cohomology and data

Lek-Heng Lim

June 2, 2016

thanks: Xiaoye Jiang, Yuan Yao, Ke Ye, Yinyu Ye, AFOSR, DARPA, NSF

cohomology

cohomology in a nutshell

- two matrices $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$ satisfying

$$AB = 0$$

- equivalently

$$\operatorname{im}(B) \subseteq \ker(A)$$

- cohomology group with respect to A and B is quotient vector space

$$\ker(A) / \operatorname{im}(B)$$

- cocycles: elements of $\ker(A)$
- coboundaries: elements of $\operatorname{im}(B)$
- cohomology classes: elements of $\ker(A) / \operatorname{im}(B)$

Hodge theory in a nutshell

- using basic linear algebra, may show

$$\ker(A)/\operatorname{im}(B) \cong \ker(A^*A + BB^*)$$

- i.e., cohomology classes are harmonic forms, solutions of

$$(A^*A + BB^*)x = 0$$

- Hodge Laplacian: $A^*A + BB^* \in \mathbb{R}^{n\times n}$
- Hodge decomposition:

$$\mathbb{R}^n = \overbrace{\operatorname{im}(A^*) \oplus \underbrace{\ker(A^*A + BB^*)}_{\ker(A)}}^{\ker(B^*)} \oplus \operatorname{im}(B)$$

$\oplus$ = orthogonal direct sum

most important example for us

$$A = \mathrm{curl}, \qquad B = \mathrm{grad}, \qquad B^* = -\mathrm{div}$$

- then
$$A^*A + BB^* = \mathrm{curl}^* \mathrm{curl} - \mathrm{grad}\,\mathrm{div} =: \Delta_1$$
- note that
$$AB = \mathrm{curl}\,\mathrm{grad} = 0$$
and that
$$\mathrm{div} = -\mathrm{grad}^*$$
- formal definitions later

beautiful mind cohomology

$$V = \{F : \mathbb{R}^3 \setminus X \to \mathbb{R}^3 : \nabla \times F = 0\}; \quad W = \{F = \nabla g\};$$

$$\dim(V/W) = ?$$

two main uses of cohomology

obstruction: quantifies obstruction from local to global

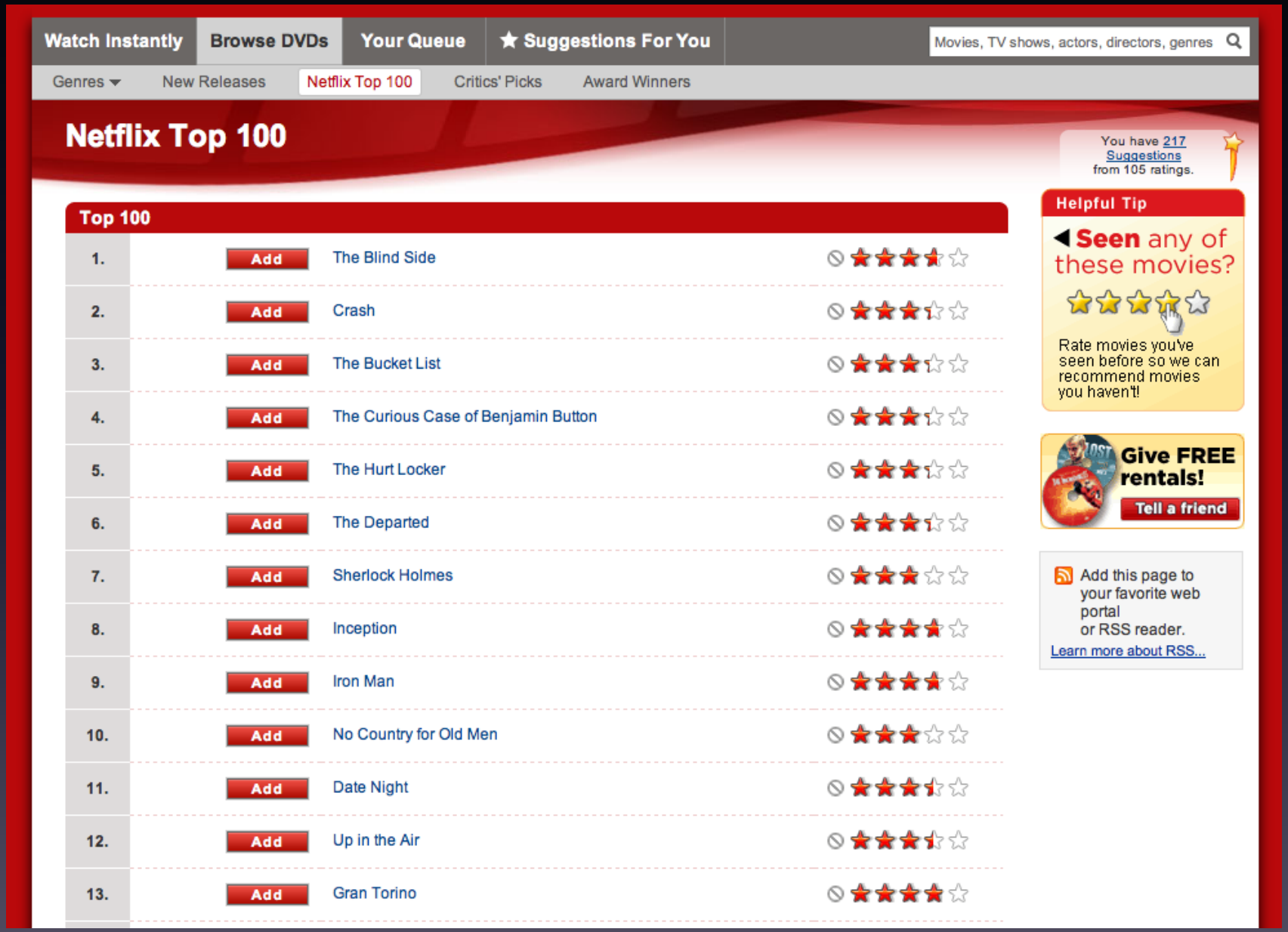
- nonexistence of Penrose tribar
- local rankings to global rankings of movies

moduli: describe a collection of mathematical objects

- parameterize line bundles on algebraic variety
- parameterize cryo-EM data sets: do two data sets give same reconstruction?

rank aggregation

global ranking



rank aggregation

- many voters, each rated a few items, want global ranking
- averaging over scores doesn't work: one movie receives single 5☆, another receives 10,000 5☆ and one 4☆
- should be invariant under monotone transformation:

$$1\text{☆}, \ldots, 5\text{☆} \quad \longrightarrow \quad 0\text{☆}, \ldots, 4\text{☆}$$

