Approximating Cycles in a Shortest Basis of the First Homology Group from Point Data

Tamal K. Dey[*] Jian Sun[†] Yusu Wang[‡]

Abstract

Inference of topological and geometric attributes of a hidden manifold from its point data is a fundamental problem arising in many scientific studies and engineering applications. In this paper we present an algorithm to compute a set of cycles from a point data that presumably sample a smooth manifold $M \subset \mathbb{R}^d$. These cycles approximate a *shortest* basis of the first homology group $\mathsf{H}_1(M)$ over coefficients in finite field $\mathbb{Z}_2$. Previous results addressed the issue of computing the rank of the homology groups from point data, but there is no result on approximating the shortest basis of a manifold from its point sample. In arriving our result, we also present a polynomial time algorithm for computing a shortest basis of $\mathsf{H}_1(\mathcal{K})$ for any finite *simplicial complex* $\mathcal{K}$ whose edges have non-negative weights.

1 Introduction

Inference of unknown structures from point data is a fundamental problem in many areas of science and engineering that has motivated wide spread research [1, 13, 23, 27, 28, 29]. Typically, this data is assumed to be sampled from a manifold sitting in a high dimensional space whose geometric and topological properties are to be derived from the data. In this work, we are particularly interested in computing a set of cycles from data which not only captures the topology but is also aware of the geometry of the sampled manifold. Specifically, we aim to approximate a shortest basis of the first homology group from the data.

Recently, a few algorithms for computing homology groups from point data have been developed. One approach would be to reconstruct the sampled space from its point data [4, 7, 12] and then apply known techniques for homology computations on triangulations [22, 24]. However, this option is not very attractive since a full-blown reconstruction with known techniques requires costly computations with Delaunay triangulations in high dimensions. Chazal and Oudot [8] showed how one can use less constrained data structures such as Rips, Čech, and witness complexes to infer the rank of the homology groups by leveraging persistence algorithms [20, 29]. Among these, the Rips complexes are the easiest to compute though they consume more space than the others, an issue which has started to be addressed [18].

All of the works mentioned above focus on computing the Betti numbers, the rank of the homology groups. Although the persistence algorithms [20, 29] also provide representative cycles of a homology basis, they remain oblivious to the geometry of the manifold. As a result, these cycles do not have nice geometric properties. A natural question is that, if the cycles of the first homology group are associated with a length under some metric, can one approximate/compute a shortest basis of the homology group in

[*] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA. Email: tamaldey@cse.ohio-state.edu

[†] Department of Computer Science, Stanford University, Palo Alto, CA 94305, USA. Current address: Mathematical Sciences Center, Tsinghua University, Bejing 100084 China, Email: jsun@math.tsinghua.edu.cn

[‡] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA. Email: yusu@cse.ohio-state.edu

polynomial time? This question has been answered in affirmative for the special case of surfaces when they are represented with triangulations [21]. In fact, considerable progress has been made for this special case on various versions of the problem. We cannot apply these techniques, mainly because we deal with point data instead of an input triangulation. Also, these works either consider a surface [5, 6, 15, 21] instead of a manifold of arbitrary dimension in an Euclidean space, or use a local measure other than the lengths of the cycles in a basis [9].

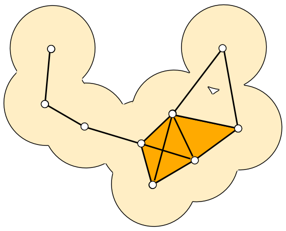
Our main result is an algorithm that can compute a set of cycles from a Rips complex of the given data with the guarantee that the lengths of the computed cycles approximate those of a shortest basis of the first homology group of the sampled manifold. In arriving at this result, we also show how to compute a shortest basis for the first homology group of any finite *simplicial complex* whose edges have non-negative weights. Given that computing a shortest basis for k-th homology groups of a simplicial complex over $\mathbb{Z}_2$ coefficients is NP-hard for $k \geq 2$ (Chen and Freedman [10]), this result settles the open case for $k = 1$.

1.1 Background and notations

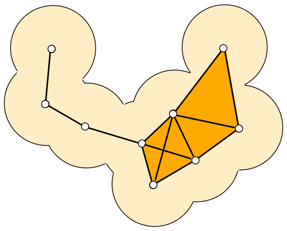
We use the concepts of homology groups, Čech and Rips complexes from algebraic topology and geodesics from differential geometry. We briefly discuss them and introduce relevant notations here; the readers can obtain the details from any standard book on the topics such as [19, 22].

Homology groups and basis: A homology group of a topological space $\mathbb{T}$ encodes its topological connectivity. We use $\mathsf{H}_i(\mathbb{T})$ to denote its i-th homology group over the coefficients in $\mathbb{Z}_2$. Since $\mathbb{Z}_2$ is a field, $\mathsf{H}_i(\mathbb{T})$ is a vector space and hence admits a basis. We are concerned with the first homology group $\mathsf{H}_1(\mathbb{T})$. The elements of $\mathsf{H}_1(\mathbb{T})$ are equivalence classes $[g]$ of 1-dimensional cycles g. A set $\{[g_1], \ldots, [g_k]\}$ generating $\mathsf{H}_1(\mathbb{T})$ is called its basis where $k = rank(\mathsf{H}_1(\mathbb{T}))$. We say $\{g_1, \ldots, g_k\}$ is a *homology cycle basis* of $\mathsf{H}_1(\mathbb{T})$ if $\{[g_1], \ldots, [g_k]\}$ is a basis of $\mathsf{H}_1(\mathbb{T})$. We also say that a set of cycles is *independent* in $\mathsf{H}_1(\mathbb{T})$ if their homology classes in $\mathsf{H}_1(\mathbb{T})$ are independent. By definition, a homology cycle basis is a maximally independent set.

We assume that each cycle g in $\mathbb{T}$ is associated with a non-negative weight $w(g)$. If $\mathbb{T}$ is a simplicial complex, the cycles are restricted to its 1-skeleton and $w(g)$ is defined to be the sum of edge weights in g which are assumed to be non-negative. If $\mathbb{T}$ is a Riemannian manifold, the weights on cycles are taken as their lengths in the Riemannian metric. The weights of the cycles define the length of a *set* of cycles $G = \{g_1, \ldots, g_k\}$ as $\mathrm{Len}(G) = \Sigma_{i=1}^{k} w(g_i)$. A *shortest basis* of $\mathsf{H}_1(\mathbb{T})$ is a homology cycle basis G of $\mathsf{H}_1(\mathbb{T})$ where $\mathrm{Len}(G)$ is minimal over all such bases. In applications, the weights could be the Euclidean lengths of the edges in which case a shortest basis would coincide with a set of cycles whose total Euclidean length is the smallest among all homology cycle bases.



Čech complex $\mathcal{C}^r(P)$



Rips complex $\mathcal{R}^r(P)$

Complexes: Let $B(p, r)$ denote an open Euclidean d-ball centered at p with radius r. For a point set $P \subset \mathbb{R}^d$, and a real $r > 0$, the Čech complex $\mathcal{C}^r(P)$ is a simplicial complex where a simplex $\sigma \in \mathcal{C}^r(P)$ if and only if $\mathrm{Vert}(\sigma)$, the vertices of σ, are in P and are the centers of d-balls of radius $r/2$ which have a non-empty common intersection, that is, $\cap_{p \in \mathrm{Vert}(\sigma)} B(p, r/2) \neq \emptyset$. Instead of common intersection, if we only require pairwise intersection among the d-balls, we get the Rips complex $\mathcal{R}^r(P)$. See the figure on right for an example, where the radius of each disk is $r/2$. Notice that the top-right triangle is in $\mathcal{R}^r(P)$ but not in $\mathcal{C}^r(P)$. It is well known that the two complexes are related by a nesting property:

Proposition 1.1 *For any finite set $P \subset \mathbb{R}^d$ and any $r \geq 0$, one has $\mathcal{C}^r(P) \subseteq \mathcal{R}^r(P) \subseteq \mathcal{C}^{2r}(P)$.*

Geodesics: The vertex set P of the simplicial complexes we consider is a dense sample of a smooth compact manifold $M \subset \mathbb{R}^d$ without boundary. Assume that M is isometrically embedded, that is, M inherits the metric from $\mathbb{R}^d$. For two points $p, q \in M$, a *geodesic* is a curve connecting p and q in M whose acceleration has no component in the tangent spaces of M. Two points may have more than one geodesic among which the ones with the minimum length are called *minimizing geodesics*. Since M is compact, any two points admit a minimizing geodesic. The lengths of minimizing geodesics induce a distance metric $d_M : M \times M \to \mathbb{R}$ where $d_M(p,q)$ is the length of a minimizing geodesic between p and q. Clearly, $d(p,q) \leq d_M(p,q)$ where $d(p,q)$ is the Euclidean distance. If $d(p,q)$ is small, Proposition 1.2 asserts that there is an upper bound on $d_M(p,q)$ in terms of $d(p,q)$. Our proof extends a result in [2] where Belkin et al. show the same result on a surface in $\mathbb{R}^3$. The *reach* $\rho(M)$ is defined as the minimum distance between M and its medial axis [16].

Proposition 1.2 *If $d(p,q) \leq \rho(M)/2$, one has*

$$d_M(p,q) \leq (1 + \frac{4d^2(p,q)}{3\rho^2(M)})d(p,q).$$

Proof: Let $\gamma(t)$ be a minimizing geodesic between p and q parameterized by length and set $l = d_M(p,q)$. By Proposition 6.3 in [27] we have that $l \leq 2d(p,q)$. Let $u_t = \dot{\gamma}(t)$ be the *unit* tangent vector of γ at t. We have $t = d_M(p, \gamma(t))$.

