defined $\mathrm{R}(\mathbb{V}) = (R_0, R_1, \dots, R_n)$:

- If $\mathbb{V}^+$ is $\mathbb{V} \xrightarrow{f_n} V_{n+1}$, then $\mathrm{R}(\mathbb{V}^+) = (f_n(R_0), f_n(R_1), \dots, f_n(R_n), V_{n+1})$.
- If $\mathbb{V}^+$ is $\mathbb{V} \xleftarrow{g_n} V_{n+1}$, then $\mathrm{R}(\mathbb{V}^+) = (0, g_n^{-1}(R_0), g_n^{-1}(R_1), \dots, g_n^{-1}(R_n))$.

To verify that $\mathrm{R}(\mathbb{V}^+)$ in the two inductive cases is a filtration of the specified form, note that $R_i \leq R_{i+1}$ implies that $f_n(R_i) \leq f_n(r_{i+1})$ in the first case, and $g_n^{-1}(R_i) \leq g_n^{-1}(R_{i+1})$ in the second case. Moreover $f_n(R_0) = f_n(0) = 0$, and $g_n^{-1}(R_n) = g_n^{-1}(V_n) = V_{n+1}$.

Example 3.2. Here are the right-filtrations for the two length-2 types:

$$\begin{aligned} \mathrm{R}(V_1 \xrightarrow{f_1} V_2) &= (0, f_1(V_1), V_2) \\ \mathrm{R}(V_1 \xleftarrow{g_1} V_2) &= (0, g_1^{-1}(0), V_2) \end{aligned}$$

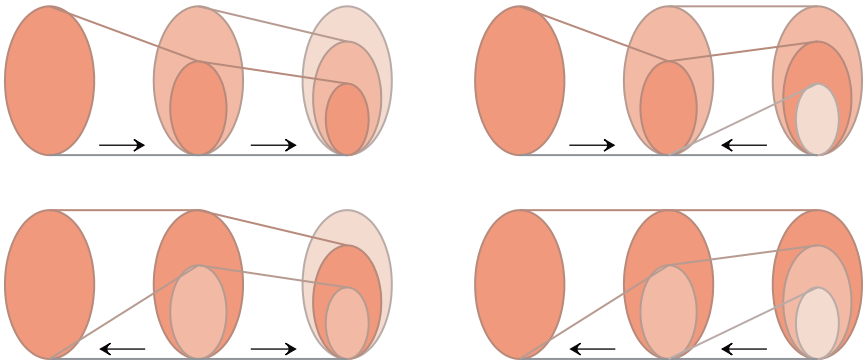


FIGURE 2. Forward propagation of the right-filtration, illustrated for the four types ff, fg, gf, gg of length 3.

Example 3.3. Here are the right-filtrations for the four length-3 types:

$$\begin{aligned}
\mathrm{R}(V_1 \xrightarrow{f_1} V_2 \xrightarrow{f_2} V_3) &= (0, f_2 f_1(V_1), f_2(V_2), V_3) \\
\mathrm{R}(V_1 \xrightarrow{f_1} V_2 \xleftarrow{g_2} V_3) &= (0, g_2^{-1}(0), g_2^{-1} f_1(V_1), V_3) \\
\mathrm{R}(V_1 \xleftarrow{g_1} V_2 \xrightarrow{f_2} V_3) &= (0, f_2 g_1^{-1}(0), f_2(V_2), V_3) \\
\mathrm{R}(V_1 \xleftarrow{g_1} V_2 \xleftarrow{g_2} V_3) &= (0, g_2^{-1}(0), g_2^{-1} g_1^{-1}(0), V_3)
\end{aligned}$$

See Figure 2 for a schematic representation.

Remark. In the examples above, it is not difficult to see that $\mathrm{R}(\mathbb{V})$ comprises all the subspaces of V_n that are naturally definable in terms of the maps p_i.

Each of the n subquotients R_i/R_{i-1} carries information dating back to some earliest V_j in the sequence of vector spaces.

Example 3.4. The module $V_1 \xrightarrow{f_1} V_2$ has right-filtration $(0, f_1(V_1), V_2)$. The first subquotient $f_1(V_1)/0 = f_1(V_1)$ corresponds to vectors born at time 1 which survive to time 2. The second subquotient $V_2/f_1(V_1)$ corresponds to vectors which appear only at time 2.

Example 3.5. The module $V_1 \xleftarrow{g_1} V_2$ has right-filtration $(0, g_1^{-1}(0), V_2)$. The first subquotient $g_1^{-1}(0)$ corresponds to vectors at time 2 which are destroyed when mapping back to time 1. The second subquotient $V_2/g_1^{-1}(0)$ is isomorphic to $g_1(V_2)$ and records those vectors which survive from time 2 back to time 1.

Definition 3.6. The **birth-time index** $\mathrm{b}(\tau) = (b_1, b_2, \ldots, b_n)$ is a vector of integers b_i which indicate the birth-times associated with the subquotients R_i/R_{i-1} of the right-filtration of a τ-module. This is defined recursively as follows.

Base case:

- If τ is empty (i.e. $\mathbb{V}$ has length 1) then $\mathrm{b}(\tau) = (1)$.

Recursive step. Suppose we have already defined $\mathrm{b}(\tau) = (b_1, b_2, \ldots, b_n)$:

- If τ^+ is τf then $\mathrm{b}(\tau^+) = (b_1, \ldots, b_n, n+1)$.
- If τ^+ is τg then $\mathrm{b}(\tau^+) = (n+1, b_1, \ldots, b_n)$.

Example 3.7. At length 2 we have $\mathrm{b}(f) = (1, 2)$ whereas $\mathrm{b}(g) = (2, 1)$. This is consonant with the discussion in Examples 3.4 and 3.5.

Example 3.8. Here are the birth-time indices for the types of length 3.

$$\mathrm{b}(ff) = (1, 2, 3), \qquad \mathrm{b}(fg) = (3, 1, 2), \qquad \mathrm{b}(gf) = (2, 1, 3), \qquad \mathrm{b}(gg) = (3, 2, 1).$$

In summary, the information in a τ-module $\mathbb{V}$ which survives to time n is encoded as a filtration $R(\mathbb{V})$ on V_n. The 'age' of the information at each level of the filtration (i.e. at each subquotient) is recorded in the birth-time index $\mathrm{b}(\tau)$.

For a simplified but precise version of this last claim, we now calculate the right-filtrations of interval τ-modules. In the filtration specified in the following lemma, $J_i/J_{i-1} = \mathbb{F}$ is the only non-zero subquotient, corresponding to the birth time b_i.

Lemma 3.9. *Let τ be a type of length n, with $\mathrm{b}(\tau) = (b_1, b_2, \ldots, b_n)$. For each $i = 1, 2, \ldots, n$, we have an isomorphism*

$$\mathrm{R}(\mathbb{I}_\tau(b_i, n)) = \mathcal{J}(i, n),$$

where $\mathcal{J}(i, n) = (J_0, J_1, \ldots, J_n)$ is the filtration on $\mathbb{F}$ defined by

$$J_0 = \cdots = J_{i-1} = 0; \quad J_i = \cdots = J_n = \mathbb{F}.$$

Remark. We refer to the $\mathcal{J}(b, n)$ also as *intervals* (but now in the category of filtered vector spaces).

Proof. This is a straightforward calculation by induction on τ. For the base case, τ is empty and $\mathrm{b}(\tau) = (1)$. Then $\mathrm{R}(\mathbb{I}(1, 1)) = (0, \mathbb{F}) = \mathcal{J}(1, 1)$ as claimed. Now suppose the result is known for τ, with $\mathrm{b}(\tau) = (b_1, \ldots, b_n)$. Suppose $\tau^+ = \tau f$ or τg. In both cases, write $\mathrm{b}(\tau^+) = (b_1^+, \ldots, b_{n+1}^+)$.

Case f: Suppose that $1 \le i \le n$; then $b_i^+ = b_i$ and therefore

$$\mathbb{I}_{\tau^+}(b_i^+, n + 1) = (\mathbb{I}_\tau(b_i, n) \xrightarrow{1} \mathbb{F}).$$

Writing $\mathrm{R}(\mathbb{I}_\tau(b_i, n)) = \mathcal{J}(i, n) = (J_0, J_1, \ldots, J_n)$, it follows that

$$\mathrm{R}(\mathbb{I}_{\tau^+}(b_i^+, n + 1)) = (J_0, J_1, \ldots, J_n, \mathbb{F}) = \mathcal{J}(i, n + 1).$$

For $i = n + 1$, we have $b_{n+1}^+ = n + 1$, and indeed

$$\mathrm{R}(\mathbb{I}_{\tau^+}(n + 1, n + 1)) = \mathrm{R}((\ldots) \xrightarrow{0} \mathbb{F}) = (0, \ldots, 0, \mathbb{F}) = \mathcal{J}(n + 1, n + 1).$$

Case g: Suppose that $2 \le i \le n + 1$; then $b_i^+ = b_{i-1}$ and therefore

$$\mathbb{I}_{\tau^+}(b_i^+, n + 1) = (\mathbb{I}_\tau(b_{i-1}, n) \xleftarrow{1} \mathbb{F}).$$

Writing $\mathrm{R}(\mathbb{I}_\tau(b_{i-1}, n)) = \mathcal{J}(i - 1, n) = (J_0, J_1, \ldots, J_n)$, it follows that

$$\mathrm{R}(\mathbb{I}_{\tau^+}(b_i^+, n + 1)) = (0, J_0, J_1, \ldots, J_n) = \mathcal{J}(i, n + 1).$$

For $i = 1$, we have $b_1^+ = n + 1$ and then

$$\mathrm{R}(\mathbb{I}_{\tau^+}(n + 1, n + 1)) = \mathrm{R}((\ldots) \xleftarrow{0} \mathbb{F}) = (0, \mathbb{F}, \ldots, \mathbb{F}) = \mathcal{J}(1, n + 1)$$

as required. □

Thus the right-filtration (with the help of the birth-time index) distinguishes the different intervals $\mathbb{I}(b, n)$. It gives no information about intervals $\mathbb{I}(b, d)$ when $d < n$, since in those cases $I_n = 0$.

