Maximum Subbarcode Matching and Subbarcode Distance

Oliver Chubet *†

Abstract

We investigate the maximum subbarcode matching problem which arises from the study of persistent homology and introduce the subbarcode distance on barcodes. A barcode is a set of intervals which correspond to topological features in data and is the output of a persistent homology computation. A barcode $\mathbb{A}$ has a subbarcode matching to $\mathbb{B}$ if each interval in $\mathbb{A}$ matches to an interval in $\mathbb{B}$ which contains it. We present an algorithm which takes two barcodes, $\mathbb{A}$ and $\mathbb{B}$, and returns a maximum subbarcode matching. The subbarcode matching algorithm we present is a generalization of the up-right matching algorithm given by Karp et al [11]. Our algorithm also works on multiset input. It has $O(n \log n)$ runtime, where n is the number of distinct intervals in the barcodes. We show that the subbarcode relation is transitive and induces a partial order on barcodes. We introduce subbarcode distance and show that the subbarcode distance is a lower bound for bottleneck distance. We also give an algorithm to compute subbarcode distance, which has expected $O(n \log^2 n)$ runtime and uses $O(n)$ space.

1 Introduction

In persistent homology the barcode is a multiset of intervals encoding topological information. There is new interest in the implications arising when one has only partial knowledge or an approximation of the barcode. For example, in recent work, Chubet et al [3] establish that one can use subbarcodes in topological data analysis to make strong claims about an unknown function given only upper and lower bounds. Having efficient subbarcode matching algorithms allows one to implement strategies suggested by these new theoretical developements. The subbarcode matching algorithm and subbarcode distance are practical tools for comparing the topological invariants of two datasets.

2 Background

A multiset $\mathbb{A} = (A, \omega_A)$ is a pair consisting of a set A and a multiplicity function $\omega_A : A \to \mathbb{N}$. The weight of $\mathbb{A}$ is the sum of the multiplicities of the elements of A, denoted, $|\mathbb{A}| = \sum_{a \in A} \omega_A(a)$.

A matching $\mathbb{M}$ between multisets $\mathbb{A} = (A, \omega_A)$ and $\mathbb{B} = (B, \omega_B)$ is a multiset $\mathbb{M} = (M, \omega)$ where $M \subset A \times B$ with multiplicity function $\omega : M \to \mathbb{N}$ such that

$$\sum_{b \in B} \omega(a, b) \leq \omega_A(a) \text{ for all } a \in A \text{ and}$$
$$\sum_{a \in A} \omega(a, b) \leq \omega_B(b) \text{ for all } b \in B.$$

A matching $\mathbb{M}$ is a maximum matching if it has maximum weight over all valid matchings. If $|\mathbb{M}| = |\mathbb{A}| = |\mathbb{B}|$ then we call $\mathbb{M}$ a perfect matching.

An interval is a pair (a_x, a_y) for $a_x, a_y \in \mathbb{R}$. See Figure 1. Given intervals $s = (s_L, s_R)$ and $b = (b_L, b_R)$, if

$$b_L \leq s_L, \quad \text{and} \quad s_R \leq b_R.$$

then b contains s, denoted $s \preceq b$. Containment of intervals defines a partial order on intervals.

A barcode $\mathbb{B} = (B, \omega_B)$ is a multiset where B is a set of intervals. A subbarcode matching from $\mathbb{S}$ to $\mathbb{B}$ is

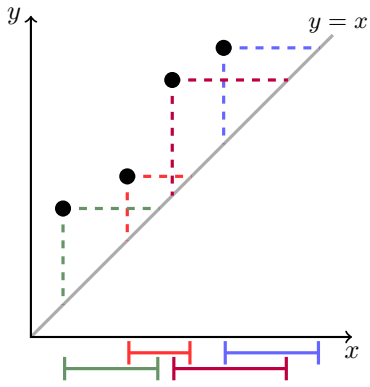

Figure 1: We may represent intervals as points in $\mathbb{R}^2$ by taking their endpoints as coordinates .

a multiset matching $\mathbb{M} = (M, \omega_M)$, of $\mathbb{S}$ and $\mathbb{B}$, where $(s, b) \in M$ implies $s \preceq b$. See Figure 2.

The maximum subbarcode matching problem is to find a subbarcode matching of maximum weight. If there exists a subbarcode matching $\mathbb{M}$ from $\mathbb{A}$ to $\mathbb{B}$ such that $|\mathbb{M}| = |\mathbb{A}|$, then we call $\mathbb{A}$ a subbarcode of $\mathbb{B}$, denoted $\mathbb{A} \sqsubseteq \mathbb{B}$.

*North Carolina State University, oliver.chubet@gmail.com

†This work was partially funded by the NSF under grant CCF-2017980.

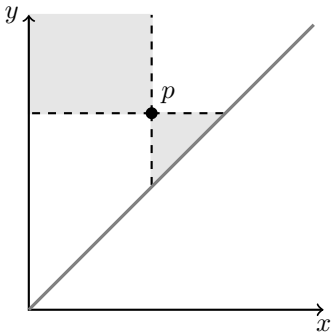

Figure 2: Any point in the upper shaded region contains p as an interval. Any point in the lower shaded region is contained in p as an interval.