- basic unit of ranking: pairwise ranking
- Netflix user-product rating matrix

$$z_{ij} = i\text{th user's rating for } j\text{th movie},$$

$Z \in \mathbb{R}^{480,189 \times 17,770}$ has 98.82% missing values

- average over pairwise rankings instead

$$y_{ij} = \text{how much } i\text{th movie is preferred over } j\text{th movie},$$

$Y \in \mathbb{R}^{17,770 \times 17,770}$ has 0.22% missing values

averaging over pairwise rankings

classical problem in statistics

linear model: average score difference between i and j

$$y_{ij} = \frac{\sum_h (z_{hj} - z_{hi})}{\#\{h : z_{hi}, z_{hj} \text{ exist}\}}$$

invariant under translation

log-linear model: log average score ratio of positive scores

$$y_{ij} = \frac{\sum_h (\log z_{hj} - \log z_{hi})}{\#\{h : z_{hi}, z_{hj} \text{ exist}\}}$$

invariant up to a multiplicative constant

averaging over pairwise rankings

linear probability model: probability j preferred to i in excess of purely random choice

$$y_{ij} = \Pr\{h : z_{hj} > z_{hi}\} - \frac{1}{2}$$

invariant under monotone transformation

Bradley–Terry model: logarithmic odd ratio (logit)

$$y_{ij} = \log \frac{\Pr\{h : z_{hj} > z_{hi}\}}{\Pr\{h : z_{hj} < z_{hi}\}}$$

invariant under monotone transformation

H. A. David, *The Method of Paired Comparisons*, 2nd Ed., Griffin's Statistical Monographs and Courses, **41**, Oxford University Press, New York, NY, 1988.

difficulties with rank aggregation

- Condorcet's paradox: majority vote intransitive $i \succeq j \succeq k \succeq i$ [Condorcet, 1785]
- Arrow/Sen's impossibility: any sufficiently sophisticated preference aggregation must exhibit intransitivity [Arrow, 1950], [Sen, 1970]
- McKelvey/Saari's chaos: almost every possible ordering can be realized by a clever choice of the order in which decisions are taken [McKelvey, 1979], [Saari, 1989]
- Kemeny optimal is NP-hard: even with just 4 voters [Dwork–Kumar–Naor–Sivakumar, 2001], quadratic assignment problem [Cook–Kress, 1984]
- empirical evidence: lack of consensus common in group decision making (e.g. US congress)

objectives

ordinal: intransitivity, $i \succeq j \succeq k \succeq i$

cardinal: inconsistency, $y_{ij} + y_{jk} + y_{ki} \neq 0$

- want global ranking of alternatives if a reasonable one exists
- want certificate of reliability to quantify validity of global ranking
- if no meaningful global ranking, analyze nature of inconsistencies

graphs

- $G = (V, E)$ undirected graph
- V vertices
- $E \subseteq \binom{V}{2}$ edges
- $T \subseteq \binom{V}{3}$ triangles or 3-cliques, i.e.,

$$\{i, j, k\} \in T \quad \text{iff} \quad \{i, j\}, \{j, k\}, \{k, i\} \in E$$

- more generally $K_k \subseteq \binom{V}{k}$ k-cliques, i.e.,

$$\{i_1, \ldots, i_k\} \in K_k \quad \text{iff} \quad \text{it is a complete subgraph of } G$$

- $K(G)$ clique complex of a graph G is an abstract simplicial complex

functions on graphs

- vertex functions: $f : V \to \mathbb{R}$
- edge flows: $X : V \times V \to \mathbb{R}$

$$X(i,j) = -X(j,i)$$

for $\{i,j\} \in E$, zero otherwise

- triangular flows: $\Phi : V \times V \times V \to \mathbb{R}$

$$\Phi(i,j,k) = \Phi(j,k,i) = \Phi(k,i,j)$$
$$= -\Phi(j,i,k) = -\Phi(i,k,j) = -\Phi(k,j,i)$$

for $\{i,j,k\} \in T$, zero otherwise

- introduce inner products: $L^2(V)$, $L^2_\wedge(E)$, $L^2_\wedge(T)$

$$\langle f, g \rangle_V = \sum_{i=1}^{n} w_i f(i) g(i), \quad \langle X, Y \rangle_E = \sum_{i<j} w_{ij} X(i,j) Y(i,j),$$
$$\langle \Phi, \Psi \rangle_T = \sum_{i<j<k} w_{ijk} \Phi(i,j,k) \Psi(i,j,k)$$

operators on functions on graphs

gradient: $\operatorname{grad} : L^2(V) \to L^2_\wedge(E)$,

$$(\operatorname{grad} f)(i,j) = f(j) - f(i)$$

curl: $\operatorname{curl} : L^2_\wedge(E) \to L^2_\wedge(T)$,

$$(\operatorname{curl} X)(i,j,k) = X(i,j) + X(j,k) + X(k,i)$$

divergence: $\operatorname{div} : L^2_\wedge(E) \to L^2(V)$,

$$(\operatorname{div} X)(i) = \sum_{j=1}^{n} w_{ij} X(i,j)$$

graph Laplacian: $\Delta_0 : L^2(V) \to L^2(V)$,

$$\Delta_0 = \operatorname{div} \operatorname{grad}$$

graph Helmholtzian: $\Delta_1 : L^2_\wedge(E) \to L^2_\wedge(E)$,

$$\Delta_1 = -\operatorname{grad} \operatorname{div} + \operatorname{curl}^* \operatorname{curl}$$

Hodge decomposition of rankings

- pairwise comparison graph $G = (V, E)$; V: set of alternatives, E: pairs of alternatives compared
- space of pairwise rankings, $L^2_\wedge(E)$, admits an orthogonal decomposition into three components

$$L^2_\wedge(E) = \overbrace{\mathrm{im}(\mathrm{curl}^*)}^{\ker(\mathrm{div})} \oplus \underbrace{\ker(\Delta_1) \oplus \mathrm{im}(\mathrm{grad})}_{\ker(\mathrm{curl})}$$

