The Universal ℓ^p-Metric on Merge Trees

Robert Cardona (§), Justin Curry, Tung Lam (§), Mike Lesnick
SUNY Albany

CG Week 2022

June 10th, 2022

Trees for Biology

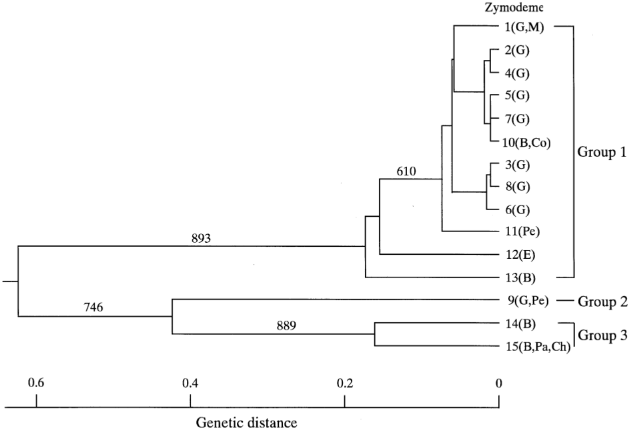
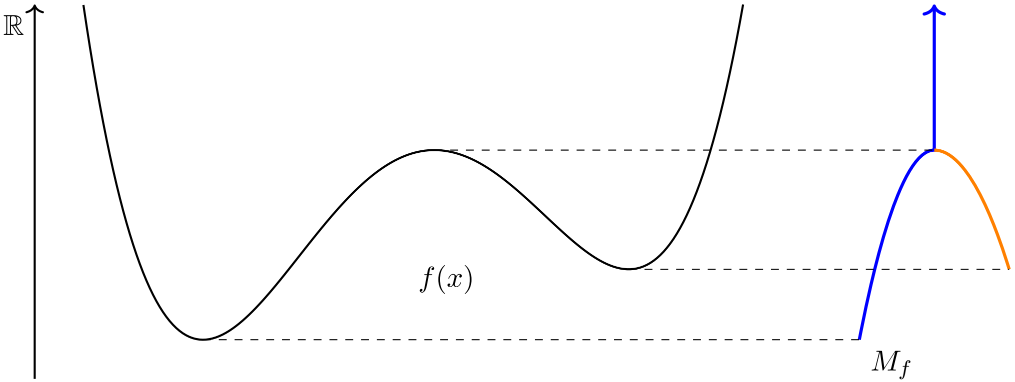
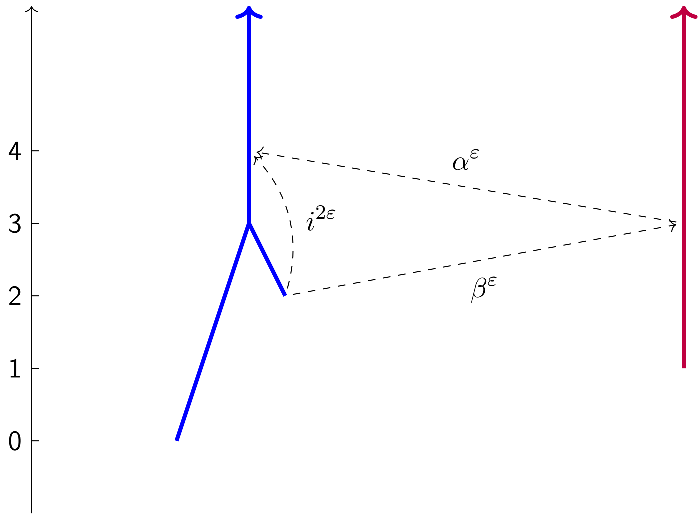