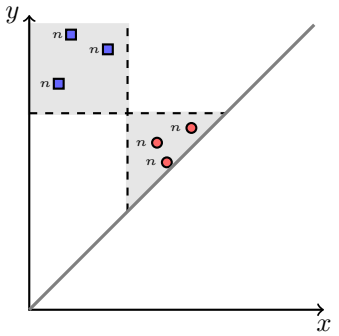

Figure 3: Two barcodes for which there exists a quadratic size subbarcode matchings.

3 Related Work

Traditionally, persistence diagrams have been compared via bottleneck distance. Bottleneck matching is an instance of the assignment problem. The traditional Hopcroft-Karp algorithm for maximum matching in bipartite graphs runs in $O(n^{\frac{5}{2}})$ [10]. However, Efrat et al [5] reduced this runtime to $O(n^{\frac{3}{2}} \log n)$ by using a geometric data structure. Kerber et al [12] also improved this algorithm for persistence diagrams, using k-d trees.

We use a sweepline approach in our subbarcode matching algorithm [1]. Our algorithm builds upon the up-right matching algorithm given by Karp et al [11]. In the case of matching finite subsets of the unit square, this algorithm has been proven to find the optimal matching. Two additional related problems include the maximum matching problem for intersecting intervals [2] and maximum matching in convex bipartite graphs [7, 13, 8]. The strategy used in these algorithms is to avoid backtracking to keep the total operations per element small.

4 Subbarcode Algorithm

We present an algorithm to compute a linear-sized maximum multiset subbarcode matching. See Figure 4.

SUBMATCH($\mathbb{A}, \mathbb{B}$):
Input Two barcodes: $\mathbb{A} = (A, \omega_A)$, $\mathbb{B} = (B, \omega_B)$
Output A subbarcode matching from $\mathbb{A}$ to $\mathbb{B}$

Sort $A \cup B$ by the x-coordinates.
Initialize $\mathtt{T}$ to be an empty balanced binary search tree to store points from B ordered by y-coordinate.
Initialize residual weights $\mathtt{r}_b = \omega_B(b)$ for each $b \in B$ and $\mathtt{r}_a = \omega_A(a)$ for each $a \in A$.
Initialize $(\mathtt{M}, \mathtt{W})$ to store the matching and multiplicities.

For each $p \in A \cup B$, where $p = (p_x, p_y)$:

If $p \in B$, insert b into $\mathtt{T}$.
Else

While $\mathtt{r}_p > 0$:
Search for $b \in \mathtt{T}$ with minimum b_y such that $b_y \geq p_y$.
If there is none, then `break`.
Let $\mathtt{r} = \min\{\mathtt{r}_p, \mathtt{r}_b\}$.
Add (p, b) to $\mathtt{M}$ and set $\mathtt{W}[(p, b)] = \mathtt{r}$,
then update the residual weights of p and b: $\mathtt{r}_p = \mathtt{r}_p - \mathtt{r}$ and $\mathtt{r}_b = \mathtt{r}_b - \mathtt{r}$.
If $\mathtt{r}_b = 0$, then remove b from $\mathtt{T}$.

Return $(\mathtt{M}, \mathtt{W})$.

When both input weight functions uniformly map all elements to 1 this algorithm reduces to the up-right matching algorithm presented by Karp et al [11]. In this case, it is clear that the output size is linear. However, in the case where we are matching multisets, it is possible for a subbarcode matching to have quadratic size.

For example, suppose there are n intervals in barcodes $\mathbb{A}$ and $\mathbb{B}$ respectively such that all intervals have multiplicity n and all intervals in $\mathbb{A}$ are subbars of all intervals in $\mathbb{B}$, as depicted in Figure 3. Then a valid matching could match each interval in A once with each of the intervals in $\mathbb{B}$. This illustrates the significance of a linear-size guarantee.

In the following lemma we prove that the output remains linear.

Lemma 1 *Let $\mathbb{A} = (A, \omega_A)$ and $\mathbb{B} = (B, \omega_B)$ be barcodes. The subbarcode matching $\mathbb{M} =$ SUBMATCH$(\mathbb{A}, \mathbb{B})$ has size $O(n)$, where $n = \#A + \#B$.*

In particular, $\#M \leq n$.

Proof. Let $(M, \omega) =$ SUBMATCH$(\mathbb{A}, \mathbb{B})$. Let $G = (V, M)$ be the weighted graph induced by taking M as the edge