Example 3.10. Consider $\tau = fgf$, so $\mathrm{b}(\tau) = (b_1, b_2, b_3, b_4) = (3, 1, 2, 4)$ and in general

$$\mathrm{R}(\, V_1 \xrightarrow{f_1} V_2 \xleftarrow{g_2} V_3 \xrightarrow{f_3} V_4\,) = (0,\ f_3 g_2^{-1}(0),\ f_3 g_2^{-1} f_1(V_1),\ f_3(V_3),\ V_4).$$

In particular,

$$\begin{array}{rclcccccccccl}
\mathbb{I}(b_2,4) & = & \mathrm{R}(\,\mathbb{F} & \xrightarrow{1} & \mathbb{F} & \xleftarrow{1} & \mathbb{F} & \xrightarrow{1} & \mathbb{F}\,) & = & (0,\,0,\,\mathbb{F},\,\mathbb{F},\,\mathbb{F}) & = & \mathcal{J}(2,4) \\
\mathbb{I}(b_3,4) & = & \mathrm{R}(\,0 & \longrightarrow & \mathbb{F} & \xleftarrow{1} & \mathbb{F} & \xrightarrow{1} & \mathbb{F}\,) & = & (0,\,0,\,0,\,\mathbb{F},\,\mathbb{F}) & = & \mathcal{J}(3,4) \\
\mathbb{I}(b_1,4) & = & \mathrm{R}(\,0 & \longrightarrow & 0 & \longleftarrow & \mathbb{F} & \xrightarrow{1} & \mathbb{F}\,) & = & (0,\,\mathbb{F},\,\mathbb{F},\,\mathbb{F},\,\mathbb{F}) & = & \mathcal{J}(1,4) \\
\mathbb{I}(b_4,4) & = & \mathrm{R}(\,0 & \longrightarrow & 0 & \longleftarrow & 0 & \longrightarrow & \mathbb{F}\,) & = & (0,\,0,\,0,\,0,\,\mathbb{F}) & = & \mathcal{J}(4,4)
\end{array}$$

which is in accordance with Lemma 3.9.

3.2. **Decompositions of filtered vector spaces.** We now consider filtered vector spaces in their own right, independently of the connection to zigzag-modules, and develop the theory of Remak decompositions. We will see later that this is the right tool for understanding Remak decompositions of zigzag modules.

A filtered vector space of depth n is a sequence $\mathcal{R} = (R_0, R_1, \ldots, R_n)$ of vector spaces, where $0 = R_0 \leq R_1 \leq \cdots \leq R_n$. The class of such objects is denoted Filt_n. The right-filtration $\mathrm{R}(\mathbb{V})$ of any zigzag module $\mathbb{V}$ of length n belongs to this class, as do the intervals $\mathcal{J}(i,n)$ defined in Lemma 3.9. Indeed, if $\mathcal{R} \in \mathrm{Filt}_n$ satisfies $\dim(R_n) = 1$, then $\mathcal{R}$ is isomorphic to some $\mathcal{J}(i,n)$.

Remark. Filt_n can be given the structure of a category in a natural way, but it is not quite an abelian category since morphisms do not generally have cokernels.

A filtered vector space $\mathcal{S} = (S_0, S_1, \ldots, S_n)$ is a **subspace** of $\mathcal{R}$ if $S_i \leq R_i$ for all i. It is appropriate to consider a stronger notion of subspace when dealing with direct-sum decompositions: $\mathcal{S}$ is an **induced subspace** of $\mathcal{R}$ if there exists a vector subspace $K \leq R_n$ such that $S_i = R_i \cap K$ for all i. In that event, we write $\mathcal{S} = \mathcal{R} \cap K$. Note that $K = R_n \cap K = S_n$.

We say that $\mathcal{R}$ is the **direct sum** of two subspaces, and write $\mathcal{R} = \mathcal{S} \oplus \mathcal{T}$, if $R_i = S_i \oplus T_i$ for all i. We claim that $\mathcal{S}, \mathcal{T}$ must be induced subspaces. Note that $S_n \cap T_n = 0$. For each i, then, $R_i \cap S_n$ is a subspace of R_i which contains S_i and meets $T_i \leq T_n$ only at 0. It follows that $R_i \cap S_n = S_i$ for all i. Thus $\mathcal{S} = \mathcal{R} \cap S_n$, and symmetrically $\mathcal{T} = \mathcal{R} \cap T_n$.

The general form of a direct-sum decomposition is therefore $\mathcal{R} = (\mathcal{R} \cap K) \oplus (\mathcal{R} \cap L)$. What are the requirements on K, L to make this a valid decomposition? The direct sum condition implies that $R_n = K \oplus L$ as a vector space. Moreover, given a vector space decomposition $R_n = K \oplus L$, the further condition

$$R_i = \mathrm{Span}\,(R_i \cap K, R_i \cap L) \text{ for all } i$$

is necessary and sufficient to guarantee $\mathcal{R} = (\mathcal{R} \cap K) \oplus (\mathcal{R} \cap L)$.

If $\mathcal{R} = \mathcal{S} \oplus \mathcal{T}$, the two subspaces $\mathcal{S}, \mathcal{T}$ are said to be complementary summands. The following fact radically simplifies the decomposition theory of filtered vector spaces.

Proposition 3.11. *Every induced subspace of a filtered vector space has a complementary summand.*

Proof. We are given $\mathcal{S} = \mathcal{R} \cap K$, and seek to construct $\mathcal{T} = (T_0, T_1, \ldots, T_n)$ such that $\mathcal{R} = \mathcal{S} \oplus \mathcal{T}$. We proceed inductively. Since $R_0 = S_0 = 0$ we take $T_0 = 0$. Now suppose we have chosen T_k so that $R_k = S_k \oplus T_k$. In particular, $T_k \cap S_k = 0$. Then

$$T_k \cap S_{k+1} \leq T_k \cap S_n = (T_k \cap R_k) \cap S_n = T_k \cap (R_k \cap S_n) = T_k \cap S_k = 0.$$

Thus T_k and S_{k+1} are independent subspaces of R_{k+1}, and so T_k can be extended to a complement T_{k+1} of S_{k+1} in R_{k+1}. This completes the induction. □

Corollary 3.12. *The indecomposables in* Filt_n *are precisely the intervals* $\mathcal{J}(i,n)$*. Thus, every filtered vector space can be decomposed as a finite direct sum of intervals.*

Proof. By Proposition 3.11, $\mathcal{R}$ has nontrivial summands if and only if R_n has nontrivial vector subspaces; this happens exactly when $\dim(R_n) > 1$. □

The dimension of $\mathcal{R} \in \mathrm{Filt}_n$ is defined to be the vector of integers

$$\dim(\mathcal{R}) = (c_1, c_2, \ldots, c_n),$$

where $c_i = \dim(R_i / R_{i-1})$ are the dimensions of the successive subquotients of the filtration.

Proposition 3.13. *Let* $\mathcal{R}$ *be a filtered vector space of depth* n*, with* $\dim(\mathcal{R}) = (c_1, c_2, \ldots, c_n)$*. For any decomposition of* $\mathcal{R}$ *into intervals, the multiplicity of* $\mathcal{J}(i,n)$ *is* c_i*. Thus:*

$$\mathcal{R} \cong \bigoplus_{1 \leq i \leq n} c_i \, \mathcal{J}(i,n).$$

Proof. Let m_i be the multiplicity of $\mathcal{J}(i,n)$. Then, for all k,

$$\dim(R_k) = m_1 + m_2 + \cdots + m_k$$

by considering the contribution of each summand, whereas

$$\dim(R_k) = c_1 + c_2 + \cdots + c_k$$

by considering the contribution of each subquotient R_i / R_{i-1}. This is possible only if $m_i = c_i$ for all i. □

This concludes our tour of the decomposition theory for filtered vector spaces. Now we must leverage this to achieve a decomposition theory for τ-modules. In one direction, the relationship is straightforward:

Proposition 3.14. *The right-filtration operation respects direct sums, in the sense that*

$$\mathrm{R}(\mathbb{V}_1 \oplus \cdots \oplus \mathbb{V}_N) = \mathrm{R}(\mathbb{V}_1) \oplus \cdots \oplus \mathrm{R}(\mathbb{V}_N)$$

for τ*-modules* $\mathbb{V}_1, \ldots, \mathbb{V}_N$*.*

Proof. This is proved by induction on τ, following the recursive structure of Definition 3.1 and using the standard facts

$$(f_1 \oplus \cdots \oplus f_N)(R_1 \oplus \cdots \oplus R_N) = f_1(R_1) \oplus \cdots \oplus f_N(R_N)$$

and

$$(g_1 \oplus \cdots \oplus g_N)^{-1}(R_1 \oplus \cdots \oplus R_N) = g_1^{-1}(R_1) \oplus \cdots \oplus g_N^{-1}(R_N)$$

from linear algebra. (For simplicity we are suppressing various indices here.) □

However, what we need is a converse to Proposition 3.14: if the filtered vector space $\mathcal{R} = \mathrm{R}(\mathbb{V})$ can be split as a direct sum $\mathcal{R} = \mathcal{R}_1 \oplus \cdots \oplus \mathcal{R}_N$, we would like to infer a corresponding splitting $\mathbb{V} = \mathbb{V}_1 \oplus \cdots \oplus \mathbb{V}_N$ of τ-modules. In the following two sections we establish such a principle for a particular class: the 'streamlined' τ-modules.

3.3. **Streamlined modules.** We introduce a special class of τ-module for which the right-filtration functor preserves all structural information.

Definition 3.15. A τ-module $\mathbb{V}$ is **(right-)streamlined** if each $\xrightarrow{f_i}$ is injective and each $\xleftarrow{g_i}$ is surjective.

Similarly, we may say that a τ-module $\mathbb{V}$ is left-streamlined if each $\xrightarrow{f_i}$ is surjective and each $\xleftarrow{g_i}$ is injective. We will not need to consider left-streamlined modules until Section 5. By default, 'streamlined' will be taken to mean 'right-streamlined'.

Example 3.16. Intervals $\mathbb{I}(b, n)$ are streamlined (but not $\mathbb{I}(b, d)$ for $d < n$). Conversely, a streamlined τ-module $\mathbb{V}$ with $\dim(V_n) = 1$ is necessarily isomorphic to some $\mathbb{I}(b, n)$. Indeed, $\dim(V_i)$ is a non-decreasing sequence and therefore comprises some $b - 1$ zeros (where $1 \leq b \leq n$) followed by $n - b + 1$ ones. The maps between the one-dimensional terms are injective or surjective, and therefore isomorphisms.

Proposition 3.17. *A direct sum $\mathbb{V} = \mathbb{V}_1 \oplus \cdots \oplus \mathbb{V}_N$ of τ-modules is streamlined if and only if each summand is streamlined.*

Proof. Each $\xrightarrow{f}$ in $\mathbb{V}$ decomposes as $f = f_1 \oplus \cdots \oplus f_N$ and is injective if and only if each f_j is injective. Each $\xleftarrow{g}$ in $\mathbb{V}$ decomposes as $g = g_1 \oplus \cdots \oplus g_N$ and is surjective if and only if each g_j is surjective. □

The proof of the following lemma appears at the end of this section.

