A Survey on the Mapper Algorithm

Ankita Sarkar `f004nwn`
COSC 249, Fall 2021, Dartmouth College

Abstract

The Mapper [SMC07] algorithm in topological data analysis is used for analyzing the shape of high-dimensional data, via techniques inspired by the Nerve Theorem in algebraic topology. In this survey, we outline the Mapper algorithm, its limitations, and some proposed mitigation of these limitations. We discuss the theoretical guarantees behind an algorithm called Multiscale Mapper [DMW16] that uses ideas from persistent homology to produce a more reliable version of Mapper. We also discuss how Mapper can be viewed [CMO18, MW16] as a discretized version of a topological notion called the Reeb space, which serves as an indicator of the correctness of Mapper.

1 Introduction

In 2007, three scientists at Stanford University – Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson – proposed a new method [SMC07] to produce meaningful low-dimensional visualizations of high-dimensional data. They called this method *Mapper*, and Singh and Carlsson founded the startup Ayasdi to apply Mapper and related topological techniques to big data. One of the most anecdotally famous instances of such applications involves Muthu Alagappan, at the time an intern at Ayasdi, and his discovery [Coh12] of a 13-position classification of basketball players that challenged the usual 5-position classification, launching Alagappan into sports analytics limelight.

Behind the scenes of trendy applications, Mapper relies on the fundamental algebraic-topology concept of the *nerve of a cover*. Given a topological space X, and a *filter* function $f : X \to \mathbb{R}^D$ for some small D, the Mapper algorithm uses a cover $\mathcal{I}$ of $f(X)$ to obtain a "pullback" cover inside X, on which it computes a nerve. It turns out that high-dimensional data can be approximated as a topological space in many ways – for example, one can construct a Rips complex on the point cloud, or view the point cloud as a sampling from an unknown topological space. Under such approximations, Mapper can be sensibly applied to data, allowing us to visualize the shape of a high-dimensional point cloud, and to compute the homology of such a shape.

The Mapper algorithm, as originally presented, relies on hand-tuning of the choices of f and $\mathcal{I}$, which requires human intervention and domain knowledge. Since the goal of Mapper is to uncover unknown topological features, there is no obvious way to judge the optimality of the parameter choices from the output that they produce. Thus, a key challenge in implementing Mapper is the automation of parameter tuning. A natural idea that had emerged was to compute the persistence features of the resulting nerve, but that raises a further question – how do we know that the homology of the nerve actually does represent the homology of the original space? Theorem 2.1 guarantees a faithful representation for a topological space X under certain nice conditions; but it is not immediate whether that guarantee continues to hold for a point cloud under possibly suboptimal choices of parameters. In fact, the guarantee is not even immediate for the one-dimensional Mapper, where $D = 1$. For the subsequent discussion, we will use the term D-Mapper for Mapper with codomain $\mathbb{R}^D$.

As outlined in the above discussion, even the "correctness" of Mapper is under question. Carrière and Oudot [CO17] addressed this for 1-Mapper by comparing its output to the Reeb graph of X with respect to f. They conclude that, as $\mathcal{I}$ gets more granular, the Mapper converges to the Reeb graph. This suggests that the optimal choice of parameters for the Mapper algorithm should produce an output that imitates the Reeb graph closely. Carrière, Michel, and Oudot [CMO18] build on that notion to optimize parameter tuning for Mapper. They study the persistence features of 1-Mapper as well as the Reeb graph under the same filter function, and compare them under the bottleneck distance. They use this measure to tune the parameters of Mapper, and establish Mapper as an optimal estimator of the Reeb graph.

We would naturally want to extend the above idea to larger D, since intuitively, f must obscure some information when it reduces the dimension. For even $D = 2$, however, there is no clear notion of optimality for

2-Mapper. In fact, the high-dimensional analogue of the Reeb graph, called the Reeb space, is not as intuitive to study, which creates challenges in replicating the above optimality results for larger D. Furthermore, even when useful nerves are produced by 2-Mapper, the choice of $\mathcal{I}$ becomes more complicated. It is not even clear whether the Reeb space is the right analogue in higher dimensions – an alternative suggested [MW16] is a category-theoretic version of the Reeb space, which produces some optimality results, but without clear computational implications at $D > 1$.

Beyond correctness, one also has to consider the notion of efficiency, which encourages us to try avoiding persistence diagrams as an additional step before we can utilize Mapper. Moreover, since a key motivation is to produce useful visuals, one would prefer a single low-dimensional simplicial complex over a collection of barcodes. One method, that directly applies a persistence-like notion over gradually varying parameters, is Multiscale Mapper [DMW16]. They present their technique alongside stability and efficiency guarantees that enable practical computation of persistence features. However, since it still requires multiple nerves to be studied at once, it is not as useful for producing a digestible visual interpretation. A subsequent work [DSK+18] suggests a tool called Multimapper, where different sets of parameters are applied to different areas of the point cloud to produce a more faithful visualization. Unlike Multiscale Mapper, however, Multimapper comes with no mathematically robust guarantees of its correctness and efficiency.

In this survey, we explain what Mapper is, and provide examples of the challenges outlined above. Then, we exhibit a theoretical stability guarantee of the Multiscale Mapper [DMW16] technique. We conclude with an outline of the category-theoretic approach [MW16] for proving that the D-Mapper converges to the D-dimensional Reeb space.

2 Preliminaries

Let us first see how Mapper works on a topological space. Figure 1 provides an example, and the formal definition is developed below.

Definition 2.1 (pullback cover). *Given a function $f : X \to \mathbb{R}^D$ from a topological space X to the D-dimensional Euclidean space, and a cover $\mathcal{I}$ of $f(X)$, define the pullback cover $f^*(\mathcal{I})$ of X to be the collection $\cup_{I\in\mathcal{I}}\mathcal{C}_I$ where $\mathcal{C}_I$ is the collection of connected components of $f^{-1}(I)$ for any $I \in \mathcal{I}$.*

Definition 2.2 (nerve of a cover). *Given a topological space X and a cover $\mathcal{J}$ of X, the nerve $N(\mathcal{J})$ of $\mathcal{J}$ is a simplicial complex constructed as follows: for each $k \in \mathbb{N}$, for each $J_1, J_2, \ldots, J_k \in \mathcal{J}$ such that $\left(\cap_{i=1}^k J_i\right) \cap X \neq \emptyset$, we introduce the $(k-1)$-simplex $\sigma([k])$.*

Definition 2.3 (Mapper of a topological space). *Given a topological space X, a* filter *function $f : X \to \mathbb{R}^D$, and a cover $\mathcal{I}$ of $f(X)$, the Mapper of X with respect to f and $\mathcal{I}$, denoted $M(X, f, \mathcal{I})$, is defined as the nerve of $f^*(\mathcal{I})$, i.e.*

$$M(X, f, \mathcal{I}) = N(f^*(\mathcal{I}))$$

When X and f are clear from context, we will drop them from the notation.

