MATH 145: SUPPLEMENTARY NOTES

VIN DE SILVA

Contents

Date: 2013–jan–25.

1. Winding numbers

Let $x \in \mathbb{R}^2$ and let γ be an oriented closed curve in $\mathbb{R}^2 - \{x\}$. Our goal in this section is to define the **winding number**:

$$\mathrm{w}(\gamma, x) = \text{“number of times } \gamma \text{ winds round } x\text{”}$$

The winding number is signed: $+$ for counterclockwise, $-$ for clockwise.

The definition must be rigorous and usable (so that we can prove theorems with it). We will apply this to:

- the fundamental theorem of algebra,
- the Brouwer fixed-point theorem,
- the sensor network coverage problem.

1.1. Angles in $\mathbb{R}^2$.

We wish to define an angle function for points in the plane. This is essentially the same as the ‘argument’ of a nonzero complex number, so we use the same function name, ‘arg’.

Let $\mathbf{a} \in \mathbb{R}^2 - \{\mathbf{0}\}$. Then

$$\arg(\mathbf{a}) = \text{“angle between the positive } x\text{-axis and } \vec{\mathbf{0a}}\text{”}$$

Example. Consider the point $(1,1)$.

$$\begin{aligned} \arg((1,1)) &= \tfrac{\pi}{4} \ \text{ or } \ \tfrac{\pi}{4} + 2\pi \ \text{ or } \ \tfrac{\pi}{4} - 2\pi \ \text{ or } \ \dots \\ &= \tfrac{\pi}{4} + 2\pi n \quad (\text{any } n \in \mathbb{Z}) \end{aligned}$$

Formally, $\arg(\mathbf{a})$ is not a real number, but an element of the quotient set $\mathbb{R}/2\pi\mathbb{Z}$. An element of this quotient set is an equivalence class of numbers:

$$[\phi] = \{\text{all numbers of the form } \phi + 2\pi n, \text{ where } n \in \mathbb{Z}\}$$

Thus $\arg((1,1)) = [\frac{\pi}{4}] = [\frac{9\pi}{4}] = [-\frac{7\pi}{4}]$ etc.

Remark. Just like $\mathbb{R}$, the quotient set $\mathbb{R}/2\pi\mathbb{Z}$ is an abelian group. You can add or subtract two equivalence classes to get another equivalence class:

$$[\phi_1] + [\phi_2] = [\phi_1 + \phi_2], \qquad [\phi_1] - [\phi_2] = [\phi_1 - \phi_2].$$

You cannot multiply by arbitrary scalars. Suppose I wish to define

$$c[\phi] = [c\phi].$$

Well, let’s try multiplying by the scalar $c = \frac{1}{2}$. By the definition,

$$\tfrac{1}{2}[0] = [\tfrac{1}{2} \cdot 0] = [0]$$

and

$$\tfrac{1}{2}[2\pi] = [\tfrac{1}{2} \cdot 2\pi] = [\pi]$$

Since $[0] = [2\pi]$ the left-hand sides of these equations are the same. But the right-hand sides are not the same. Our definition is inconsistent.

Notation. The degree symbol is defined $^\circ = \frac{\pi}{180}$. For example, $360^\circ = 360 \cdot \frac{\pi}{180} = 2\pi$.

Suppose we want an angle function that takes values in $\mathbb{R}$ rather than in the quotient set $\mathbb{R}/2\pi\mathbb{Z}$. Let us call that function $\overline{\arg}$ to distinguish it from arg. It is required to satisfy $\arg(\mathbf{a}) = [\overline{\arg}(\mathbf{a})]$ for any $\mathbf{a}$ in its domain. We must compromise in one of two ways:

- We can require $\overline{\arg}(\mathbf{a})$ to be a real number in a suitable range such as $(-\pi, \pi]$ or $[0, 2\pi)$.
- We can select a ray
$$R_\phi = \left\{\mathbf{a} \in \mathbb{R}^2 - \{\mathbf{0}\} \mid \arg(\mathbf{a}) = [\phi]\right\} \cup \{\mathbf{0}\}$$
and remove it, defining
$$\overline{\arg}_\phi : \mathbb{R}^2 - R_\phi \to (\phi, \phi + 2\pi).$$

With the first approach, $\overline{\arg}$ is discontinuous somewhere (on R_π if the range is $(-\pi, \pi]$, on R_0 if the range is $[0, 2\pi)$, for example). With the second approach, the function is continuous on its domain, but the domain is smaller.