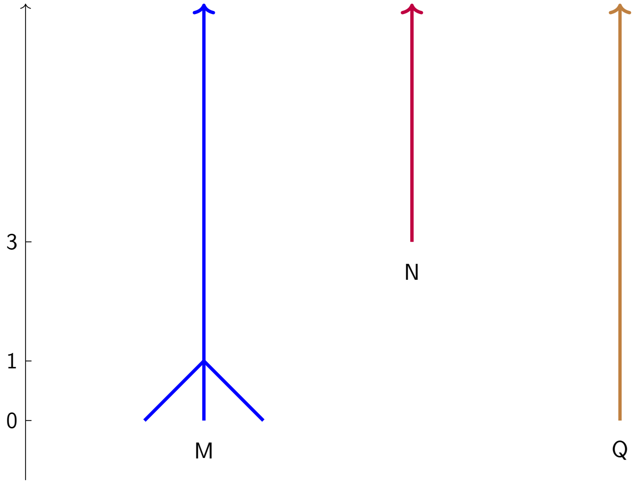
Fig. 1. Dendrogram drawn based on the matrix of genetic distances among 15 zymodemes of *Trypanosoma cruzi* using UPGMA. The figures on branches indicate the number of times that the branch was observed in 1000 bootstraps. Bootstrap values below 600 are not given. Abbreviations: B, Brazil; Ch, Chile; Co, Colombia; E, Ecuador; G, Guatemala; M, Mexico; Pa, Paraguay; Pe, Peru.

Trees for Scalar Data



Morozov, Beketayev, and Weber introduced *the interleaving distance* d_I on merge trees [4].

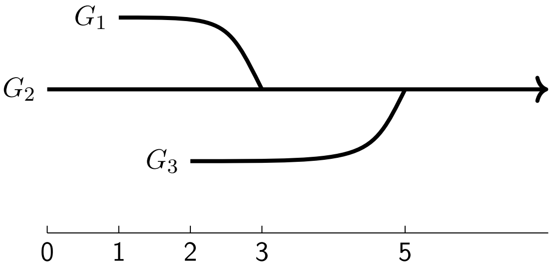




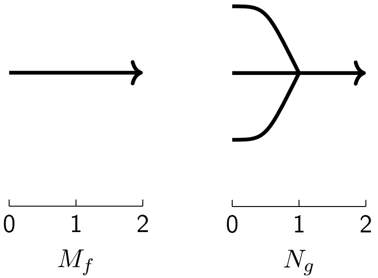
N.B. $d_I(M,N) = d_I(Q,N) = 3$, but intuitively Q is “closer” to N.

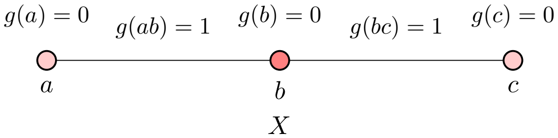
Cophenetic vectors

- Our vector summaries are subtly different from cophenetic vectors, i.e. the LCA matrix [2, 5, 3], as the length of our vectors is $2n - 1$ versus ${}_nC_2 = O(n^2)$.
- In particular, the p-cophenetic distance is not Lipschitz stable for $p \neq \infty$.



$$
\begin{array}{cc}
 & \begin{array}{ccc} G_1 & G_2 & G_3 \end{array} \\
\begin{array}{c} G_1 \\ G_2 \\ G_3 \end{array} & \begin{pmatrix} 1 & 3 & 5 \\ * & 0 & 5 \\ * & * & 2 \end{pmatrix}
\end{array}
$$



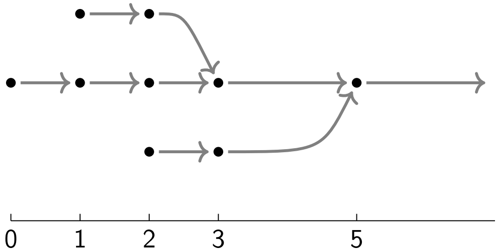


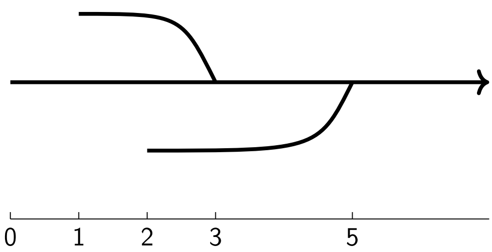
Here $\|f - g\|_1 = 2$, while ℓ^1-cophenetic distance is 3.
Instead, we mimic a construction by Bjerkevik and Lesnick [1].