Here, it is useful to see how the choice of $\mathcal{I}$ affects the output. Figure 2 shows how a suboptimal $\mathcal{I}$ can obscure important features of the shape of X. Notice that the same change can be effected by slightly changing f instead of $\mathcal{I}$: in the hand shape of Figure 1, the dark green interval ends very close to the base of the thumb, and a slight perturbation of f could create the situation where the inverse image of the dark green interval has a single connected component.

The definition of Mapper can be translated almost exactly for a point cloud $\mathbb{X}$, but we need a more sensible notion in the place of connected components since $\mathbb{X}$ is discrete. Hence, we use a clustering algorithm, which becomes one of the parameters of Mapper, to identify points in each $f^{-1}(I)$ that are close together.

Definition 2.4 (Mapper of a point cloud). *Given a point cloud $\mathbb{X}$, a filter function $f : \mathbb{X} \to \mathbb{R}^D$, a cover $\mathcal{I}$ of $f(\mathbb{X})$, and a clustering algorithm $\mathcal{C}$ that returns a collection of clusters, the Mapper of $\mathbb{X}$ with respect to f, $\mathcal{I}$, and $\mathcal{C}$, denoted $M(\mathbb{X}, f, \mathcal{I}, \mathcal{C})$, is defined as the nerve of the cover $\cup_{I\in\mathcal{I}}\mathcal{C}(f^{-1}(I))$.*

As before, we will drop f and $\mathbb{X}$ from the notation whenever clear from context. For the purposes of this discussion, we will not focus on optimizing the choice of clustering algorithm, so we will also drop $\mathcal{C}$ from our

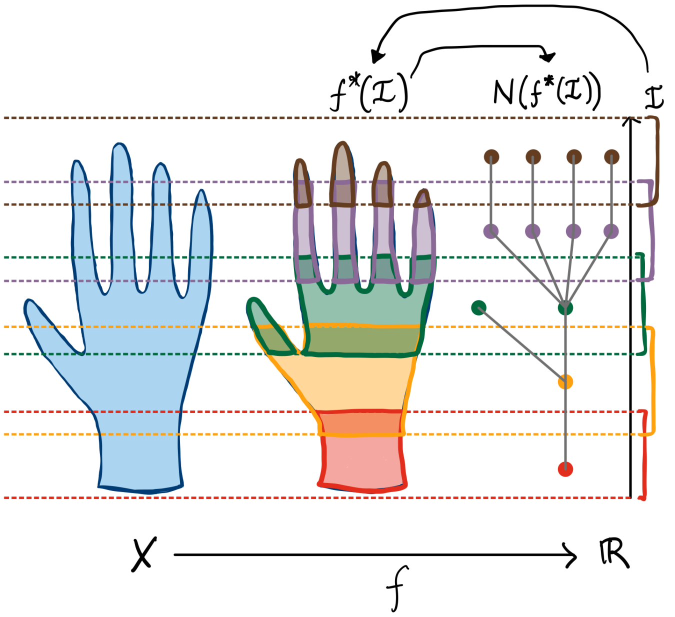


Figure 1: 1-Mapper on a hand shape with respect to the height function

notation hereafter. By abuse of notation, we will also refer to the pullback cover in the point cloud context, defining

$$f^*(\mathcal{I}) := \cup_{I\in\mathcal{I}}\mathcal{C}(f^{-1}(I))$$

Moving to the point cloud context introduces further optimization issues with the choice of $\mathcal{I}$. Earlier, we saw how issues arise when $\mathcal{I}$ is not granular enough. In Figure 3, we see how issues can arise in the point cloud setting when $\mathcal{I}$ is *too granular*.

To establish some notion of correctness, Mapper is compared to the Reeb space, which is defined as follows:

Definition 2.5 (Reeb space)**.** *Given a topological space X and a continuous function $f : X \to \mathbb{R}^D$, we define the equivalence relation $\sim_f$ on X by saying that $x \sim_f y$ if $f(x) = f(y)$, and x and y lie in the same connected component of $f^{-1}(f(x))$. The Reeb space of X with respect to f is the space ${}^{X}\!/\!_{\sim_f}$.*

Definition 2.6 (Reeb graph)**.** *When $D = 1$, the Reeb space is called the Reeb graph.*

Comparing Figure 2, Figure 1, and Figure 4 gives us the intuitive idea of how, with finer and finer covers being selected, 1-Mapper indeed converges to the Reeb graph. This notion can be formalized [CMO18].

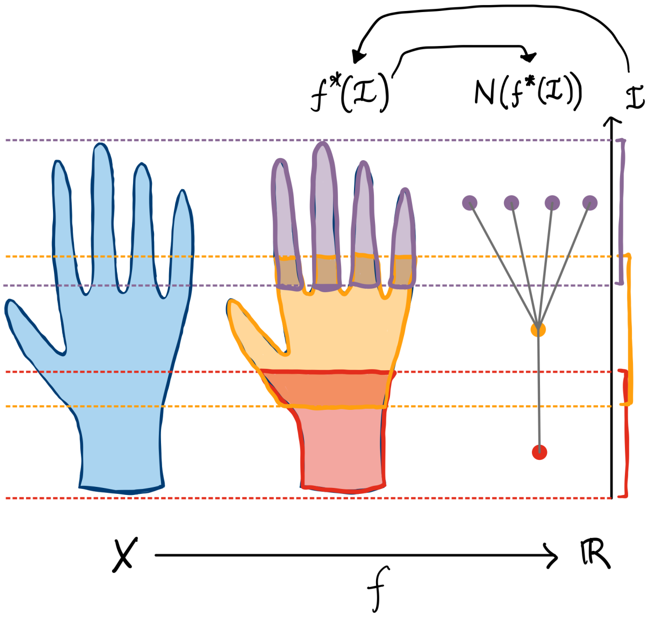


Figure 2: 1-Mapper on a hand with $\mathcal{I}$ such that the thumb is not detected