Remark. How do we prove that $\overline{\arg}_\phi$ is continuous on $\mathbb{R}^2 - R_\phi$?

- Here is a geometric argument. Consider a point $\mathbf{a}$ and a nearby point $\mathbf{a} + \Delta\mathbf{a}$, both in the domain. Suppose we wish $|\overline{\arg}_\phi(\mathbf{a} + \Delta\mathbf{a}) - \overline{\arg}_\phi(\mathbf{a})| < \epsilon$. By trigonometry, we can ensure this by requiring that $|\Delta\mathbf{a}|$ is smaller than $|\mathbf{a}| \sin\epsilon$ and smaller than the distance between $\mathbf{a}$ and the ray R_ϕ.
- Here is an algebraic argument. Consider $\overline{\arg}_{-\pi}$, which is defined everywhere except the negative x-axis and $\mathbf{0}$, and which takes values in $(-\pi, \pi)$. By calculating the complex square root of $x + yi$, we can show that
$$\overline{\arg}_{-\pi}((x, y)) = 2 \arctan\left(\frac{y}{x + \sqrt{x^2 + y^2}}\right).$$
Given this explicit formula, the standard rules of analysis imply that the function is continuous over the region where $x + \sqrt{x^2 + y^2}$ is nonzero, which is precisely $\mathbb{R}^2 - R_\pi$. The continuity of $\overline{\arg}_{-\pi}$ implies the continuity of each $\overline{\arg}_\phi$, for instance by rotating the plane.

There are some easy explicit formulas for arg that are valid on different half-planes.

$$\arg((x, y)) = \begin{cases} [\arctan(y/x)] & \text{if } x > 0 \\ [\arctan(y/x) + \pi] & \text{if } x < 0 \\ [-\arctan(x/y) + \frac{\pi}{2}] & \text{if } y > 0 \\ [-\arctan(x/y) - \frac{\pi}{2}] & \text{if } y < 0 \end{cases}$$

The formulas inside the square brackets can be thought of as versions of $\overline{\arg}$ defined over the specified half-planes.

1.2. **The angle subtended by a line segment.** Given two points $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$ we let $[\mathbf{a}, \mathbf{b}]$ denote the directed line segment from $\mathbf{a}$ to $\mathbf{b}$. We can parametrize this

$$\gamma : [0,1] \to \mathbb{R}^2; \ \gamma(t) = \mathbf{a} + t(\mathbf{b} - \mathbf{a})$$

and we let

$$|[\mathbf{a}, \mathbf{b}]| = \{\mathbf{a} + t(\mathbf{b} - \mathbf{a}) \mid t \in [0,1]\}$$

denote the set of points on the line segment; this is a subset of $\mathbb{R}^2$.

Definition (angle cocycle)**.** Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$ and suppose $\mathbf{0} \notin |[\mathbf{a}, \mathbf{b}]|$. Define

$$\theta([\mathbf{a}, \mathbf{b}], \mathbf{0}) = \text{“the unique } \theta \in (-\pi, \pi) \text{ such that } [\theta] = \arg(\mathbf{b}) - \arg(\mathbf{a})\text{”}$$

Despite the quotation marks, the definition is precise. There is certainly such a number θ in the range $(-\pi, \pi]$; and $\theta = \pi$ is ruled out by the condition that $\mathbf{0}$ is not directly between $\mathbf{a}$ and $\mathbf{b}$.