Lemma 3.18 (Decomposition Lemma)**.** *Let $\mathbb{V}$ be a streamlined τ-module and let $\mathcal{R} = \mathrm{R}(\mathbb{V})$. For any decomposition $\mathcal{R} = \mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_N$, there exists a unique decomposition $\mathbb{V} = \mathbb{W}_1 \oplus \cdots \oplus \mathbb{W}_N$ such that $\mathcal{S}_i = \mathrm{R}(\mathbb{W}_j)$ for all j.*

Theorem 3.19 (Interval decomposition for streamlined modules)**.** *Let $\mathbb{V}$ be a streamlined τ-module of length n, and write $\dim(\mathrm{R}(\mathbb{V})) = (c_1, c_2, \ldots, c_n)$ and $\mathrm{b}(\tau) = (b_1, b_2, \ldots, b_n)$. Then there is an isomorphism of τ-modules*

$$\mathbb{V} \cong \bigoplus_{1 \leq i \leq n} c_i \, \mathbb{I}(b_i, n).$$

Proof. Let $\mathcal{R} = \mathrm{R}(\mathbb{V})$. By Proposition 3.13, there is a decomposition $\mathcal{R} = \mathcal{J}_1 \oplus \cdots \oplus \mathcal{J}_N$, where the $\mathcal{J}_j$ are a collection of $N = c_1 + \cdots + c_n$ intervals with $\mathcal{J}(i, n)$ occuring with multiplicity c_i. Lemma 3.18 produces a decomposition $\mathbb{V} = \mathbb{I}_1 \oplus \cdots \oplus \mathbb{I}_N$, with $\mathrm{R}(\mathbb{I}_j) = \mathcal{J}_j$ for all j. Each $\mathbb{I}_j$ is streamlined (Proposition 3.17) with maximum dimension $\dim((I_j)_n) = 1$, and is therefore isomorphic to some $\mathbb{I}(b, n)$ (Example 3.16). By Lemma 3.9, we must have $\mathbb{I}_j = \mathbb{I}(b_i, n)$ whenever $\mathcal{J}_j = \mathcal{J}(i, n)$. It follows that the $\mathbb{I}_j$ are a collection of $N = c_1 + \cdots + c_n$ intervals with $\mathbb{I}(b_i, n)$ occuring with multiplicity c_i. □

We complete this chapter with a proof of the Decomposition Lemma.

Proof of Lemma 3.18. We may assume that $N = 2$, since the general case follows by iteration. Accordingly, suppose that $\mathcal{R} = \mathrm{R}(\mathbb{V})$ can be written in the form $\mathcal{R} = \mathcal{S} \oplus \mathcal{T}$; we must show that there is a corresponding decomposition $\mathbb{V} = \mathbb{W} \oplus \mathbb{X}$. We will argue by induction on $n = \mathrm{length}(\tau)$.

The first step is to determine the splitting $V_n = W_n \oplus X_n$. In fact, the stipulation that $\mathcal{S} = \mathrm{R}(\mathbb{W})$ and $\mathcal{T} = \mathrm{R}(\mathbb{X})$ forces $W_n = S_n$ and $X_n = T_n$. If $n = 1$, then we are done. Otherwise, let $\hat{\mathbb{V}}$ denote the truncation of $\mathbb{V}$ to the indices $\{1, \ldots, n-1\}$ and let $\hat{\mathcal{R}} = \mathrm{R}(\hat{\mathbb{V}})$. We will shortly establish that $\mathcal{R} = \mathcal{S} \oplus \mathcal{T}$ induces a unique compatible decomposition $\hat{\mathcal{R}} = \hat{\mathcal{S}} \oplus \hat{\mathcal{T}}$. The inductive hypothesis will then provide $\hat{\mathbb{V}} = \hat{\mathbb{W}} \oplus \hat{\mathbb{X}}$, which combines with $V_n = W_n \oplus X_n$ to produce the desired decomposition $\mathbb{V} = \mathbb{W} \oplus \mathbb{X}$. That will complete the proof.

Write $\mathcal{R} = (R_0, R_1, \ldots, R_n)$. There are two cases.

Case $\xrightarrow{f_{n-1}}$, injective. We can identify V_{n-1} with the subspace $f_{n-1}(V_{n-1}) = R_{n-1}$ of V_n. Thereupon we have

$$\hat{\mathcal{R}} = (R_0, R_1, \ldots, R_{n-1}).$$

The unique splitting of V_{n-1} compatible with $V_n = W_n \oplus X_n$ is

$$V_{n-1} = (R_{n-1} \cap W_n) \oplus (R_{n-1} \cap X_n) = S_{n-1} \oplus T_{n-1}.$$

We must now verify that the induced subspaces $\hat{\mathcal{S}} = \hat{\mathcal{R}} \cap S_{n-1}$ and $\hat{\mathcal{T}} = \hat{\mathcal{R}} \cap T_{n-1}$ give a valid decomposition $\hat{\mathcal{R}} = \hat{\mathcal{S}} \oplus \hat{\mathcal{T}}$ of filtered vector spaces. This follows because $\hat{S}_i = R_i \cap S_{n-1} = R_i \cap S_n = S_i$ and similarly $\hat{T}_i = T_i$, for all $i < n$; so $R_i = S_i \oplus T_i = \hat{S}_i \oplus \hat{T}_i$ as required.

Case $\xleftarrow{g_{n_1}}$, surjective. Here we identify V_{n-1} as the quotient $V_n / \ker(g_{n-1}) = R_n / R_1$. Under this identification,

$$\hat{\mathcal{R}} = (R_1/R_1, R_2/R_1, \ldots, R_n/R_1).$$

In splitting $V_{n-1} = W_{n-1} \oplus X_{n-1}$ we are compelled to take

$$W_{n-1} = g_{n-1}(W_n) = S_n/S_1, \qquad X_{n-1} = g_{n-1}(X_n) = T_n/T_1,$$

which induce

$$\hat{S}_i = g_{n-1}(S_{i+1}) = S_{i+1}/S_1, \qquad \hat{T}_i = g_{n-1}(T_{i+1}) = T_{i+1}/T_1,$$

for the purported splitting $\hat{\mathcal{R}} = \hat{\mathcal{S}} \oplus \hat{\mathcal{T}}$. To confirm that this is a genuine decomposition, we note from linear algebra that the twin facts

$$R_{i+1} = S_{i+1} \oplus T_{i+1}, \qquad R_1 = S_1 \oplus T_1 = (S_{i+1} \cap R_1) \oplus (T_{i+1} \cap R_1)$$

imply that

$$R_{i+1}/R_1 = (S_{i+1}/S_1) \oplus (T_{i+1}/T_1)$$

as required. □

Remark. There is a high-level proof of Lemma 3.18 which in some sense is the natural explanation for the result. We outline this proof now. The first observation is that the transformation $\mathbb{V} \to \mathrm{R}(\mathbb{V})$ is a functor from $\tau\mathrm{Mod}$ to Filt_n: a morphism $\alpha : \mathbb{V} \to \mathbb{W}$ induces a morphism $\mathrm{R}(\alpha) : \mathrm{R}(\mathbb{V}) \to \mathrm{R}(\mathbb{W})$. Indeed, $\mathrm{R}(\alpha)$ is defined to be $\alpha_n : V_n \to W_n$; one must check that this respects the filtrations on V_n and W_n. Being a functor, R defines a ring homomorphism $\mathrm{End}(\mathbb{V}) \to \mathrm{End}(\mathrm{R}(\mathbb{V}))$. The second key fact is that this homomorphism is an isomorphism if $\mathbb{V}$ is streamlined (in general it is surjective). It is well known that direct-sum decompositions of a module can be extracted from the structure of its endomorphism

ring: direct summands correspond to idempotent elements of the ring. It follows that $\mathbb{V}$ and $\mathrm{R}(\mathbb{V})$ have the same decomposition structure.

4. The Interval Decomposition Algorithm

Here we describe the algorithm for determining the indecomposable factors of a τ-module. We give three versions of the 'algorithm'.

The first version, in Section 4.1, is not an algorithm but a proof that every τ-module decomposes as a sum of interval modules (Theorem 2.5). Moreover, the structure of the proof makes it clear how to compute the interval decomposition (Theorem 4.1). The algorithms in the subsequent sections build on this.

In Section 4.2 we describe an abstract form of the decomposition algorithm, using the language of vector spaces and linear maps. No consideration is given to how the spaces and maps are described and manipulated in practice.

In Section 4.3 we suppose that the maps f_i, g_i are presented concretely as matrices M_i, N_i with respect to a choice of bases for the vector spaces V_i. We describe an algorithm which takes these matrices as input and returns the interval decomposition.

4.1. The interval decomposition theorem.

Our present goal is to give a somewhat constructive proof of Theorem 2.5, which asserts that any τ-module $\mathbb{V}$ is isomorphic to a direct sum of intervals $\mathbb{I}(b, d)$. We prove a stronger, more precise result, which explicitly determines the multiplicity of each interval within $\mathbb{V}$.

Some notation will help with the theorem statement. If

$$\mathbb{V} = (V_1 \xleftrightarrow{p_1} \dots \xleftrightarrow{p_{n-1}} V_n)$$

then let

$$\mathbb{V}[k] = (V_1 \xleftrightarrow{p_1} \dots \xleftrightarrow{p_{k-1}} V_k)$$

denote the truncation of $\mathbb{V}$ to length k, and let $\tau[k]$ denote its type (which is a truncation of τ).

Theorem 4.1 (Interval decomposition)**.** *Let $\mathbb{V}$ be a τ-module. For $1 \le k \le n$, define*

$$(b_1^k, b_2^k, \dots, b_k^k) = \mathrm{b}(\tau[k]).$$

Writing $\mathcal{R}_k = \mathrm{R}(\mathbb{V}[k])$, define

$$(c_1^k, c_2^k, \dots, c_k^k) = \begin{cases} \dim(\mathcal{R}_k \cap \mathrm{Ker}(f_k)) \\ \dim(\mathcal{R}_k) - \dim(\mathcal{R}_k \cap \mathrm{Im}(g_k)) \end{cases}$$

(whichever is applicable) when $k \neq n$, and

$$(c_1^n, c_2^n, \dots, c_n^n) = \dim(\mathcal{R}_n).$$

Then

$$\mathbb{V} \cong \bigoplus_{1 \le i \le k \le n} c_i^k \, \mathbb{I}(b_i^k, k).$$

Addendum 4.2. *In the situation of Theorem 4.1, write*

$$(r_1^k, \dots, r_k^k) = \dim(\mathcal{R}_k)$$

for $k = 1, \ldots, n$, and conventionally define $r_i^{n+1} = 0$ for all i. Then

$$c_i^k = \begin{cases} r_i^k - r_i^{k+1} & \text{case } \xrightarrow{f_k} \\ r_i^k - r_{i+1}^{k+1} & \text{case } \xleftarrow{g_k} \end{cases}$$

for $1 \le i \le k \le n$.

The decomposition strategy begins with the following lemma. The idea is to proceed from left to right along the complex, removing streamlined summands at each step. Having done this, the Remak decompositions of those summands can be determined by counting dimensions, as prescribed in Theorem 3.19.

Lemma 4.3. *Let $\mathbb{V} = V_1 \xleftrightarrow{p_1} \ldots \xleftrightarrow{p_{n-1}} V_n$ be an irreducible τ-module of length n. Then there exists a direct-sum decomposition*

$$\mathbb{V} = \mathbb{V}^1 \oplus \mathbb{V}^2 \oplus \cdots \oplus \mathbb{V}^n$$

where each $\mathbb{V}^k$ is supported over the indices $\{1, 2, \ldots, k\}$ and is right-streamlined over that range.