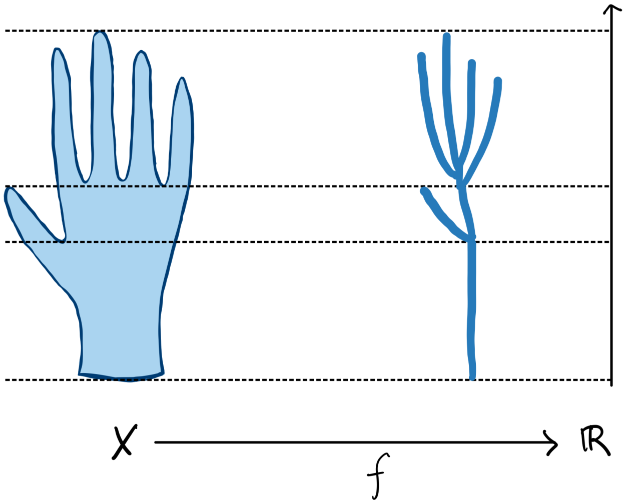
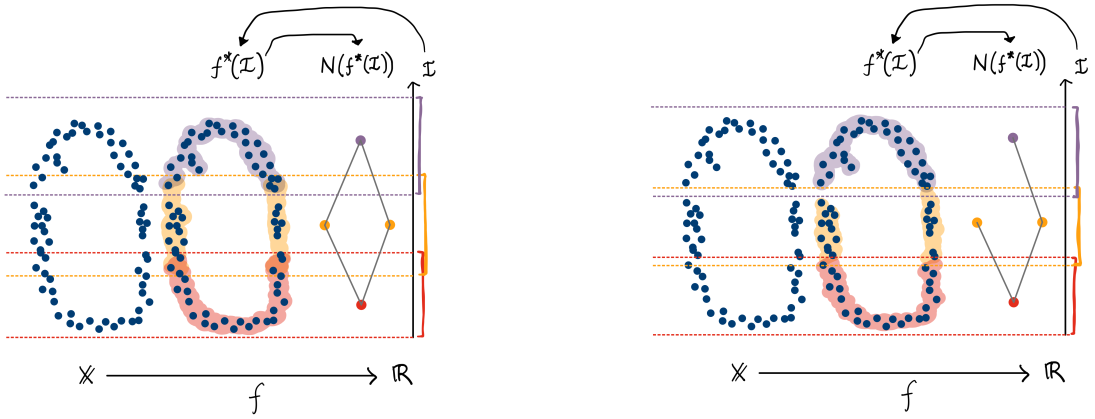


Figure 4: Example of a Reeb graph

Before we proceed, we state the theorem that provides intuition for Mapper being a suitable representation of the original data.

Theorem 2.1 (Nerve Theorem [Hat02]). *If $\mathcal{J}$ is an open cover of a paracompact space X such that every nonempty intersection of finitely many sets in $\mathcal{J}$ is contractible, then X is homotopy equivalent to the nerve $N(\mathcal{J})$.*



(a) Good choice of $\mathcal{I}$

(b) Bad choice of $\mathcal{I}$

Figure 3: 1-Mapper on ring-shaped point cloud comparing different choices of $\mathcal{I}$

3 Multiscale Mapper [DMW16]

Dey, Mémoli, and Wang [DMW16] suggest a technique called Multiscale Mapper that combines the Mapper algorithm with a persistence-like idea, constructing a series of Mappers at varying scales and setting up a sequence of maps to compare subsequent Mappers on this scale. Let us first see, in Figure 5, an intuitive example of this construction, which should make clear its relationship to persistence.

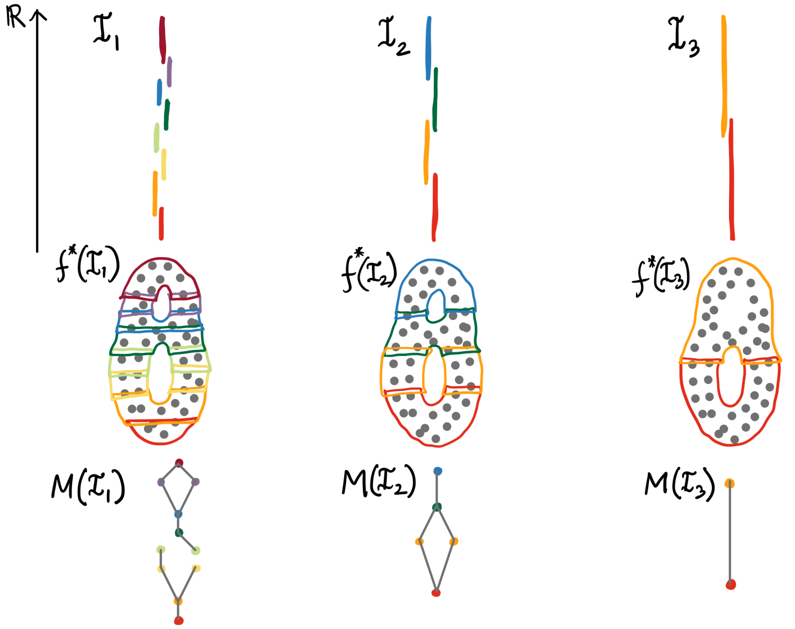


Figure 5: Example of Multiscale Mapper showing appearance and disappearance of holes

To formally set up the Multiscale Mapper construction, we need the following definitions.

Definition 3.1 (Simplicial map [DMW16])**.** *Let K and L be two simplicial complexes over the vertex sets V_K and V_L respectively. A set map $\phi : V_K \to V_L$ is a* simplicial map *if $\phi(\sigma) \in L$ for each $\sigma \in K$.*

For the Multiscale Mapper construction, it is assumed that all covers chosen are open and path-connected, and that the filter function chosen is piecewise linear (which ensures that it is Morse – a standard requirement [CO17] in Mapper literature). In order to study how Mappers vary with variations in the choice of cover, we set up maps between covers, and corresponding maps between the resulting nerves.

Definition 3.2 (Maps between covers and nerves [DMW16])**.** *Given two covers $\mathcal{I} = \{I_a\}_{a \in A}$, $\mathcal{J} = \{J_b\}_{b \in B}$ of the same space X, a* map of covers *is a map $\xi : A \to B$ such that for each $a \in A$, $I_a \subseteq J_{\xi(a)}$. For simplicity, we abuse notation and call the map of covers ξ as well.*

A map of covers ξ induces a simplicial map $N(\xi) : N(\mathcal{I}) \to N(\mathcal{J})$, which is given on vertices by ξ and then extended to the higher-dimensional simplices.

The relationship between the two kinds of maps above is preserved under composition. If we have maps of covers $\mathcal{I} \xrightarrow{\xi} \mathcal{I}' \xrightarrow{\zeta} \mathcal{I}''$, then we have $N(\zeta \circ \xi) = N(\zeta) \circ N(\xi)$. Figure 6 exhibits this correspondence intuitively.