Let $B : T_{\gamma(t)} \times T_{\gamma(t)} \to T^{\perp}_{\gamma(t)}$ be the second fundamental form associated with the manifold M. Since γ is a geodesic, $du_t/dt = B(u_t, u_t) = \ddot{\gamma}(t)$. Write $\rho = \rho(M)$ and $d = d(p,q)$ for convenience. From Proposition 6.1 in [27], we have

$$\|\ddot{\gamma}(t)\| \leq 1/\rho$$

since the norm of the second fundamental form is bounded by $1/\rho$ in all directions, and thus $\|du_t/dt\| \leq 1/\rho$. Hence we have that

$$\begin{aligned} \|u_t - u_p\| &= \|\int_{[0,t]} du_y\| &\leq& \int_{[0,t]} \frac{1}{\rho} dy = \frac{t}{\rho} \\ \Rightarrow \quad \sin\frac{\angle(u_p, u_t)}{2} && \leq& \frac{t}{2\rho}. \end{aligned}$$

Furthermore, let $u \cdot v$ denote the scalar-product between vectors u and v. Then we have that

$$\begin{aligned} \int_{[0,l]} u_t \cdot u_p \, dt &= \int_{[0,l]} \cos\angle(u_t, u_p) \, dt = \int_{[0,l]} (1 - 2\sin^2\frac{\angle(u_t, u_p)}{2})dt \\ &\geq \int_{[0,l]} \left(1 - \frac{t^2}{2\rho^2}\right) dt = l - \frac{l^3}{6\rho^2} \end{aligned}$$

On the other hand, observe that $\int_{[0,l]} u_t \cdot u_p \, dt$ measures the length of the (signed) projection of γ along the direction u_p. That is,

$$\int_{[0,l]} u_t \cdot u_p \, dl_t = (q - p) \cdot u_p.$$

Hence we have that

$$\begin{aligned} & d = \|p-q\| \geq (q-p)\cdot u_p \geq l - \tfrac{l^3}{6\rho^2} \\ \Rightarrow \qquad & l \;\leq\; d + \tfrac{l^3}{6\rho^2} \;\leq\; d + \tfrac{4d^3}{3\rho^2}\ . \end{aligned}$$

The last inequality follows from the fact that $l \leq 2d$. This proves the lemma. ■

Convexity radius and sampling: For a point $p \in M$, the set of all points q with $d_M(p,q) < r$ forms p's *geodesic ball* $B_M(p,r)$ of radius r. It is known that there is a positive real r_p for each point $p \in M$ so that $B_M(p,r)$ is *convex* for $r \leq r_p$. It means that, for $r \leq r_p$, any two points in $B_M(p,r)$ admit a *unique* minimizing geodesic that lies in $B_M(p,r)$. The *convexity radius* of M is $\rho_c(M) = \inf_{p\in M} r_p$. Intuitively, a geodesic ball centered anywhere on M with convexity radius is guaranteed to be convex in the sense that any two points within it has a unique minimizing geodesic. We use Euclidean distances to define the sampling density. We say a discrete set $P \subset M$ is an ε-sample[1] of M if $B(x,\varepsilon) \cap P \neq \emptyset$ for each point $x \in M$.

1.2 Main results

We compute a set of cycles $G = \{g_1, \ldots, g_k\}$ from an ε-sample P of M whose total length, denoted $\mathrm{Len}(G)$, is within a factor of the total length of a shortest basis in $\mathsf{H}_1(M)$. Recall that the length of a cycle g in M is defined as the length of g in the Riemannian metric associated with M. The factor depends on ε, $\rho(M)$, and an input parameter $r > 0$.

Theorem 1.3 *Let $M \subset \mathbb{R}^d$ be a smooth, closed manifold with ℓ as the length of a shortest basis of $\mathsf{H}_1(M)$. Given an ε-sample $P \subset M$ of n points and $4\varepsilon \leq r \leq \min\{\frac{1}{2}\sqrt{\frac{3}{5}}\rho(M), \rho_c(M)\}$, one can compute a set of cycles G in $\mathbb{R}^d$ where:*

i.
$$\frac{1}{1+\frac{4r^2}{3\rho^2(M)}}\ell \leq \mathrm{Len}(G) \leq (1+\frac{4\varepsilon}{r})\ell.$$

ii. Treating G as a 1-complex, there is a map $h\colon G \to M$ so that $h(G)$ is a homology cycle basis of $\mathsf{H}_1(M)$ and the Hausdorff distance between the underlying space of g and $h(g)$ is at most $r/2$ for each $g \in G$.

iii. The cycles in G can be computed in $O(n(n+n_e)^2(n_e+n_t))$ time where n_e and n_t are the numbers of edges and triangles respectively in the Rips complex $\mathcal{R}^{2r}(P)$.

The above result suggests that $\lim_{\frac{\varepsilon}{r}, r\to 0} \mathrm{Len}(G) \to \ell$. To make $\frac{\varepsilon}{r}$ and r simultaneously approach 0, one may take $r = O(\sqrt{\varepsilon})$ and let $\varepsilon \to 0$. We note that $n_e = O(n^2)$ and $n_t = O(n^3)$ giving an $O(n^8)$ worst-case complexity for the algorithm. However, if $r = \Theta(\varepsilon)$ and points in P have $\Omega(\varepsilon)$ pairwise distance, n_e and n_t reduce to $O(n)$ by a result of [8]. In this case we get a time complexity of $O(n^4)$. In arriving at Theorem 1.3, we also prove the following result which is of independent interest.

Theorem 1.4 *Let $\mathcal{K}$ be a finite simplicial complex with non-negative weights on edges. A shortest basis for $\mathsf{H}_1(\mathcal{K})$ can be computed in $O(n^4)$ time where n is the size of $\mathcal{K}$.*

[1] Here ε-sample is not defined relative to reach or feature size as commonly done in reconstruction literature [1, 7, 12].

2 Algorithm description

The algorithm that we propose proceeds as follows. We compute a Rips complex $\mathcal{R}^{2r}(P)$ out of the given point cloud $P \subset M$. Next, we compute the rank k of $\mathsf{H}_1(M)$ by considering the persistent homology group

$$\mathsf{H}_1^{r,2r}(\mathcal{R}(P)) = \text{ image } \iota_*$$

where the inclusion $\iota : \mathcal{R}^r(P) \hookrightarrow \mathcal{R}^{2r}(P)$ induces the homomorphism $\iota_* : \mathsf{H}_1(\mathcal{R}^r(P)) \to \mathsf{H}_1(\mathcal{R}^{2r}(P))$. As a homology group over $\mathbb{Z}_2$, $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ is a vector space and the rank of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ coincides with that of $\mathsf{H}_1(M)$ for appropriate r, see Proposition 3.5.

A basis of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ is formed by the classes of a maximal set of cycles in $\mathcal{R}^r(P)$ whose classes remain independent in $\mathsf{H}_1(\mathcal{R}^{2r}(P))$ under the map ι_*. We show that a shortest basis of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ approximates a shortest basis of $\mathsf{H}_1(M)$. Therefore, we aim to compute a shortest basis of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ from $\mathcal{R}^r(P)$ and $\mathcal{R}^{2r}(P)$. To accomplish this, the algorithm augments $\mathcal{R}^{2r}(P)$ by putting a weight $w(e)$ on each edge $e \in \mathcal{R}^{2r}(P)$. The weights are of two types: either they are the lengths of the edges, or a very large value W which is larger than k times the total weight of $\mathcal{R}^r(P)$. Precisely we set

$$w(e) = \begin{cases} \text{length of } e & \text{if } e \in \mathcal{R}^r(P) \\ W & \text{if } e \in \mathcal{R}^{2r}(P) \setminus \mathcal{R}^r(P). \end{cases}$$

Let the complex $\mathcal{R}^{2r}(P)$ augmented with weights be denoted as $\mathcal{R}^{2r+}(P)$. A shortest basis of $\mathsf{H}_1(\mathcal{R}^{2r+}(P))$ does not necessarily form a shortest basis of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$. However, the first k cycles sorted according to lengths in a shortest basis of $\mathsf{H}_1(\mathcal{R}^{2r+}(P))$ form a shortest basis of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$. We give an algorithm to compute a shortest basis for any simplicial complex which we apply to $\mathcal{R}^{2r+}(P)$.

Since we are interested in computing a homology cycle basis of the first homology group, it is sufficient to consider all simplices up to dimension two, that is, only vertices, edges, and triangles in the simplicial complexes that we deal with. Henceforth, we assume that all complexes that we consider have simplices up to dimension two.

2.1 Computing cycles

We will prove later that a shortest basis for $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ indeed approximates a shortest basis for $\mathsf{H}_1(M)$. The algorithm SHORTCYCLE computes them.

Algorithm 1 SHORTCYCLE (P, r)

```
1: Compute the Rips complex $\mathcal{R}^{2r}(P)$ and a weighted complex $\mathcal{R}^{2r+}(P)$ from it as described.
2: Compute the rank $k$ of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ by the persistence algorithm.
3: Compute a shortest basis for $\mathsf{H}_1(\mathcal{R}^{2r+}(P))$.
4: Return the first $k$ smallest cycles from this shortest basis.
```

Theorem 2.1 *The algorithm* SHORTCYCLE(P, r) *computes a shortest basis for the persistent homology group* $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$.