- cohomology group is

$$\ker(\Delta_1) = \ker(\mathrm{curl}) \cap \ker(\mathrm{div})$$

Hodge decomposition

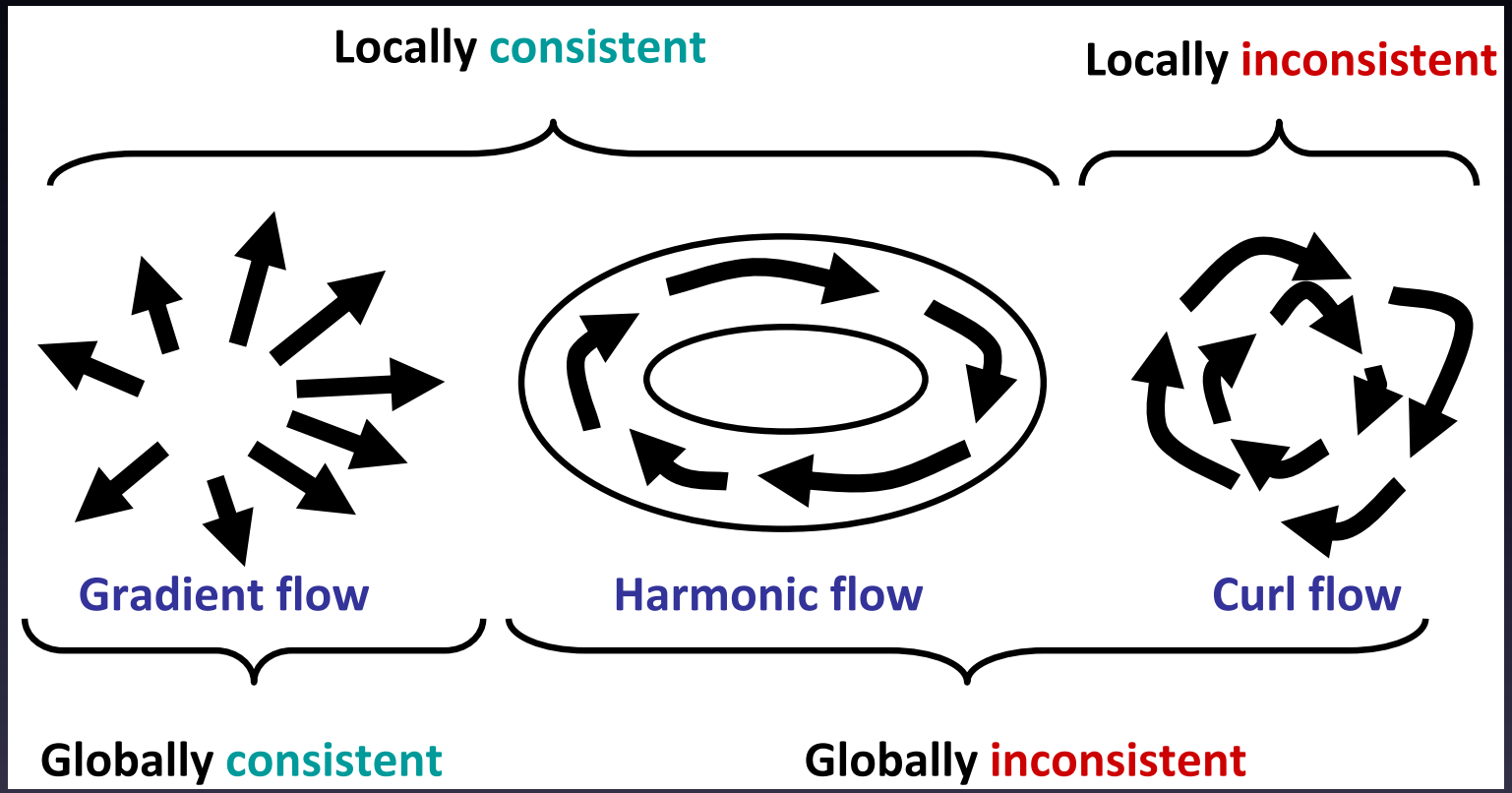


Figure: cartoon courtesy of Pablo Parrilo

HodgeRank

- Hodge decomposition of ranking:

$$\textit{aggregate pairwise ranking} = \textit{consistent} \oplus \textit{locally inconsistent} \oplus \textit{globally inconsistent}$$

- consistent component gives global ranking
- total size of inconsistent components gives certificate of reliability
- local and global inconsistent components tell us about nature of inconsistencies
- quantifies Condorcet paradox, Arrow's impossibility, McKelvey chaos, etc
- numerical, not combinatorial, so not NP-hard

analyzing inconsistencies

- locally inconsistent rankings should be acceptable
  - inconsistencies in items ranked closed together but not in items ranked far apart
  - ordering of 4th, 5th, 6th ranked items cannot be trusted but ordering of 4th, 50th, 600th ranked items can
  - e.g. no consensus for hamburgers, hot dogs, pizzas, and no consensus for caviar, foie gras, truffle, but clear preference for latter group
- globally inconsistent rankings might be rare

Theorem (Kahle, 2007)

Erdős-Rényi $G(n,p)$*, n alternatives, comparisons occur with probability p, clique complex* $K(G)$ *almost always have zero* 1*-homology, unless*

$$\frac{1}{n^2} \ll p \ll \frac{1}{n}.$$

relates to Kemeny optimum

- ranking data lives on pairwise comparison graph $G = (V, E)$; V: set of alternatives, E: pairs compared
- optimize over model class $\mathcal{M}$

$$\min_{X \in \mathcal{M}} \sum_{\alpha,i,j} w_{ij}^{\alpha} (x_{ij} - y_{ij}^{\alpha})^2$$

- Y_{ij}^{α} measures preference of i over j of voter α
- w_{ij}^{α} metric; 1 if α made comparison for $\{i, j\}$, 0 otherwise
- Kemeny optimization:

$$\mathcal{M}_K = \{X \in \mathbb{R}^{n \times n} : x_{ij} = \text{sign}(f_j - f_i),\ f : V \to \mathbb{R}\}$$

- relaxed version

$$\mathcal{M}_G = \{X \in \mathbb{R}^{n \times n} : x_{ij} = f_j - f_i,\ f : V \to \mathbb{R}\}$$

- rank-constrained regression on skew-symmetric matrices
- solution is precisely consistent component in HodgeRank

top Netflix movies by HodgeRank

Linear Full	Linear 30	Bradley–Terry Full
Greatest Story Ever ...	LOTR III: Return ...	LOTR III: Return ...
Terminator 3	LOTR I: The Fellowship ...	LOTR II: The Two ...
Michael Flatley	LOTR II: The Two ...	LOTR I: The Fellowship ...
Hannibal [Bonus]	Star Wars VI: Return ...	Star Wars V: Empire ...
Donnie Darko [Bonus]	Star Wars V: Empire ...	Raiders of the Lost Arc
Timothy Leary's ...	Star Wars IV: A New Hope	Star Wars IV: A New Hope
In Country	LOTR III: Return ...	Shawshank Redemption
Bad Boys II [Bonus]	Raiders of the Lost Arc	Star Wars VI: Return ...
Cast Away [Bonus]	The Godfather	LOTR III: Return ...
Star Wars: Ewok ...	Saving Private Ryan	The Godfather