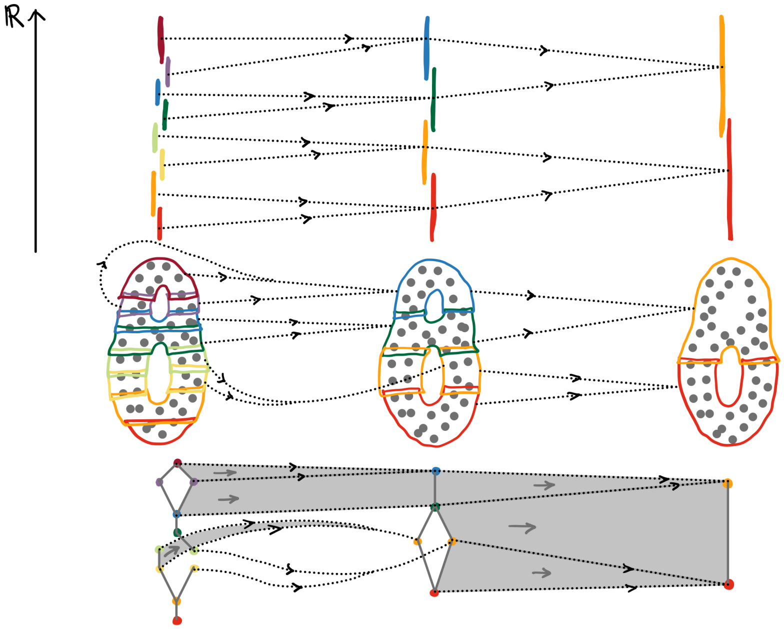


Figure 6: Example showing correspondence of tower of covers, tower of pullback covers, and tower of nerves

Now we start defining Multiscale Mapper formally.

Definition 3.3 (Tower [DMW16])**.** *A* tower $\mathfrak{I}$ *of covers with* resolution $\text{res}(\mathfrak{I}) = r \in \mathbb{R}$ *is a collection $\{\mathcal{I}_\varepsilon\}_{\varepsilon \geq r}$ of covers together with maps $\xi_{\varepsilon,\varepsilon'} : \mathcal{I}_\varepsilon \to \mathcal{I}_{\varepsilon'}$ so that $\xi_{\varepsilon,\varepsilon} = \text{id}$, and for $r \leq \varepsilon \leq \varepsilon' \leq \varepsilon''$, $\xi_{\varepsilon,\varepsilon''} = \xi_{\varepsilon',\varepsilon''} \circ \xi_{\varepsilon,\varepsilon'}$.*

We write $\mathfrak{I} = \{\mathcal{I}_\varepsilon \xrightarrow{\xi_{\varepsilon,\varepsilon'}} \mathcal{I}_{\varepsilon'}\}_{r \leq \varepsilon \leq \varepsilon'}$ to denote a tower of covers.

A tower of simplicial complexes or vector spaces is also defined in the exact same way, with the intervening maps being simplicial maps or linear maps respectively.

One can verify that, given $f : X \to \mathbb{R}^n$ and a tower $\mathfrak{I}$ of covers of $\mathbb{R}^n$, the pullback $f^*(\mathfrak{I}) := \{f^*(\mathcal{I})\}_{\mathcal{I} \in \mathfrak{I}}$ is a tower of covers of X. We can intuitively see this in Figure 6.

On a tower of covers $\mathfrak{I} = \{\mathcal{I}_\varepsilon \xrightarrow{\xi_{\varepsilon,\varepsilon'}} \mathcal{I}_{\varepsilon'}\}_{r \leq \varepsilon \leq \varepsilon'}$, we can naturally define the induced tower of simplicial complexes $N(\mathfrak{I}) := \{N(\mathcal{I}_\varepsilon) \xrightarrow{N(\xi_{\varepsilon,\varepsilon'})} N(\mathcal{I}_{\varepsilon'})\}_{r \leq \varepsilon \leq \varepsilon'}$. With this in mind, we are ready to understand Multiscale Mapper formally.

Definition 3.4 (Multiscale Mapper, [DMW16])**.** *With respect to a filter function $f : X \to \mathbb{R}^n$ and a tower $\mathfrak{I}$ of covers of $\mathbb{R}^n$, the Multiscale Mapper $MM(X, f, \mathfrak{I})$ is defined to be the tower of nerves induced on X by* *the pullback of* $\mathfrak{I}$*, i.e.*

$$MM(X, f, \mathfrak{I}) = N(f^*(\mathfrak{I}))$$

As we did for Mapper, we will drop X and f whenever clear from context.

The tower of nerves induces a tower of vector spaces over $\mathbb{Z}$ when we compute the homology of each nerve in the tower. Computing barcodes on this tower allows us to identify the covers for which Mapper gives a faithful visualization of the original space. Since barcodes are computed for a continuously varying parameter, we need a notion of gradual change to develop Multiscale Mapper. This is achieved by using only those towers of covers where successive covers vary gradually with the change in index. In this discussion, such a property (Assumption 1) is assumed for towers of covers hereafter. Our assumption is a simplification of a more rigorous goodness condition that is presented in the original analysis [DMW16].

Assumption 1 (Good tower)**.** *A tower of covers* $\mathfrak{I}$ *with resolution* s *has the following properties for each* $\varepsilon \geq s$*:*

- *If* $I \in \mathcal{I}_\varepsilon$*, then* $\operatorname{diam}(I) \leq 2^\varepsilon$
- *For any* $J \subseteq \mathbb{R}^D$ *with* $\operatorname{diam}(J) \leq 2^\varepsilon$*,* $\exists I \in \mathcal{I}_\varepsilon$ *such that* $J \subseteq I$*.*

The above property captures the desired notion of gradualness because it ensures that the diameters of the cover elements are bounded by a function of the indices of the tower. We will also assume that, when we use a filter function f, we will only use towers of covers with resolution at most $\log \operatorname{diam}(f(X))$.

We saw earlier that small variations in $\mathcal{I}$ or f can alter the output of Mapper drastically. We will now see how Multiscale Mapper is more stable in this sense.

3.1 Stability under perturbation of f

Dey, Mémoli and Wang [DMW16] exhibit that, if we begin from the same tower $\mathfrak{I}$ of covers, and use two different filter functions f and g that are almost the same, then the resulting towers $MM(f, \mathfrak{I})$ and $MM(g, \mathfrak{I})$ are *interleaved*, i.e. the levels of one tower sandwiches the levels of the other tower. This indicates that they are a small perturbation of each other.

Below, we see a simplified version of the result. The precise result requires more machinery, which we omit here in the interest of focusing on the overview. In particular, where the original result uses the *contiguity* notion to denote that two simplicial maps are similar, we will pretend that those simplicial maps are equal.