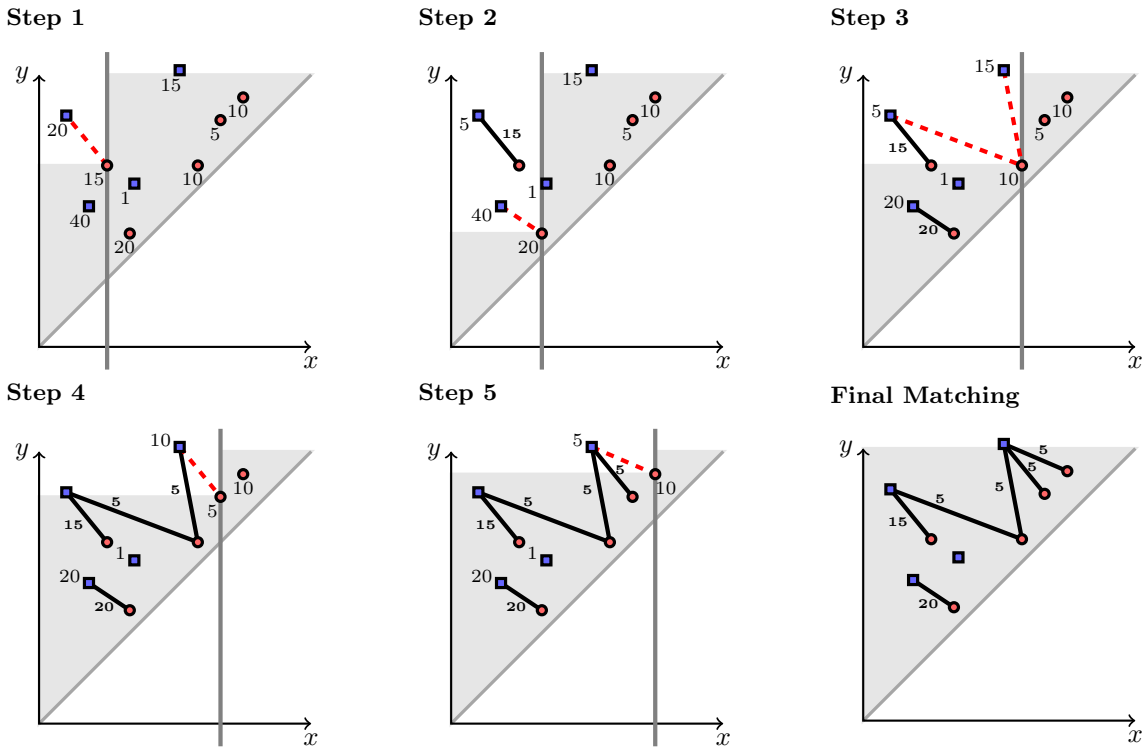

Figure 4: We find a maximum subbarcode matching from $\mathbb{A}$ to $\mathbb{B}$ (circles and squares respectively) labeled by their multiplicities. We iterate through $\mathbb{A}$ in order of x-coordinate and match to the point in $\mathbb{B}$ with lowest y-coordinate. Each edge represents the match labeled with the multiplicity, and the residual multiplicities are updated for $\mathbb{A}$ and $\mathbb{B}$ accordingly.

set with weights given by ω. All edges $(a, b) \in M \subseteq A \times B$. Let $m = \#M$, and $n = \#V$.

We know $m = \frac{1}{2}\sum_{v \in V} \deg(v)$ to be a property of all graphs. Because G is bipartite, it is sufficient to consider only the degrees of elements in A. We partition A into high and low degree nodes,

$$H = \{a \in A \mid \deg(a) \geq 2\} \text{ and } L = A \setminus H.$$

Then, $m = \sum_{a \in H} \deg(a) + \sum_{a \in L} \deg(a)$.

For $a \in H$, consider the sequence $(b_0, \ldots, b_r)$ of all points in B adjacent to a in G, where b_0 is the first point to match to a and b_r is the last point to match to a. Then for $b_i \in \{b_0, \ldots, b_{r-1}\}$ we know that a is the last point to match to b_i, because the algorithm does not proceed to matching b_{i+1} until the remaining multiplicity of b_i is matched.

Each point in B can only have one point being the last to match to it, so

$$\sum_{a \in H} \deg(a) \leq \#H + \#B \text{ and } \sum_{a \in L} \deg(a) \leq \#L.$$

Therefore, $m \leq \#H + \#L + \#B \leq n$. □

Theorem 2 SUBMATCH *uses $O(n)$ space.*

Proof. The only structures maintained during SUBMATCH are the input, the output, and the search tree. The input and search tree are linear size. By Lemma 1, the output of SUBMATCH is linear size as well. Thus total space used is $O(n)$. □

Theorem 3 *The matching from* SUBMATCH *is maximum.*

Proof. First consider the case where $\mathbb{A} = (A, \omega_A)$ and $\mathbb{B} = (B, \omega_B)$ with $\omega_A \equiv \omega_B \equiv 1$. Then SUBMATCH reduces to the up-right matching algorithm given by Karp et al [11], which has previously been shown to be optimal.

If we consider two barcodes $\mathbb{A} = (A, \omega_A)$ and $\mathbb{B} = (B, \omega_B)$, we can construct $\mathbb{A}' = (A', \omega'_A)$, $\mathbb{B}' = (B', \omega'_B)$ such that $\forall a \in A$ we have $\omega_A(a)$ distinct copies $a^{(i)}$ of a in A', for $i \in \{1, \ldots, \omega_A(a)\}$. Similarly, $b^{(j)} \in B'$ for $j \in \{1, \ldots, \omega'_B(b)\}$. Then we have reduced the input to the first case described above. □

Theorem 4 SUBMATCH *computes a linear-sized maximum subbarcode matching in $O(n \log n)$ time.*

Proof. Let $\mathbb{A} = (A, \omega_A)$ and $\mathbb{B} = (B, \omega_B)$ be barcodes. Let $\mathtt{T}$ be search tree constructed in SUBMATCH and let

$\mathbb{M} = (M, \omega)$ be the output matching. Let $G = (V, M)$ be the graph induced by taking M as the edge set. Given $a \in A$ each time we search $\mathtt{T}$ either we find a match or we don't. We find a match $\deg(a)$ times, and we don't find a match at most once. It follows, the number of searches is at most $\sum_{a \in A}(\deg(a)+1) = \#M + \#A$. In Lemma 1 we proved that $\#M$ is linear size. Thus, the number of searches is $O(n)$. Furthermore, there are $O(n)$ insertions and deletions and $\mathtt{T}$ is balanced, so each search operation takes $O(\log n)$. Therefore, the runtime is $O(n \log n)$. □

5 Subbarcode Transitivity

For intervals a and b, recall that $a \preceq b$ if b contains a.