Definition. More generally, suppose $\mathbf{x} \notin |[\mathbf{a}, \mathbf{b}]|$. Define $\theta([\mathbf{a}, \mathbf{b}], \mathbf{x}) = \theta([\mathbf{a} - \mathbf{x}, \mathbf{b} - \mathbf{x}], \mathbf{0})$.

We think of $\theta([\mathbf{a}, \mathbf{b}], \mathbf{x})$ as the signed angle subtended at $\mathbf{x}$ by $[\mathbf{a}, \mathbf{b}]$. You can check that the sign is positive if $\mathbf{x}, \mathbf{a}, \mathbf{b}$ are arranged counterclockwise, and negative if they are arranged clockwise.

Remark. Note that $\theta([\mathbf{a}, \mathbf{b}], \mathbf{x})$ is a real number whereas $\arg(\mathbf{a})$ is an equivalence class of real numbers. This is important. This whole concept of winding number is built on the ‘tension’ between $\mathbb{R}$ and $\mathbb{R}/2\pi\mathbb{Z}$.

Proposition 1.1. *Fix $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$. The function $\mathbf{x} \mapsto \theta([\mathbf{a}, \mathbf{b}], \mathbf{x})$ is continuous over its domain $\mathbb{R}^2 - |[\mathbf{a}, \mathbf{b}]|$.*

Proof. Think of $\mathbf{a}, \mathbf{b}, \mathbf{x}$ as complex numbers. Consider $\mathbf{t} = (\mathbf{b} - \mathbf{x})/(\mathbf{a} - \mathbf{x})$. Because of the way complex multiplication works, we have

$$\arg(\mathbf{t}) = \arg(\mathbf{b} - \mathbf{x}) - \arg(\mathbf{a} - \mathbf{x})$$

On the other hand, because $\mathbf{x} \notin |[\mathbf{a}, \mathbf{b}]|$ we know that $\mathbf{t}$ does not lie on the ray R_π, so

$$[\overline{\arg}_{-\pi}(\mathbf{t})] = \arg(\mathbf{b} - \mathbf{x}) - \arg(\mathbf{a} - \mathbf{x}).$$

Since $\overline{\arg}_{-\pi}(\mathbf{t})$ lies in $(-\pi, \pi)$, it must be “the unique θ” that we seek. It follows that we have the explicit formula

$$\theta([\mathbf{a}, \mathbf{b}], \mathbf{x}) = \overline{\arg}_{-\pi}\left(\frac{\mathbf{b} - \mathbf{x}}{\mathbf{a} - \mathbf{x}}\right).$$

This is the composite of continuous functions, and hence is continuous over the domain where it is defined, that is over $\mathbb{R}^2 - |[\mathbf{a}, \mathbf{b}]|$. $\square$

1.3. **The discrete winding number.** Let $\gamma = P(\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_n)$ denote the polygon whose vertices are $\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^2$ and whose edges are the directed line segments

$$[\mathbf{a}_0, \mathbf{a}_1], [\mathbf{a}_1, \mathbf{a}_2], \ldots, [\mathbf{a}_{n-1}, \mathbf{a}_n].$$

We say that γ is a 'directed' polygon. It is **closed** if $\mathbf{a}_0 = \mathbf{a}_n$. The **support** of γ is the set of points

$$|\gamma| = \bigcup_{k=1}^{n} |[\mathbf{a}_{k-1}, \mathbf{a}_k]|$$

on the edges of γ. This is a subset of $\mathbb{R}^2$. We will increasingly often think of γ as the sum of its edges:

$$\gamma = [\mathbf{a}_0, \mathbf{a}_1] + [\mathbf{a}_1, \mathbf{a}_2] + \cdots + [\mathbf{a}_{n-1}, \mathbf{a}_n] = \sum_{k=1}^{n} [\mathbf{a}_{k-1}, \mathbf{a}_k]$$

Since addition is supposed to be commutative, this notation suggests that the order of the edges is not important.