The following picture illustrates the decomposition.

$$\mathbb{V} = \begin{cases} \mathbb{V}^1 & = & V_1^1 & & & & & & \\ \oplus & & & & & & & & \\ \mathbb{V}^2 & = & V_1^2 & \xleftrightarrow{p_1} & V_2^2 & & & & \\ \oplus & & & & & & & & \\ \mathbb{V}^3 & = & V_1^3 & \xleftrightarrow{p_1} & V_2^3 & \xleftrightarrow{p_2} & V_3^3 & & \\ \oplus & & & & & & & & \\ \vdots & & & & & & & & \\ \oplus & & & & & & & & \\ \mathbb{V}^n & = & V_1^n & \xleftrightarrow{p_1} & V_2^n & \xleftrightarrow{p_2} & V_3^n & \xleftrightarrow{p_3} \cdots \xleftrightarrow{p_{n-1}} & V_n^n \end{cases}$$

Each row (i.e. summand) is right-streamlined, and therefore amenable to analysis via the right-filtration functor.

Proof. We proceed by induction on the length of $\mathbb{V}$. The inductive statement is that

$$\mathbb{V}[k] = \mathbb{V}^1 \oplus \cdots \oplus \mathbb{V}^{k-1} \oplus \mathbb{W}^k$$

where the $\mathbb{V}^i$ are as in the theorem statement, and $\mathbb{W}^k$ is itself right-streamlined.

For the base case $k = 1$, there is nothing to prove: take $\mathbb{W}^1 = \mathbb{V}[1]$. Now suppose the inductive statement is established for k, and consider $\mathbb{V}[k+1]$. This can be written

$$\begin{aligned} V[k+1] &= (\mathbb{V}^1 \oplus \cdots \oplus \mathbb{V}^{k-1} \oplus \mathbb{W}^k) \xleftrightarrow{p_k} V_{k+1} \\ &= \mathbb{V}^1 \oplus \cdots \oplus \mathbb{V}^{k-1} \oplus (\mathbb{W}^k \xleftrightarrow{p_k} V_{k+1}) \end{aligned}$$

where the rebracketing is permissible because all of the $\mathbb{V}^i$ terms terminate before time k, and therefore do not interact with $\xleftrightarrow{p_k}$. The goal now is to rewrite $(\mathbb{W}^k \xleftrightarrow{p_k} V_{k+1})$ as $\mathbb{V}^k \oplus \mathbb{W}^{k+1}$, where $\mathbb{V}^k$ terminates at time k and both $\mathbb{V}^k$ and $\mathbb{W}^{k+1}$ are right-streamlined. The rightmost term of $\mathbb{W}^k$ is V_k, so $\mathrm{R}(\mathbb{W}^k)$ is a filtration on V_k.

Case f: $\mathbb{W}^k \xrightarrow{f_k} V_{k+1}$. In other words $f_k : V_k \to V_{k+1}$. Let $\mathcal{S} = \mathrm{R}(\mathbb{W}^k) \cap \mathrm{Ker}(f_k)$. Proposition 3.11 implies that $\mathcal{S}$ has a complement in $\mathrm{R}(\mathbb{W}^k)$; say $\mathrm{R}(\mathbb{W}^k) = \mathcal{S} \oplus \mathcal{T}$. This

corresponds (Lemma 3.18) to a direct sum decomposition $\mathbb{W}^k = \mathbb{V}^k \oplus \hat{\mathbb{W}}^k$, where both summands are streamlined (Proposition 3.17). This defines $\mathbb{V}^k$, and we set $\mathbb{W}^{k+1} = (\hat{\mathbb{W}}^k \xrightarrow{f_k} V_{k+1})$. To check that this works, note that f_k is zero on $(\mathbb{V}^k)_k = \mathrm{Ker}(f_k)$ and is injective on the complementary subspace $(\hat{\mathbb{W}}^k)_k$. Thus $\mathbb{V}^k$ is a summand of $\mathbb{V}[k+1]$ terminating at time k, and $\mathbb{W}^{k+1}$ is streamlined.

Case g: $\mathbb{W}^k \xleftarrow{g_k} V_{k+1}$. In other words $g_k : V_{k+1} \to V_k$. Let $\mathcal{S} = \mathrm{R}(\mathbb{W}^k) \cap \mathrm{Im}(g_k)$. Proposition 3.11 implies that $\mathcal{S}$ has a complement in $\mathrm{R}(\mathbb{W}^k)$; say $\mathrm{R}(\mathbb{W}^k) = \mathcal{S} \oplus \mathcal{T}$. This corresponds (Lemma 3.18) to a direct sum decomposition $\mathbb{W}^k = \hat{\mathbb{W}}^k \oplus \mathbb{V}^k$, where both summands are streamlined (Proposition 3.17). This defines $\mathbb{V}^k$, and we set $\mathbb{W}^{k+1} = (\hat{\mathbb{W}}^k \xleftarrow{g_k} V_{k+1})$. To check that this works, note that g_k is surjective onto $(\hat{\mathbb{W}}^k)_k = \mathrm{Im}(g_k)$ and misses the complementary subspace $(\mathbb{V}^k)_k$. Thus $\mathbb{V}^k$ is a summand of $\mathbb{V}[k+1]$ terminating at time k, and $\mathbb{W}^{k+1}$ is streamlined.

This establishes the inductive step, so eventually

$$\mathbb{V} = \mathbb{V}[n] = \mathbb{V}^1 \oplus \cdots \oplus \mathbb{V}^{n-1} \oplus \mathbb{W}^n$$

and we set $\mathbb{V}^n = \mathbb{W}^n$ to finish the proof. □

Proof of Theorem 4.1. Write $\mathbb{V} = \mathbb{V}^1 \oplus \cdots \oplus \mathbb{V}^n$ according to Lemma 4.3. We now calculate the decomposition of each $\mathbb{V}^k$ into intervals $\mathbb{I}(b,k)$. Note that

$$\mathbb{V}[k] = \mathbb{V}^k \oplus \mathbb{V}^{k+1}[k] \oplus \cdots \oplus \mathbb{V}^n[k].$$

We can write $\mathbb{W}^k = \mathbb{V}^{k+1} \oplus \cdots \oplus \mathbb{V}^n$, so then

$$\mathcal{R}_k = \mathrm{R}(\mathbb{V}^k \oplus \mathbb{W}^k[k]) = \mathrm{R}(\mathbb{V}^k) \oplus \mathrm{R}(\mathbb{W}^k[k])$$

(using Proposition 3.14). This is a filtration on $V_k^k \oplus W_k^k$.

Suppose $k < n$. We note that $\mathbb{W}^k$ is streamlined up to time $k+1$, whereas $\mathbb{V}^k$ is zero at time $k+1$. The next map in the sequence is

$$V_k^k \oplus W_k^k \xrightarrow{f_k} W_{k+1}^k \qquad \text{or} \qquad V_k^k \oplus W_k^k \xleftarrow{g_k} W_{k+1}^k.$$

In the first case, it follows that $V_k^k = \mathrm{Ker}(f_k)$ and therefore $\mathrm{R}(\mathbb{V}^k) = \mathcal{R}_k \cap \mathrm{Ker}(f_k)$. In the second case, V_k^k is a complement to $\mathrm{Im}(g_k)$ in V_k, so $\mathcal{R}_k = \mathrm{R}(\mathbb{V}^k) \oplus (\mathcal{R}_k \cap \mathrm{Im}(g_k))$. Thus

$$\dim(\mathrm{R}(\mathbb{V}^k)) = \left\{ \begin{array}{l} \dim(\mathcal{R}_k \cap \mathrm{Ker}(f_k)) \\ \dim(\mathcal{R}_k) - \dim(\mathcal{R}_k \cap \mathrm{Im}(g_k)) \end{array} \right\} = (c_1^k, \ldots, c_k^k)$$

(whichever middle term is applicable). When $k = n$, moreover, we have

$$\dim(\mathrm{R}(\mathbb{V}^n)) = \dim(\mathcal{R}_n) = (c_1^n, \ldots, c_n^n).$$

Thus, at last,

$$\mathbb{V} = \bigoplus_{1 \le k \le n} \mathbb{V}^k \cong \bigoplus_{1 \le k \le n} \left\{ \bigoplus_{1 \le i \le k} c_i^k \, \mathbb{I}(b_i^k, k) \right\}$$

using Theorem 3.19 to decompose the $\mathbb{V}^k$. □

Proof of Addendum 4.2. Write $(w_1^k, \ldots, w_k^k) = \dim(\mathrm{R}(\mathbb{W}^k[k]))$. Since $\mathcal{R}_k = \mathrm{R}(\mathbb{V}^k) \oplus \mathrm{R}(\mathbb{W}^k[k])$ we can take dimensions and obtain the formula

$$(r_1^k, \ldots, r_k^k) = (c_1^k, \ldots, c_k^k) + (w_1^k, \ldots, w_k^k).$$

Note also that $\mathcal{R}_{k+1} = \mathrm{R}(\mathbb{V}[k+1]) = \mathrm{R}(\mathbb{W}^k[k+1])$. Moreover, $\mathbb{W}^k$ is streamlined up to time $k+1$. It follows that

$$(r_1^{k+1}, \ldots, r_{k+1}^{k+1}) = \dim(\mathrm{R}(\mathbb{W}^k[k+1])) = \begin{cases} (w_1^k, \ldots, w_k^k, ?) & \text{case } f \\ (?, w_1^k, \ldots, w_k^k) & \text{case } g \end{cases}$$

and therefore

$$c_i^k = r_i^k - w_i^k = \begin{cases} r_i^k - r_i^{k+1} & \text{case } f \\ r_i^k - r_{i+1}^{k+1} & \text{case } g \end{cases}$$

which is the desired formula. □

4.2. **Abstract vector spaces.** We now transcribe Theorem 4.1 as an abstract algorithm for determining the interval structure of a τ-module $\mathbb{V}$ of length n. This algorithm will serve as a skeleton for the more concrete algorithms developed later.

Algorithm 4.4. We proceed through $k = 1, 2, \ldots, n$, computing the filtration $\mathcal{R}_k = \mathrm{R}(\mathbb{V}[k])$, the birth-time index $\mathrm{b}(\tau[k])$, and the dimensions c_i^k iteratively.

BEGIN

Initialisation $(k = 1)$:

(1) $\mathcal{R}_1 = (0, V_1)$.

(2) $\mathrm{b}(\tau[1]) = (1)$.