Merge Trees as Persistent Sets

A **merge tree** is a functor $M\colon \mathbb{R} \to \mathbf{Set}$ that is

- **constructible**, i.e. $\exists\, \tau := \{s_0 < s_1 < \cdots < s_n\} \subset \mathbb{R}$, such that
  - (i) $M(s) = \emptyset$ for all $s < s_0$, and
  - (ii) $M(s \leq t)$ is an isomorphism whenever $s, t \in [s_i, s_{i+1})$, and also for $s, t \in [s_n, \infty)$.
- and where $|M(t)| = 1$ for t sufficiently large.



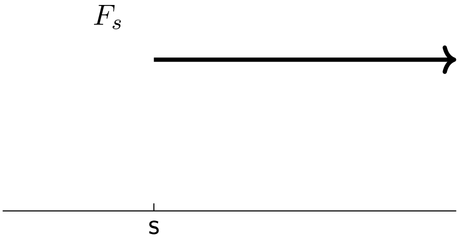


Building Blocks for Merge Trees

A **strand** is a merge tree $F_s : \mathbb{R} \to \mathbf{Set}$, for $s \in \mathbb{R}$, defined by

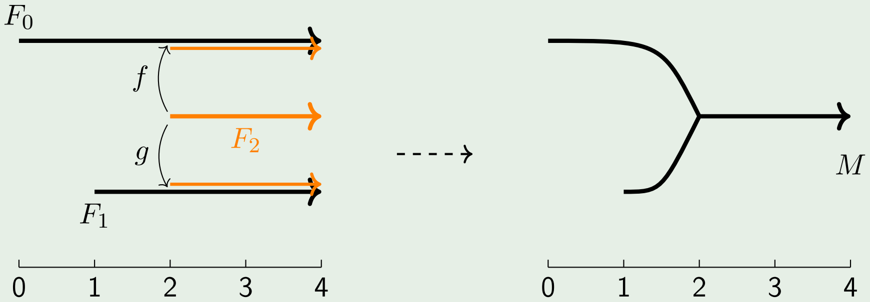
$$F_s(t) := \begin{cases} \emptyset & \text{if } t < s, \\ \{*\} & \text{if } t \geq s, \end{cases}$$

with the structure maps all inclusions. We call s birth time of the branch F_s



Example Presentation of a Merge Tree

Any merge tree M can be constructed via gluing strands pairwise together.



$$F_2 \overset{f}{\underset{g}{\rightrightarrows}} F_0 \sqcup F_1 \dashrightarrow M,$$

Presentation of a Merge Tree

A **presentation** of a merge tree M consists of

- generators G_i's and relations R_j's that are strands;
- together with pairs of underlying merge functions $f_j, g_j : R_j \to \sqcup_i G_i$ that choose explicit strands for merging.

$$\bigsqcup_{j=1}^{l} R_j \overset{f}{\underset{g}{\rightrightarrows}} \bigsqcup_{i=1}^{k} G_i \dashrightarrow M,$$

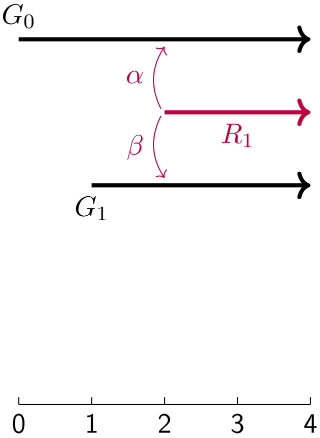
Presentation Matrix and Label Vector

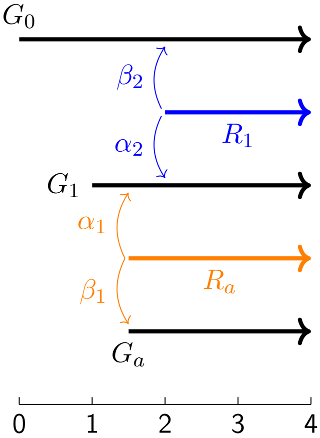
To a presentation P_M we have a **presentation matrix** where