Definition (winding number)**.** Let $\gamma = P(\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_n)$ be a directed polygon, and let $\mathbf{x} \in |\gamma|$. The winding number of γ about $\mathbf{x}$ is defined:

$$\mathrm{w}(\gamma, \mathbf{x}) = \frac{1}{2\pi} \sum_{k=1}^{n} \theta([a_{k-1}, a_k], \mathbf{x})$$

Remark. If we do think of γ as the sum of its edges, then this formula is linear.

Example. In class, we considered an example with γ being the four edges of a square, oriented counterclockwise, and three choices for $\mathbf{x}$. It seemed that inside the square the winding number of γ is 1, and outside the square the winding number is 0.

Theorem 1.2. *Let γ be a* **closed** *directed polygon, and let $\mathbf{x} \in \mathbb{R}^2 - |\gamma|$. Then $\mathrm{w}(\gamma, \mathbf{x})$ is an integer.*

Proof. For each vertex $\mathbf{a}_k$, select a real number $\phi_k \in \arg(\mathbf{a}_k - \mathbf{x})$. In other words, find a real number such that $\arg(\mathbf{a}_k - \mathbf{x}) = [\phi_k]$. Arrange also that $\phi_0 = \phi_n$. Notice that for each k we can write

$$\theta([\mathbf{a}_{k-1}, \mathbf{a}_k], \mathbf{x}) = \phi_k - \phi_{k-1} + 2\pi m_k$$

where $m_k \in \mathbb{Z}$. Summing from $k = 1, \ldots n$ we get

$$2\pi\, \mathrm{w}(\gamma, \mathbf{x}) = 2\pi(m_1 + \cdots + m_n)$$

since all the ϕ_k terms cancel out. Thus $\mathrm{w}(\gamma, \mathbf{x}) = m_1 + \cdots + m_n$ is an integer. $\square$

Remark. From the proof we see that each edge of γ contributes m_k to the winding number. We discern this contribution only when the polygon is closed, because that is when the ϕ_k terms cancel out.

1.4. **Oriented 1-cycles.** Which other configurations of directed edges give rise to an integer-valued winding number?

Consider points $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ and $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ in $\mathbb{R}^2$. Let γ be the formal sum

$$\gamma = [\mathbf{a}_1, \mathbf{b}_1] + [\mathbf{a}_1, \mathbf{b}_2] + \cdots + [\mathbf{a}_n, \mathbf{b}_n] = \sum_{k=1}^{n} [\mathbf{a}_k, \mathbf{b}_k].$$

Such a thing is called a **1-chain**, the '1' referring to the fact the edges are 1-dimensional.

We say that γ is **closed** if every point in $\mathbb{R}^2$ occurs equally often in the collection $(\mathbf{a}_k)$ as in the collection $(\mathbf{b}_k)$. In other words, if every point in the plane is a 'head' equally as often as it is a 'tail'. In other words, if the multi-set $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$ is equal to the multi-set $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$.

A closed 1-chain is generally called a **1-cycle** or sometimes just a **cycle**.

(We did some examples in class.)

Theorem 1.3. *Let γ be a 1-cycle, and let $\mathbf{x} \in \mathbb{R}^2 - |\gamma|$. Then $\mathrm{w}(\gamma, \mathbf{x})$ is an integer.*

Proof. The proof is the same as for Theorem 1.2. Let $\overline{\arg}$ be a fixed choice of real-valued angle function defined on $\mathbb{R}^2 - \mathbf{0}$ (is is not, of course, continuous). Then for every edge we have

$$\theta([\mathbf{a}_k, \mathbf{b}_k], \mathbf{x}) = \overline{\arg}(\mathbf{b}_k - \mathbf{x}) - \overline{\arg}(\mathbf{a}_k - \mathbf{x}) + 2\pi m_k$$

for some $m_k \in \mathbb{Z}$. Summing over all edges, the $\overline{\arg}$ terms cancel and we are left with $\mathrm{w}(\gamma, \mathbf{x}) = m_1 + \cdots + m_n$. $\square$

1.5. **Formal properties of the winding number.** We collect some basic results on the winding number.