Iterative step $(k = 1, 2, \ldots, n-1)$:

(3) Calculate $\mathcal{R}_{k+1}$ from $\mathcal{R}_k = (R_0^k, R_1^k, \ldots, R_k^k)$ using Definition 3.1:

$$(R_0^{k+1}, R_1^{k+1}, \ldots, R_{k+1}^{k+1}) = \begin{cases} (f_k(R_0^k), f_k(R_1^k), \ldots, f_k(R_k^k), V_{k+1}) & \text{case } f \\ (0, g_k^{-1}(R_0^k), g_k^{-1}(R_1^k), \ldots, g_k^{-1}(R_k^k)) & \text{case } g \end{cases}$$

(4) Calculate $\mathrm{b}(\tau[k+1]))$ from $\mathrm{b}(\tau[k]) = (b_1^k, b_2^k, \ldots, b_k^k)$ using Definition 3.6:

$$(b_1^{k+1}, \ldots, b_{k+1}^{k+1}) = \begin{cases} (b_1^k, \ldots, b_k^k, k+1) & \text{case } f \\ (k+1, b_1^k, \ldots, b_k^k) & \text{case } g \end{cases}$$

(5) Calculate $(c_1^k, \ldots, c_k^k)$ using the formula in Theorem 4.1:

$$(c_1^k, c_2^k, \ldots, c_k^k) = \begin{cases} \dim(\mathcal{R}_k \cap \mathrm{Ker}(f_k)) & \text{case } f \\ \dim(\mathcal{R}_k) - \dim(\mathcal{R}_k \cap \mathrm{Im}(g_k)) & \text{case } g \end{cases}$$

Alternatively, use the formula in Addendum 4.2:

$$c_i^k = \begin{cases} r_i^k - r_i^{k+1} & \text{case } f \\ r_i^k - r_{i+1}^{k+1} & \text{case } g \end{cases}$$

Here $(r_1^k, \ldots, r_k^k) = \dim(\mathcal{R}_k)$.

Terminal step $(k = n)$:

(6) Calculate $(c_1^n, \ldots, c_n^n) = \dim(\mathcal{R}(\mathbb{V}))$.

Print results:

(7) For $1 \le i \le k \le n$, the interval $\mathbb{I}(b_i^k, k)$ occurs with multiplicity c_i^k.

END

Note that steps (3–5) have an 'f' verson and a 'g' version, depending on the direction of the map p_k.

This abstract algorithm does not specify how the filtered vector spaces $\mathrm{R}(\mathbb{V}[k+1])$ are stored, nor how the maps f_k or g_k (which are used in steps (3) and (5)) are represented. In any concrete setting, it is necessary to specify data structures. A good choice will facilitate the calculations in steps (3) and (5). In the next section, we work out the details in a simple scenario.

4.3. **Concrete vector spaces.** In this section we describe an algorithm to solve the following concrete problem. Let τ be a type of length n. We specify a τ-module $\mathbb{V}$ as follows. Set $V_i = \mathbb{F}^{a_i}$ for integers $a_i \geq 0$. For each i, the map f_i is defined by an a_{i+1}-by-a_i matrix M_i or else the map g_i is defined by an a_i-by-a_{i+1} matrix N_i. We are to determine $\mathrm{Pers}(\mathbb{V})$, given τ and the matrices M_i or N_i.

We follow Algorithm 4.4. The substantial task is to calculate the sequence of right-filtrations $\mathcal{R}_k = \mathrm{R}(\mathbb{V}[k])$, for step (3). Everything else is book-keeping: the birth-time indices b_i^k are calculated according to step (4); and the filtration dimensions r_i^k (and hence the c_i^k) will be easy to read off from the stored description of the filtrations.

Basis transformations. The algorithm operates on two levels. On the *conceptual* level, we proceed by modifying the bases of the spaces V_i by elementary basis transformations. Initially each basis $\mathcal{B}_i$ is the standard basis of $\mathbb{F}^{a_i}$. We perform modifications on $\mathcal{B}_2, \mathcal{B}_3, \ldots, \mathcal{B}_{n-1}$ in sequence. On the *pragmatic* level, what we actually do is apply elementary row and column operations to the matrices M_i or N_i. We make no attempt to track the bases themselves; instead we implement the effect of those changes on the matrices.

Suppose we apply elementary basis transformations to $\mathcal{B}_{k+1}$ on the conceptual level. On the pragmatic level, we must perform

$$\text{row operations on } M_k \qquad \text{or} \qquad \text{column operations on } N_k$$

and simultaneously perform

$$\text{column operations on } M_{k+1} \qquad \text{or} \qquad \text{row operations on } N_{k+1}$$

to enact those transformations. Thus, at every stage we must make parallel changes to two matrices simultaneously. Usually we are working to put M_k or N_k in a particular form, and while doing so the changes have to be mirrored in M_{k+1} or N_{k+1} (paying no attention yet to the structure of that matrix).

We now make this precise. The **elementary transformation** $E_i(p, q, \lambda)$ is defined as follows. On the conceptual level, this is a modification of $\mathcal{B}_i = (\beta_1, \ldots, \beta_{a_i})$ involving basis vectors β_p and β_q:

$$\begin{aligned} \beta_p &\leftarrow \beta_p \\ \beta_q &\leftarrow \beta_q + \lambda\beta_p \end{aligned}$$

On the pragmatic level, if L is a matrix representing a linear map $V_i \to W$ for some W (this will be N_{i-1} or M_i in our situation), then we modify the columns of L accordingly:

$$\begin{aligned} \mathrm{Column}_p &\leftarrow \mathrm{Column}_p \\ \mathrm{Column}_q &\leftarrow \mathrm{Column}_q + \lambda\,\mathrm{Column}_p \end{aligned}$$

Else, if L represents a linear map of the form $W \to V_i$ (this will be M_{i-1} or N_i in our situation) then we must apply the dual transformation to the rows of L:

$$\begin{aligned} \mathrm{Row}_p &\leftarrow \mathrm{Row}_p - \lambda\,\mathrm{Row}_q \\ \mathrm{Row}_q &\leftarrow \mathrm{Row}_q \end{aligned}$$

In spirit, we right-multiply by the matrix $\begin{bmatrix} 1 & \lambda \\ 0 & 1 \end{bmatrix}$ to modify columns, or else left-multiply by the inverse matrix $\begin{bmatrix} 1 & -\lambda \\ 0 & 1 \end{bmatrix}$ to modify rows.

Besides the elementary transformations $E_i(p, q, \lambda)$, it is sometimes appropriate to permute the basis elements. The operation $P_i(p, q)$ of interchanging β_p with β_q is realised pragmatically by interchanging Column_p with Column_q, or Row_p with Row_q, as appropriate.

Filtrations. The filtration $\mathcal{R}_k = \mathrm{R}(\mathbb{V}[k])$ on V_k is to be represented as follows. We require the basis $\mathcal{B}_k = (\beta_1, \ldots, \beta_{a_i})$ to be compatible with the filtration, in a sense that will become clear. Assuming such a basis, the filtration $\mathcal{R}_k = (R_0, R_1, \ldots, R_k)$ is represented as a non-decreasing function

$$\phi_k : \{1, 2, \ldots, a_i\} \to \{1, \ldots, k\}$$

so that

$$R_i = \mathrm{Span}\,\{\beta_p \mid \phi_k(p) \le i\}$$

for $i = 1, \ldots, k$. In other words: the first few basis elements (those β_p with $\phi_k(p) = 1$) form a basis for R_1; the next few basis elements extend this to a basis for R_2, and so on. The dimension $r_i^k = \dim(R_i/R_{i-1})$ can be read off as the cardinality of $\phi_k^{-1}(i)$.

Gaussian elimination. Step (3) boils down to the following task. Suppose that $\mathcal{B}_k$ and ϕ_k together represent the filtration $\mathcal{R}_k$; then modify $\mathcal{B}_{k+1}$ and determine ϕ_{k+1} to represent $\mathcal{R}_{k+1}$. We now explain how to do this.

Case M: the matrix M_k represents a linear map $V_k \to V_{k+1}$. We assume that $\mathcal{B}_k$ is compatible with the filtration $\mathcal{R}_k$, and that ϕ_k identifies the filtration. This gives a block structure

$$M_k = \begin{bmatrix} K_1 & K_2 & \cdots & K_k \end{bmatrix}$$

where K_i gathers together the columns q with $\phi_k(q) = i$. Using row operations only, put M_k into (unreduced) row echelon form. This means:

- Each of the top r rows contains a 1 (the *pivot*) as its leftmost nonzero entry.
- Each pivot lies strictly to the left of the pivots of the rows below it.
- The lowest $a_{k+1} - r$ rows are entirely zero.

These row operations correspond to elementary operations $E_{k+1}(p, q, \lambda)$, and the effect of these operations is felt on the next matrix M_{k+1} or N_{k+1}, which must be modified accordingly. We now define ϕ_{k+1} as follows:

$$\phi_{k+1}(p) = \begin{cases} \phi_k(q) & \text{if row } p \text{ has a pivot in column } q, \\ k+1 & \text{if row } p \text{ has no pivot.} \end{cases}$$

See Figure 3. It is evident in the figure that R_i^k maps onto R_i^{k+1} for all i.

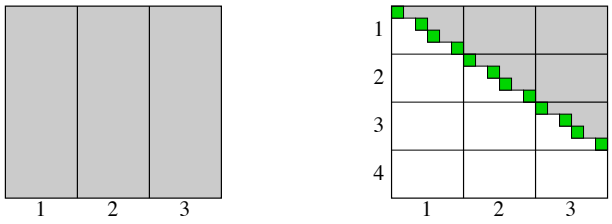


FIGURE 3. Using row echelon form to compute $\mathcal{R}_{k+1}$ from $\mathcal{R}_k$.

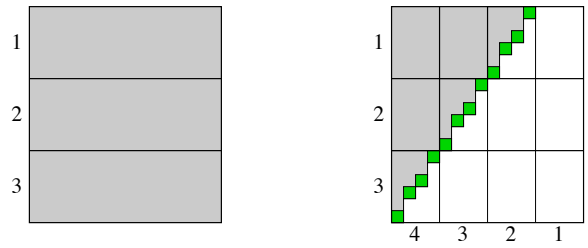


FIGURE 4. Using column echelon form to compute $\mathcal{R}_{k+1}$ from $\mathcal{R}_k$.

Case N: the matrix N_k represents a linear map $V_{k+1} \to V_k$. We assume that $\mathcal{B}_k$ is compatible with the filtration $\mathcal{R}_k$, and that ϕ_i identifies the filtration. This time we have a vertical block structure

$$N_k = \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_k \end{bmatrix}$$

where L_i gathers together the rows q with $\phi_k(q) = i$. Using column operations only, put N_k into the column echelon form defined as follows (this echelon form begins on the bottom left):

- Each of the leftmost r columns contains a 1 (the pivot) as its *lowest* nonzero entry.
- Each pivot lies strictly lower than the pivots of the columns to the right of it.
- The rightmost $a_{k+1} - r$ rows are entirely zero.

These column operations correspond to elementary operations $E_{k+1}(p, q, \lambda)$, and the effect of these operations is felt on the next matrix M_{k+1} or N_{k+1}, which must be modified accordingly. We now define ϕ_{k+1} as follows:

$$\phi_{k+1}(p) = \begin{cases} \phi_k(q) + 1 & \text{if column } p \text{ has a pivot in row } q, \\ 1 & \text{if column } p \text{ has no pivot.} \end{cases}$$

See Figure 4. It is evident in the figure that R_{i+1}^{k+1} is the largest subspace which maps into R_i^k, for all i.