Transitivity of set matching follows easily by composing the matchings. However, functions over multisets do not have a well-defined composition. In 1957, Ford and Fulkerson showed that Hall's Theorem for systems of representitives could equivalently be expressed in terms of flow networks [6, 9]. We use this approach to show the existence of a subbarcode matching is transitive.

Lemma 5 (Transitivity) *If* $\mathbb{A} \sqsubseteq \mathbb{B}$ *and* $\mathbb{B} \sqsubseteq \mathbb{C}$ *then* $\mathbb{A} \sqsubseteq \mathbb{C}$.

Proof. Given barcodes $\mathbb{A} = (A, \omega_A)$, $\mathbb{B} = (B, \omega_B)$, and $\mathbb{C} = (C, \omega_C)$ with subbarcode matchings (M, ω_M) for $\mathbb{A} \sqsubseteq \mathbb{B}$ and (T, ω_T) for $\mathbb{B} \sqsubseteq \mathbb{C}$, there is a corresponding network, $\mathsf{Net}(G)$, where $G = (A \sqcup B \sqcup C,\ M \sqcup T)$ is a digraph [6, 4]. See Figure 5.

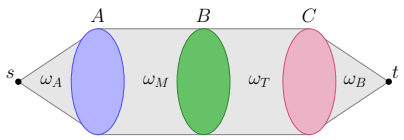

Figure 5: An (s,t)-flow f in $\mathsf{Net}(G)$ corresponds to a matching of $\mathbb{A}$ and $\mathbb{C}$.

If f is a max-flow in $\mathsf{Net}(G)$, then the corresponding matching is maximum and the value of the flow, $|f|$, is equal to the weight of the matching [4]. In Appendix A

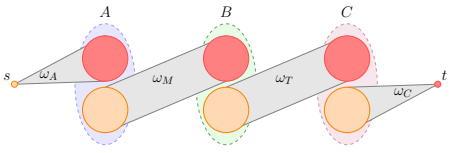

Figure 6: The the capacity of an arbitrary cut, $(L, \overline{L})$ of $\mathsf{Net}(G)$.

we show that $c(L, \overline{L}) \geq |\mathbb{A}|$ for any cut $(L, \overline{L})$ of $\mathsf{Net}(G)$. See Figure 6. Therefore $\mathbb{A} \sqsubseteq \mathbb{C}$. □

Corollary 6 *The relation* $\sqsubseteq$ *defines a partial order on barcodes.*

We call the poset of barcodes $(\mathbf{Bar}, \sqsubseteq)$.

6 Shifted Subbarcodes

There are cases where the maximum matching is not sufficient. Rather, one prefers to know "how far off" two barcodes are from having a subbarcode matching. For example, if we have only an approximation to the input, we can consider the maximum matching after shifting one set by distance δ. There are cases when only a small shift is needed to obtain a subbarcode matching.

If $\mathbb{A} \not\sqsubseteq \mathbb{B}$, we can determine the minimum shift of $\mathbb{A}$ such that the translation results in a subbarcode of $\mathbb{B}$. We use this minimum shift to define a metric on barcodes.

A δ-shift of $\mathbb{A} = (A, \omega)$ is a barcode $\mathbb{A}^\delta$ where

$$\mathbb{A}^\delta := (\delta(A), \omega \circ \delta^{-1}) \text{ and}$$
$$\delta(a) := (a_x + \delta, a_y - \delta).$$

Let $\mathbb{A}$ and $\mathbb{B}$ be barcodes such that $|\mathbb{A}| = |\mathbb{B}|$. The subbarcode distance is

$$\mathbf{d}_S(\mathbb{A}, \mathbb{B}) := \max\{\min_{\delta \geq 0} \mathbb{A}^\delta \sqsubseteq \mathbb{B}, \min_{\delta \geq 0} \mathbb{B}^\delta \sqsubseteq \mathbb{A}\}.$$

The subbarcode distance is similar to Hausdorff distance in that it is bidirectional and asymmetric in nature.

Lemma 7 (Approximation is additive.) *If* $\mathbb{A}^\delta \sqsubseteq \mathbb{B}$ *and* $\mathbb{B}^\varepsilon \sqsubseteq \mathbb{C}$ *then* $\mathbb{A}^{\delta+\varepsilon} \sqsubseteq \mathbb{C}$.

Proof. Let $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$ be barcodes such that $\mathbb{A}^\delta \sqsubseteq \mathbb{B}$ and $\mathbb{B}^\varepsilon \sqsubseteq \mathbb{C}$. Consider intervals, a and b.

If $a \preceq b$, then $a_x \geq b_x$ and $a_y \leq b_y$.

Then, $a_x + \delta \geq b_x + \delta$ and $a_y - \delta \leq b_y - \delta$.

Thus, $\delta(a) \preceq \delta(b)$.