Proof: Let $g_1, \ldots, g_a$ be the set of cycles sorted according to the non-decreasing lengths which are computed in step 3. They form a homology cycle basis of $\mathsf{H}_1(\mathcal{R}^{2r+}(P))$. Out of these cycles the algorithm outputs the first k cycles $g_1, \ldots, g_k$. Since k is the rank of $\mathsf{H}_1^{r,2r}(P)$ there are k independent cycles in $\mathsf{H}_1(\mathcal{R}^r(P))$ which remain independent in $\mathsf{H}_1(\mathcal{R}^{2r+}(P))$. We claim that the cycles $g_1, \ldots, g_k$ reside in $\mathcal{R}^r(P)$. For if they do not, the sum of their lengths would be more than W which is k times larger than the total weight of $\mathcal{R}^r(P)$. Then, we can argue that any independent set of k cycles from $\mathcal{R}^r(P)$ which remain independent in $\mathsf{H}_1(\mathcal{R}^{2r+}(P))$ can replace $g_1, \ldots, g_k$ to have a smaller length so that $g_1, \ldots, g_a$ could not be a shortest basis of $\mathsf{H}_1(\mathcal{R}^{2r+}(P))$.

The above argument implies that $g_1, \ldots, g_k$ is a homology cycle basis of $\mathsf{H}_1^{r,2r}(P)$. If it is not a shortest basis, it can be replaced by a shorter one so that again we would have a homology cycle basis of $\mathsf{H}_1(\mathcal{R}^{2r+}(P))$ which is shorter than the one computed. This is a contradiction. ■

It remains to show how to compute a shortest basis of $\mathsf{H}_1(\mathcal{R}^{2r+}(P))$ in step 3 of SHORTCYCLE.

2.2 Shortest basis

Let $\mathcal{K}$ be any finite simplicial complex embedded in $\mathbb{R}^d$ whose edges have non-negative weights. To compute a shortest basis for $\mathsf{H}_1(\mathcal{K})$ we make use of the fact that $\mathsf{H}_1(\mathcal{K})$ is a vector space as we restrict ourselves to $\mathbb{Z}_2$ coefficients. For such cases, Erickson and Whittlesey [21] observed that if a set of cycles $\mathcal{L}$ in $\mathcal{K}$ contains a shortest basis, then the greedy set G chosen from $\mathcal{L}$ is a shortest basis. The greedy set G of $\mathcal{L}$ is an *ordered* set of cycles $\{g_1, \ldots g_k\}$, $k = \mathrm{rank}\, \mathsf{H}_1(\mathcal{K})$, satisfying the following condition. The first element g_1 is the shortest cycle in $\mathcal{L}$ which is nontrivial in $\mathsf{H}_1(\mathcal{K})$. Suppose $g_1, \ldots, g_i$ have already been defined in the set G. The next chosen cycle g_{i+1} is the shortest cycle in $\mathcal{L}$ which is independent of $g_1, \ldots, g_i$, that is, $[g_{i+1}]$ cannot be written as a linear combination of $[g_1]$, ..., $[g_i]$. The check for independence is a costly step in this greedy algorithm which we aim to reduce. We construct a set of *canonical* cycles which contains a homology cycle basis of $\mathsf{H}_1(\mathcal{K})$. This set is pruned by a persistence based algorithm before applying the greedy algorithm.

2.2.1 Canonical cycles

We start with citing a result of Erickson and Whittlesey [21]. A simple cycle L is *tight* if it contains a shortest path between every pair of points in L.

Proposition 2.2 *With non-negative weights, every cycle in a shortest basis of* $\mathsf{H}_1(\mathcal{K})$ *is tight.*

To collect all tight cycles, we consider the canonical cycles defined as follows. Let T be a *shortest path tree* in $\mathcal{K}$ rooted at p. Notice that we are not assuming T to be unique, but it is fixed once computed. For any two nodes $q_1, q_2 \in P$, let $\Pi_T(q_1, q_2)$ denote the unique path from q_1 to q_2 in T. Let E_T be the set of edges in T. Given a non-tree edge $e = (q_1, q_2) \in E \setminus E_T$, define the *canonical cycle* of e with respect to p, $c_p(e)$ in short, as the cycle formed by concatenating $\Pi_T(q_1, q_2)$ and e, that is,

$$c_p(e) = \Pi_T(q_1, q_2) \circ e.$$

Let C_p be the set of all canonical cycles with respect to p, i.e., $C_p = \{c_p(e) : e \in E \setminus E_T\}$. Then we have the following easy consequence.

Proposition 2.3 $\cup_{p \in P} C_p$ *contains all tight cycles.*

For convenience, we treat $\cup_{p \in P} C_p$ as a multiset, that is, a cycle appears as many times as it is considered a canonical cycle for a point in P. The arguments and the algorithms to follow can easily be modified to eliminate this assumption. By Proposition 2.3, $\cup_{p \in P} C_p$ is a set of cycles from which the greedy set can be selected. However, $\cup_{p \in P} C_p$ can be a very large set containing possibly many trivial cycles which result into many unnecessary independence checks. To remedy this, we identify the greedy set G_p of C_p and choose the greedy set from the union $\cup_{p \in P} G_p$ instead of $\cup_{p \in P} C_p$. We shall show that that G_p can be computed by a persistence based algorithm thereby avoiding explicit independence checks.

If the lengths of the cycles in C_p are distinct, the greedy set G_p is unique. However, in presence of equal length cycles we need a mechanism to break ties. For this we introduce the notion of *canonical order*. We assign a unique number $\nu(e)$ between 1 to m to each non-tree edge e if there are m of them. For any two non-tree edges e and e', let $e < e'$ if and only if either $\mathrm{Len}(c_p(e)) < \mathrm{Len}(c_p(e'))$, or $\mathrm{Len}(c_p(e)) = \mathrm{Len}(c_p(e'))$ and $\nu(e) < \nu(e')$. The total order imposed by '$<$' provides the canonical order

$$e_1 < e_2 < \ldots < e_m.$$

Based on this canonical order, we form the greedy set G_p of C_p as described in the beginning of Section 2.2.

Below we argue that $\cup_{p \in P} G_p$ is good for our purpose and each set G_p can be computed based on the persistence algorithm. Again, we treat $\cup_{p \in P} G_p$ as a multiset for convenience.

Proposition 2.4 *The greedy set chosen from* $\cup_{p \in P} G_p$ *is a shortest basis of* $\mathsf{H}_1(\mathcal{K})$.

Proof: We show that $\cup_{p \in P} G_p$ contains a shortest basis of $\mathsf{H}_1(\mathcal{K})$. Then, the proposition follows by the argument as delineated at the beginning of section 2.2.

Consider all canonical cycles $\cup_{p \in P} C_p$. Sort them in non-decreasing order of their lengths. If two cycles have equal lengths and if there are points $p_i \in P$ for which both of them are in C_{p_i}, break the tie using the canonical order applied to the canonical cycles for any such one point. Otherwise, break the tie arbitrarily. Based on this order let G be the greedy set from $\cup_{p \in P} C_p$. Proposition 2.2 and Proposition 2.3 imply that $\cup_{p \in P} C_p$ contains a shortest basis of $\mathsf{H}_1(\mathcal{K})$ and thus G is a shortest basis. Consider any cycle L in G. It is a canonical cycle with respect to some $q \in P$ for which all cycles appearing before L in the canonical order precede it in the sorted sequence. The cycle L is independent of the cycles in $\cup_{p \in P} C_p$ appearing before L, in particular independent of the cycles in C_q appearing before L in the canonical order, which means $L \in G_q$. Therefore $\cup_{p \in P} G_p$ contains a shortest basis G of $\mathsf{H}_1(\mathcal{K})$. The proposition follows. ■

Motivated by the above observations, we formulate an algorithm CANONGEN that computes the greedy set G_p of C_p. We note that, very recently, Chen and Freedman [9] proposed a similar algorithm which computes an *approximation* of a shortest basis of a simplicial complex rather than an optimal one.

Algorithm 2 CANONGEN (p, $\mathcal{K}$)

```
Construct a shortest path tree $T$ in $\mathcal{K}$ with $p$ as the root. Let $E_T$ denote the set of tree edges.
For each non-tree edge $e = (q_1, q_2) \in E \setminus E_T$, let $c_p(e)$ be the canonical cycle of $e$.
Perform the persistence algorithm based on the following filtration of $\mathcal{K}$: all the vertices in $P = \mathrm{Vert}(\mathcal{K})$, followed by all tree edges in $T$, followed by non-tree edges in the canonical order, and followed by all the triangles in $\mathcal{K}$. There are $k = \mathrm{rank}(H_1(\mathcal{K}))$ number of edges unpaired after the algorithm, and each of them is necessarily a non-tree edge. Return the set of canonical cycles associated with them.
```

Proposition 2.5 CANONGEN $(p, \mathcal{K})$ *outputs the greedy set* G_p *chosen from* C_p.

Proof: Let $\{e_1, e_2 \cdots, e_m\}$ be the set of non-tree edges for the shortest path tree T listed in the canonical order. Let

$$G_p = \{c_p(e_1^*), c_p(e_2^*), \cdots, c_p(e_k^*)\}.$$

It suffices to show that $\{e_1^*, e_2^* \cdots, e_k^*\}$ is the set of unpaired edges. Observe that for any e_i^*, $c_p(e_i^*)$ is independent of any subset of $\{c_p(e_j) : e_j < e_i^*\}$.

We prove the proposition by contradiction. Assume some e_i^* gets paired by a triangle t in the persistence algorithm. Let $\mathcal{K}_t$ denote the complex in the filtration right before t is added. Let $f : \mathcal{K}_t \hookrightarrow \mathcal{K}$ be the inclusion map; it induces a homomorphism $f_* = \mathsf{H}_1(\mathcal{K}_t) \rightarrow \mathsf{H}_1(\mathcal{K})$. Let $[L]_t$ denote the homology class in $\mathcal{K}_t$ carried by the cycle L. The boundary ∂t uniquely determines a subset of unpaired positive edges

$e'_1 < \cdots < e'_s$ in $\mathcal{K}_t$ such that $[\partial t]_t = [c_p(e'_1)]_t + \cdots + [c_p(e'_s)]_t$. The persistence algorithm [20] picks the youngest one from this subset to pair with t, i.e., $e^*_i = e'_s$. On the other hand, we have

$$\begin{aligned}&[c_p(e'_1)] + \cdots + [c_p(e'_{s-1})] + [c_p(e^*_i)]\\ &= f_*([c_p(e'_1)]_t + \cdots + [c_p(e'_{s-1})]_t + [c_p(e^*_i)]_t)\\ &= f_*([\partial t]_t) = 0\end{aligned}$$

which means that $c_p(e^*_i)$ is dependent on a subset of $\{c_p(e_j) : e_j < e^*_i\}$. We reach a contradiction. ■

All previous results put together provide a greedy algorithm for computing a shortest basis of $\mathsf{H}_1(\mathcal{K})$.

