Persistent Homology: An Introduction and a New Text Representation for Natural Language Processing

Xiaojin Zhu University of Wisconsin-Madison, USA.
jerryzhu@cs.wisc.edu

Group Theory

Definition
A group $\langle G, * \rangle$ is a set G with a binary operation $*$ such that
1. *(associative)* $a*(b*c) = (a*b)*c$ *for all* $a, b, c \in G$.
2. *(identity)* $\exists e \in G$ *so that* $e*a = a*e = a$ *for all* $a \in G$.
3. *(inverse)* $\forall a \in G, \exists a' \in G$ *where* $a*a' = a'*a = e$.

- Examples: $\langle \mathbb{Z}, + \rangle$, $\langle \mathbb{R}, + \rangle$, $\langle \mathbb{R}_+, \times \rangle$, $\langle \mathbb{R}\backslash\{0\}, \times \rangle$.
- $\mathbb{Z}_2$

$+_2$	0	1
0	0	1
1	1	0

- All our groups G are abelian: $\forall a, b \in G, a*b = b*a$.

Definition
A subset $H \subseteq G$ of a group $\langle G, * \rangle$ is a subgroup of G if $\langle H, * \rangle$ is itself a group.

Definition
Given a subgroup H of an abelian group G, for any $a \in G$, the set $a*H = \{a*h \mid h \in H\}$ is the coset of H represented by a.

Definition
A map $\phi : G \mapsto G'$ is a homomorphism if $\phi(a*b) = \phi(a) \star \phi(b)$ for $\forall a, b \in G$.

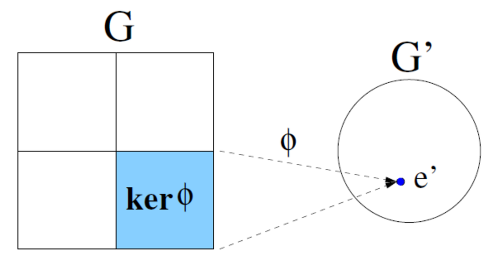
- $\langle \mathbb{R}_+, \times \rangle$ to $\langle \mathbb{Z}_2, +_2 \rangle$: trivial homomorphism $\phi(a) = 0, \forall a \in \mathbb{R}_+$
- negation in natural language: G_N

$*$	⊔	not
⊔	⊔	not
not	not	⊔

homomorphism (isomorphism) from G_N to $\mathbb{Z}_2$: $\phi(\sqcup) = 0, \phi(\text{not}) = 1$.

Definition
The kernel of a homomorphism $\phi : G \mapsto G'$ is $\mathbf{ker}\phi = \{a \in G \mid \phi(a) = e'\}$.

- In the $\phi : G_N \mapsto \mathbb{Z}_2$ example, $\mathbf{ker}\phi = \{\sqcup\}$.
- Another example: $\phi : \langle \mathbb{R}\backslash\{0\}, \times \rangle \mapsto G_N$ by $\phi(a) = \sqcup$ if $a > 0$ and "not" if $a < 0$. $\mathbf{ker}\phi = \mathbb{R}_+$
- For any homomorphism $\phi : G \mapsto G'$, $\mathbf{ker}\phi$ is a subgroup of G.
- Cosets $a * \mathbf{ker}\phi$ partition G

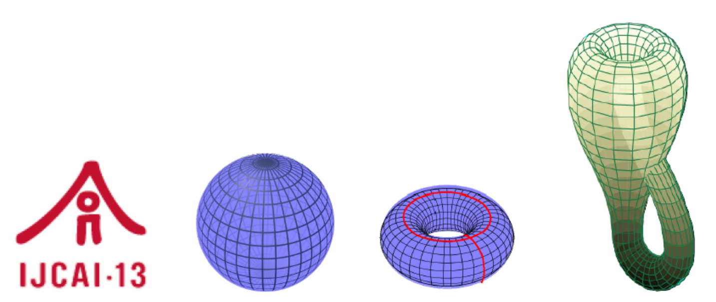


- Let $\langle H, * \rangle$ be a subgroup of an abelian group $\langle G, * \rangle$.
- A new operation on the cosets of H: $(a*H) \star (b*H) = (a*b)*H, \forall a, b \in G$.

Definition
The cosets $\{a*H \mid a \in G\}$ under the operation $\star$ form a group, called the quotient group G/H.