Winding number is additive. It is clear from the definition that if γ_1, γ_2 are 1-cycles then so is $\gamma_1 + \gamma_2$. By linearity of the defining formula, it follows that if $\mathbf{x} \in \mathbb{R}^2 - |\gamma_1| - |\gamma_2|$ then

$$\mathrm{w}(\gamma, \mathbf{x}) = \mathrm{w}(\gamma_1, \mathbf{x}) + \mathrm{w}(\gamma_2, \mathbf{x}).$$

Winding number is locally constant. For a fixed γ, consider the domain $\mathbb{R}^2 - |\gamma|$ of the function $\mathbf{x} \mapsto \mathrm{w}(\gamma, \mathbf{x})$. This is in general a finite union of connected open sets (each set is sometimes called a 'chamber').

Theorem 1.4. *The function $f(\mathbf{x}) = \mathrm{w}(\gamma, \mathbf{x})$ is constant on each connected component of $\mathbb{R}^2 - |\gamma|$.*

Proof. We know that f is continuous (being the sum of continuous functions $\frac{1}{2\pi}\theta([\mathbf{a}_k, \mathbf{b}_k], \mathbf{x})$) and integer valued (by Theorem 1.3) so it must be locally constant. $\square$

The ray escape formula. Consider a directed line segment $[\mathbf{a}, \mathbf{b}]$ and a ray

$$R_\phi(\mathbf{x}) = \big\{\mathbf{p} \in \mathbb{R}^2 - \{\mathbf{x}\} \mid \arg(\mathbf{p} - \mathbf{x}) = [\phi]\big\} \cup \{\mathbf{x}\}$$

(This is the ray originating at $\mathbf{x}$ at angle ϕ.)

Suppose $\mathbf{x} \notin |[\mathbf{a}, \mathbf{b}]|$ and $\mathbf{a}, \mathbf{b} \notin R_\phi(\mathbf{x})$. Then we can talk about the **crossing number**:

$$\mathrm{X}([\mathbf{a}, \mathbf{b}], R_\phi(\mathbf{x})) = \begin{cases} +1 & \text{if } [\mathbf{a}, \mathbf{b}] \text{ crosses the ray in counterclockwise direction} \\ -1 & \text{if } [\mathbf{a}, \mathbf{b}] \text{ crosses the ray in clockwise direction} \\ 0 & \text{if } [\mathbf{a}, \mathbf{b}] \text{ is disjoint from the ray} \end{cases}$$

We can make this a little more formal using the real-valued angle function $\overline{\arg}_\phi : \mathbb{R}^2 - R_\phi \to (\phi, \phi + 2\pi)$. By considering different cases, one sees that

$$\begin{aligned} \text{counterclockwise crossing} \;\Leftrightarrow\; -2\pi &< \overline{\arg}_\phi(\mathbf{b} - \mathbf{x}) - \overline{\arg}_\phi(\mathbf{a} - \mathbf{x}) < -\pi \\ \text{no crossing} \;\Leftrightarrow\; -\pi &< \overline{\arg}_\phi(\mathbf{b} - \mathbf{x}) - \overline{\arg}_\phi(\mathbf{a} - \mathbf{x}) < \pi \\ \text{clockwise crossing} \;\Leftrightarrow\; \pi &< \overline{\arg}_\phi(\mathbf{b} - \mathbf{x}) - \overline{\arg}_\phi(\mathbf{a} - \mathbf{x}) < 2\pi \end{aligned}$$

In more detail. To get a clockwise crossing, we need $\mathbf{b}$ to lie clockwise of the ray, $\mathbf{a}$ to lie counterclockwise of the ray, and the angle between them on the ray side to be less than 180°. This is perfectly expressed by the inequality

$$\overline{\arg}_\phi(\mathbf{b} - \mathbf{x}) - \overline{\arg}_\phi(\mathbf{a} - \mathbf{x}) > \pi.$$