Definition 3.5 (Interleaved towers [DMW16])**.** *Two towers of simplicial complexes with equal resolution,* $\mathfrak{S} = \left\{\mathcal{S}_\varepsilon \xrightarrow{s_{\varepsilon,\varepsilon'}} \mathcal{S}_{\varepsilon'}\right\}_{r \leq \varepsilon \leq \varepsilon'}$ *and* $\mathfrak{T} = \left\{\mathcal{T}_\varepsilon \xrightarrow{t_{\varepsilon,\varepsilon'}} \mathcal{T}_{\varepsilon'}\right\}_{r \leq \varepsilon \leq \varepsilon'}$*, are said to be* η*-interleaved if there are collections of maps* $\{\phi_\varepsilon : \mathcal{S}_\varepsilon \to T_{\varepsilon+\eta}\}_{\varepsilon \geq r}$ *and* $\{\psi_\varepsilon : \mathcal{T}_\varepsilon \to S_{\varepsilon+\eta}\}_{\varepsilon \geq r}$ *such that the following diagrams [DMW16] commute* up to contiguity*.*

$$\begin{tikzcd}
\mathcal{S}_\varepsilon \arrow[rr, "s_{\varepsilon,\varepsilon+2\eta}"] \arrow[dr, "\varphi_\varepsilon"] & & \mathcal{S}_{\varepsilon+2\eta} \\
& \mathcal{T}_{\varepsilon+\eta} \arrow[ur, "\psi_{\varepsilon+\eta}"] &
\end{tikzcd}
\qquad
\begin{tikzcd}
& \mathcal{S}_{\varepsilon+\eta} \arrow[dr, "\varphi_{\varepsilon+\eta}"] & \\
\mathcal{T}_\varepsilon \arrow[ur, "\psi_\varepsilon"] \arrow[rr, "t_{\varepsilon,\varepsilon+2\eta}"] & & \mathcal{T}_{\varepsilon+2\eta}
\end{tikzcd}$$

$$\begin{tikzcd}
\mathcal{S}_\varepsilon \arrow[rr, "s_{\varepsilon,\varepsilon'}"] \arrow[dr, "\varphi_\varepsilon"] & & \mathcal{S}_{\varepsilon'} \arrow[dr, "\varphi_{\varepsilon'}"] & \\
& \mathcal{T}_{\varepsilon+\eta} \arrow[rr, "t_{\varepsilon+\eta,\varepsilon'+\eta}"'] & & \mathcal{T}_{\varepsilon'+\eta}
\end{tikzcd}$$

$$\begin{tikzcd}
& \mathcal{S}_{\varepsilon+\eta} \arrow[rr, "s_{\varepsilon+\eta,\varepsilon'+\eta}"] & & \mathcal{S}_{\varepsilon'+\eta} \\
\mathcal{T}_\varepsilon \arrow[ur, "\psi_\varepsilon"] \arrow[rr, "t_{\varepsilon,\varepsilon'}"'] & & \mathcal{T}_{\varepsilon'} \arrow[ur, "\psi_{\varepsilon'}"] &
\end{tikzcd}$$

Interleaved towers of covers are defined similarly, with the maps ϕ_ε and ψ_ε being maps of covers.

Via our simplifying assumption that contiguity is equality, we can simply pretend that the diagrams above commute for interleaved towers of simplicial complexes.

The following statement shows that similar filter functions lead to similar pullback covers.

Proposition 3.1 (Simplified from Proposition 4.5, [DMW16])**.** *Consider two filter functions $f, g : X \to \mathbb{R}^D$ such that $\max_{x \in X} d(f(x), g(x)) \leq \delta$. For a good tower $\mathfrak{I}$ of covers of $\mathbb{R}^D$ with resolution s such that*

$$1 \leq s \leq \min(\log \operatorname{diam}(f(X)), \log \operatorname{diam}(g(X)), \delta)$$

the corresponding towers of pullback covers $f^*(\mathfrak{I})$ and $g^*(\mathfrak{I})$ are η-interleaved for

$$\eta := \log(2\delta + 1)$$

Proof. First, for some $J \subseteq \mathbb{R}^D$, consider $x \in f^{-1}(J)$, so we have $d(f(x), J) = 0$ and hence

$$d(g(x), J) \leq d(f(x), J) + d(f(x), g(x)) \leq \delta$$

which means that $f^{-1}(J) \subseteq g^{-1}(J^\delta)$ where

$$J^\delta = \{y \in \mathbb{R}^D \mid d(y, J) \leq \delta\}$$

Generalizing, the above means that, $\forall J \subseteq \mathbb{R}^D$

$$f^{-1}(J) \subseteq g^{-1}(J^\delta) \tag{1}$$

Now, for an index ε, consider $U \in f^*(\mathcal{I}_\varepsilon)$. Let $I \in \mathcal{I}_\varepsilon$ be such that U is a connected component of $f^{-1}(I)$. By (1), $f^{-1}(I) \subseteq g^{-1}(I^\delta)$. We know by Assumption 1 and the definition of maps of covers that, since $\operatorname{diam}\left(I^\delta\right) \leq 2\delta + \operatorname{diam}(I) \leq 2\delta + 2^\varepsilon \leq (2\delta + 1)2^\varepsilon = 2^{\eta+\varepsilon}$, there is a set $I' \in \mathcal{I}_{\varepsilon+\eta}$ containing I^δ, and hence containing I. Hence U lies in some connected component of $g^{-1}(I')$. Call this connected component V. We can construct a map of covers $i_\varepsilon : f^*(\mathcal{I}_\varepsilon) \to g^*(\mathcal{I}_\varepsilon)$ such that for each $U \in f^*(\mathcal{I}_\varepsilon)$, $i_\varepsilon(U) = V$ as per the above construction.

The map of covers in the other direction is constructed similarly. ■

The above result for $D = 1$ is illustrated in Figure 7. The interleaving maps of covers induce simplicial maps between the corresponding nerves, which are also η-interleaved since the sets in the pullback covers correspond directly with the 0-simplices. This gives us the notion of stability of Multiscale Mapper under perturbation of the filter function f.

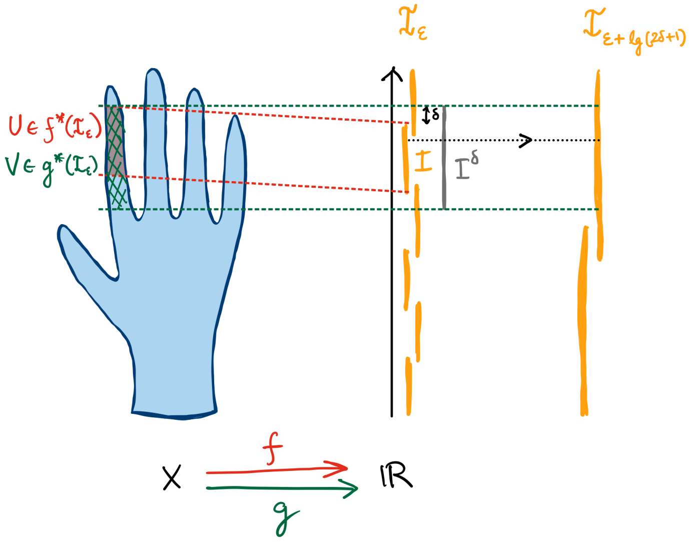

$\mathcal{I}_\varepsilon$
$\mathcal{I}_{\varepsilon+\lg(2\delta+1)}$
$U \in f^*(\mathcal{I}_\varepsilon)$
$V \in g^*(\mathcal{I}_\varepsilon)$
δ
I
I^δ
X
f
g
$\mathbb{R}$

Figure 7: Sketch of Proposition 3.1 for $D = 1$

3.2 Further stability results

A proof similar to that of Proposition 3.1 can be used [DMW16] to show that Multiscale Mapper is stable to simultaneous perturbations of both the filter function f and the tower of covers $\mathfrak{I}$. Under Assumption 1, we state a simplified version of that generalization.