By extension, if $\mathbb{A} \sqsubseteq \mathbb{B}$, then $\mathbb{A}^\delta \sqsubseteq \mathbb{B}^\delta$. By assumption, $\mathbb{A}^\delta \sqsubseteq \mathbb{B}$, so it follows, $\mathbb{A}^{\delta+\varepsilon} \sqsubseteq \mathbb{B}^\varepsilon$. Thus by transitivity of subbarcodes (Lemma 5), $\mathbb{A}^{\delta+\varepsilon} \sqsubseteq \mathbb{C}$. □

Lemma 8 (Triangle Inequality)

$$\mathbf{d}_S(\mathbb{A}, \mathbb{C}) \leq \mathbf{d}_S(\mathbb{A}, \mathbb{B}) + \mathbf{d}_S(\mathbb{B}, \mathbb{C})$$

Proof. Let $\mathbb{A}, \mathbb{B}$, and $\mathbb{C}$ be barcodes. Suppose $\mathbf{d}_S(\mathbb{A}, \mathbb{B}) = \delta$ and $\mathbf{d}_S(\mathbb{B}, \mathbb{C}) = \varepsilon$. Then by definition,

$$\mathbb{A}^\delta \sqsubseteq \mathbb{B},\ \mathbb{B}^\delta \sqsubseteq \mathbb{A},\ \mathbb{B}^\varepsilon \sqsubseteq \mathbb{C}, \text{ and } \mathbb{C}^\varepsilon \sqsubseteq \mathbb{B}.$$

By Lemma 7, it follows $\mathbb{A}^{\delta+\varepsilon} \sqsubseteq \mathbb{C}$ and $\mathbb{C}^{\delta+\varepsilon} \sqsubseteq \mathbb{A}$. Therefore $\mathbf{d}_S(\mathbb{A}, \mathbb{C}) \leq \delta + \varepsilon$. □

The remaining metric properties are easily verified, so we may conclude the following theorem.

Theorem 9 *The subbarcode distance is a metric on barcodes.*

7 Subbarcode Distance Computation

In this section we present algorithms which allow us to compute the subbarcode distance. The goal is to compute the minimum shift such that we have a subbarcode matching. To find this shift it is useful to determine cases in which we may easily recognize that we have shifted by an excesse amount.

Lemma 10 *For a subbarcode matching, (M,ω), of $\mathbb{A}^{\Delta} \sqsubseteq \mathbb{B}$, let*

$$\gamma = \min_{(\Delta(a),b)\in M} \min\{a_x + \Delta - b_x, b_y - a_y + \Delta\}.$$

Then $\mathbb{A}^{\Delta-\gamma} \sqsubseteq \mathbb{B}$.

Proof. For all $(\Delta(a),b) \in M$, $a_e + \Delta - b_x \geq \gamma$, and $b_y - a_y + \Delta \geq \gamma$. So, $a_x + (\Delta - \gamma) \geq b_x$, and $b_y \geq a_y - (\Delta - \gamma)$. Therefore $\mathbb{A}^{\Delta-\gamma} \sqsubseteq \mathbb{B}$. □

We can think of γ as an excess shift of $\mathbb{A}$. That is, we could have shifted $\mathbb{A}$ by a distance γ less than we did and the corresponding matching is still be a valid matching. So intuitively, if the shift is the subbarcode distance, then $\gamma = 0$ because there can be no excess shift.

In the next lemma we prove that the subbarcode distance, similar to Hausdorff distance and bottleneck distance, is determined by a pair from A and B. This motivates us to devise a search method to find this pair.

Lemma 11 *For some $(a,b) \in A \times B$,*

$$\mathbf{d}_S(\mathbb{A},\mathbb{B}) = \min\{a_x - b_x, b_y - a_y\}.$$

Proof. Let $\Delta = \mathbf{d}_S(\mathbb{A},\mathbb{B})$. Then there is a subbarcode matching (M,ω) for $\mathbb{A}^{\Delta} \sqsubseteq \mathbb{B}$. By Lemma 10, $\mathbb{A}^{\Delta-\gamma} \sqsubseteq \mathbb{B}$ for $\gamma = \min_{(\Delta(a),b)\in M} \min\{a_x - b_x, b_y - a_y\}$. It follows that $\gamma = 0$ because Δ is minimum. So $\Delta = \min\{a_x - b_x, b_y - a_y\}$ for some $(a,b) \in A \times B$. □

Lemma 11 enables us to compute $\mathbf{d}_S$ by finding the correct pair in $A \times B$. There are n^2 possibilities, however, we search these possibilities efficiently by taking a uniform sample of the endpoints for which the difference is within given upper and lower bounds.

For barcodes (A,ω_A) and (B,ω_B), define:

$$\text{UB} := \max\{(\max_{b\in B} b_x - \min_{a\in A} a_x), (\max_{a\in A} a_y - \min_{b\in B} b_y), 0\}$$
$$\text{LB} := \max\{(\max_{b\in B} b_x - \max_{a\in A} a_x), (\min_{a\in A} a_y - \min_{b\in B} b_y), 0\}.$$

Here, the upper bound UB is simply the distance between the farthest corners of the minimum bounding rectangles of A and B. The lower bound LB is the distance between the bottom right corners of the minimum bounding rectangles. These may be replaced with any suitable upper and lower bounds.

In MINSHIFT we use these bounds to perform a binary search through all pairs of coordinate differences in order to find the points that give us the exact subbarcode distance.