Given that $\overline{\arg}_\phi$ takes values in $(\phi, \phi + 2\pi)$, the other inequality

$$\overline{\arg}_\phi(\mathbf{b} - \mathbf{x}) - \overline{\arg}_\phi(\mathbf{a} - \mathbf{x}) < 2\pi$$

is always satisfied. The counterclockwise case is the same with $\mathbf{a}, \mathbf{b}$ interchanged, and the no-crossing case is everything that's left.

Since $\theta([\mathbf{a}, \mathbf{b}], \mathbf{x})$ is required lie in the interval $(-\pi, \pi)$, we can deduce what multiple of 2π must be added to each quantity above to put it in the range $(-\pi, +\pi)$. Specifically:

$$\theta([\mathbf{a}, \mathbf{b}], \mathbf{x}) = \overline{\arg}_\phi(\mathbf{b} - \mathbf{x}) - \overline{\arg}_\phi(\mathbf{a} - \mathbf{x}) + \begin{cases} 2\pi & \text{counterclockwise crossing} \\ 0 & \text{no crossing} \\ -2\pi & \text{clockwise crossing} \end{cases}$$

We can summarize this as

$$\boxed{\theta([\mathbf{a}, \mathbf{b}], \mathbf{x}) = \overline{\arg}_\phi(\mathbf{b} - \mathbf{x}) - \overline{\arg}_\phi(\mathbf{a} - \mathbf{x}) + 2\pi\, \mathrm{X}([\mathbf{a}, \mathbf{b}], R_\phi(\mathbf{x}))} \tag{1.5}$$

whenever the crossing number is defined.

Remark. The formula (1.5) can be regarded either as a theorem about the crossing number, or as a formal definition of the crossing number. Which logic do you prefer?

Corollary 1.6 (ray escape formula)**.** *Let $\gamma = \sum_{k=1}^n [\mathbf{a}_k, \mathbf{b}_k]$ be a 1-cycle. Let $\mathbf{x} \in \mathbb{R}^2 - |\gamma|$, and let $R_\phi(\mathbf{x})$ be a ray originating at $\mathbf{x}$ which meets none of the vertices of γ. Then*

$$\mathrm{w}(\gamma, \mathbf{x}) = \sum_{k=1}^n \mathrm{X}([\mathbf{a}, \mathbf{b}], R_\phi(\mathbf{x}))$$

Proof. Sum (1.5) over the edges of γ. The $\overline{\arg}_\phi$ terms cancel since γ is a 1-cycle. $\square$

Using the results in this section, it is straightforward to determine the function $\mathbf{x} \mapsto \mathrm{w}(\gamma, \mathbf{x})$ for any explicitly given 1-cycle γ.

1.6. The fundamental theorem of algebra.

Theorem 1.7 (The fundamental theorem of algebra)**.** *The field of complex numbers $\mathbb{C}$ is algebraically closed. In other words, given a complex polynomial of degree $d \geq 1$*

$$p(z) = z^d + \sum_{k=0}^{d-1} a_k z^k \qquad \text{where } a_0, \dots, a_{d-1} \in \mathbb{C}$$

there exist complex numbers $\lambda_1, \dots, \lambda_d$ such that

$$p(z) \equiv (z - \lambda_1)(z - \lambda_2) \cdots (z - \lambda_d).$$

The λ_k are called the 'roots' of p, since $p(\lambda_k) = 0$ for all k.

The following equivalent version is easier to prove.

Theorem 1.8. *Any complex polynomial of degree ≥ 1 has at least one root.*

This implies Theorem 1.7 because if λ is a root of p then the long division of $p(z)$ by $(z - \lambda)$ leaves remainder zero. Therefore $p(z) = (z - \lambda)\, q(z)$ for some polynomial $q(z)$, and we can repeat (i.e. use induction) to factorize completely.