Theorem 3.1 (Simplified from Theorem 4.7, [DMW16])**.** *Consider two filter functions $f, g : X \to \mathbb{R}^D$ such that $\max_{x\in X} d(f(x), g(x)) \leq \delta$. Also consider two η'-interleaved good towers $\mathfrak{I}, \mathfrak{J}$ of covers of $\mathbb{R}^D$ with $\operatorname{res}(\mathfrak{I}) = \operatorname{res}(\mathfrak{J}) = s$ such that*

$$1 \leq s \leq \min(\operatorname{diam}(f(X)), \operatorname{diam}(g(X)), \delta)$$

Then the towers of pullback covers $f^*(\mathfrak{I})$ and $g^*(\mathfrak{J})$ are η-interleaved for

$$\eta := \log(2\delta + 1 + \eta')$$

4 Mapper as a discretized Reeb space

Earlier, we provided a pictorial idea of how the 1-Mapper converges to the Reeb graph. This idea can be formalized [CMO18] in the setting where the point cloud is assumed to be an unknown sampling from an unknown topological space.

For $D > 1$, Munch and Wang [MW16] propose the use of category-theoretic tools to set up similar convergence results of the D-mapper to the Reeb space at dimension D. We outline their result below.

For the purposes of this discussion, we will think of a category as follows:

Definition 4.1 (category)**.** *A* category *C is a collection of* objects $\mathrm{Ob}(C)$ *with* arrows *or* morphisms $\mathrm{Hom}(C)$ *between them.*

The arrows are called $\mathrm{Hom}(C)$ because some common categories are formed by groups, rings, etc. algebraic structures as objects and the relevant homomorphisms as arrows. Some examples of categories are as follows. Many of them will be useful in our discussion.

Category	Objects	Arrows
Set	sets	set maps
Top	topological spaces	continuous maps
$\mathbf{Open}(\mathbb{R}^D)$	open sets in $\mathbb{R}^D$	inclusion maps
Vect	vector spaces	linear maps
$\mathbf{Cell}(K)$	simplices of a simplicial complex K	$\sigma \to \tau$ if σ is a face of τ
Poset category	poset $(P, \prec)$	$a \to b$ if $a \prec b$
$\mathbb{R}^D$-**Top**	$(X, f : X \to \mathbb{R}^D)$ where $X \in \mathrm{Ob}(\mathbf{Top})$ and f is continuous	$(X, f) \to (Y, g)$ induced by $\nu : X \to Y$ such that $g \circ \nu = f$
Opposite category C^{op}	$x \in \mathrm{Ob}(C)$	$f^{op} : y \to x$ for each $f : x \to y$ in $\mathrm{Hom}(C)$

We will need two further notions: a way to relate categories to one another, called *functors*, and a way to relate functors to one another, called *natural transformations*.

Definition 4.2 (functor)**.** *A* functor *between two categories C and D, denoted $F : C \to D$, sends objects to objects and arrows to arrows in a manner that respects identity and composition laws, i.e.*

- $\forall x \in \mathrm{Ob}(C)$, $F[\mathrm{id}_x] = \mathrm{id}_{F(x)}$*, and*
- $\forall f, g \in \mathrm{Hom}(C)$, $F[g \circ f] = F[g] \circ F[f]$

Here are some examples of functors that will be useful in our subsequent discussion. Notice that these constructions are somewhat reminiscent of our earlier discussion about pullbacks and maps of covers.

$F : C \to D$	on $\mathrm{Ob}(C)$	on $\mathrm{Hom}(C)$
$\pi_0 : \mathbf{Top} \to \mathbf{Set}$	Set of connected components of $X \in \mathrm{Ob}(\mathbf{Top})$	Set map induced by f from $\pi_0(X)$ to $\pi_0(Y)$
Thickening functor $T_\varepsilon : \mathbf{Open}(\mathbb{R}^D) \to \mathbf{Open}(\mathbb{R}^D)$	$X^\varepsilon = \{x \in \mathbb{R}^D \mid d(x, X) \leq \varepsilon\}$	$X \hookrightarrow X^\varepsilon$

Definition 4.3 (natural transformation)**.** *Given two functors $F, G : C \to D$, a* natural transformation *between them is a family of arrows $\phi = \{\phi_x : F(x) \to G(x)\}_{x \in \mathrm{Ob}(C)}$ in D such that the following diagram [MW16] commutes:*

$$\begin{tikzcd}
F(x) \arrow[r, "\varphi_x"] \arrow[d, "F[f]"'] & G(x) \arrow[d, "G[f]"] \\
F(y) \arrow[r, "\varphi_y"] & G(y)
\end{tikzcd}$$

We can think of natural transformations as arrows between objects that are themselves functors. This motivates the following definition:

Definition 4.4. *For two categories C and D, D^C is the category such that* $\mathrm{Ob}(D^C)$ *is the collection of functors $C \to D$, and* $\mathrm{Hom}(D^C)$ *is the collection of all natural transformations between all such functors.*

Equipped with these notions, we can now describe the Reeb space in category theory terms.

Definition 4.5 (categorical Reeb Space, [MW16])**.** *For $(X, f : X \to \mathbb{R}^D) \in \mathrm{Ob}(\mathbb{R}^D\text{-}\mathbf{Top})$, the* categorical Reeb space *is the functor $\pi_0 f^{-1} : \mathbf{Open}(\mathbb{R}^D) \to \mathbf{Set}$ such that*

- *For an open set $U \subseteq \mathbb{R}^D$, $\pi_0 f^{-1}(U)$ is the set of connected components of $f^{-1}(U)$.*
- *For an inclusion map $U \hookrightarrow V$, $\pi_0 f^{-1}[U \hookrightarrow V]$ is the set map induced by f between the connected components of $f^{-1}(U)$ and $f^{-1}(V)$.*

The above notion can be reconciled with our usual definition of a Reeb space by observing that if all possible pullback covers are defined, then our Reeb space is also indirectly defined. We also notice that, by the definition, a categorical Reeb space is an object in the category $\mathbf{Set}^{\mathbf{Open}(\mathbb{R}^D)}$. So the construction of the Reeb space from the original space is a functor $\mathcal{C} : \mathbb{R}^D\text{-}\mathbf{Top} \to \mathbf{Set}^{\mathbf{Open}(\mathbb{R}^D)}$.