Algorithm 3 SPGEN ($\mathcal{K}$)

```
1:  For each $p \in P = \mathrm{Vert}(\mathcal{K})$ compute $G_p :=$CANONGEN ($p$,$\mathcal{K}$). Let $k = |G_p|$.
2:  Sort all cycles in $\cup_p G_p$ by their lengths in the increasing order. Let $g_1, \ldots, g_{k|P|}$ be this sorted list.
3:  Initialize $G := \{g_1\}$.
4:  for $i := 2$ to $k|P|$, do
5:      if $|G| = k$, then
6:          Exit the for loop.
7:      else if $g_i$ is independent of all cycles in $G$, then
8:          Add $g_i$ to $G$.
9:      end if
10: end for
11: Return $G$.
```

2.2.2 Checking independence

In step 7 of SPGEN we need to determine if a cycle g is independent of all cycles $g'_1, \ldots, g'_s$ so far selected in G. Suppose we obtain g from running persistence algorithm on a shortest path tree based filtration for a point p in step 3 of CANONGEN. At the end of this persistence algorithm we must have gotten an unpaired edge, say e, where $c_p(e) = g$. To determine if g is independent of all cycles selected so far we adopt a sealing technique proposed in [9]. We fill $g'_1 \ldots g'_s$ with triangles. The filling is done only combinatorially by choosing a dummy vertex, say v, and adding triangles vv_iv_{i+1} for each edge v_iv_{i+1} of the cycles to be filled. Let $\mathcal{K}'$ be the new complex after adding these triangles and their edges to $\mathcal{K}$. In effect, these triangles and edges make the cycles $g'_1, \ldots, g'_s$ trivial in $\mathsf{H}_1(\mathcal{K})$. They make the cycle g trivial as well if and only if g is dependent on $g'_1, \ldots, g'_s$. Since we are sealing according to the greedy order, the proof of Lemma 4.4 in [9] applies to establish this fact. Whether g is rendered trivial or not can be determined as follows. We continue the persistence algorithm corresponding to the vertex p with the addition of the simplices in $\mathcal{K}' \setminus \mathcal{K}$ and check if e is now paired or not.

Let n_v, n_e, and n_t denote the number of vertices, edges, and triangles respectively in $\mathcal{K}$. Notice that we add at most n_e edges and triangles for sealing since the dummy vertex is added to at most n_e edges to create new triangles in $\mathcal{K}'$.

2.3 Time complexity

First, we analyze the time complexity of CANONGEN. Shortest path tree computation in step 1 of CANONGEN takes $O(n_v \log n_v + n_e)$ time. The persistence algorithm for CANONGEN can be implemented using matrix reductions [14] in time $O((n_v + n_e)^2(n_e + n_t))$. This is because there are $n_v + n_e$ rows in this matrix and each insertion of $n_e + n_t$ simplices can be implemented in $O(n_v + n_e)$ column operations each taking $O(n_v + n_e)$ time. Therefore, CANONGEN takes $O(n_v \log n_v + (n_v + n_e)^2(n_e + n_t))$ time.

Step 1 of SPGEN calls CANONGEN n_v times. Therefore, step 1 of SPGEN takes $O(n_v^2 \log n_v + n_v(n_v + n_e)^2(n_e + n_t))$ time. Step 2 of SPGEN can be performed in $O(n_v k \log n_v k)$ time where $k = O(n_e)$ is the rank of $\mathsf{H}_1(\mathcal{K})$. The time complexity for independence check in step 7 is dominated by the persistence algorithm which is continued on $\mathcal{K}$ to accommodate simplices in $\mathcal{K}'$. Since we add $O(n_e)$ new simplices in $\mathcal{K}'$, it has the same asymptotic complexity as for running the persistence algorithm on $\mathcal{K}$. We conclude that SPGEN spends $O(n_v(n_v + n_e)^2(n_e + n_t))$ time in total. If we take $n = |\mathcal{K}|$, this gives an $O(n^4)$ time complexity.

Now, we analyze the time complexity of SHORTCYCLE which is the main algorithm. Let n_e and n_t be the number of edges and triangles in $\mathcal{R}^{2r}(P)$ created out of n points. Step 1 takes at most $O(n + n_e + n_t)$ time since we only compute edges and triangles of $\mathcal{R}^{2r}(P)$ out of n points. Accounting for the persistence algorithm in step 2 and the time complexity of step 3 we get that SHORTCYCLE takes

$$O(n(n + n_e)^2(n_e + n_t)) \text{ time.}$$

The procedure SPGEN($\mathcal{K}$) computes canonical sets G_p which is ensured by Proposition 2.5. Then, it forms a greedy set from these canonical sets which is a shortest basis for $\mathsf{H}_1(\mathcal{K})$ by Proposition 2.4. This and the time analysis for SPGEN establish Theorem 1.4.

3 Approximation for M

The algorithm SPGEN is used in SHORTCYCLE to produce a shortest basis for the persistent homology group $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$. Proposition 3.5 in this section shows that a shortest basis of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ coincides with a shortest basis in $\mathsf{H}_1(\mathcal{C}^r(P))$. Therefore, if we show that a shortest basis in $\mathsf{H}_1(\mathcal{C}^r(P))$ approximates a shortest basis in $\mathsf{H}_1(M)$, we have the approximation result of Theorem 1.3.

3.1 Connecting M, Čech complex, and Rips complex

First, we note the following result established in [27] which connects M with the union of the balls $P^r = \cup_{p \in P} B(p, r)$.

Proposition 3.1 *Let $P \subset M$ be an ε-sample. If $2\varepsilon \le r \le \sqrt{\frac{3}{5}}\rho(M)$, there is a deformation retraction from P^r to M so that the corresponding retraction $t : P^r \to M$ has $t(B) \subset B$ for any ball $B \in \{B(p, r)\}_{p \in P}$.*

Recall that $\mathcal{C}^{2r}(P)$ is the nerve of the cover $\{B(p, r)\}_{p \in P}$ of the space P^r. By a result of Leray [25], it is known that P^r and $\mathcal{C}^{2r}(P)$ are homotopy equivalent. The next proposition follows from examining the specific equivalence maps used to prove the Nerve Lemma in Hatcher [22]. In particular, the simplices of the Čech complex are mapped to a subset of the union of the balls centered at their vertices, see Appendix for its proof.

Proposition 3.2 *There exists a homotopy equivalence $f \colon \mathcal{C}^{2r}(P) \to P^r$ such that for each simplex $\sigma \in \mathcal{C}^{2r}(P)$, one has $f(\sigma) \subset \cup_{p \in \mathrm{Vert}(\sigma)} B(p, r)$ and $f(p) = p$ for any $p \in P$.*

The two propositions above together provide the connection between M and the Čech complex:

Proposition 3.3 *Let $P \subset M$ be an ε-sample. If $2\varepsilon \le r \le \sqrt{\frac{3}{5}}\rho(M)$, there is a homotopy equivalence map $h = t \circ f : \mathcal{C}^{2r}(P) \to M$ such that $h(\sigma) \subset M \cap (\cup_{p \in \mathrm{Vert}(\sigma)} B(p, r))$ and $h(p) = p$ for any $p \in P$.*

Now we establish a connection between Čech complex and Rips complexes which helps proving Proposition 3.5.

Proposition 3.4 *Let $P \subset M$ be an ε-sample. Then, for $4\varepsilon \leq r \leq \frac{1}{2}\sqrt{\frac{3}{5}}\rho(M)$, we have the following isomorphisms*

$$\mathsf{H}_1^{r,2r}(\mathcal{R}(P)) \approx \mathsf{H}_1(\mathcal{C}^r(P)) \overset{j_{1*}}{\approx} \mathsf{H}_1(\mathcal{C}^{2r}(P)) \overset{j_{2*}}{\approx} \mathsf{H}_1(\mathcal{C}^{4r}(P)),$$

where j_{1*} and j_{2*} are induced by the inclusion maps j_1 and j_2 respectively. Moreover, if

$$\mathcal{C}^r(P) \overset{i_1}{\hookrightarrow} \mathcal{R}^r(P)) \overset{i_2}{\hookrightarrow} \mathcal{C}^{2r}(P)) \overset{i_3}{\hookrightarrow} \mathcal{R}^{2r}(P)) \overset{i_4}{\hookrightarrow} \mathcal{C}^{4r}(P),$$

then $j_1 = i_2 \circ i_1$, and $j_2 = i_4 \circ i_3$ and $\mathsf{H}_1^{r,2r}(\mathcal{R}(P)) = image\ (\iota_*)$ where $\iota_* : \mathsf{H}_1(\mathcal{R}^r(P)) \to \mathsf{H}_1(\mathcal{R}^{2r}(P))$ is induced by the inclusion $\iota = i_3 \circ i_2$.