The field of real numbers $\mathbb{R}$ is not algebraically closed, since $p(x) = x^2 + 1$ has no real root. It does have the following property, which comes close.[1]

Proposition 1.9. *However, let $p(x)$ be a real polynomial of degree d where d is odd. Then $p(x)$ has a root.*

Proof. The leading term x^d "dominates" when $|x|$ is large. Let R be a large positive number. Then:

$$\begin{aligned} p(R) &\approx R^d & &> 0 \\ p(-R) &\approx (-R)^d = -R^d & &< 0 \qquad \text{(since } d \text{ is odd).} \end{aligned}$$

The intermediate value theorem now guarantees a root $\lambda \in (-R, R)$. □

Details. Let us be precise about how large R has to be. Suppose

$$R > |a_0| + |a_1| + \cdots + |a_{d-1}| \qquad \text{and} \qquad R \geq 1.$$

[1] How close? There is a Galois theory proof of FTA which uses only the fact that $\mathbb{C}$ is a quadratic extension of a field (namely $\mathbb{R}$) in which every odd-degree polynomial has a root.

Then

$$\begin{aligned}
p(R) &= R^d + \sum_{k=0}^{d-1} a_k R^k \\
&\geq R^d - \sum_{k=0}^{d-1} |a_k| R^k \\
&\geq R^d - \sum_{k=0}^{d-1} |a_k| R^{d-1} && \text{since } R \geq 1 \\
&= R^{d-1} \left(R - (|a_0| + |a_1| + \cdots + |a_{d-1}|) \right)
\end{aligned}$$

which is positive by choice of R. A similar calculation shows that $p(-R) < 0$. □

Remark. The IVT is a 1-dimensional topological theorem. We will prove the FTA by mimicking this argument in 2D. We replace the IVT by a topological theorem involving winding numbers.

Strategy for proof of FTA. We will evaluate $p(z)$ on circles $z = re^{2\pi it}$, for varying values of r, and we consider the winding number of the curves $p(re^{2\pi it})$ about zero.

- When r is very large, the z^d term dominates and $\mathrm{w}(p(re^{2\pi it}), 0) = d$,
- When $r = 0$, the curve is constant and the winding number is zero.
- Therefore, for some intermediate value of r, the curve $p(re^{2\pi it})$ must cross zero.

To make this work, we must rigorously define $\mathrm{w}(f, 0)$ and establish the necessary properties.

1.7. **Loops and homotopies.** Previously, we defined winding numbers for 1-cycles. Now we will define winding numbers for closed curves. To begin with, we must talk about **loops**, as well as an equivalence relation between loops called **homotopy**. This equivalence relation allows us to use loops to detect the topology of a region D in the plane.

Definition. Let $D \subseteq \mathbb{R}^2$. A loop in D is a continuous function

$$f : [0,1] \to D$$

such that $f(0) = f(1)$. The set of all loops in D is denoted $\mathrm{Loops}(D)$.

Definition. Two loops $f, g \in \mathrm{Loops}(D)$ are **homotopic** (in D) if there exists a continuous function on the unit square

$$H : [0,1] \times [0,1] \to D$$

such that

- $H(0,t) = f(t)$ for all $t \in [0,1]$,
- $H(1,t) = g(t)$ for all $t \in [0,1]$,
- $H(s,0) = H(s,1)$ for all $s \in [0,1]$.

If such an H exists, we write $f \simeq g$ or $f \simeq_H g$.

Remark. In other words, H equals f when restricted to the left-hand edge of the square, and g when restricted to the right-hand edge. In between, the restriction to each vertical segment of the square must also be a loop.