- LOTR III shows up twice because of the two DVD editions
- full model has many "bonus" discs that Netflix rents; these are items enjoyed by only a few people

cryo-electron microscopy

group-valued cohomology

$\mathbb{R}$-valued

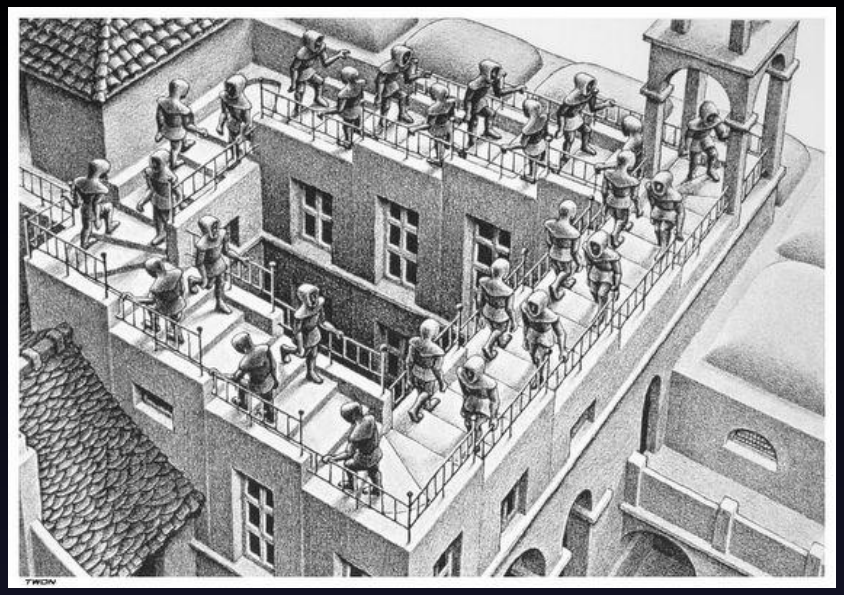
- $x_{ij} + x_{ji} = 0$
- $\varphi_{ijk} + \varphi_{ikj} = \varphi_{ijk} + \varphi_{kji}$
  $= \varphi_{ijk} + \varphi_{jik} = 0$
- $(\operatorname{grad} f)_{ij} = f_j - f_i$
- $(\operatorname{curl} X)_{ijk} = x_{ij} + x_{jk} + x_{ki}$
- $x_{ij} + x_{jk} + x_{ki} = 0$

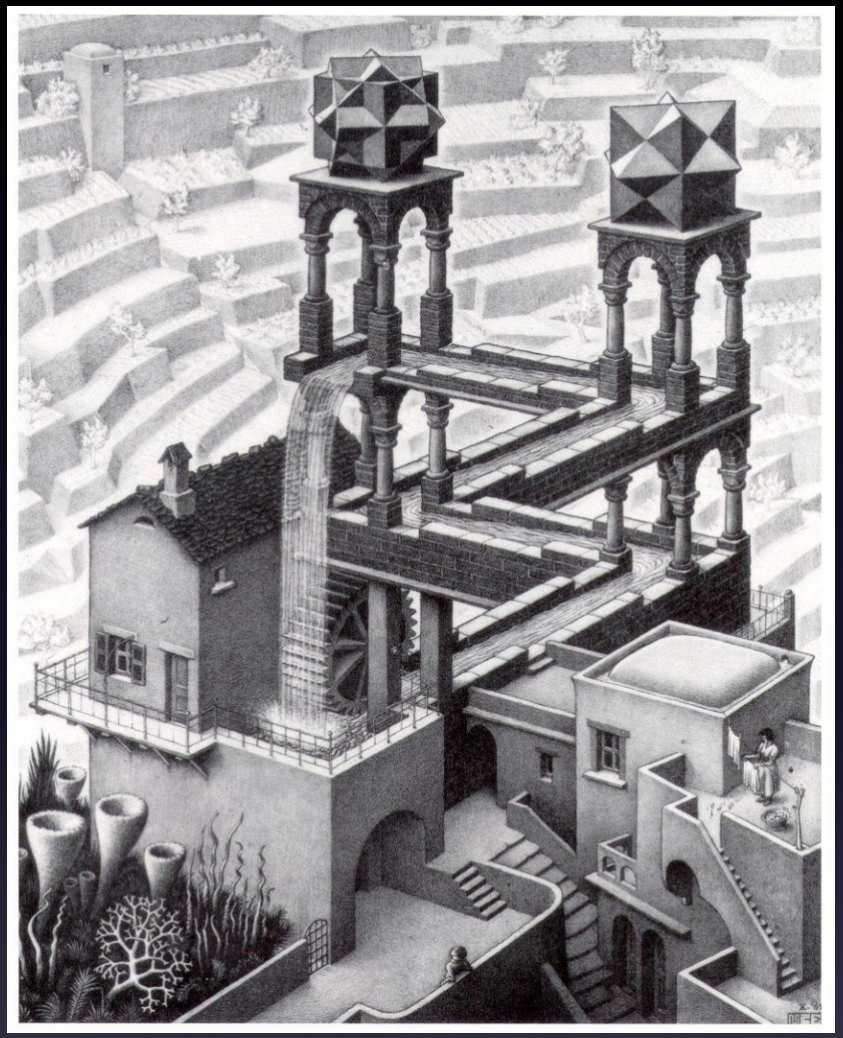
G-valued

- $g_{ij} g_{ji} = 1$
- $g_{ijk} g_{ikj} = g_{ijk} g_{kji}$
  $= g_{ijk} g_{jik} = 1$
- $\left(\delta_0 (g_\alpha)_{\alpha \in I}\right)_{ij} = g_j g_i^{-1}$
- $\left(\delta_1 (g_{\alpha\beta})_{\alpha,\beta \in I}\right)_{ijk} = g_{ij} g_{jk} g_{ki}$
- $g_{ij} g_{jk} g_{ki} = 1$

cryo-EM application: $G = SO(2)$ and $G = SO(2)_d$, i.e., $SO(2)$ with discrete topology