We will use the above machinery to introduce the notion of *interleaved* categorical Reeb spaces. Recall the proof of Proposition 3.1; we mimic that idea here, in the following steps. Since the Reeb space is a functor $F : \mathbf{Open}(\mathbb{R}^D) \to \mathbf{Set}$, we can compose it with the functor $T_\varepsilon : \mathbf{Open}(\mathbb{R}^D) \to \mathbf{Open}(\mathbb{R}^D)$ in the obvious way: for each open set, start with its contribution to the Reeb space, and go to its thickening's contribution to the Reeb space. Formally:

$$S_\varepsilon(F) := FT_\varepsilon$$

By the above, $F[I \hookrightarrow I^{2\varepsilon}]$ induces a natural transformation $n_F : F \Rightarrow S_{2\varepsilon}F$. In Figure 8: the yellow shading shows how $F = \pi_0 f^{-1}$ behaves, the green shading shows how $S_{2\varepsilon}(F)$ behaves, and the red shadings show the natural transformation between these two Reeb spaces.

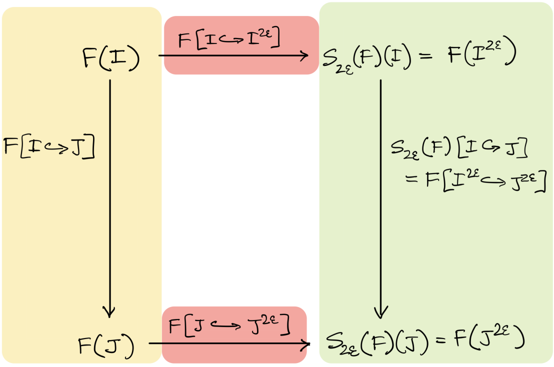


Figure 8: Natural transformation between Reeb spaces induced by ε-thickening

This interleaving produces the notion of a distance between two Reeb spaces.

Definition 4.6 (Interleaving Reeb spaces [MW16])**.** *An ε-interleaving between Reeb spaces $F, G : \mathbf{Open}(\mathbb{R}^D) \to \mathbf{Set}$ are a pair of natural transformations $\phi : F \Rightarrow S_\varepsilon(G)$ and $\psi : G \Rightarrow S_\varepsilon(F)$ such that the following diagrams [MW16] commute, where $\eta = n_F, \tau = n_G$.*

$$\begin{tikzcd}
\mathcal{F} \arrow[r, Rightarrow, "\varphi"] \arrow[dr, Rightarrow, "\eta"'] & \mathcal{S}_\varepsilon(\mathcal{G}) \arrow[d, "\mathcal{S}_\varepsilon[\psi]"] \\
& \mathcal{S}_{2\varepsilon}(\mathcal{F})
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathcal{G} \arrow[r, Rightarrow, "\psi"] \arrow[dr, Rightarrow, "\tau"'] & \mathcal{S}_\varepsilon(\mathcal{F}) \arrow[d, "\mathcal{S}_\varepsilon[\varphi]"] \\
& \mathcal{S}_{2\varepsilon}(\mathcal{G})
\end{tikzcd}$$

This induces the following notion of interleaving distance *between two Reeb spaces:*

$$d_I(F,G) = \inf\{\varepsilon \in [0,\infty) \mid F, G \textit{ are } \varepsilon\textit{-interleaved}\}$$

The convention above is that the infimum of an empty set is ∞. The above notion of distance is indeed a sensible distance between Reeb graphs:

Theorem 4.1 (Theorem 5.2, [MW16])**.** *d_I is an extended pseudometric on* $\mathrm{Ob}(\mathbf{Set}^{\mathbf{Open}(\mathbb{R}^D)})$.

Now, we will set up a categorical definition of Mapper. We will assume that we have a good cover with regards to the Nerve Theorem conditions, i.e.

Assumption 2. *A cover $\mathcal{I}$ of $\mathbb{R}^D$ and its pullback $f^*(\mathcal{I})$ satisfy the hypotheses of Theorem 2.1.*

Fix a cover $\mathcal{I}$ of $\mathbb{R}^D$, and set $K := M(\mathcal{I})$. Then, consider an arrow $\sigma \leq \tau$ in $\mathbf{Cell}(K)$. The corresponding intersection of cover elements is in the opposite direction, i.e. is of the form $U_\tau \subseteq U_\sigma$, where U_σ is the intersection of elements of $f^*(\mathcal{I})$ corresponding to σ. Hence, we will use the opposite category $\mathbf{Cell}(K)^{op}$ to construct the categorical Mapper.

Definition 4.7 (categorical Mapper, [MW16])**.** *The categorical Mapper on (X,f) is the functor $\mathcal{C}_K^f$: $\mathbf{Cell}(K)^{op} \to \mathbf{Set}$ given by*

$$\mathcal{C}_K^f(\sigma) = \pi_0 f^{-1}(I_\sigma)$$

where I_σ is the intersection of elements of $\mathcal{I}$ corresponding to U_σ. Hence the process of constructing the categorical Mapper is a functor $\mathcal{C}_k : \mathbb{R}^D\text{-}\mathbf{Top} \to \mathbf{Set}^{\mathbf{Open}(\mathbb{R}^D)}$.

With the above definitions of the categorical Reeb space and Mapper, we are ready to state (a simplified statement of) the main convergence result in Munch and Wang's [MW16] paper. The result relies on the existence of a functor $\mathcal{P}_K : \mathbf{Set}^{\mathbf{Cell}(K)^{op}} \to \mathbf{Set}^{\mathbf{Open}(\mathbb{R}^D)}$, which creates a continuous version of the Mapper, thus pushing it into the category $\mathbf{Set}^{\mathbf{Open}(\mathbb{R}^D)}$ and enabling it to be compared to the Reeb space. The construction of $\mathcal{P}_K$ is the main technical contribution in the paper [MW16]. Below, we state their main result in terms of $\mathcal{P}_K$, and then provide an overview of how $\mathcal{P}_K$ is constructed.

In the following statement, the resolution of a cover $\mathcal{I}$ of $\mathbb{R}^D$ is defined as

$$\mathrm{res}(\mathcal{I}) = \sup_{I \in \mathcal{I}} \mathrm{diam}\,(I)$$

Theorem 4.2 (Simplified from Theorem 4.1, [MW16])**.** *Given $(X,f) \in \mathrm{Ob}(\mathbb{R}^D\text{-}\mathbf{Top})$ and a cover $\mathcal{I}$ of $f(X) \subseteq \mathbb{R}^D$, let $K = M(X,f,\mathcal{I})$. Then*

$$d_I(\mathcal{C}(X,f), \mathcal{P}_K\mathcal{C}_K(X,f)) \leq \mathrm{res}(\mathcal{I})$$

The functor $\mathcal{P}_K$ is set up in terms of a category theoretic construction knows as the *colimit* of a functor. This is defined as follows.