This concludes our treatment of the concrete form of the zigzag algorithm.

5. Further Algebraic Techniques

5.1. **Localization at a single index.** Let $\mathbb{V}$ be a zigzag module of length n and let $1 \le k \le n$. We consider the problem of determining the set of intervals in $\mathrm{Pers}(\mathbb{V})$ which contain k, without necessarily computing $\mathrm{Pers}(\mathbb{V})$ itself. We shall see that all the necessary information is contained in a pair of filtrations on the vector space V_k.

Definition 5.1. Let $\mathbb{V}$ be a zigzag module of length n. The **left-filtration** of $\mathbb{V}$ is a filtration on V_1 of depth n, defined as

$$\mathrm{L}(\mathbb{V}) = \mathrm{R}(\bar{\mathbb{V}})$$

where $\bar{\mathbb{V}}$ is the reversal of $\mathbb{V}$; so $\bar{V}_i = V_{n+1-i}$, with maps $\bar{f}_i = g_{n-i}$ or $\bar{g}_i = f_{n-i}$.

For any k we therefore have two natural filtrations on V_k:

$$\begin{array}{rcccl} \mathcal{R}_k & = & (R_0, R_1, \ldots, R_k) & = & \mathrm{R}(\mathbb{V}[1,k]), \\ \mathcal{L}_k & = & (L_0, L_1, \ldots, L_{n+1-k}) & = & \mathrm{L}(\mathbb{V}[k,n]); \end{array}$$

the right-filtration over the index set $\{1, \ldots, k\}$ and the left-filtration over the index set $\{k, \ldots, n\}$. We also have birth-time and death-time indices

$$\begin{array}{rcccl} \mathrm{b}_k & = & (b_1, \ldots, b_k) & = & \mathrm{b}(\tau[1,k]) \\ \mathrm{d}_k & = & (d_1, \ldots, d_{n+1-k}) & = & n+1-\mathrm{b}(\bar{\tau}[k,n]) \end{array}$$

which indicate the birth and death times associated with the respective subquotients of $\mathcal{R}_k$ and $\mathcal{L}_k$. These depend on the type τ of $\mathbb{V}$.

Example 5.2. Consider the zigzag module

$$\mathbb{V} \;=\; (\, V_1 \xrightarrow{f_1} V_2 \xrightarrow{f_2} V_3 \xleftarrow{g_3} V_4 \,).$$

At $k = 2$, for instance, we have

$$\begin{array}{rcl} \mathcal{R}_2 & = & (0, f_1(V_1), V_2) \\ \mathcal{L}_2 & = & (0, f_2^{-1}(0), f_2^{-1} g_3(V_4), V_2) \end{array}$$

and

$$\begin{array}{rcl} \mathrm{b}_2 & = & (1,2) \\ \mathrm{d}_2 & = & (2,4,3). \end{array}$$

We can now state the main theorem of this section.

Theorem 5.3 (Localization at index k)**.** *Let $\mathbb{V}$ be a zigzag module of length n and let $1 \le k \le n$. Let $\mathcal{R}_k, \mathcal{L}_k$ denote the right- and left-filtrations at k, and let $\mathrm{b}_k, \mathrm{d}_k$ denote the birth-time and death-time indices at k. Then, for all i, j in the range $1 \le i \le k$, $1 \le j \le n+1-k$, the multiplicity of $[b_i, d_j]$ in $\mathrm{Pers}(\mathbb{V})$ is equal to*

$$c_{ij} = \dim(R_i \cap L_j) - \dim(R_{i-1} \cap L_j) - \dim(R_i \cap L_{j-1}) + \dim(R_{i-1} \cap L_{j-1}).$$

Remark. Equivalently, $c_{ij} = \dim((R_i \cap L_j)/((R_{i-1} \cap L_j) + (R_i \cap L_{j-1})))$, the dimension of the (i,j)-th bifiltration subquotient.

This theorem answers the original question, because every interval containing k can be written as $[b_i, d_j]$ for some choice of i, j. We now work towards a proof of Theorem 5.3.

Proposition 5.4. *It is sufficient to prove Theorem 5.3 in the special case where $\mathbb{V}$ is right-streamlined over $\{1, \ldots, k\}$ and left-streamlined over $\{k, \ldots, n\}$.*

Proof. It is clear from Lemma 4.3 that we can write $\mathbb{V} = \mathbb{U} \oplus \mathbb{W}$ where $\mathbb{U}$ is supported in $\{1, \dots, k-1\}$ and $\mathbb{W}$ is right-streamlined over $\{1, \dots, k\}$. Indeed, take $\mathbb{U} = \mathbb{V}^1 \oplus \cdots \oplus \mathbb{V}^{k-1}$ and $\mathbb{W} = \mathbb{V}^k \oplus \cdots \oplus \mathbb{V}^n$. Moreover, it is sufficient to prove Theorem 5.3 for $\mathbb{W}$, because the filtrations $\mathcal{R}_k, \mathcal{L}_k$ remain unchanged from $\mathbb{V}$, and the discarded term $\mathbb{U}$ decomposes into intervals which do not contain k. Thus, we may assume that $\mathbb{V}$ is right-streamlined over $\{1, \dots, k\}$.

Repeating this argument from the other side, we may further assume that $\mathbb{V}$ is left-streamlined over $\{k, \dots, n\}$. □

Proof of Theorem 5.3. Assume that $\mathbb{V}$ satisfies the condition in Proposition 5.4. It follows that every interval in $\mathrm{Pers}(\mathbb{V})$ contains k: any other interval in the decomposition would cause a failure of the streamline condition. We can therefore write the interval decomposition of $\mathbb{V}$ as

$$\mathbb{V} = \bigoplus_{a \in A} \mathbb{I}_a \cong \bigoplus_{a \in A} \mathbb{I}(b_{p(a)}, d_{q(a)})$$

where A indexes the summands, and $p : A \to \{1, \dots, k\}$ and $q : A \to \{1, \dots, n-k+1\}$ identify the interval type of each summand in terms of the birth-time and death-time indices. It is apparent from this formulation that

$$c_{ij} = \#\{a \in A \mid p(a) = i,\, q(a) = j\}$$

and it remains to compute this in terms of the dimensions $\dim(R_i \cap L_j)$.

The interval decomposition restricts at index k to a direct sum decomposition of V_k into 1-dimensional subspaces U_a, generated by elements x_a, say. Then

$$\mathcal{R}_k = \bigoplus_{a \in A} \mathrm{R}(\mathbb{I}_a[1,k]) = \bigoplus_{a \in A} \mathcal{R}_k \cap U_a \cong \bigoplus_{a \in A} \mathcal{J}(p(a), k)$$

where the final isomorphism comes from Lemma 3.9. Now, the filtration subspace R_i is spanned by the terms isomorphic to $\mathcal{J}(p,k)$ with $p \le i$. In other words, for $i = 1, \dots, k$ we have

$$R_i = \mathrm{Span}\left\{x_a \mid p(a) \le i\right\}.$$

A similar argument proceeding from the other direction gives the analogous formula

$$L_j = \mathrm{Span}\left\{x_a \mid q(a) \le j\right\},$$

for $j = 1, \dots, n+1-k$. Since the x_a are independent, these formulas give bases for R_i, L_j.

We now claim that

$$R_i \cap L_j = \mathrm{Span}\{x_a \mid p(a) \le i,\, q(a) \le j\}$$

for all i, j. The inclusion $\mathrm{Span} \subseteq R_i \cap L_j$ is obvious, because each of the spanning vectors x_a belongs to both R_i and L_j. In the other direction, if $x \in R_i \cap L_j$ then write $x = \sum_{a \in A} \lambda_a x_a$. Since $x \in R_i$, all the coefficients λ_a with $p(a) > i$ must be zero. Since $x \in L_j$, all the coefficients λ_a with $q(a) > j$ must be zero. Thus $x \in \mathrm{Span}\{x_a \mid p(a) \le i,\, q(a) \le j\}$. This establishes the reverse inclusion $R_i \cap L_j \subseteq \mathrm{Span}$ and hence the equality.

Then

$$\dim(R_i \cap L_j) = \#\{x_a \mid p(a) \le i,\, q(a) \le j\} = \sum_{p=1}^{i} \sum_{q=1}^{j} c_{pq}$$

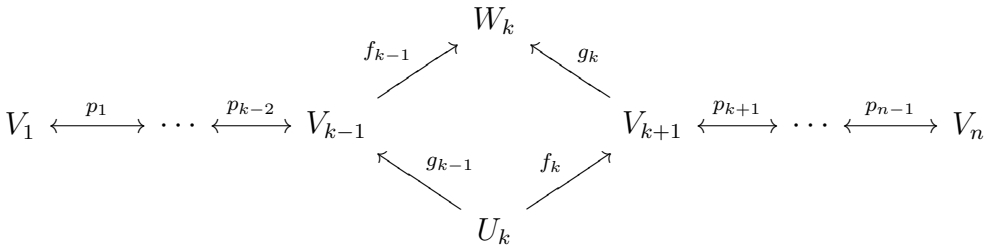
for all i, j. The formula in the theorem follows easily from this. □

Remark. The salient fact behind this result is that it is possible to find a direct sum decomposition of V_k which simultaneously decomposes the filtered spaces $\mathcal{R}_k, \mathcal{L}_k$ into intervals within their respective categories Filt_k, Filt_{n+1-k}. Here we achieved this by appealing to the interval decomposition of $\mathbb{V}$, but this can also be proved directly for an arbitrary pair of filtrations on a single vector space. The analogous statement for a triple of filtrations is false. For example

$$(0, \mathbb{F} \oplus 0, \mathbb{F}^2), \qquad (0, 0 \oplus \mathbb{F}, \mathbb{F}^2), \qquad (0, \Delta, \mathbb{F}^2),$$

(where $\Delta = \{(x, x) \mid x \in \mathbb{F}\}$) cannot be simultaneously decomposed into intervals.

5.2. **The Diamond Principle.** Consider the following diagram:



Let $\mathbb{V}^+$ and $\mathbb{V}^-$ denote the two zigzag modules contained in the diagram:

$$\begin{aligned} \mathbb{V}^+ &= (\, V_1 \longleftrightarrow \ldots \longleftrightarrow V_{k-1} \xrightarrow{f_{k-1}} W_k \xleftarrow{g_k} V_{k+1} \longleftrightarrow \ldots \longleftrightarrow V_n \,) \\ \mathbb{V}^- &= (\, V_1 \longleftrightarrow \ldots \longleftrightarrow V_{k-1} \xleftarrow{g_{k-1}} U_k \xrightarrow{f_k} V_{k+1} \longleftrightarrow \ldots \longleftrightarrow V_n \,) \end{aligned}$$

We wish to compare $\mathrm{Pers}(\mathbb{V}^+)$ with $\mathrm{Pers}(\mathbb{V}^-)$, particularly with respect to intervals that meet $\{k-1, k, k+1\}$. This requires a favourable condition on the four maps in the middle diamond.