Proof: Based on Proposition 3.3, it can be proved by following the idea in [8] of intertwined Čech and Rips complexes. ■

By definition the set of edges in $\mathcal{C}^r(P)$ is same as the set of edges in $\mathcal{R}^r(P)$. This means a set of cycles in $\mathcal{R}^r(P)$ also forms a set of cycles in $\mathcal{C}^r(P)$. In light of Proposition 3.4, this implies:

Proposition 3.5 *Let $P \subset M$ be an ε-sample and $4\varepsilon \leq r \leq \frac{1}{2}\sqrt{\frac{3}{5}}\rho(M)$. Then $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ and $\mathsf{H}_1(M)$ are isomorphic and a basis for $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ is shortest if and only if it is shortest for $\mathsf{H}_1(\mathcal{C}^r(P))$.*

Proof: From Proposition 3.3 and Proposition 3.4, we have the following isomorphisms:

$$\mathsf{H}_1^{r,2r}(\mathcal{R}(P)) \approx \mathsf{H}_1(\mathcal{C}^r(P)) \approx \mathsf{H}_1(M).$$

Let $A = \{a_1, \cdots, a_k\}$ be a shortest basis for $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$. Each a_i is a cycle in $\mathcal{R}^r(P)$ and hence in $\mathcal{C}^r(P)$. Obviously A is a homology cycle basis of $\mathsf{H}_1(\mathcal{C}^r(P))$ as the inclusion map from $\mathcal{C}^r(P)$ to $\mathcal{R}^r(P)$ induces a homomorphism. Thus, a shortest basis for $\mathsf{H}_1(\mathcal{C}^r(P))$ must be no longer than that of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$. Similarly if $A = \{a_1, \cdots, a_k\}$ is a shortest basis of $\mathsf{H}_1(\mathcal{C}^r(P))$, then each a_i must be in $\mathcal{R}^r(P)$ and survive in $\mathcal{R}^{2r}(P)$ as it must survive in $\mathcal{C}^{4r}(P)$. Thus A is a homology cycle basis for $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ and hence a shortest basis of $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ is no longer than that of $\mathsf{H}_1(\mathcal{C}^r(P))$. This proves the proposition. ■

3.2 Bounding the lengths

Our idea is to argue that a shortest basis of $\mathsf{H}_1(\mathcal{C}^r(P))$ can be mapped to a homology cycle basis of $\mathsf{H}_1(M)$ by the map h of Proposition 3.3. We argue that the lengths of the homology cycle basis cannot change too much in the process.

Let g be any closed curve in M. Following [3], we define a procedure to approximate g by a cycle $\hat{g}$ in the 1-skeleton of $\mathcal{C}^r(P)$. This procedure called *Decomposition method* is not part of our algorithm, but is used in our argument about length approximations of cycles in M.

Decomposition method If $\ell = \mathrm{Len}(g) > r - 2\varepsilon > 0$, we can write $\ell = \ell_0 + (\ell_1 + \ell_1 + \ldots + \ell_1) + \ell_0$ where $\ell_1 = r - 2\varepsilon$ and $r - 2\varepsilon > \ell_0 \geq (r - 2\varepsilon)/2$. Starting from an arbitrary point, say x, split g into pieces whose lengths coincide with the decomposition of ℓ. This produces a sequence of points $x = x_0, x_1, \ldots, x_m = x$ along g which divide it according to the lengths constraints. Because of our sampling condition, each point x_i has a point $p_i \in P$ within ε distance. We define a cycle $\hat{g} = \{p_0 p_1 \ldots p_m\}$ with consecutive points joined by line segments. Proposition 3.6 shows that $\hat{g}$ resides in the 1-skeleton of $\mathcal{C}^r(P)$.

Proposition 3.6 *Given a closed curve g on M with* $\mathrm{Len}(g) > r - 2\varepsilon > 0$, Decomposition method *finds a cycle $\hat{g}$ from the* 1*-skeleton of $\mathcal{C}^r(P)$ such that:* $\mathrm{Len}(\hat{g}) \leq \frac{r}{r-2\varepsilon}\mathrm{Len}(g)$.

Proof: From the construction and sampling condition, it follows that, for $1 \leq i \leq m-2$,

$$\begin{aligned} d(p_i, p_{i+1}) &\leq d(x_i, p_i) + d(x_i, x_{i+1}) + d(x_{i+1}, p_{i+1}) \\ &< 2\varepsilon + \ell_1 = r = \frac{r}{(r-2\varepsilon)}\ell_1 \end{aligned}$$

Similarly,

$$d(p_0, p_1) \leq \frac{r}{r-2\varepsilon}\ell_0 \text{ and } d(p_{m-1}, p_0) \leq \frac{r}{r-2\varepsilon}\ell_0.$$

Since $\frac{r}{r-2\varepsilon}\ell_0 < r$, each edge p_ip_{i+1} belongs to $\mathcal{C}^r(P)$. Therefore, we obtain a cycle $\hat{g} = p_0p_1 \ldots p_m$ in the 1-skeleton of $\mathcal{C}^r(P)$ whose length satisfies:

$$\mathrm{Len}(\hat{g}) = \Sigma_{i=0}^{m-1} d(p_i, p_{i+1}) \leq \frac{r}{r-2\varepsilon}\mathrm{Len}(g).$$

■

Consider a homology cycle basis of $\mathsf{H}_1(M)$ where each cycle is a closed geodesic on M. For a smooth, compact manifold such a basis always exists by a well known result in differential geometry [19]. Let $G = \{g_1, \ldots, g_k\}$ be this set of geodesic cycles. By Proposition 3.6, we claim that there is a set of cycles $\hat{G} = \{\hat{g}_1, \ldots, \hat{g}_k\}$ in $\mathcal{C}^r(P)$ whose length is within a small factor of the length of G. However, we need to show that $\hat{G}$ indeed a homology cycle basis of $\mathsf{H}_1(\mathcal{C}^r(P))$. We show this by mapping each $\hat{g}_j \in \hat{G}$ to M by the homotopy equivalence h (Proposition 3.3) and arguing that $[h(\hat{g}_j)] = [g_j]$ in $\mathsf{H}_1(M)$. Since h is a homotopy equivalence map, it follows that the isomorphism $h_* : \mathsf{H}_1(\mathcal{C}^r(P)) \to \mathsf{H}_1(M)$ maps the class $[\hat{g}_j]$ to $[g_j]$. This implies that $\hat{G}$ is a homology cycle basis of $\mathsf{H}_1(\mathcal{C}^r(P))$.

To prove that $h(\hat{g}_j)$ is a representative of the class $[g_j]$, we consider a tubular neighborhood of g_j of radius r which is smaller than the convexity radius $\rho_c(M)$. Then, we show that each segment p_ip_{i+1} of $\hat{g}_j$ is mapped to a curve $h(p_ip_{i+1})$ which lies within this tubular neighborhood. Because of this containment, $h(p_ip_{i+1})$ must be homotopic to a geodesic segment of g_j. All these homotopies together provide a homotopy between $h(g_j)$ and g_j. First we show that the tubular neighborhood of a segment of g_j that we consider is indeed simply connected.

Proposition 3.7 *Let $\gamma = \gamma(p, q)$ be a minimizing geodesic between two points $p, q \in M$. Consider its tubular neighborhood $\mathrm{Tub}_s(\gamma)$ on M that consists of the points on M within a geodesic distance s from γ, i.e., $\mathrm{Tub}_s(\gamma) = \{x \in M : \min_{y\in\gamma} d_M(x, y) < s\}$. Then if $s < \rho_c(M)$, $\mathrm{Tub}_s(\gamma)$ is contractible, in particular, $\mathrm{Tub}_s(\gamma)$ is simply connected.*

Proof: We show that $\mathrm{Tub}_s(\gamma)$ deformation retracts to γ. For any point $x \in \mathrm{Tub}_s(\gamma)$, consider an open geodesic ball B of radius s. We claim that $\gamma \cap B$ has a unique point x_m which is at a minimum geodesic distance from x. Suppose not, that is, there is another minimum x'_m. The geodesic segment $\gamma(x_m, x'_m)$ on γ goes outside the open geodesic ball $B' = B_M(x, d_M(x, x_m))$. Since $s < \rho_c(M)$, B' has a radius less than the convexity radius. It follows that there is a unique minimizing geodesic between x_m and x'_m lying in B'. Then, we have two distinct minimizing geodesics between x_m and x'_m, one lying in B' and another going outside B' though both of which lie in B. This is impossible since B also has a radius less than the convexity radius.

Consider the retraction map $t : \mathrm{Tub}_s(\gamma) \to \gamma$ where $t(x) = x_m$. One can construct a deformation retraction that deforms the identity on $\mathrm{Tub}_s(\gamma)$ to t by moving each point x along the minimizing geodesic path that connect x to x_m in γ. ■

Proposition 3.8 *Let $P \subset M$ be an ε-sample and $4\varepsilon \leq r \leq \min\{\frac{1}{2}\rho(M), \rho_c(M)\}$. If $\hat{g}$ is the cycle on $\mathcal{C}^r(P)$ constructed from a geodesic cycle g in M by* Decomposition method, *then $[h(\hat{g})] = [g]$ where h is the homotopy equivalence defined in Proposition 3.3.*

Proof: Since g is a geodesic cycle, it follows from standard results in differential geometry [19] that $\mathrm{Len}(g) > 2\rho_c(M)$. Thus $\hat{g}$ can be constructed from a geodesic cycle g using *Decomposition method*. Each vertex p_i of $\hat{g}$ is within an ε Euclidean distance from the point x_i in g. Next, notice that, since $\mathcal{C}^r(P)$ uses balls of radius $r/2$, the stated range of r satisfies the condition of Proposition 3.3. By Proposition 3.3, for any point y on the segment p_ip_{i+1}, $h(y)$ is within $r/2$ Euclidean distance to either p_i or p_{i+1}. This implies that $h(y)$ is within $r/2+\varepsilon$ Euclidean distance, and hence, by Proposition 1.2, within r geodesic distance to either x_i or x_{i+1}. In addition, since the sub-curve of the geodesic cycle g between x_i and x_{i+1}, denoted $\gamma(x_i, x_{i+1})$, is of length $\ell_1 = r - 2\varepsilon < \rho_c(M)$, $\gamma(x_i, x_{i+1})$ is a minimizing geodesic between x_i and x_{i+1}. Therefore $h(p_ip_{i+1}) \in \mathrm{Tub}_r(\gamma(x_i, x_{i+1}))$. In particular, there are minimizing geodesics $\gamma(x_i, h(p_i))$ and $\gamma(x_{i+1}, h(p_{i+1}))$ that reside in $\mathrm{Tub}_r(\gamma(x_i, x_{i+1}))$.