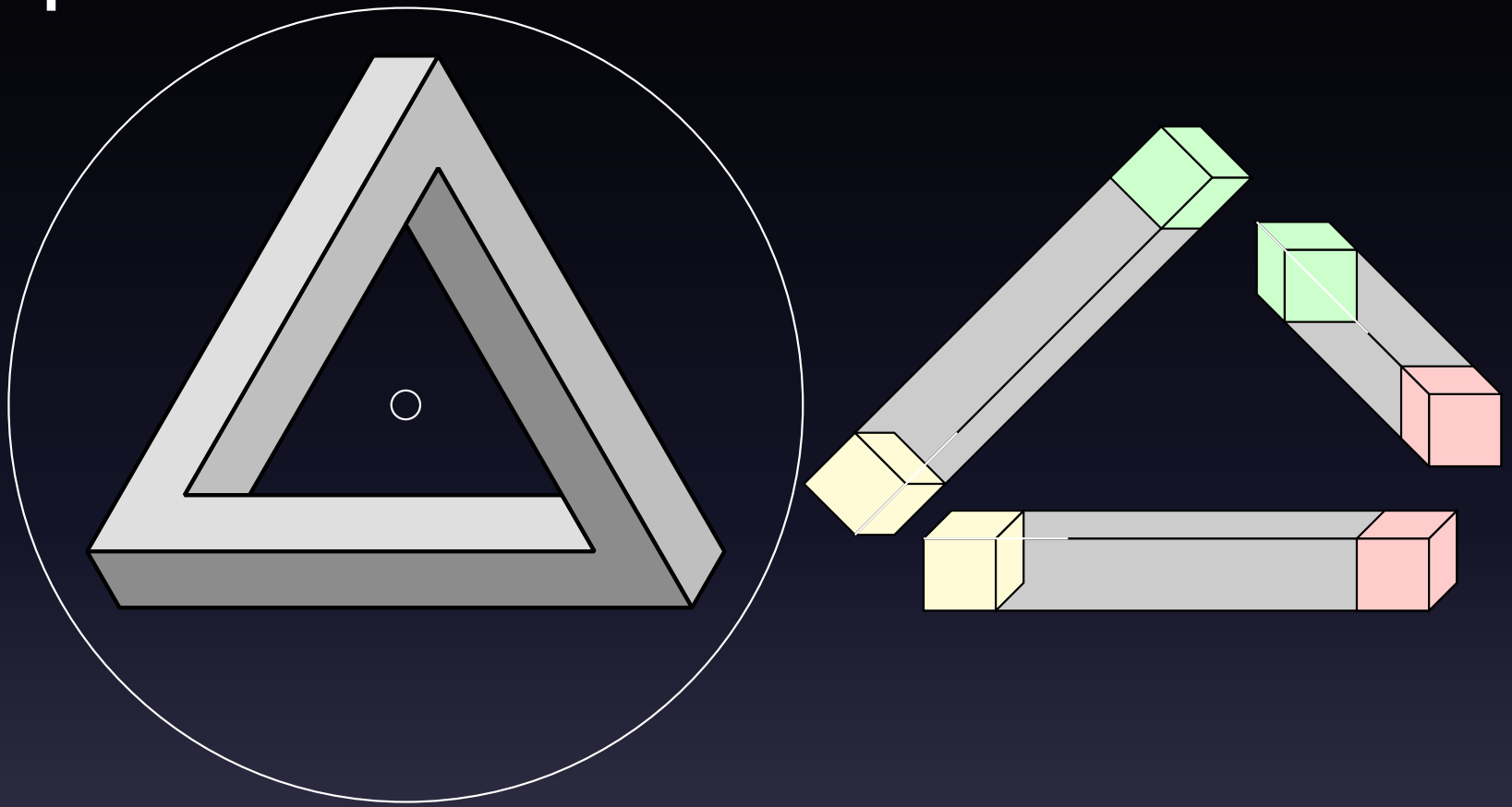
full story: Čech cohomology with G coefficients $\check{H}^1(K, G)$





- M. C. Escher's optical illusions
- all based on L. S. Penrose and R. Penrose's tribar