- the i-th row corresponds to the i-th generator G_i, labelled by the birth time of G_i; and
- the j-th column corresponds to the j-th relation R_j, labelled by the birth time of R_j.
- The (i, j)-entry is 1 if G_i is in the image of R_j (under f or g) and 0 otherwise.

The **label vector** $L(P_M)$ of a $k \times l$ presentation matrix is the $(k + l)$-vector where

- the first k entries are the row labels, i.e. heights of leaf nodes, and
- the last l entries are column labels, i.e. the heights of internal nodes.





$$P_M : \begin{array}{cc} & 2 \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} 1 \\ 1 \end{pmatrix} \end{array} \quad L(P_M) = [0, 1; 2]$$

versus

$$P'_M : \begin{array}{cc} & \begin{array}{cc} 2 & a \end{array} \\ \begin{array}{c} 0 \\ 1 \\ a \end{array} & \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \end{array} \quad L(P'_M) = [0, 1, a; 2, a]$$

Compatible Presentations

Two presentations P_M, P_N are **compatible** if their presentation matrices have the same underlying matrix, after forgetting row and column labels.

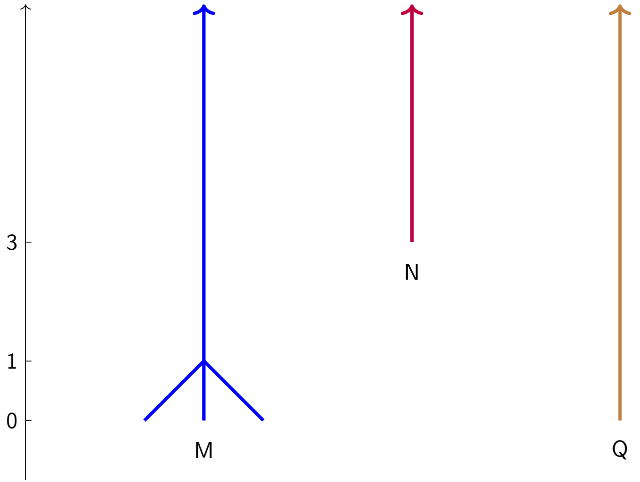
Lemma

Every pair of merge trees M and N, have compatible presentations P_M and P_N.

Definition

Given $p \in [1, \infty]$, the ***p*-presentation semi-distance** between merge trees M and N is

$$\hat{d}_I^p(M, N) = \inf\{\|L(P_M) - L(P_N)\|_p : P_M \text{ and } P_N \text{ are compatible.}\}$$



$$
\begin{aligned}
\hat{d}_I^1(N,Q) &= \|[3]-[0]\|_1 = 3 \\
\hat{d}_I^1(M,Q) &= \|[0,0,0;1,1]-[0,0,0;0,0]\|_1 = 2 \\
\hat{d}_I^1(M,N) &= \|[0,0,0;1,1]-[3,3,3;3,3]\|_1 = 13
\end{aligned}
$$

p-Presentation distance

We see $\hat{d}_I^p$ does not satisfy the triangle inequality. Fortunately there is a universal fix.

Definition

The ***p*-presentation distance** between M and N is

$$d_I^p(M,N) := \inf \sum_{i=0}^{n-1} \hat{d}_I^p(Q_i, Q_{i+1}),$$

where we infimize over all finite sequences of merge trees $M = Q_0, \ldots, Q_n = N$.