Definition 5.5. We say that the diagram

$$\begin{array}{ccc} V_{k+1} & \xrightarrow{g_k} & W_k \\ {\scriptstyle f_k}\uparrow & & \uparrow{\scriptstyle f_{k-1}} \\ U_k & \xrightarrow{g_{k-1}} & V_{k-1} \end{array}$$

is **exact** if $\mathrm{Im}(D_1) = \mathrm{Ker}(D_2)$ in the following sequence

$$U_k \xrightarrow{D_1} V_{k-1} \oplus V_{k+1} \xrightarrow{D_2} W_k$$

where $D_1(u) = g_{k-1}(u) \oplus f_k(u)$ and $D_2(v \oplus v') = f_{k-1}(v) - g_k(v')$.

Theorem 5.6 (The Diamond Principle)**.** *Given $\mathbb{V}^+$ and $\mathbb{V}^-$ as above, suppose that the middle diamond is exact. Then there is a partial bijection of the multisets $\mathrm{Pers}(\mathbb{V}^+)$ and $\mathrm{Pers}(\mathbb{V}^-)$, with intervals matched according to the following rules:*

- *Intervals of type $[k, k]$ are unmatched.*
- *Type $[b, k]$ is matched with type $[b, k-1]$ and vice versa, for $b \leq k-1$.*
- *Type $[k, d]$ is matched with type $[k+1, d]$ and vice versa, for $d \geq k+1$.*
- *Type $[b, d]$ is matched with type $[b, d]$, in all other cases.*

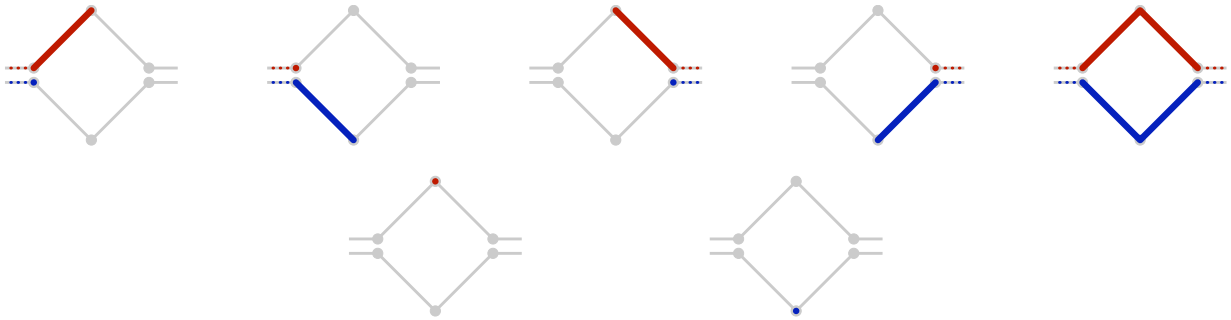


FIGURE 5. Interval matching between $\mathrm{Pers}(\mathbb{V}^+)$ and $\mathrm{Pers}(\mathbb{V}^-)$: (top row) the five cases where matching occurs; (bottom row) unmatched intervals $[k,k]$.

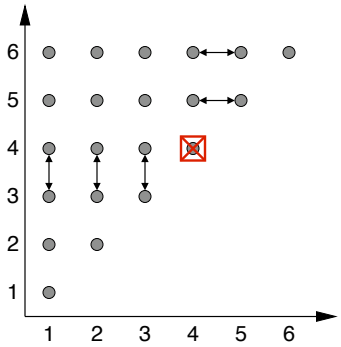


FIGURE 6. From $\mathrm{Pers}(\mathbb{V}^+)$ to $\mathrm{Pers}(\mathbb{V}^-)$, for $n = 6$, $k = 4$: points in the persistence plane move according to the arrows; the multiplicity of the point marked ⊠ changes unpredictably.

It follows that the restrictions $\mathrm{Pers}(\mathbb{V}^+)|_K$, $\mathrm{Pers}(\mathbb{V}^-)|_K$ *to the set* $K = \{1,\dots,n\} \setminus \{k\}$ *are equal.*

Figures 5 and 6 illustrate Theorem 5.6 in terms of barcodes and persistence diagrams, respectively.

Remark. The $\mathbb{I}(k,k)$ summands in $\mathrm{Pers}(\mathbb{V}^+)$ span the cokernel of D_2, whereas the $\mathbb{I}(k,k)$ summands in $\mathrm{Pers}(\mathbb{V}^-)$ span the kernel of D_1. The hypothesis of Theorem 5.6 does not bring about any relation between these spaces (which is why the $[k,k]$ intervals are unmatched). In Section 5.3, however, we consider a situation in which the $[k,k]$ intervals can be tracked.

We use the localization technique of Section 5.1 to prove Theorem 5.6. We begin with birth- and death-time indices.

Proposition 5.7. *Let* τ^+, τ^- *denote the zigzag types of* $\mathbb{V}^+, \mathbb{V}^-$ *respectively. If we write*

$$(b_1,\dots,b_{k-1}) = \mathrm{b}(\tau^+[1,k-1]) = \mathrm{b}(\tau^-[1,k-1])$$

for the birth-time index up to time $k-1$*, then*

$$\begin{aligned} \mathrm{b}(\tau^+[1,k+1]) &= (k+1, b_1, \dots, b_{k-1}, k). \\ \mathrm{b}(\tau^-[1,k+1]) &= (k, b_1, \dots, b_{k-1}, k+1), \end{aligned}$$

Similarly, if we write

$$(d_1, \ldots, d_{n-k}) = \mathrm{d}(\tau^+[k+1,n]) = \mathrm{d}(\tau^-[k+1,n])$$

for the death-time index from time $k+1$, then

$$\begin{aligned} \mathrm{d}(\tau^+[k-1,n]) &= (k-1, d_1, \ldots, d_{n-k}, k). \\ \mathrm{d}(\tau^-[k-1,n]) &= (k, d_1, \ldots, d_{n-k}, k-1), \end{aligned}$$

Proof. This is immediate from the recursive definition of birth-time index. If we write $\tau_0 = \tau^+[1,k-1] = \tau^-[1,k-1]$ then $\tau^+[1,k+1] = \tau_0 fg$ and $\tau^-[1,k+1] = \tau_0 gf$. The death-time index is treated similarly. □

Here is the crux of the matter:

Lemma 5.8. *In the situation of Theorem 5.6, the following filtrations are equal:*

$$\begin{aligned} \mathrm{R}(\mathbb{V}^+[1,k+1]) &= \mathrm{R}(\mathbb{V}^-[1,k+1]) \\ \mathrm{L}(\mathbb{V}^+[k-1,n]) &= \mathrm{L}(\mathbb{V}^-[k-1,n]) \end{aligned}$$

Proof. Write $(R_0, R_1, \ldots, R_{k-1}) = \mathrm{R}(\mathbb{V}^+[1,k-1]) = \mathrm{R}(\mathbb{V}^-[1,k-1])$. By the recursive formula (Definition 3.1),

$$\mathrm{R}(\mathbb{V}^+[1,k+1]) = (0, g_k^{-1} f_{k-1}(R_0), \ldots, g_k^{-1} f_{k-1}(R_{k-1}), V_{k+1})$$

and

$$\mathrm{R}(\mathbb{V}^-[1,k+1]) = (0, f_k g_{k-1}^{-1}(R_0), \ldots, f_k g_{k-1}^{-1}(R_{k-1}), V_{k+1}).$$

Thus we can prove the first statement of the lemma by showing that

$$f_k g_{k-1}^{-1}(R) = g_k^{-1} f_{k-1}(R)$$

for any subspace $R \leq V_{k-1}$. We use first-order logic. Let $x \in V_{k+1}$. We have the following chain of equivalent statements.

$$\begin{aligned} & x \in f_k g_{k-1}^{-1}(R) \\ \Leftrightarrow\ & (\exists z \in R)\,(\exists y \in U_k)\,((g_{k-1}(y) = z)\,\&\,(f_k(y) = x)) \\ \Leftrightarrow\ & (\exists z \in R)\,(\exists y \in U_k)\,(D_1(y) = z \oplus x) \\ \Leftrightarrow\ & (\exists z \in R)\,(z \oplus x \in \mathrm{Im}(D_1)) \end{aligned}$$

On the other hand:

$$\begin{aligned} & x \in g_k^{-1} f_{k-1}(R) \\ \Leftrightarrow\ & (\exists z \in R)\,(f_{k-1}(z) = g_k(x)) \\ \Leftrightarrow\ & (\exists z \in R)\,(z \oplus x \in \mathrm{Ker}(D_2)) \end{aligned}$$

Since $\mathrm{Im}(D_1) = \mathrm{Ker}(D_2)$ by hypothesis, it follows that $f_k g_{k-1}^{-1}(R) = g_k^{-1} f_{k-1}(R)$.

This proves the first equality. The second equality follows symmetrically. □

Proof of Theorem 5.6. We adopt the notation of Section 5.1, and consider the right- and left-filtrations at V_{k+1}, for both $\mathbb{V}^+$ and $\mathbb{V}^-$. Since $\mathbb{V}^+[k+1,n] = \mathbb{V}^-[k+1,n]$ we have

$$\mathcal{L}_{k+1}^+ = \mathcal{L}_{k+1}^- \qquad \text{and} \qquad \mathrm{d}_{k+1}^+ = \mathrm{d}_{k+1}^-,$$

and by Lemma 5.8 we have

$$\mathcal{R}_{k+1}^+ = \mathcal{R}_{k+1}^-.$$

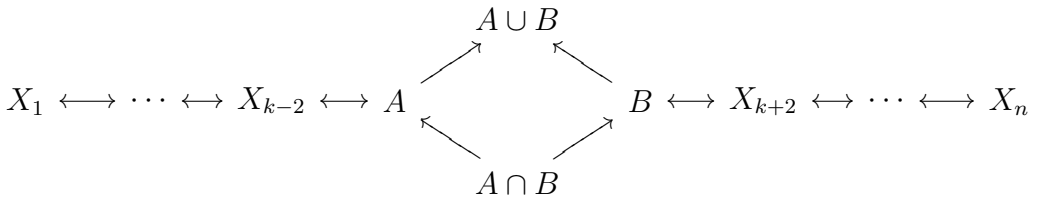
Finally, b^+_{k+1} agrees with b^-_{k+1} except that $k, k+1$ are interchanged, according to Proposition 5.7. Thus, when we use Theorem 5.3 to calculate the multiplicity of $[b,d]$ for $b \leq k+1 \leq d$, there is perfect agreement between $\mathbb{V}^+$ and $\mathbb{V}^-$ except that we must interchange $k, k+1$ when they occur as birth-times.

A symmetrical argument can be made, localizing at V_{k-1}. When we compute the multiplicity of $[b,d]$ for $b \leq k-1 \leq d$, there is perfect agreement between $\mathbb{V}^+$ and $\mathbb{V}^-$ except that we must interchange $k, k-1$ when they occur as death-times.