Consider the cycle formed by the three geodesic segments $\gamma(x_i, x_{i+1})$, $\gamma(x_i, h(p_i))$, $\gamma(x_{i+1}, h(p_{i+1}))$, and the curve $h(p_ip_{i+1})$. From Proposition 3.7, this cycle is contractible in M as it resides in $\mathrm{Tub}_r(\gamma(x_i, x_{i+1}))$. In fact, there is a homotopy H_i that takes $h(p_ip_{i+1})$ to $\gamma(x_i, x_{i+1})$ while H_i keeps $h(p_i)$ and $h(p_{i+1})$ on the geodesics $\gamma(x_i, p_i)$ and $\gamma(x_{i+1}, p_{i+1})$ respectively. We can combine all homotopies H_i for $0 \le i \le m$ to define a homotopy between $h(\hat{g})$ and g. It follows that $[h(\hat{g})] = [g]$. ■

Proposition 3.9 *Let $P \subset M$ be an ε-sample and $4\varepsilon \le r \le \min\{\frac{1}{2}\rho(M), \rho_c(M)\}$. If $G = \{g_1, \ldots, g_k\}$ and $G' = \{g'_1, \ldots, g'_k\}$ are the cycles of a shortest basis of $\mathsf{H}_1(M)$ and $\mathsf{H}_1(\mathcal{C}^r(P))$ respectively, then we have $\mathrm{Len}(G') \le (1 + \frac{4\varepsilon}{r})\mathrm{Len}(G)$.*

Proof: It is obvious that any g_i must be a geodesic cycle. Let $\hat{g}_i$ be the cycle constructed by *Decomposition method* in the 1-skeleton of $\mathcal{C}^r(P)$. Thus, we have a set $\hat{G} = \{\hat{g}_1, \cdots, \hat{g}_k\}$. By Proposition 3.8, there is a homotopy equivalence $h : \mathcal{C}^r(P) \to M$ so that $[h(\hat{g}_j)] = [g_i]$, which means that $\hat{G}$ is also a homology cycle basis of $\mathsf{H}_1(\mathcal{C}^r(P))$. By Proposition 3.6,

$$\mathrm{Len}(G') \le \mathrm{Len}(\hat{G}) \le \frac{r}{r-2\varepsilon}\mathrm{Len}(G) \le (1 + \frac{4\varepsilon}{r})\mathrm{Len}(G).$$

■

We now consider the opposite direction, and provide a lower bound for the total length of a shortest basis of $\mathsf{H}_1(\mathcal{C}^r(P))$ in terms of the length of a shortest basis of $\mathsf{H}_1(M)$.

Proposition 3.10 *Let $P \subset M$ be an ε-sample and $4\varepsilon \le r \le \min\{\frac{1}{2}\rho(M), \rho_c(M)\}$. Let G and G' be defined as in Proposition 3.9. We have $\mathrm{Len}G \le (1 + \frac{4r^2}{3\rho^2(M)})\mathrm{Len}(G')$. Moreover, there exists a map $h\colon G' \to M$ so that $h(G')$ is a homology cycle basis of $\mathsf{H}_1(M)$ and the Hausdorff distance between each cycle $g \in G'$ and $h(g')$ is at most $\frac{r}{2}$.*

Proof: We construct a set of cycles in M from G'. First, we show that the length of these cycles is at most $(1 + \frac{4r^2}{3\rho^2(M)})$ times the length of G'. Next, we show that the constructed cycles form a homology cycle basis of $\mathsf{H}_1(M)$.

For each cycle $g' \in G'$, we construct $\bar{g}$ as follows. The vertices and edges of g' are vertices and edges of $\mathcal{C}^r(P)$. For an edge $e = pq \in g'$, $p, q \in P$ thus $p, q \in M$. We connect p and q by a minimizing geodesic $\gamma(p, q)$ on M, and map e to this geodesic. Mapping each edge in g' on M, we obtain $\bar{g}$. Thus we obtain a set $\bar{G} = \{\bar{g}_1, \cdots, \bar{g}_k\}$. By Proposition 1.2, $d_M(p, q) \le (1 + \frac{4d^2(p,q)}{3\rho^2(M)})d(p, q) \le (1 + \frac{4r^2}{3\rho^2(M)})d(p, q)$. Hence the length bound follows.

We now show that the set $\bar{G}$ is a homology cycle basis for $\mathsf{H}_1(M)$. Consider mapping $g'_j \in G'$ to M by the homotopy equivalence h. Each edge $e = pq \in g'_j$ is mapped to a curve $h(pq)$. From Proposition 3.3, we have that $h(p) = p$ and $h(q) = q$ and each point of $h(pq)$ is within $r/2$ Euclidean distance and hence r geodesic distance to either p or q. This implies that $h(pq) \subset \mathrm{Tub}_r(\gamma(p, q))$. Then, by using similar argument as in Proposition 3.7, we claim that $\gamma(p,q)$ and $h(pq)$ are homotopic. Combining all homotopies for each edge of g'_j, we get that $h(g'_j)$ is homotopic to $\bar{g}_j$. Since h is a homotopy equivalence, $h(G')$ and hence $\bar{G} = \{\bar{g}_1, \ldots, \bar{g}_k\}$ are a homology cycle basis of $\mathsf{H}_1(M)$. Therefore,

$$\mathrm{Len}(G) \leq \mathrm{Len}(\bar{G}) \leq (1 + \frac{4r^2}{3\rho^2(M)})\mathrm{Len}(G').$$

The cycles in $h(G')$ form a homology cycle basis of $\mathsf{H}_1(M)$ and each cycle $g' \in G'$ has a Hausdorff distance of $r/2$ with $h(g')$ satisfying the last claim. ■

Thanks to Proposition 3.5, shortest bases in $\mathcal{C}^r(P)$ and $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ are same for an appropriate range of r.

Theorem 3.11 *Let $P \subset M$ be an ε-sample and r be a real positive with $4\varepsilon \leq r \leq \min\{\frac{1}{2}\sqrt{\frac{3}{5}}\rho(M), \rho_c(M)\}$. Let G and G' be a shortest basis of $\mathsf{H}_1(M)$ and $\mathsf{H}_1^{r,2r}(\mathcal{R}(P))$ respectively. We have*

i. $\frac{1}{1+\frac{4r^2}{3\rho^2(M)}}\mathrm{Len}(G) \leq \mathrm{Len}(G') \leq (1 + \frac{4\varepsilon}{r})\mathrm{Len}(G)$.

ii. There is a map $h : G' \to M$ so that $h(G')$ is a homology cycle basis of $\mathsf{H}_1(M)$ and the Hausdorff distance between the underlying space of g' and $h(g')$ is at most $r/2$ for each $g' \in G'$.

Theorem 1.3 follows from Theorem 3.11, Theorem 2.1, and the time complexity analysis in section 2.3.

4 Conclusions

We have given a polynomial time algorithm for approximating a shortest basis of the first homology group of a smooth manifold from a point data. We have also presented an algorithm to compute a shortest basis for the first homology of any finite simplicial complex with non-negative weights on its edges.

We use Rips complexes for computations and use Čech complexes for analysis. One may observe that Čech complexes can be used directly in the algorithm. Since we know that $\mathcal{C}^r(P)$ is homotopy equivalent to M for an appropriate range of r, we can compute a shortest basis for $\mathsf{H}_1(\mathcal{C}^r(P))$ which can be shown to approximate a shortest basis for $\mathsf{H}_1(M)$ using our analysis. In technical terms, this will get rid of the weighting in step 1 and also step 4 of SHORTCYCLE algorithm, and make Theorem 2.1 and Proposition 3.5 redundant. Although the time complexity does not get affected in the worst-case sense, computing the triangles for Čech complexes becomes harder numerically in high dimensions than those for the Rips complexes. This is why we chose to describe an algorithm using the Rips complexes.

Recently, new persistence algorithms based on matrix multiplications [26, 11] have been proposed which have improved time complexity. It will be interesting to see how the time complexity of our algorithm can be improved using similar techniques.

Computing a shortest basis for other homology groups with $\mathbb{Z}_2$ coefficients has been shown to be NP-hard by Chen and Freedman [10]. A related topic that has been addressed in the literature is the problem of homology localization which asks for computing a shortest cycle in a given homology class. The problem has been shown to be NP-hard for a large number of cases [6, 10] with $\mathbb{Z}_2$ coefficient. Interestingly, it is shown in [17] that the problem is polynomial time solvable for a class of spaces when the homology is defined with $\mathbb{Z}$ instead of $\mathbb{Z}_2$. Does similar disparity exist for the shortest basis problem between different coefficient rings?