MINSHIFT($\mathbb{A},\mathbb{B}$, LB, UB):
Input: Barcodes $\mathbb{A}$,$\mathbb{B}$, and upper and lower bounds $\text{LB} \leq \mathbf{d}_S(\mathbb{A},\mathbb{B}) \leq \text{UB}$
Output: The subbarcode distance, Δ

Let X, Y be the sorted x- and y-coordinates of $A \cup B$.
Δ = SAMPLE(X, Y, LB, UB)
While Δ exists:

  (M,ω) = SUBMATCH($\mathbb{A}^{\Delta}$, $\mathbb{B}$)
  If (M,ω) is a perfect matching, set UB = Δ.
  Else LB = Δ
  Δ = SAMPLE(X, Y, LB, UB)

Return UB

A binary search is made possible by using SAMPLE to obtain a uniform random sample of all pairs with coordinate differences contained within the given bounds. In a linear scan of the sets of x- and y-coordinates we determine the prevalence of each coordinate in the set suitable pairs. We then sample a pair from this set and return the minimum coordinate difference. See Figure 7 and Figure 8.

SAMPLE(X, Y, LB, UB):
Input: Sorted lists X and Y, and bounds LB and UB
Output: A uniform random sample

In a linear scan of X, find indices, l_i and u_i, such that

$$\text{X}[l_i - 1] \leq \text{X}[i] + \text{LB} < \text{X}[l_i],$$
$$\text{and } \text{X}[u_i] < \text{X}[i] + \text{UB} \leq \text{X}[u_i + 1].$$

Similarly, scanning Y, find indices, l'_i and u'_i, such that

$$\text{Y}[u'_i - 1] \leq \text{Y}[i] - \text{UB} < \text{Y}[u'_i],$$
$$\text{and } \text{Y}[l'_i] < \text{Y}[i] - \text{LB} \leq \text{Y}[l'_i + 1].$$

If $l_i = u_i$ and $l'_i = u'_i$ for all i, return nothing.
Otherwise, sample an index i with probability proportional to $(u_i - l_i) + (l'_i - u'_i)$.
Sample endpoint e uniformly from $\text{X}[l_i : u_i] \sqcup \text{Y}[u'_i : l'_i]$.
If e is from X then return $e - \text{X}[i]$. Otherwise return $\text{Y}[i] - e$.

Theorem 12 MINSHIFT *computes the subbarcode distance with an expected $O(n \log^2 n)$ time.*

Proof. Using SAMPLE to get a uniform sample of all pairwise distances of endpoints, MINSHIFT reduces to a

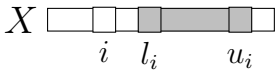

Figure 7: If $X[i]$ is an endpoint and $X[j]$ is from the range $X[l_i : u_i]$ then $\text{LB} < X[j] - X[i] < \text{UB}$.

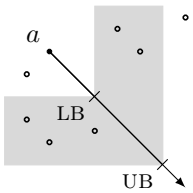

Figure 8: Depicted above are the points considered by SAMPLE for a single point $a \in A$. The points in the shaded region form a subset of B for which the minimum coordinate differences are within the bounds given.

randomized binary search over n^2 elements. Thus there is an expected $O(\log n)$ iterations, where each iteration is $O(n \log n)$. Therefore MINSHIFT has expected runtime $O(n \log^2 n)$. □

8 Persistence Diagrams

In topological data analysis it is common to compare persistence diagrams rather than barcodes. In this section we show that, with slight modification, the algorithms presented in Section 7 also apply in this setting.

The diagonal of $\mathbb{R}$ is the set $\mathcal{D} = \{(x, x) \mid x \in \mathbb{R}\}$. A persistence diagram for a barcode $\mathbb{B} = (B, \omega_B)$ is a multiset $\mathsf{PD}(\mathbb{B}) := (B \cup \mathcal{D}, \omega)$, where

$$\omega(x) = \begin{cases} \omega_B(x), & x \in B \\ \infty, & x \in \mathcal{D}. \end{cases}$$

We have added the diagonal of $\mathbb{R}$ with infinite multiplicity.

Let $\mathsf{PD}(\mathbb{A}) = (A \cup \mathcal{D}, \omega)$, be a persistence diagram. Note that shifting this diagram by δ gives us the multiset

$$\mathsf{PD}(\mathbb{A})^\delta = (\delta(A \cup \mathcal{D}),\ \omega \circ \delta^{-1}).$$

It is useful to refer to only the points above the diagonal, because points which have been shifted below $y = x$ can now match to the diagonal. We denote this as $[\mathbb{X}^\delta]$, where $\mathbb{X}$ is a barcode.

Lemma 13 *Let $\mathbb{A} = (A, \omega_A)$ and $\mathbb{B} = (B, \omega_B)$ be barcodes. Then*

$$\mathsf{PD}(\mathbb{A})^\delta \sqsubseteq \mathsf{PD}(\mathbb{B}) \quad \textit{if and only if} \quad [\mathbb{A}^\delta] \sqsubseteq \mathbb{B}.$$

Proof. Let (M, ω) be a subbarcode matching for $\mathsf{PD}(\mathbb{A})^\delta \sqsubseteq \mathsf{PD}(\mathbb{B})$. Consider $a \in [\delta(A)]$. Note that if $(a, b) \in M$, then $b \notin \mathcal{D}$,so we can restrict M to $M \cap ([\delta(A)] \times B)$ to obtain a matching for $[\mathbb{A}^\delta] \sqsubseteq \mathbb{B}$.