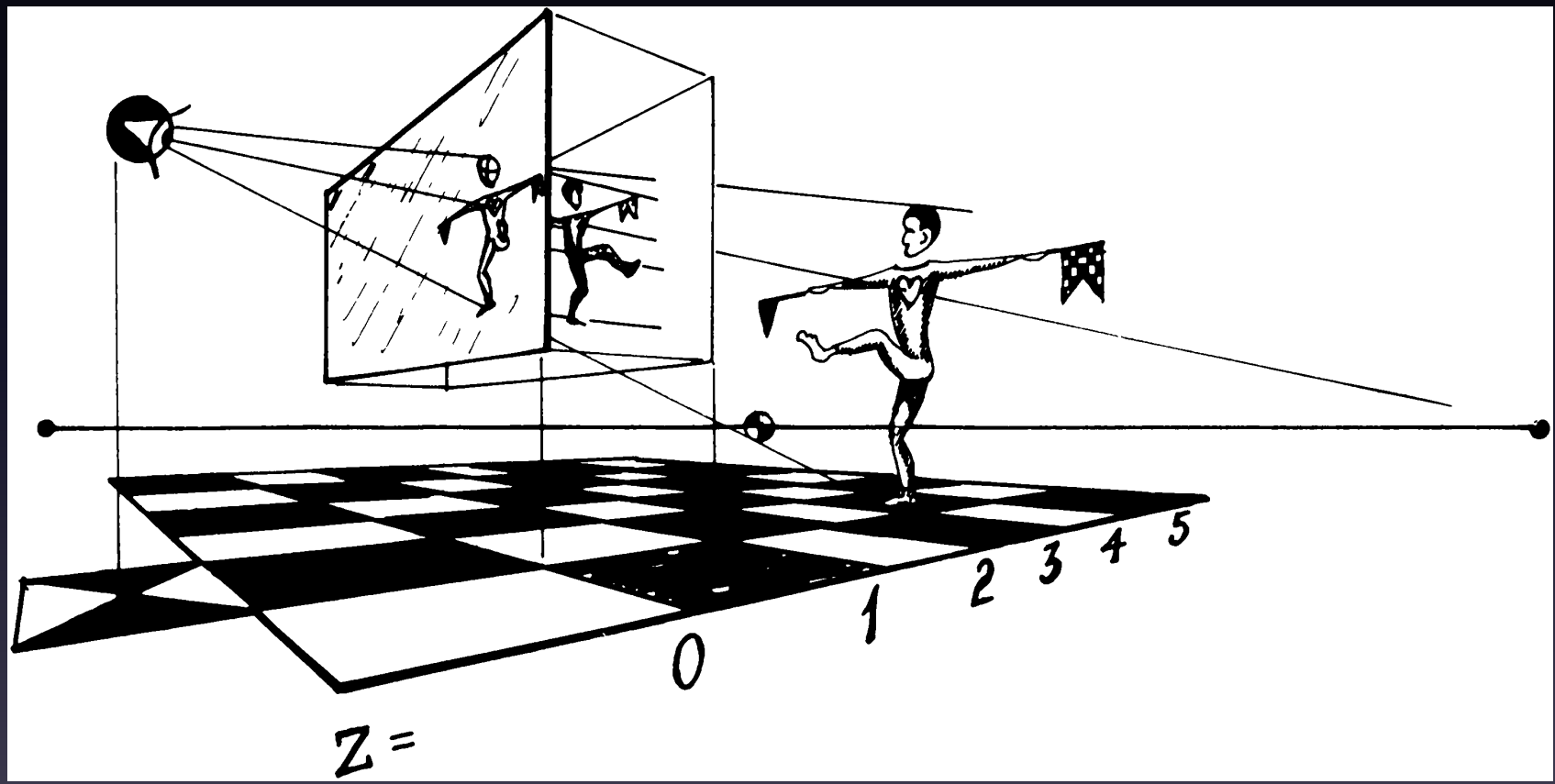
inspiration: Penrose tribar



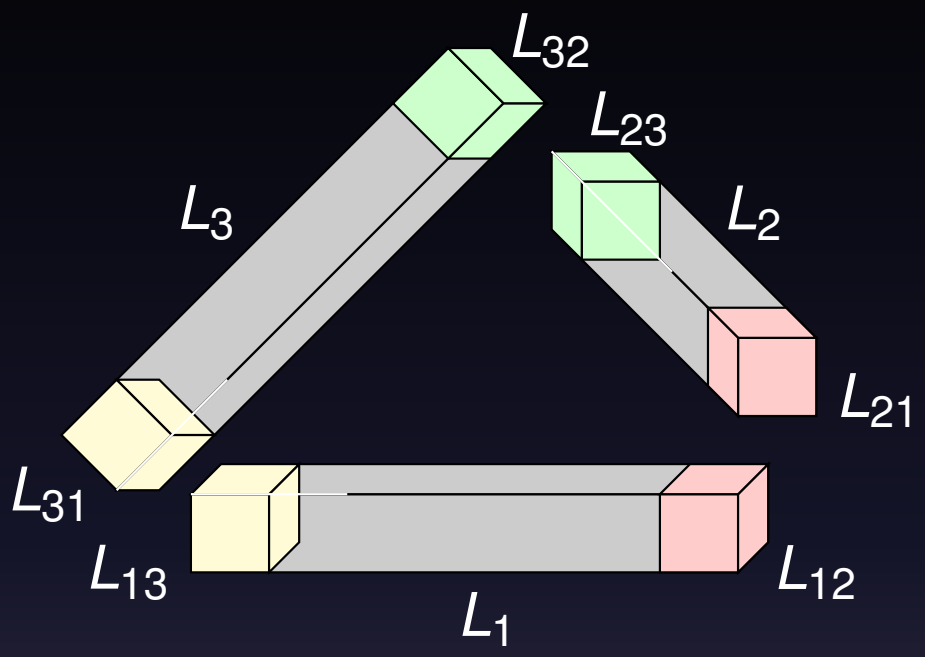
- $\Delta =$ 2D figure on left, embedded in $Q =$ annulus in $\mathbb{R}^2$
- appears to be a projection of a (nonexistent) 3D tribar

R. Penrose, “On the cohomology of impossible figures,” *Structural Topology*, **17** (1991), pp. 11–16.

technique: perspective



Penrose's example

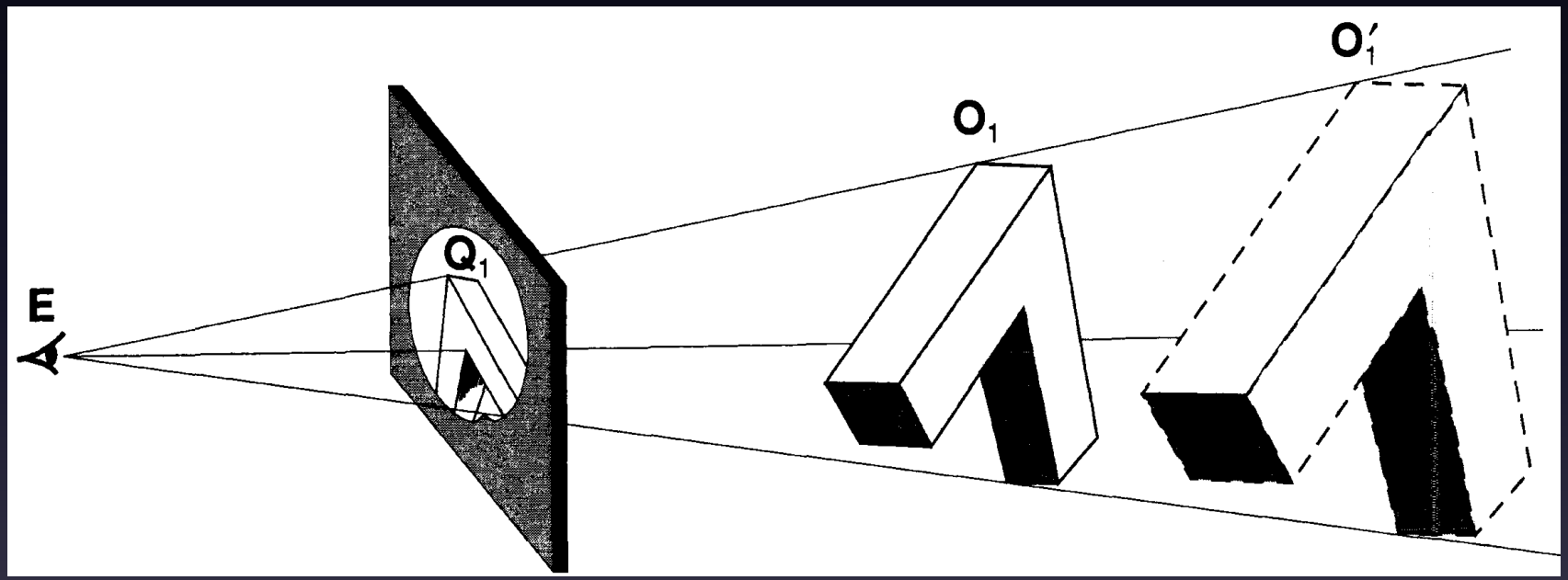


- E and L_1, L_2, L_3 on opposite sides of hyperplane $H \subseteq \mathbb{R}^3$
- $d_{ij} \in \mathbb{R}_+$ distance from E to center of L_{ij}

$$g_{ij} = \frac{d_{ij}}{d_{ji}}, \qquad g = \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix} \in \mathbb{R}_+^{3\times 3}$$

- $g_{ij}^{-1} = g_{ji}$, $g_{ii} = 1$

picture



$\mathbb{R}^+$-valued cohomology

- may move L_1, L_2, L_3 independently along viewing direction so that projection onto H always give Δ
- results in scaling by a factor $g_i \in \mathbb{R}_+$: $d'_{ij} = d_{ij}/g_i$

$$g'_{ij} = \frac{d'_{ij}}{d'_{ji}} = \frac{d_{ij}/g_i}{d_{ji}/g_j} = g_{ij}\frac{g_j}{g_i}$$