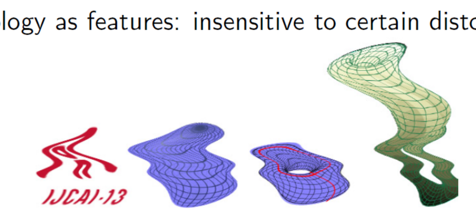
- Example: $G = \mathbb{R}\backslash\{0\}$ and $\mathbf{ker}\phi = \mathbb{R}_+$, two cosets: $\mathbb{R}_+$ and $\mathbb{R}_-$.
- The quotient group $(\mathbb{R}\backslash\{0\})/\mathbb{R}_+$ has the two coset elements.
- $R_- \star \mathbb{R}_- = (-1 \times R_+) \star (-1 \times R_+) = (-1 \times -1) \times R_+ = 1 \times R_+ = R_+$.
- This quotient group $(\mathbb{R}\backslash\{0\})/\mathbb{R}_+$ is isomorphic to $\mathbb{Z}_2$.

Homology

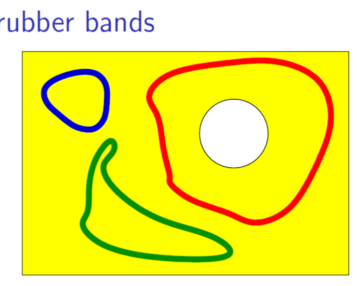


Betti_0 (clusters)	11	1	1	1
Betti_1 (holes)	2	0	2	2
Betti_2 (voids)	0	1	1	1

Homology as features: insensitive to certain distortions



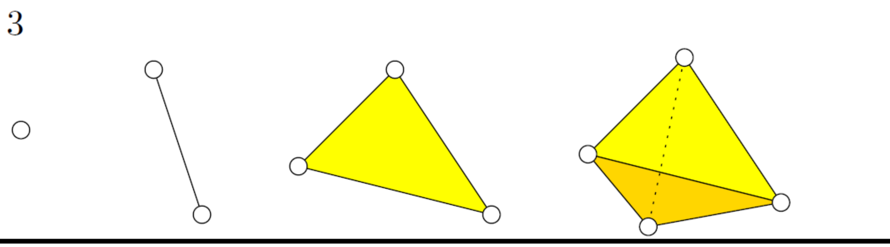
The group of rubber bands



quotient group "all rubber bands"/"uninteresting rubber bands"

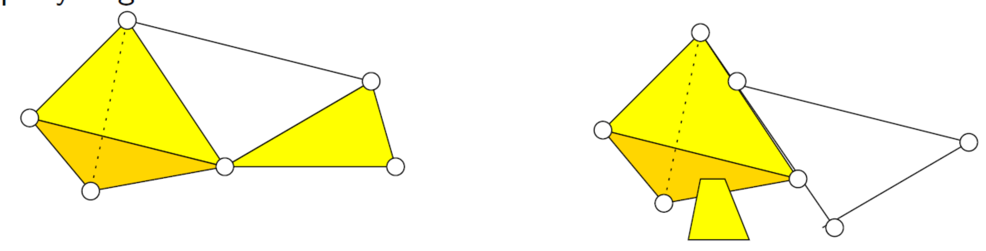
Definition
A p-simplex σ is the convex hull of $p+1$ affinely independent points $x_0, x_1, \ldots, x_p \in \mathbb{R}^d$. We denote $\sigma = \text{conv}\{x_0, \ldots, x_p\}$. The dimension of σ is p.

- $p = 0, 1, 2, 3$



Definition
A simplicial complex K is a finite collection of simplices such that $\sigma \in K$ and τ being a face of σ implies $\tau \in K$, and $\sigma, \sigma' \in K$ implies $\sigma \cap \sigma'$ is either empty or a face of both σ and σ'.

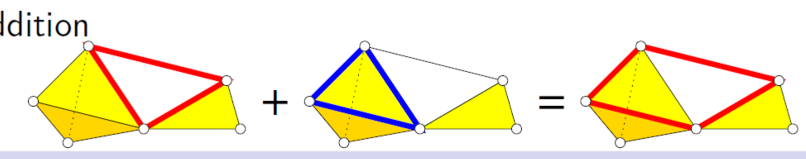
- Properly aligned



Definition
A p-chain is a subset of p-simplices in a simplicial complex K.