Acknowledgments. We thank anonymous reviewers for thoughtful comments on the paper. We acknowledge the helpful comments of Steve Oudot on an earlier draft of this paper. We also acknowledge Oleksiy

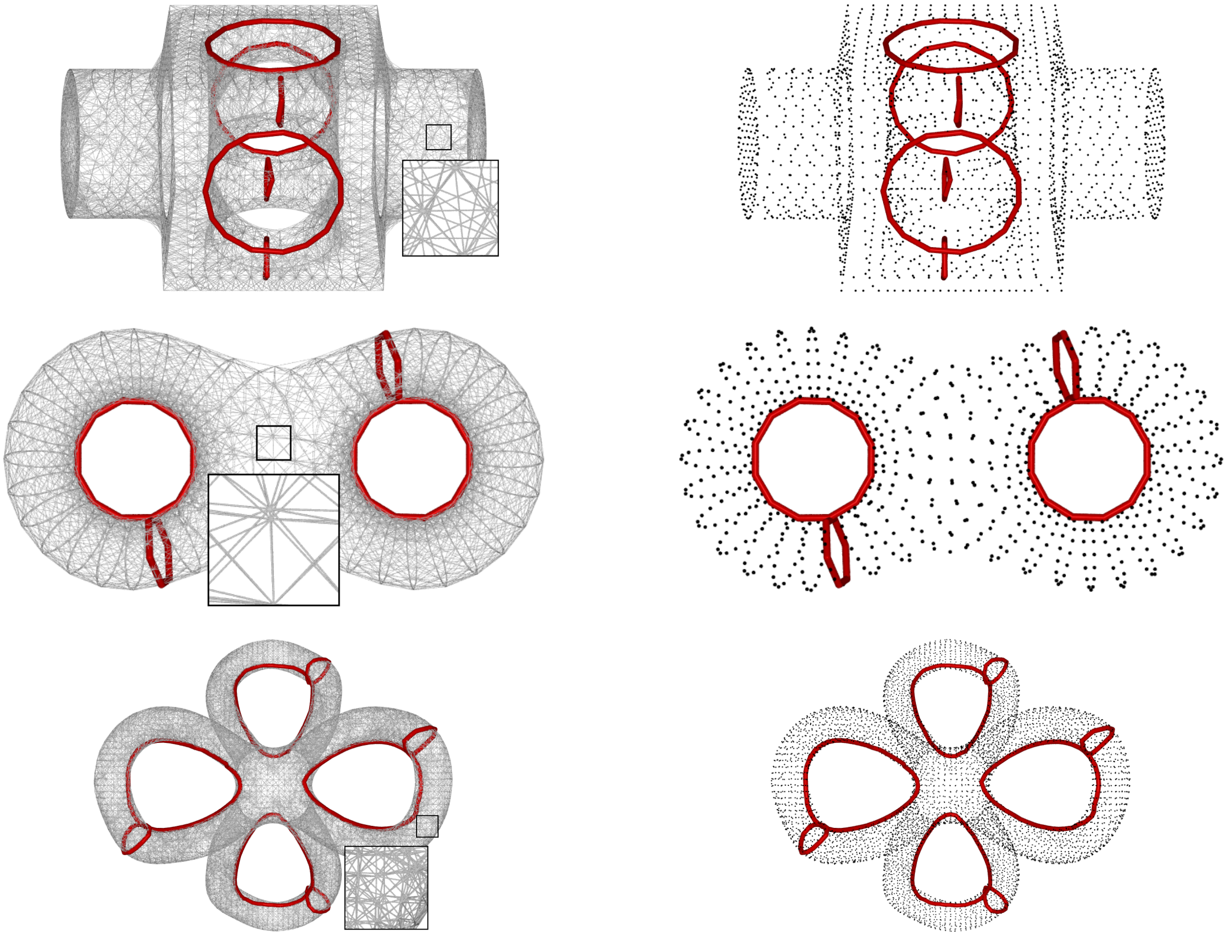


Figure 1: Cycles in a shortest basis computed in Rips complexes (left column) constructed out of point data (right column).

Busaryev who implemented SHORTCYCLE and produced the pictures in Figure 1. Tamal Dey acknowledges the support of NSF grants CCF-0830467 and CCF-0915996, Jian Sun acknowledges the support of NSF grant ITR 0205671 and DARPA grant HR0011-05-1-0007, and Yusu Wang acknowledges the support of NSF grants CCF-0747082 and DBI-0750891 .

References

[1] N. Amenta and M. Bern. Surface reconstruction by Voronoi filtering. *Discrete Comput. Geom.* **22** (1999), 481–504.

[2] M. Belkin, J. Sun, and Y. Wang. Discrete Laplace operator for meshed surfaces. *24th. Ann. Sympos. Comput. Geom.* (2008), 278–287.

[3] M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum. Graph approximations to geodesics on embedded manifolds. *Tech Report*, Dept. Psychology, Stanford University, USA, 2000. Available at http://isomap.stanford.edu/BdSLT.pdf

[4] J.-D. Boissonnat, L. J. Guibas, and S. Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Proc. 23rd Ann. Sympos. Comput. Geom.* (2007), 194–203.

[5] E. W. Chambers, J. Erickson, and A. Nayyeri. Homology flows, cohomology cuts. *41st Ann. ACM Sympos. Theory Comput.* (2009), 273–282.

[6] E. W. Chambers, J. Erickson, and A. Nayyeri. Minimum cuts and shortest homologous cycles. *25th Ann. Sympos. Comput. Geom.* (2009), 377–385.

[7] F. Chazal and A. Lieutier. Topology guaranteeing manifold reconstruction using distance function to noisy data. *Proc. 22nd Ann. Sympos. Comput. Geom.* (2006), 112–118.

[8] F. Chazal and S. Oudot. Towards persistence-based reconstruction in Euclidean spaces. *Proc. 24th Ann. Sympos. Comput. Geom.* (2008), 232–241.

[9] C. Chen and D. Freedman. Measuring and localizing homology classes. arXiv:0705.3061v2[cs.CG] (2007).

[10] C. Chen and D. Freedman. Hardness results for homology localization. *ACM/SIAM Sympos. Discrete Algorithms* (2010), 1594–1604.

[11] C. Chen and M. Kerber. An output-sensitive algorithm for persistent homology. *Proc. 27th. Ann. Sympos. Comput. Geom.* (2011), 207–216.

[12] S.-W. Cheng, T. K. Dey, and E. A. Ramos. Manifold reconstruction from point samples. *Proc. 16th Sympos. Discrete Algorithms* (2005), 1018–1027.

[13] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete Comput. Geom.* **37** (2007), 103-120.

[14] D. Cohen-Steiner, H. Edelsbrunner, and D. Morozov. Vines and vineyards by updating persistence in linear time. *Proc. 22nd Ann. Sympos. Comput. Geom.* (2006), 119–134.

[15] É. Colin de Verdière and J. Erickson. Tightening non-simple paths and cycles on surfaces. *Proc. ACM-SIAM Sympos. Discrete Algorithms* (2006), 192–201.

[16] T. K. Dey. Curve and Surface Reconstruction : Algorithms with Mathematical Analysis. Cambridge University Press, New York, 2007.

[17] T. K. Dey, A. Hirani, and B. Krishnamoorthy. Optimal homologous cycles, total unimodularity, and linear programming. *Proc. 42nd ACM Sympos. Theory Computing* (2010), 221–230.

[18] T. K. Dey and K. Li. Cut locus and topology from surface point data. *25th Ann. Sympos. Comput. Geom.*(2009), 125–134.

[19] M. P. do Carmo. Riemannian geometry. Birkhäuser, Boston, 1992.

[20] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.* **28** (2002), 511–533.

[21] J. Erickson and K. Whittlesey. Greedy optimal homotopy and homology generators. *Proc. ACM-SIAM Sympos. Discrete Algorithms* (2005), 1038–1046.

[22] A. Hatcher. Algebraic Topology. Cambridge U. Press, New York, 2002.

[23] I. T. Jollife. Principal Component Analysis. Springer series in statistics, Springer, NY, 2002.

[24] T. Kaczynski, K. Mischaikow, and M. Mrozek. Computing homology. *Homology, Homotopy and Applications.* **5** (2003), 233–256.

[25] J. Leray. Sur la forme des espaces topologiques et sur les points fixes des représentations. *J. Math. Pure Appl.* **24** (1945), 95–167.

[26] N. Milosavljević, D. Morozov, and P. Škraba. Zigzag persistent homology in matrix multiplication time. *Proc. 27th Ann. Sympos. Comput. Geom.* (2011), 216–225.

[27] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** (2008), 419–441.

[28] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science* **290** (2000), 2319–2323.

[29] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.* **33** (2005), 249–274.

Appendix

Proof of Proposition 3.2.

The proof is based on that of Nerve Lemma in [22] (Chapter 4.G). Let Γ be the barycentric subdivision of $\mathcal{C}^{2r}(P)$. We consider the following sequence — we will describe the space ΔP^r and maps involved in this sequence shortly.

$$\mathcal{C}^{2r}(P) \xrightarrow{h} \Gamma \xrightarrow[\Delta p]{\Delta q} \Delta P^r \xrightarrow{\pi} P^r. \tag{1}$$

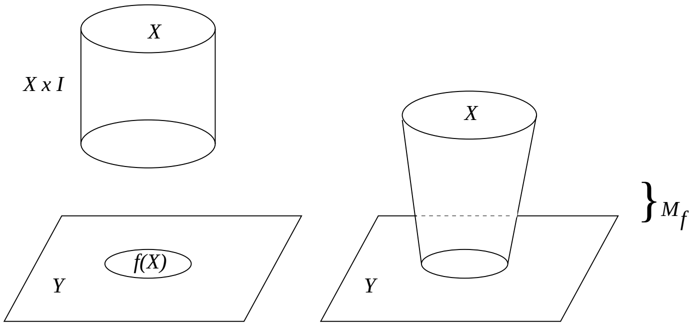
We prove the proposition by defining $f = \pi \circ \Delta q \circ h$ and showing that f is a homotopy equivalence meeting the requirements stated in the proposition.