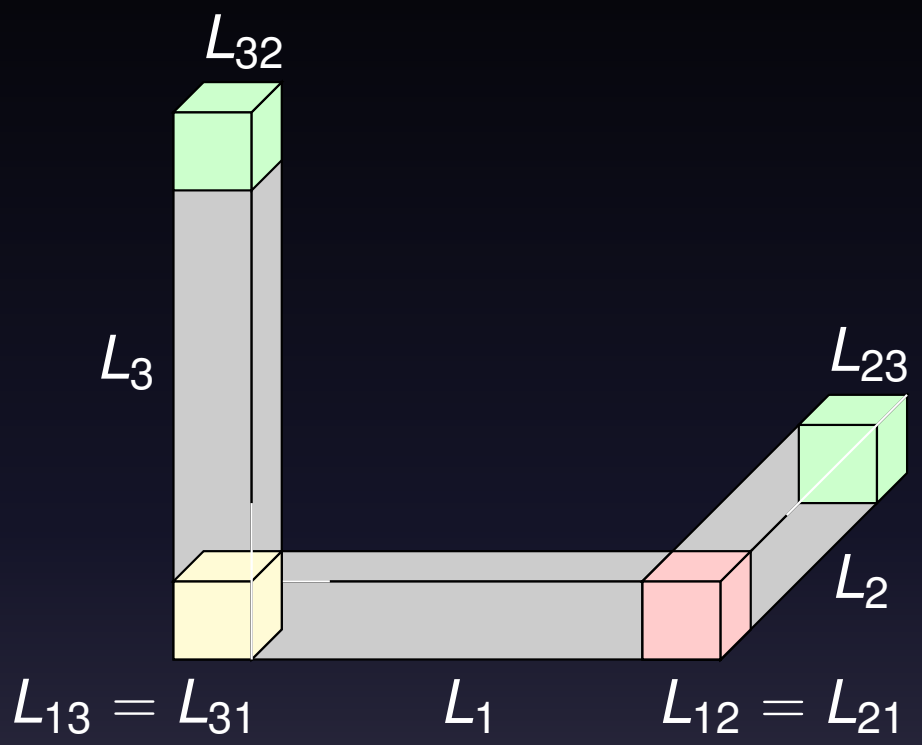
- if L_1, L_2, L_3 can be moved to form tribar, then centers of L_{ij} and L_{ji} coincide and so

$$d'_{ij} = d'_{ji}, \qquad g'_{ij} = 1$$

- i.e., g is coboundary,

$$g_{ij} = \frac{g_i}{g_j}$$

contradiction



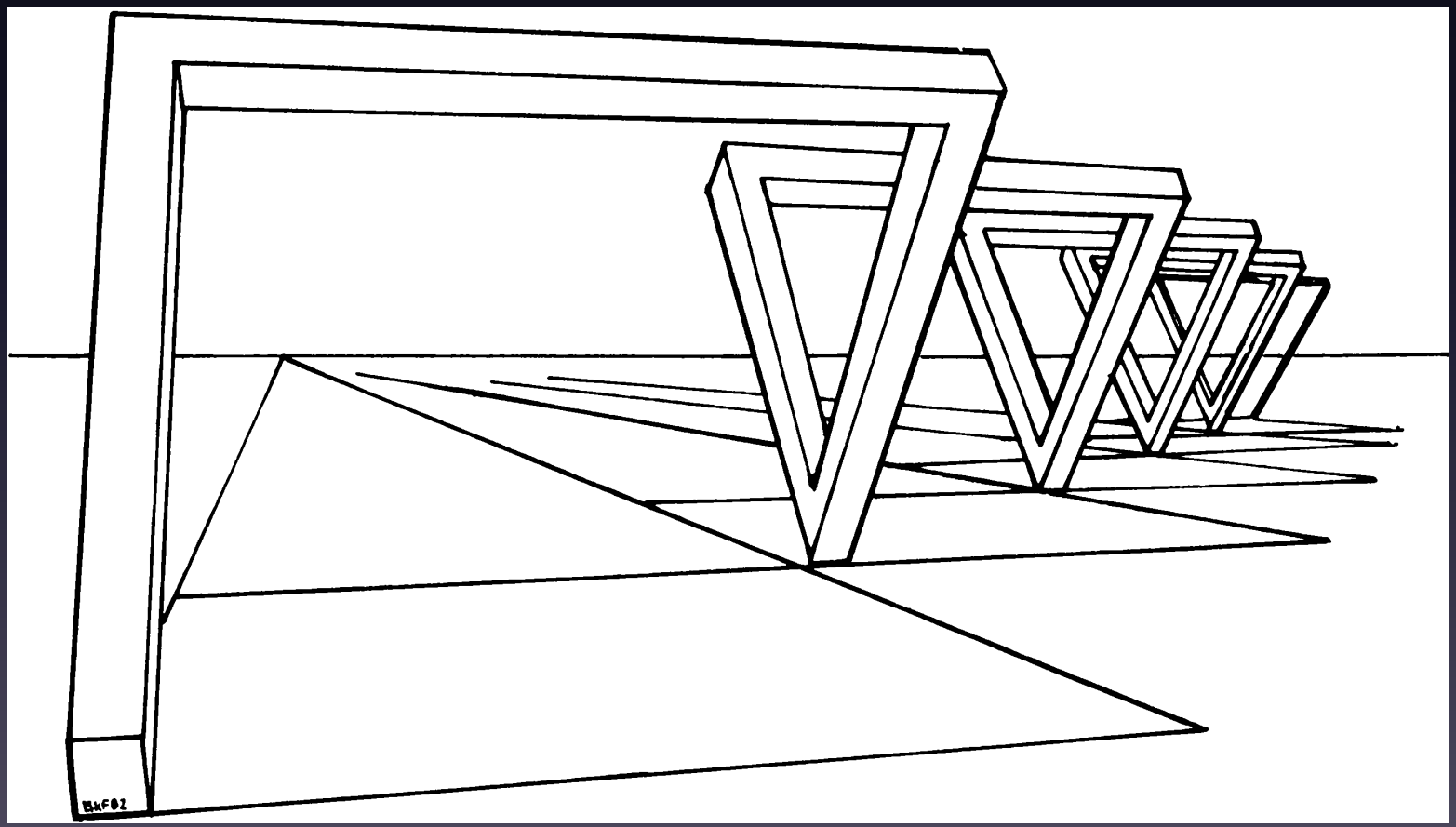
- figure projects to Δ on H

$$g = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & g_{23} \\ 1 & g_{32} & 1 \end{bmatrix}$$

- if tribar exists, then g is coboundary, i.e., $g_{ij} = g_i / g_j$ for some $g_i, g_j \in \mathbb{R}^+$
- so $g_1 = g_2 = g_3$ and so $g_{23} = 1$
- contradiction: L_{23} does not intersect L_{32} and so $g_{23} \neq 1$

other embeddings

- tribar does not exist as a 3D object, i.e., cannot be embedded in $\mathbb{R}^3$
- embeddable in 3-manifold $\mathbb{R}^3/\mathbb{Z} \cong \mathbb{S}^1 \times \mathbb{R}^2$ [Francis, 1987]