Definition
The set of p-chains of a simplicial complex K form a p-chain group C_p.

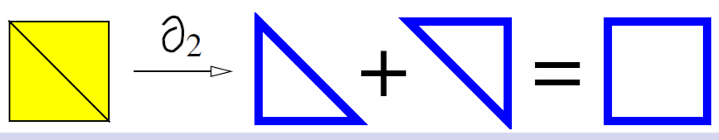
- Mod-2 addition



Definition
The boundary of a p-simplex is the set of $(p-1)$-simplices faces.

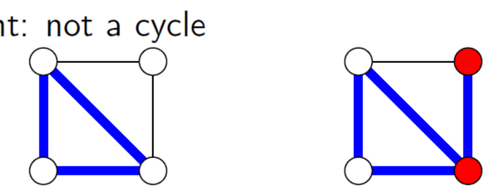
Definition
The boundary of a p-chain is the Mod-2 sum of the boundaries of its simplices. Taking the boundary is a group homomorphism ∂_p from C_p to C_{p-1}.

- Faces shared by an even number of p-simplices in the chain will cancel out:

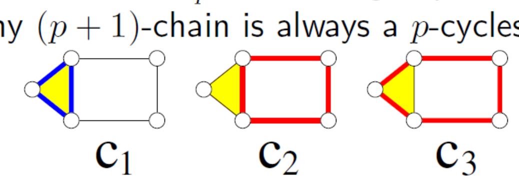


Definition
A p-cycle c is a p-chain with empty boundary: $\partial_p c = 0$ (the identity in C_{p-1}).

- Discrete p-dimensional "rubber bands"
- Left: a 1-cycle; Right: not a cycle

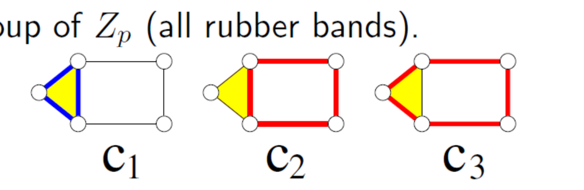


- Z_p= all p-cycles (all rubber bands)
- $\partial_p Z_p = 0$: Z_p is the kernel $\mathbf{ker}\partial_p$ and a subgroup of C_p.
- The boundary of any $(p+1)$-chain is always a p-cycles



Definition
A p-boundary-cycle is a p-cycle that is also the boundary of some $(p+1)$-chain.

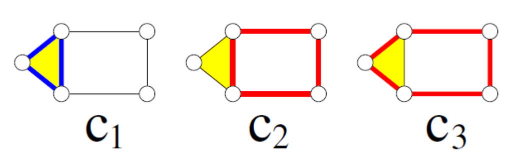
- Let $B_p = \partial_{p+1} C_{p+1}$, the p-boundary-cycles.
- B_p are the uninteresting rubber bands (e.g., $B_1 = \{0, c_1\}$)
- B_p is a subgroup of Z_p (all rubber bands).



- c_2 and c_3 in Z_1 but not in B_1
- We can drag rubber band c_2 over the yellow triangle to make c_3
- Formally, $c_3 = c_2 + c_1$.
- c_2 and c_3 are equivalent in the hole they surround.
- The equivalence class: $c + B_p$

Definition
The p-th homology group is the quotient group $H_p = Z_p / B_p$.

- Example:



- All the 1-cycles : $Z_1 = \{0, c_1, c_2, c_3\}$.
- The uninteresting 1-cycles: $B_1 = \{0, c_1\}$, a subgroup of Z_1.
- The interesting 1-cycles: $c_2 + B_1 = c_3 + B_1 = \{c2, c3\}$
- The homology group $H_1 = Z_1 / B_1$ isomorphic to $\mathbb{Z}_2$

Definition
The p-th Betti number is the rank of the homology group: $\beta_p = \text{rank}(H_p)$.