Theorem (Cardona, C., Lam, Lesnick '21)

- $d_I^\infty = d_I$, *i.e., the* ∞*-presentation distance equals the interleaving distance.*
- *For* $p \in [1,\infty]$, d_I^p *is a pseudometric.*

Wasserstein Stability

We extend a lower bound on the interleaving distance due to Morozov et al.

Proposition (CCLL'21)

For $p \in [1, \infty]$ *and merge trees* M, N*:*

$$d_{\mathcal{W}}^{p}(\mathcal{B}(M), \mathcal{B}(N)) \leq d_{I}^{p}(M, N).$$

Here $d_{\mathcal{W}}^{p}$ the denotes p-Wasserstein distance between barcodes.

Monotone Cellular Functions

Let X be a finite CW-complex.

- We say $f : X \to \mathbb{R}$ is **monotone** if for any face τ of σ, one has $f(\tau) \leq f(\sigma)$.
- We can define $\|f\|_p$ by identifying f with an element of $\mathbb{R}^{|\mathsf{Cell}(X)|}$.

Theorem (Skraba and Turner, 20')

Let $f, g : X \to \mathbb{R}$ be monotone cellular functions. Then

$$d_{\mathcal{W}}^p(\mathcal{B}(f), \mathcal{B}(g)) \leq \|f - g\|_p.$$

Here $\mathcal{B}(f)$ is the persistence barcode for the sublevel set filtration of f.

ℓ^p-stability & Universality

We provide an analogue of the interleaving stability for p-presentation distances.

Theorem (ℓ^p-Stability, CCLL'21)

For any monotone cellular functions $f, g : X \to \mathbb{R}$.

$$d_I^p(M_f, M_g) \leq ||f - g||_p,$$

Here $M_f = \pi_0 \circ S^{\uparrow}(f)$.

Theorem (Universality, CCLL'21)

If d is any distance on merge trees satisfying the above stability property, then $d \leq d_I^p$.

Final Thoughts

(i) The approach of Bjerkevik and Lesnick seems to generalize to a much broader class of objects. Anything with a notion of presentation where generators and relations have gradings in a metric space should work.

(ii) However, these metrics feel very complex; NP-most likely.

(iii) Geometry and stratification theory should guide when the infimum—when passing from the semi-distance to the actual distance—is actually obtained.

Final Thoughts

(i) The approach of Bjerkevik and Lesnick seems to generalize to a much broader class of objects. Anything with a notion of presentation where generators and relations have gradings in a metric space should work.

(ii) However, these metrics feel very complex; NP-most likely.

(iii) Geometry and stratification theory should guide when the infimum—when passing from the semi-distance to the actual distance—is actually obtained.

Thank you for your attention!

References I

[BL21] Håvard Bakke Bjerkevik and Michael Lesnick. *ℓ^p-Distances on Multiparameter Persistence Modules*. 2021. arXiv: `2106.13589 [math.AT]`.

[Car+13] Gabriel Cardona et al. "Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf". In: *BMC bioinformatics* 14.1 (2013), pp. 1–13.

[Gas+19] Ellen Gasparovic et al. "Intrinsic Interleaving Distance for Merge Trees". working paper or preprint. Dec. 2019. URL: `https://hal.inria.fr/hal-02425600`.

[MBW13] Dmitriy Morozov, Kenes Beketayev, and Gunther Weber. "Interleaving distance between merge trees". In: *Discrete and Computational Geometry* 49.22-45 (2013), p. 52.

References II

[MS19] Elizabeth Munch and Anastasios Stefanou[‡]. "The ℓ^∞-Cophenetic Metric for Phylogenetic Trees As an Interleaving Distance". In: *Research in Data Science*. Association for Women in Mathematics Series. Springer International Publishing, 2019, pp. 109–127. DOI: 10.1007/978-3-030-11566-1_5. arXiv: 1803.07609.

[ST21] Primoz Skraba and Katharine Turner. *Wasserstein Stability for Persistence Diagrams*. 2021. arXiv: 2006.16824 [math.AT].