Now let (N, ω) be a matching for $[\mathbb{A}^\delta] \sqsubseteq \mathbb{B}$. For any $a \in \delta(A \cup \mathcal{D}) \setminus [\delta(A)]$, there is $d = (a_x, a_x) \in \mathcal{D}$ such that $a \preceq d \in \mathcal{D}$. Because d has infinite multiplicity in $\mathsf{PD}(\mathbb{B})$, we can add (a, d) to N and set $\omega(a, d) = \omega_A \circ \delta^{-1}(a)$. Thus N is a subbarcode matching. □

This result allows us to compute a subbarcode matching of persistence diagrams $\mathsf{PD}(\mathbb{A}^\delta)$ and $\mathsf{PD}(\mathbb{B})$ by computing $\text{SUBMATCH}([\mathbb{A}^\delta], \mathbb{B})$. Additionally, we can compute $\mathbf{d}_S(\mathsf{PD}(\mathbb{A}), \mathsf{PD}(\mathbb{B}))$ by modifying MINSHIFT slightly. Rather than returning the minimum Δ such that $\mathbb{A}^\Delta \sqsubseteq \mathbb{B}$, we return the minimum Δ such that $[\mathbb{A}^\Delta] \sqsubseteq \mathbb{B}$.

Note that because persistence diagrams fall under our definition of barcodes, the subbarcode distance is also a metric on persistence diagrams.

9 Subbarcode Distance and Bottleneck Distance

In this section we establish the relationship between the subbarcode distance and bottleneck distance.

Let $\mathbb{A}$ and $\mathbb{B}$ be barcodes such that $|\mathbb{A}| = |\mathbb{B}|$. Let $\mathcal{M}$ be the set of all possible perfect matchings between $\mathbb{A}$ and $\mathbb{B}$. The bottleneck distance is

$$\mathbf{d}_B(\mathbb{A}, \mathbb{B}) := \min_{(M,\omega) \in \mathcal{M}} \left\{ \max_{(a,b) \in M} ||a - b||_\infty \right\}$$

A bottleneck matching between barcodes $\mathbb{A}$ and $\mathbb{B}$ is a matching $\mathbb{M} = (M, \omega)$ where

$$\max_{(a,b) \in M} ||a - b||_\infty = \mathbf{d}_B(\mathbb{A}, \mathbb{B}).$$

Theorem 14 *For any two barcodes $\mathbb{A}$ and $\mathbb{B}$, where $|\mathbb{A}| = |\mathbb{B}|$,*

$$\mathbf{d}_S(\mathbb{A}, \mathbb{B}) \leq \mathbf{d}_B(\mathbb{A}, \mathbb{B}).$$

Proof. Let $\mathbb{M} = (M, \omega_M)$ be a bottleneck matching between $\mathbb{A}$ and $\mathbb{B}$. Let $\beta := \mathbf{d}_B(\mathbb{A}, \mathbb{B})$. Then for any edge $(a, b) \in M$, $||a - b||_\infty \leq \beta$. Moreover, $|b_x - a_x| \leq \beta$ and $|b_y - a_y| \leq \beta$. It follows that $b_x \leq a_x + \beta$ and $a_y - \beta \leq b_y$, implying $\beta(a) \preceq b$ for each $(a, b) \in M$. We can then construct a matching as follows: Let $\mathbb{T} = (T, \omega_T)$, where

$$T = \{(\beta(a), b) \mid (a, b) \in M\} \text{ and}$$

$$\omega_T(\beta(a), b) := \omega_M(a, b).$$

Then $\mathbb{T}$ is a subbarcode matching. We note that $|\mathbb{T}| = |\mathbb{M}|$ and $|\mathbb{A}| = |\mathbb{A}^\beta|$. Additionally, $\mathbb{M}$ is a perfect matching, so $|\mathbb{M}| = |\mathbb{A}| = |\mathbb{B}|$. It follows that $\mathbb{A}^\beta \sqsubseteq \mathbb{B}$. By a similar argument we may also show that $\mathbb{B}^\beta \sqsubseteq \mathbb{A}$. Therefore, $\mathbf{d}_S(\mathbb{A}, \mathbb{B}) \leq \beta = \mathbf{d}_B(\mathbb{A}, \mathbb{B})$. □

10 Conclusion

We have given an efficient method for computing maximum subbarcode matchings and subbarcode distance. We have shown that barcodes are a poset under the subbarcode relation, and that subbarcode distance is a metric on persistence diagrams. Subbarcodes present efficient methods of comparison for persistence diagrams.

Acknowledgement

Thank you to Don Sheehy, my advisor for invaluable discussions and feedback, especially regarding the sampling algorithm.

References

[1] J. L. Bentley and T. A. Ottmann. Algorithms for reporting and counting geometric intersections. *IEEE Transactions on computers*, 28(09):643–647, 1979.

[2] D. Z. Chen, X. S. Hu, and X. Wu. Maximum red/blue interval matching with application. In *International Computing and Combinatorics Conference*, pages 150–158. Springer, 2001.

[3] O. A. Chubet, K. P. Gardner, and D. R. Sheehy. A theory of sub-barcodes. *arXiv preprint arXiv:2206.10504*, 2022.

[4] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2022.

[5] A. Efrat, A. Itai, and M. J. Katz. Geometry helps in bottleneck matching and related problems. *Algorithmica*, 31(1):1–28, 2001.

[6] L. R. Ford and D. Fulkerson. Network flow and systems of representatives. *Canadian Journal of Mathematics*, 10:78–84, 1958.

[7] G. Gallo. An o (n log n) algorithm for the convex bipartite matching problem. *Operations Research Letters*, 3(1):31–34, 1984.

[8] F. Glover. Maximum matching in a convex bipartite graph. *Naval research logistics quarterly*, 14(3):313–316, 1967.