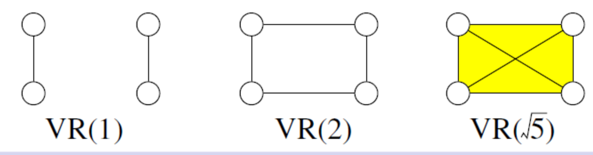
- In our example, $\beta_1 = \text{rank}(\mathbb{Z}_2) = 1$ (one 1st-order hole)
- β_p is the number of independent p-th holes.
- A tetrahedron has $\beta_0 = 1$ (connected), $\beta_1 = \beta_2 = 0$ (no holes or voids)
- A hollow tetrahedron has $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$
- Removing the four triangle faces, the edge skeleton has $\beta_0 = 1$, $\beta_1 = 3$ (one is the sum of the other three), $\beta_2 = 0$ (no more void).
- Removing the edges, $\beta_0 = 4$ (4 vertices) and $\beta_1 = \beta_2 = 0$.

From data to simplicial complex

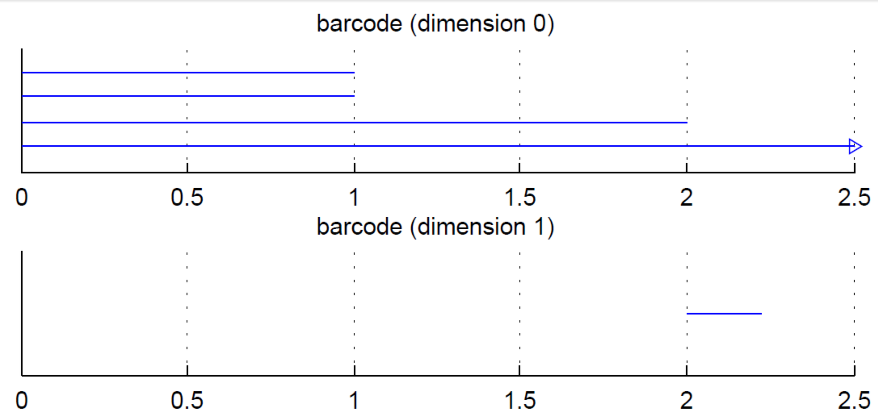
- Given data $x_1, \ldots, x_n \in \mathbb{R}^d$.
- If any subset of $p+1$ points are within diameter ϵ, we add a p-simplex generated by those points.

Definition
A Vietoris-Rips complex of diameter ϵ is the simplicial complex $VR(\epsilon) = \{\sigma \mid \text{diam}(\sigma) \leq \epsilon\}$.

- Example

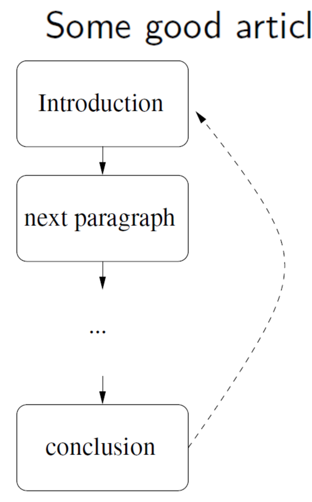


Definition
An increasing sequence of ϵ produces a filtration, i.e., a sequence of increasing simplicial complexes $VR(\epsilon_1) \subseteq VR(\epsilon_2) \subseteq \ldots$, with the property that a simplex enters the sequence no earlier than all its faces.



Applications to natural language processing

Some good articles "tie back." Capture such loops with homology.



Example: Itsy bitsy spider

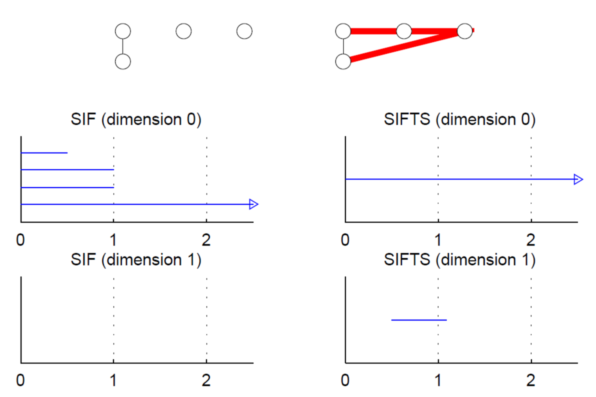
The Itsy Bitsy Spider climbed up the water spout
Down came the rain and washed the spider out
Out came the sun and dried up all the rain
And the Itsy Bitsy Spider climbed up the spout again

Similarity Filtration (SIF)