Definition 4.8 (cocone)**.** *The* cocone *of a functor $F : C \to D$ is a pair (N,ψ) where $N \in \mathrm{Ob}(D)$, and $\psi = \{\psi_x : F(x) \to N\}_{x \in \mathrm{Ob}(C)}$ is a family of arrows in D, such that $\forall f : x \to y$ in $\mathrm{Hom}(C)$*

$$\psi_y \circ F[f] = \psi_x$$

Definition 4.9. *A cocone (N,ψ) is said to* factor through *the cocone (L,ϕ) if $\exists u \in \mathrm{Hom}(D)$ such that $\forall x \in \mathrm{Ob}(C), u \circ \phi_x = \psi_x$.*

Definition 4.10 (colimit)**.** *(L,ψ) is said to be the* colimit *of $F : C \to D$ if it is a cocone of F such that any other cocone of F factors through it.*

The following diagram [MW16] provides the visual definition of the colimit.

$$\begin{tikzcd}
F(x) \arrow[dd, "F[f]"'] \arrow[dr, "\varphi_x"'] \arrow[drr, bend left, "\psi_x"] & & \\
& L \arrow[r, dashed, "u"] & N \\
F(y) \arrow[ur, "\varphi_y"] \arrow[urr, bend right, "\psi_y"'] & &
\end{tikzcd}$$

To define $\mathcal{P}_K : \mathbf{Set}^{\mathbf{Cell}(K)^{op}} \to \mathbf{Set}^{\mathbf{Open}(\mathbb{R}^D)}$, we need a functor $\mathcal{P}_K(F) : \mathbf{Open}(\mathbb{R}^D) \to \mathbf{Set}$ for any functor $F : \mathbf{Cell}(K)^{op} \to \mathbf{Set}$. This definition is made as follows. For any open set $A \subseteq \mathbb{R}^D$, K_A be the set of simplices σ in K such that I_σ intersects A. Then, for an open set $A \subseteq \mathbb{R}^D$,

$$P_K(F)(A) = \operatorname{colim}_{\sigma \in K_A} F(\sigma)$$

and the arrows are the natural transformations induced by the above.

To understand $\mathcal{P}_K$ intuitively, note that F in the definition of $\mathcal{P}_K$ is the Mapper, so $F(\sigma)$ is the set of pullback elements that contribute to U_σ. Since the target category is **Set**, the L in the colimit is formed by the union; so by taking the colimit, we are able to consider all connected components in the pullback that intersect with $f^{-1}(A)$. The proof in the paper [MW16] uses this to identify the part of the Reeb space that corresponds to A. In Figure 9, we sketch this intuition in the $D = 1$ case, with the gray-shaded regions corresponding to the construction of $\mathcal{P}_K(F)(A)$. The striped portion in the sketch is an example of $F(\sigma)$ as computed from a simplex $\sigma \in K_A$, which in this example is the top right edge in the Mapper.

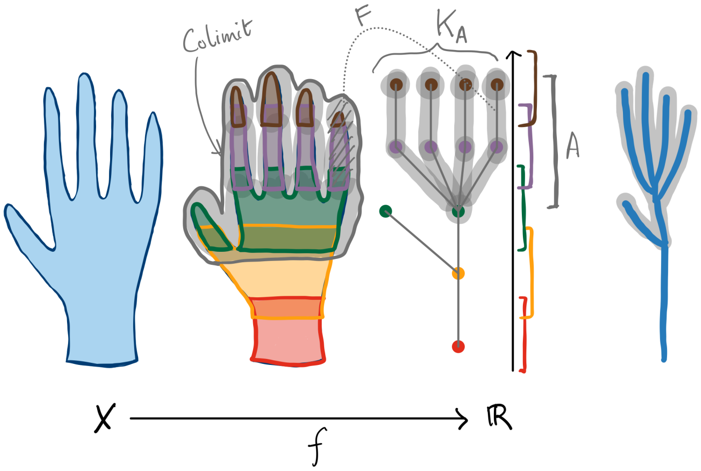


Figure 9: Sketch of construction of $\mathcal{P}_K : \mathbf{Set}^{\mathbf{Cell}(K)^{op}} \to \mathbf{Set}^{\mathbf{Open}(\mathbb{R}^D)}$

We conclude with an alternate visual overview of the main result Theorem 4.2. Consider the following diagram [MW16]:

$$\begin{tikzcd}
\mathbb{R}^d\text{-}\mathbf{Top} \arrow[d, "\mathcal{C}_K"] \arrow[dr, bend left, "\mathcal{C}"] & \\
\mathbf{Set}^{\mathbf{Cell}(K)^{\mathrm{op}}} \arrow[r, "\mathcal{P}_K"] & \mathbf{Set}^{\mathbf{Open}(\mathbb{R}^d)}
\end{tikzcd}$$

$$\begin{tikzcd}
\text{Data} \arrow[dd] \arrow[dr] & \\
& \text{Categorical Reeb space} \\
\text{Categorical mapper} \arrow[r] & \text{Continuous categorical mapper}
\end{tikzcd}$$

The above diagram *does not* commute, but the main result of the paper [MW16] bounds how far this diagram is from being commutative.

References

[CMO18] Mathieu Carrière, Bertrand Michel, and Steve Oudot. Statistical analysis and parameter selection for mapper. *Journal of Machine Learning Research*, 19(12):1–39, 2018.

[CO17] Mathieu Carrière and Steve Oudot. Structure and stability of the one-dimensional mapper. *Foundations of Computational Mathematics*, 18(6):1333–1396, 2017.

[Coh12] Ben Cohen. Muthuball: How to build an NBA championship team - the mit sloan sports analytics conference, Jun 2012.

[DMW16] Tamal K. Dey, Facundo Mémoli, and Yusu Wang. *Multiscale Mapper: Topological Summarization via Codomain Covers*, pages 997–1013. 2016.

[DSK$^+$18] Bishal Deb, Ankita Sarkar, Nupur Kumari, Akash Rupela, Piyush Gupta, and Balaji Krishnamurthy. Multimapper: Data density sensitive topological visualization. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1054–1061, 2018.

[Hat02] Allen Hatcher. *Algebraic topology.* Cambridge University Press, 2002.

[MW16] E. Munch and Bei Wang. Convergence between categorical representations of reeb space and mapper. In *SoCG*, 2016.

[SMC07] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007.