[9] P. Hall†. *On Representatives of Subsets*, pages 58–62. Birkhäuser Boston, Boston, MA, 1987.

[10] J. E. Hopcroft and R. M. Karp. An nˆ5/2 algorithm for maximum matchings in bipartite graphs. *SIAM Journal on computing*, 2(4):225–231, 1973.

[11] R. M. Karp, M. Luby, and A. Marchetti-Spaccamela. A probabilistic analysis of multidimensional bin packing problems. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 289–298, 1984.

[12] M. Kerber, D. Morozov, and A. Nigmetov. Geometry helps to compare persistence diagrams, 2017.

[13] G. Steiner and J. S. Yeomans. A linear time algorithm for maximum matchings in convex, bipartite graphs. *Computers & Mathematics with Applications*, 31(12):91–96, 1996.

A Subbarcode Transitivity

Lemma 15 (Transitivity) *If* $\mathbb{A} \sqsubseteq \mathbb{B}$ *and* $\mathbb{B} \sqsubseteq \mathbb{C}$ *then* $\mathbb{A} \sqsubseteq \mathbb{C}$.

Proof. Let $\mathbb{A} = (A, \omega_A)$, $\mathbb{B} = (B, \omega_B)$, and $\mathbb{C} = (C, \omega_C)$ be barcodes such that $\mathbb{A} \sqsubseteq \mathbb{B}$ and $\mathbb{B} \sqsubseteq \mathbb{C}$. Then there exists subbarcode matchings, (M, ω_M) from $\mathbb{A}$ to $\mathbb{B}$ and (T, ω_T) from $\mathbb{B}$ to $\mathbb{C}$.

Let $\mathsf{Net}(G)$ be the corresponding network to find the the maximum subbarcode matching from $\mathbb{A}$ to $\mathbb{C}$, as described in Section 5. Let $(L, \overline{L})$ be a cut of $\mathsf{Net}(G)$. Then $L = X \sqcup Y \sqcup Z$ and $\overline{L} = \overline{X} \sqcup \overline{Y} \sqcup \overline{Z}$ for

$$X = A \cap L \qquad Y = B \cap L \qquad Z = C \cap L$$
$$\overline{X} = A \setminus X \qquad \overline{Y} = B \setminus Y \qquad \overline{Z} = C \setminus Z$$

We examine $c(L, \overline{L})$:

$$\begin{aligned} \mathsf{c}(L, \overline{L}) &= \mathsf{c}(X \sqcup Y \sqcup Z,\ \overline{X} \sqcup \overline{Y} \sqcup \overline{Z}) \\ &= \mathsf{c}(s, \overline{X}) + \mathsf{c}(X, \overline{Y}) + \mathsf{c}(Y, \overline{Z}) + \mathsf{c}(Z, t) \end{aligned}$$

We now evaluate each term:

$$\mathsf{c}(s, \overline{X}) = \sum_{a \in \overline{X}} \omega_A(a) \qquad \mathsf{c}(X, \overline{Y}) = \sum_{a \in X} \sum_{b \in \overline{Y}} \omega_M(a, b)$$
$$\mathsf{c}(Z, t) = \sum_{c \in Z} \omega_C(c) \qquad \mathsf{c}(Y, \overline{Z}) = \sum_{b \in Y} \sum_{c \in \overline{Z}} \omega_T(b, c)$$

Notice (T, ω_T) is a subbarcode matching, so by necessity ω_C is greater than the marginals of ω_T for each $c \in C$. Similarly, ω_B is greater than the marginals of ω_M for each $b \in B$.

$$\begin{aligned} \sum_{c \in Z} \omega_C(c) &\geq \sum_{c \in Z} \sum_{b \in B} \omega_T(b, c) \\ &= \sum_{c \in Z} \sum_{b \in Y} \omega_T(b, c) + \sum_{c \in Z} \sum_{b \in \overline{Y}} \omega_T(b, c) \end{aligned}$$

It follows,

$$\begin{aligned} \mathsf{c}(Y, \overline{Z}) + \mathsf{c}(Z, t) &\geq \sum_{y \in Y} \sum_{c \in C} \omega_T(b, c) = \sum_{b \in Y} \omega_B(b) \\ &\geq \sum_{b \in Y} \sum_{a \in X} \omega_M(a, b). \end{aligned}$$

Then,

$$\begin{aligned} &\mathsf{c}(X, \overline{Y}) + \mathsf{c}(Y, \overline{Z}) + \mathsf{c}(Z, t) \\ &\qquad \geq \sum_{a \in X} \sum_{b \in \overline{Y}} \omega_M(a, b) + \sum_{a \in X} \sum_{b \in Y} \omega_M(a, b) \\ &\qquad = \sum_{a \in X} \omega_A(a). \end{aligned}$$

Finally,

$$\begin{aligned} &\mathsf{c}(s, \overline{X}) + \mathsf{c}(X, \overline{Y}) + \mathsf{c}(Y, \overline{Z}) + \mathsf{c}(Z, t) \\ &\qquad \geq \sum_{a \in \overline{X}} \omega_A(a) + \sum_{a \in X} \omega_A(a) \\ &\qquad = \sum_{a \in A} \omega_A(a) = |\mathbb{A}|. \end{aligned}$$

Thus, $c(L, \overline{L}) \geq |\mathbb{A}|$ for any cut $(L, \overline{L})$ of $\mathsf{Net}(G)$. Therefore $\mathbb{A} \sqsubseteq \mathbb{C}$. □