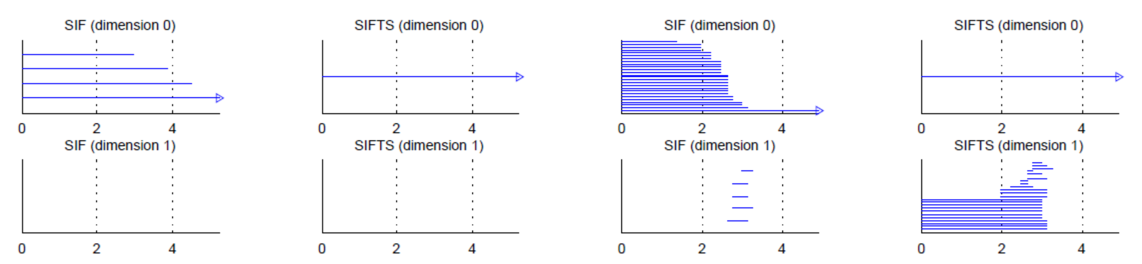
$D_{max} = \max D(x_i, x_j), \forall i, j = 1 \ldots n$
FOR $m = 0, 1, \ldots M$
  Add $VR(\frac{m}{M} D_{max})$ to the filtration
END
Compute persistent homology on the filtration

Similarity Filtration with Time Skeleton (SIFTS)

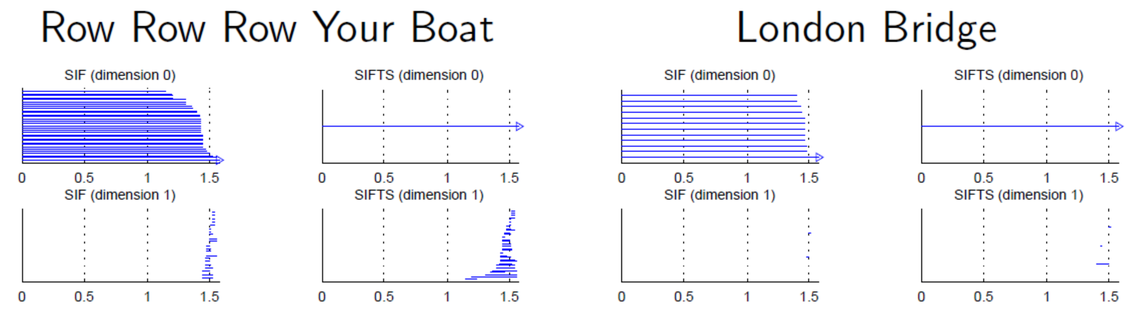
$D(x_i, x_{i+1}) = 0$ for $i = 1, \ldots, n-1$
$D_{max} = \max D(x_i, x_j), \forall i, j = 1 \ldots n$
FOR $m = 0, 1, \ldots M$
  Add $VR(\frac{m}{M} D_{max})$ to the filtration
END
Compute persistent homology on the filtration



On Nursery Rhymes and Other Stories



Row Row Row Your Boat   London Bridge



Little Red-Cap   Alice in Wonderland

- London Bridge: "My fair Lady" repeats 12 times.
- Little Red-Cap: "The better to see you with, my dear" and "The better to eat you with!"

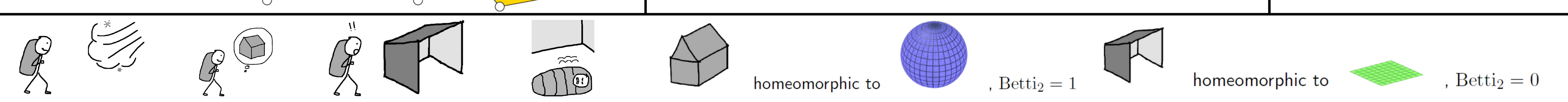
On Child and Adolescent Writing

LUCY corpus: children (ages 9–12, 150 essays), undergraduates

- holes?: what percentage of articles have H_1 holes
- $|H_1|$: number of holes in the article
- ϵ^*: the smallest ϵ when the first hole in H_1 forms.

	child	adolescent	adol. trunc.
holes?	87%	100%*	98%*
$\lvert H_1 \rvert$	3.0 (±0.2)	17.6 (±0.9)*	3.9 (±0.2)*
ϵ^*	1.35 (±.02)	1.27 (±.02)*	1.38 (±.01)

*: statistically significantly different from "child"



Acknowledgment: NSF IIS-0953219, IIS-1216758, IIS-1148012, IIS-0916038