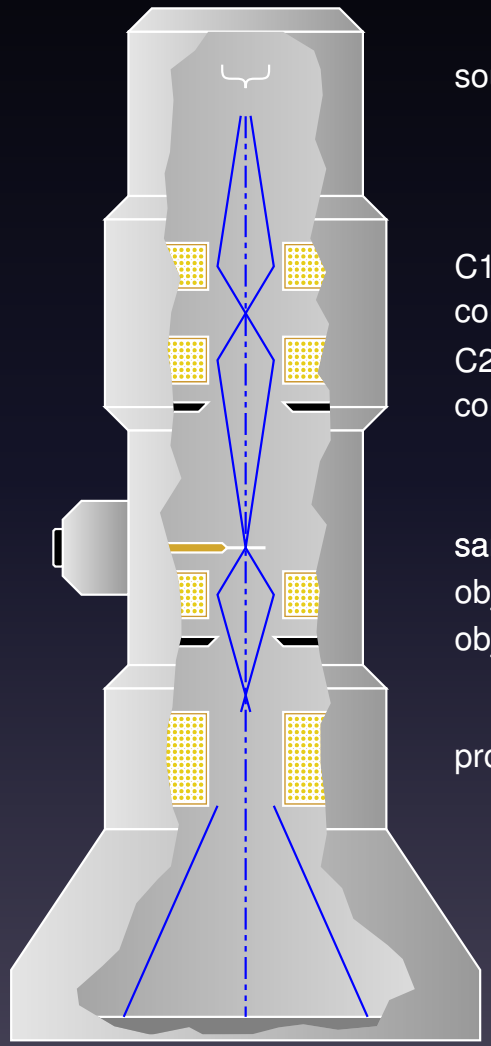


notes

- depends on cohomology of $H^1(Q, \mathbb{R}_+)$ being non-trivial
- not homotopy invariant, unlike embedding Klein bottle in $\mathbb{R}^3$
- tribar homotopy equivalent to torus and trivially embeddable in $\mathbb{R}^3$
- like cryo-EM: construct 3D structure of molecule exactly, not up to homotopy

cryo-electron microscopy

- immobilize many identical copies of molecule in ice
- each copy of molecule frozen in some unknown orientation
- electron microscope produces 2D images
- each 2D image is projection of molecule from an unknown viewing direction
- 2D image shows (i) shape of molecule in plane of viewing direction, (ii) density of molecule, captured in intensity of pixel
- goal: construct 3D structure of molecule from set of 2D projected images



mathematical model

- molecule described by potential function $\varphi : \mathbb{R}^3 \to \mathbb{R}$
- viewing direction described by a point on $\mathbb{S}^2$
- orientation of image is described by 3×3 matrix $A = [a, b, c] \in SO(3)$
- orthonormal column vectors a, b, c such that $\mathrm{span}\{a, b\}$ is projection plane and c viewing direction
- projected image ψ of molecule φ by A given by function $\psi : \mathbb{R}^2 \to \mathbb{R}$

$$\psi(x, y) = \int_{z \in \mathbb{R}} \varphi(xa + yb + zc)\, dz$$

- ψ describes density of molecule along viewing direction

R. Hadani, A. Singer, “Representation theoretic patterns in three-dimensional cryo-electron microscopy I,” *Ann. Math.*, **174** (2011), no. 2, pp. 1219–1241.

cryo-EM data

- $\psi_1, \ldots, \psi_n$ projected images of molecule
- $SO(2)$ action

$$(g \cdot \psi)(x, y) = \psi(g^{-1}(x, y))$$

- distance between images

$$d(\psi_i, \psi_j) = \min_{g \in SO(2)} \|g \cdot \psi_i - \psi_j\|$$

- get

$$g_{ij} = \operatorname*{argmin}_{g \in SO(2)} \|g \cdot \psi_i - \psi_j\|$$

- clearly $g_{ii} = 1$ and $g_{ij} g_{ji} = 1$
- discrete cryo-EM data set

$$D = \{g_{ij} \in SO(2) : i, j = 1, \ldots, n\}$$

cryo-EM complex

- $G_\varepsilon = (V, E)$: $V = \{1, \ldots, n\}$ corresponds to images

$$\{i, j\} \in E \quad \text{if and only if} \quad d(\psi_i, \psi_j) \leq \varepsilon$$

- let K_ε be 2-clique complex of G_ε
  - 0-simplices are vertices in V
  - 1-simplices are edges in E
  - 2-simplices are triangles $\{i, j, k\}$ with $\{i, j\}, \{i, j\}, \{k, i\} \in E$
- for $\varepsilon > 0$ small enough [Singer et al, 2011]

$$g_{ij} g_{jk} g_{ki} = 1$$

- for $\varepsilon > 0$ small enough, get cocycle

$$z_\varepsilon^d = \{g_{ij} \in SO(2) : \{i, j\} \in K_\varepsilon\}$$

classification

- every discrete cryo-EM data set on K_ε is a Čech 1-cocycle z_ε^d on K_ε
- each z_ε^d determines a flat oriented circle bundle on K_ε
- z_ε^d and $z_\varepsilon'^d$ determine isomorphic flat oriented circle bundles if and only if

$$g'_{ij} = g_{ij} g_i g_j^{-1}$$

for some $g_i, g_j \in SO(2)$, $\{i,j\} \in K_\varepsilon$

$$\begin{aligned}\check{H}^1\left(K_\varepsilon, SO(2)_d\right) &= \{\text{cohomologically equivalent} \\ &\qquad \text{discrete cryo-EM data sets on } K_\varepsilon\} \\ &= \{\text{isomorphism classes of} \\ &\qquad \text{flat oriented circle bundles on } K_\varepsilon\}\end{aligned}$$

references

- simple introduction to cohomology
  - L.-H. Lim, “Hodge Laplacians on graphs,” S. Mukherjee (Ed.), *Geometry and Topology in Statistical Inference*, *Proc. Sympos. Appl. Math.*, **73**, AMS, Providence, RI, 2015.
- ranking
  - X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, “Statistical ranking and combinatorial Hodge theory,” *Math. Program.*, **127** (2011), no. 1, pp. 203–244.
  - D. Gleich and L.-H. Lim, “Rank aggregation via nuclear norm minimization,” *Proc. ACM SIGKDD Conf. Knowledge Discovery and Data Mining* (KDD), **17** (2011), pp. 60–68.
  - A. Rajkumar, S. Ghoshal, L.-H. Lim, and S. Agarwal, “Ranking from stochastic pairwise preferences: recovering Condorcet winners and tournament solution sets at the top,” *Proc. Int. Conf. Mach. Learn.* (ICML), **37** (2015), pp. 665–673.
- cryo-EM microscopy
  - K. Ye and L.-H. Lim, “Cohomology of cyro-electron microscopy,” `arXiv:1604.01319`, (2016).