We have covered all cases of the theorem except for intervals which meet neither $k-1$ nor $k+1$. Intervals contained in $[1, k-2]$ are automatically the same for $\mathbb{V}^+$ and $\mathbb{V}^-$ because they can be computed by restricting to $\mathbb{V}^+[1,k-1]$ and $\mathbb{V}^-[1,k-1]$, which are equal. Similarly, intervals contained in $[k+2, n]$ are the same for $\mathbb{V}^+$ and $\mathbb{V}^-$, by restricting to $\mathbb{V}^+[k+1,n] = \mathbb{V}^-[k+1,n]$.

Finally, consider intervals $[k,k]$. Nothing can be said about those. □

5.3. **The Strong Diamond Principle.** The Diamond Principle can usefully be applied to the following diagram of topological spaces and continuous maps. The four maps in the central diamond are inclusion maps, and the remaining maps $\leftrightarrow$ are arbitrary.



Let $\mathbb{X}^+, \mathbb{X}^-$ denote the upper and lower zigzag diagrams contained in this picture; so $\mathbb{X}^+$ passes through $A \cup B$ and $\mathbb{X}^-$, passes through $A \cap B$.

Theorem 5.9 (The Strong Diamond Principle)**.** *Given $\mathbb{X}^+$ and $\mathbb{X}^-$ as above, there is a (complete) bijection between the multisets $\mathrm{Pers}(H_*(\mathbb{X}^+))$ and $\mathrm{Pers}(H_*(\mathbb{X}^-))$. Intervals are matched according to the following rules:*

- *$[k,k] \in \mathrm{Pers}(H_{\ell+1}(\mathbb{X}^+))$ is matched with $[k,k] \in \mathrm{Pers}(H_\ell(\mathbb{X}^-))$.*

In the remaining cases, the matching preserves homological dimension:

- *Type $[b,k]$ is matched with type $[b,k-1]$ and vice versa, for $b \leq k-1$.*
- *Type $[k,d]$ is matched with type $[k+1,d]$ and vice versa, for $d \geq k+1$.*
- *Type $[b,d]$ is matched with type $[b,d]$, in all other cases.*

Proof. For any ℓ, apply the homology functor H_ℓ to the diagram. The central diamond

$$\begin{array}{ccc} H_\ell(A) & \longrightarrow & H_\ell(A \cup B) \\ \uparrow & & \uparrow \\ H_\ell(A \cap B) & \longrightarrow & H_\ell(B) \end{array}$$

is exact by virtue of the Mayer–Vietoris theorem, according to which

$$\ldots \longrightarrow H_\ell(A \cap B) \xrightarrow{D_1} H_\ell(A) \oplus H_\ell(B) \xrightarrow{D_2} H_\ell(A \cup B) \longrightarrow \ldots$$

is an exact sequence. The Diamond Principle therefore applies to $H_\ell(\mathbb{X}^+)$ and $H_\ell(\mathbb{X}^-)$, and we have a partial bijection which accounts for all intervals except those of type $[k,k]$.

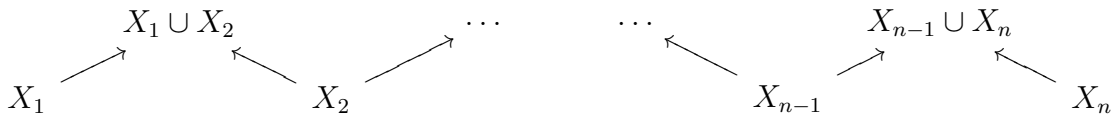
Now consider the connecting homomorphism in the same Mayer–Vietoris sequence:

$$\ldots \xrightarrow{D_2} H_{\ell+1}(A \cup B) \xrightarrow{\partial} H_\ell(A \cap B) \xrightarrow{D_1} \ldots$$

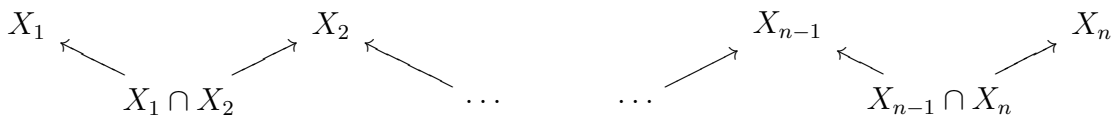
By exactness, ∂ induces an isomorphism between the cokernel of D_2 and the kernel of D_1. But the $[k,k]$ summands of $\mathrm{Pers}(H_{\ell+1}(\mathbb{X}^+))$ precisely span $\mathrm{Coker}(D_2)$, whereas the $[k,k]$ summands of $\mathrm{Pers}(H_\ell(\mathbb{X}^-))$ span $\mathrm{Ker}(D_1)$. This establishes the claimed bijection between the $[k,k]$ intervals. □

Example 5.10. Let $\mathbb{X} = (X_1, \ldots, X_n)$ be a sequence of simplicial complexes defined on a common vertex set. Suppose these have arisen in some context where each transition X_i to X_{i+1} is regarded as being a 'small' change. There are two natural zigzag sequences linking the X_i.

The union zigzag, $\mathbb{X}_\cup$:



The intersection zigzag, $\mathbb{X}_\cap$:



We can think of these as being indexed by the half-integers $\{1, 1\frac{1}{2}, 2, 2\frac{1}{2}, \ldots, n\}$.

We can apply the Strong Diamond Principle $n-1$ times to derive the following relationship between the zigzag persistence of the two sequences $\mathrm{Pers}(H_\ell(\mathbb{X}_\cap))$ and $\mathrm{Pers}(H_\ell(\mathbb{X}_\cup))$. Restricting to the integer indices, there is a coarse equality:

$$\mathrm{Pers}(H_\ell(\mathbb{X}_\cup))|_{\{1,\ldots,n\}} = \mathrm{Pers}(H_\ell(\mathbb{X}_\cap))|_{\{1,\ldots,n\}}$$

More finely, there is a partial bijection between $\mathrm{Pers}(H_\ell(\mathbb{X}_\cup))$ and $\mathrm{Pers}(H_\ell(\mathbb{X}_\cap))$. Intervals $[k\frac{1}{2}, k\frac{1}{2}]$ shift homological dimension by $+1$ (from the intersection sequence to the union sequence). Otherwise $[b,d] \leftrightarrow [b',d']$ where $\{b,b'\}$ is an unordered pair of the form $\{k\frac{1}{2}, k+1\}$ and $\{d,d'\}$ is an unordered pair of the form $\{k, k\frac{1}{2}\}$; dimension is preserved. Figure 7 illustrates the complete correspondence as a transformation of the persistence diagram, for $n = 5$.

Concluding Remarks

We have presented the foundations of a theory of zigzag persistence which, we believe, considerably extends and enriches the well known and highly successful theory of persistent homology. Zigzag persistence originates in the work of Gabriel and others in the theory of quiver representations. One of our goals has been to bridge the gap between the quiver literature (which is read largely by algebraists) and the current language of applied and computational topology. To this end, we have presented an algorithmic form of Gabriel's structure theorem for A_n quivers, and have indicated the first steps towards integrating these ideas into tools for applied topology.

There are several ways in which this work is incomplete. The most significant omission is an algorithm for computing zigzag persistence in a homological setting (as distinct from

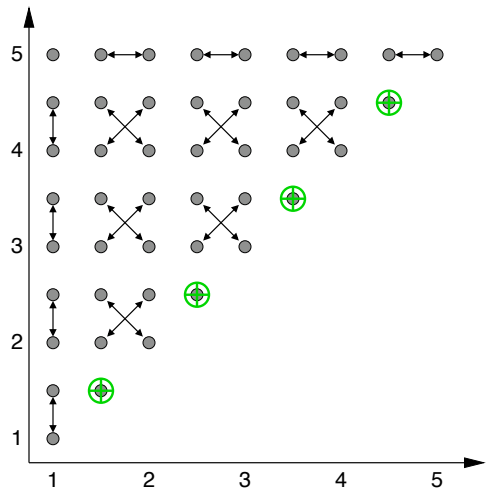


FIGURE 7. From $\mathrm{Pers}(H_*(\mathbb{X}_\cap))$ to $\mathrm{Pers}(H_*(\mathbb{X}_\cup))$, for $n = 5$: points in the persistence plane move according to the arrows; points marked $\oplus$ stay fixed and increase homological dimension by 1.

the somewhat sanitised vector space algorithm described in Section 4.3). We address this gap in a forthcoming paper with Dmitriy Morozov [2], where we present an algorithm for computing the zigzag persistence intervals of a 1-parameter family of simplicial complexes on a fixed vertex set.

We have made no effort in this paper to flesh out the applications suggested in Section 1. There is often a substantial gap between the concrete world of point-cloud data sets and the ideal world of simplicial complexes and topological spaces. We intend to develop some of these applications in future work. Meanwhile, we have given priority to establishing the theoretical language and tools. The Diamond Principle is particularly powerful. In the manuscript with Morozov [2], we show that the Diamond Principle can be used to establish isomorphisms between several different classes of persistence invariants of a space with a real-valued (e.g. Morse) function. In particular, we use zigzag persistence to resolve an open conjecture concerning extended persistence [5]. This supports our prejudice that zigzag persistence provides the appropriate level of generality and power for understanding the heuristic concept of persistence in its many manifestations.

Acknowledgements. The authors wish to thank Greg Kuperberg, Konstantin Mischaikow and Dmitriy Morozov for helpful conversations and M. Khovanov for helpful correspondence. The authors gratefully acknowledge support from DARPA, in the form of grants HR0011-05-1-0007 and HR0011-07-1-0002. The second author wishes to thank Pomona College and Stanford University for, respectively, granting and hosting his sabbatical during late 2008.

REFERENCES

[1] M. F. Atiyah. On the Krull–Schmidt theorem with application to sheaves. *Bulletin de la S. M. F.*, 84:307–317, 1956.

[2] Gunnar Carlsson, Vin de Silva, and Dmitriy Morozov. Zigzag persistent homology and real-valued functions. Manuscript, December 2008.

[3] Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, January 2008.

[4] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

[5] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Extending persistence using Poincaré and Lefschetz duality. *Foundations of Computational Mathematics*, 2008.

[6] Vin de Silva and Gunnar Carlsson. Topological estimation using witness complexes. In M. Alexa and S. Rusinkiewicz, editors, *Eurographics Symposium on Point-Based Graphics*, ETH, Zürich, Switzerland, 2004.

[7] Harm Derksen and Jerzy Weyman. Quiver representations. *Notices of the American Mathematical Society*, 52(2):200–206, February 2005.

[8] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002.

[9] Herbert Edelsbrunner and Ernst P. Mücke. Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13(1):43–72, 1994.

[10] P. Gabriel. Unzerlegbare darstellungen I. *Manuscripta Mathematica*, 6:71–103, 1972.

[11] V. G. Kac. Infinite root systems, representations of graphs and invariant theory. *Inventiones Mathematicae*, 56(1):57–92, 1980.

[12] Serge Lang. *Algebra*. Graduate Texts in Mathematics. Springer-Verlag, 3rd edition, 2005.

[13] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.