We first introduce the concept of mapping cylinder. For a map $f : X \to Y$, the mapping cylinder M_f is the quotient space of the disjoint union $(X \times I) \bigsqcup Y$ with $(x, 1)$ identified with $f(x) \in Y$, denoted $M_f = X \bigsqcup_f Y$, see Figure 2(a). It is obvious that M_f retracts to Y under a deformation retraction. Let e_Y be the retraction from M_f to Y. It is well-known (e.g., Corollary 0.21 in [22]) that f is a homotopy equivalence map if and only if M_f retracts to X under a deformation retraction. See Figure 2(b). In fact, if e_X in Figure 2(b) is a retraction under a deformation retraction, the map $g = e_X \circ i_Y$ is a homotopy equivalence map from Y to X. We will use this fact later in the proof to define the map $\Delta q : \Delta P^r \to \Gamma$.

We are now ready to explain each map in the composition of the map f. Since Γ is the barycentric subdivision of $\mathcal{C}^{2r}(P)$, we can take h as an identity map between the underlying spaces of $\mathcal{C}^{2r}(P)$ and Γ. Index the points in $P = \{p_i\}_{i=1}^m$ arbitrarily. Let $B_i = B(p_i, r)$. To facilitate the argument, label the vertices in Γ using B_i's and their finite intersections, see Figure 3. Under such labeling, the vertex set of any k-simplex Δ^k in Γ can be ordered as

$$\Delta^k = \left(B_{i_0} \cap \cdots \cap B_{i_n}, B_{i_0} \cap \cdots \cap B_{i_{n-1}}, \ldots, B_{i_0} \cap \cdots \cap B_{i_{n-k}} \right), \tag{2}$$

where the size of the index set in the label of each vertex decreases from $n+1$ to $n-k$ for some n. Each edge (1-simplex) in Γ is associated with an inclusion map between the labels of its vertices. This induces



(a)

$$\begin{tikzcd}
& Y \arrow[dl, "g"] \arrow[dd, shift left, "i_Y"] \\
X \arrow[ur, "f"] \arrow[dr, "i_X"] & \\
& M_f = X \sqcup_f Y \arrow[ul, "e_X"] \arrow[uu, shift left, "e_Y"]
\end{tikzcd}$$

(b)

Figure 2: (a) the mapping cylinder $M_f = X \bigsqcup_f Y$ [22]; (b) the maps among X, Y and M_f : i_X and i_Y are the inclusion maps from X and Y into M_f respectively, e_Y is a retraction from M_f to Y given by a deformation retraction, and e_X is a retraction from M_f to X given by a deformation retraction provided f is a homotopy equivalence.

the following sequence of inclusion maps by considering only edges between two consecutive vertices in Eqn (2) for any k-simplex Δ^k in Γ:

$$\begin{aligned}(B_{i_0} \cap \cdots \cap B_{i_n}) &\hookrightarrow (B_{i_0} \cap \cdots \cap B_{i_{n-1}}) \\ &\hookrightarrow \cdots \hookrightarrow (B_{i_0} \cap \cdots \cap B_{i_{n-k}}).\end{aligned} \tag{3}$$

We now give a brief account for the construction of ΔP^r used in the sequence of maps given in Eqn (1). The readers are referred to [22] for more details. ΔP^r is realized using the concept of iterated mapping cylinder defined over a sequence of maps. Specifically, the sequence of inclusion maps as shown in Eqn (3) associated with any k-simplex Δ^k in Γ induces an iterated mapping cylinder over Δ^k. We obtain ΔP^r by gluing these iterated mapping cylinders over all simplices in Γ, see the top right most picture in Figure 3. There is a canonical projection $\Delta p : \Delta P^r \to \Gamma$ induced by projecting each finite intersection to its corresponding vertex in Γ.

To define the map Δq in Eqn (1), consider the mapping cylinder $M_{\Delta p}$. In Chapter 4.G of [22], the Nerve Lemma was proved by showing that $M_{\Delta p}$ retracts to ΔP^r under a deformation retraction; let $e_{\Delta P^r} : M_{\Delta p} \to \Delta P^r$ be the associated retraction. We set $\Delta q := e_{\Delta P^r} \circ i_\Gamma$ where i_Γ is the inclusion map from Γ into $M_{\Delta p}$. From our earlier discussion about Figure 2 (b), Δq is a homotopy equivalence (setting $X = \Delta P^r$ and $Y = \Gamma$ in the diagram of Figure 2 (b)). Furthermore, [22] showed that the retraction $e_{\Delta P^r}$ in fact maps a simplex $\Delta^k \in \Gamma$ to the iterated mapping cylinder defined over the same Δ^k, implying that Δq maps a simplex $\Delta^k \in \Gamma$ into the iterated mapping cylinder defined by the sequence of inclusion maps associated with Δ^k.

On the other hand, ΔP^r can also be considered as the quotient space of the disjoint union of all the products $B_{i_0} \cap \cdots \cap B_{i_n} \times \Delta^n$, as the subscripts range over set of $n+1$ distinct indices and any $n \geq 0$,

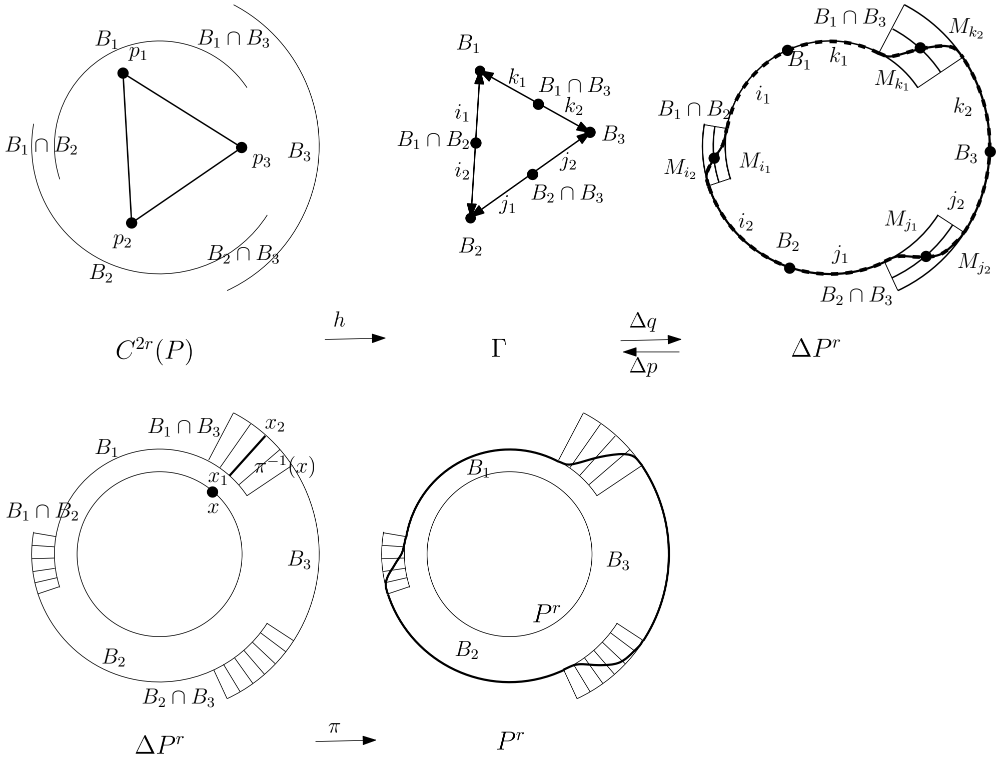


Figure 3: Illustration of the maps and the spaces involved in Eq. 1.

with the identifications over the faces of Δ^n using inclusions $B_{i_0} \cap \cdots \cap B_{i_n} \hookrightarrow B_{i_0} \cap \cdots \cap \hat{B}_{i_j} \cap \cdots \cap B_{i_n}$ where $\hat{}$ means the corresponding term is missing. From this viewpoint, any point $x \in P^r$ has a fiber $\pi^{-1}(x)$ in ΔP^r defined as follows. Let x_i be a copy of x in B_i for those B_i containing x and define $\pi^{-1}(x) = \{\sum_i t_i x_i : \sum_i t_i = 1 \text{ and } t_i \geq 0\}$, see the bottom left most picture in Figure 3. It is easy to see that P^r can be embedded into ΔP^r as a section of ΔP^r, denoted $s(P^r)$. Since points in $\pi^{-1}(x)$ can move linearly along line segments to $s(x)$, $s(P^r)$ is a retract of ΔP^r. Taking π as the retraction from ΔP^r to $s(P^r)$, we have π as a homotopy equivalence. Thus, $f = \pi \circ \Delta q \circ h$ is a homotopy equivalence.

Observe that a point y in an iterated mapping cylinder over a simplex $\Delta^k = (B_{i_0} \cap \cdots \cap B_{i_n}, \cdots, B_{i_0} \cap \cdots \cap B_{i_{n-k}})$ in Γ is in the fiber $\pi^{-1}(x)$ for some x in B_{i_0}. This means that if an l-simplex Δ^l is in the closure of the star of a vertex $p \in P$ in Γ, then any point y in the iterated mapping cylinder over Δ^l is in the fiber of a point $x \in B(p, r)$. Indeed, this follows from the previous statement by considering any simplex Δ^k containing p and Δ^l. Now consider a simplex $\sigma \in \mathcal{C}^{2r}(P)$. Any simplex in the barycentric subdivision of σ must be in the closure of the star of some vertex of σ. Thus σ, under the map $\Delta q \circ h$, is mapped into the union of the iterated mapping cylinders defined over the simplices in the barycentric subdivision of σ, and thus , its image, under the map π, is further mapped into $\cup_{p \in \mathrm{Vert}(\sigma)} B(p, r)$.

In addition, it is clear that it is possible to choose f so that it fixes each vertex in $\mathcal{C}^{2r}(P)$. This proves the proposition.