Remark. It follows that if $f \simeq_H g$ then we can think of the homotopy H as specifying a path γ_H in $\mathrm{Loops}(D)$ from f to g. At each $s \in [0,1]$, the loop $\gamma_H(s)$ is $(t \mapsto H(s,t))$. It is possible to define a topology on $\mathrm{Loops}(D)$ such that the continuity of

$$H : [0,1] \times [0,1] \to D$$

is equivalent to the continuity of

$$\gamma_H : [0,1] \to \mathrm{Loops}(D).$$

Powerful things happen when this is achieved. But please don't worry about it too much.

Proposition 1.10. *The relation $\simeq$ on $\mathrm{Loops}(D)$ is an equivalence relation:*

- $f \simeq f$
- $f \simeq g$ *implies* $g \simeq f$
- $f \simeq g$ *and* $g \simeq h$ *imply* $f \simeq h$

for all $f, g, h \in \mathrm{Loops}(D)$.

Proof. See homework. □

In order to exploit this equivalence relation, we need techniques for showing that $f \simeq g$ and also techniques for showing that $f \not\simeq g$. The simplest way to construct a homotopy is to interpolate linearly between f and g. This works provided that the interpolation remains within the domain D:

Proposition 1.11. *Let $D \subseteq \mathbb{R}^2$ and let $f, g \in \mathrm{Loops}(D)$. Suppose that for every $t \in [0,1]$ the line segment $|[f(t), g(t)]|$ is contained in D. Then $f \simeq g$ in D.*

Proof. The function $H(s,t) = (1-s)f(t) + sg(t)$ is a homotopy from f to g. Indeed, it is continuous because products and sums of continuous functions are continuous, and the three bulleted conditions are easily verified. □

The function H in the proof is called the **straight-line homotopy** from f to g.

Corollary 1.12. *Any two loops $f, g \in \mathrm{Loops}(\mathbb{R}^2)$ are homotopic.* □

Example 1.13. Not all f, g in $\mathrm{Loops}(\mathbb{R}^2 - \{0\})$ are homotopic. Consider:

$$\begin{aligned} f(t) &= e^{2\pi i t} \\ g(t) &= (-1+i) + \tfrac{1}{10} e^{2\pi i t} \end{aligned}$$

The straight-line homotopy doesn't work, because at $t = \frac{3}{2}\pi$ the interval $[f(t), g(t)]$ passes through 0. Intuitively, it seems that there should be no way to devise a homotopy between f and g. But how are we to prove this?

1.8. **The continuous winding number.**

Theorem 1.14. *There exists a function* $\mathrm{Loops}(\mathbb{R}^2 - \{0\}) \to \mathbb{Z}$*, expressed by the notation* $f \mapsto \mathrm{w}(f, 0)$*, such that*

- *if* $f \simeq g$ *in* $\mathrm{Loops}(\mathbb{R}^2 - \{0\})$ *then* $\mathrm{w}(f, 0) = \mathrm{w}(g, 0))$*, and*
- *if* $f = e^{2\pi i d t}$ *then* $\mathrm{w}(f, 0) = d$*.*

Definition. For any point $x \in \mathbb{R}^2$ and $f \in \mathrm{Loops}(\mathbb{R}^2 - \{x\})$ we set $\mathrm{w}(f, x) = \mathrm{w}(f - x, 0)$.

Remark. One of the strengths of this theorem is that it can be used very effectively as a 'black box'. The existence of a function w with these properties is a powerful mathematical fact with many consequences. The details of the proof (and there are many proofs) are encapsulated and kept separate from the uses of the theorem.

Solution to Example 1.13. Loop g is homotopic to the constant loop 1 by a straight-line homotopy. Indeed, g is contained in the strict interior of the second quadrant, so the line segments $[g(t), 1]$ all avoid 0. It follows that

$$\mathrm{w}(g, 0) = \mathrm{w}(1, 0) = \mathrm{w}(e^{2\pi i 0 t}, 0) = 0.$$

On the other hand

$$\mathrm{w}(f, 0) = \mathrm{w}(e^{2\pi i t}, 0) = 1.$$

Since $\mathrm{w}(f, 0) \neq \mathrm{w}(g, 0)$ it follows that $f \not\simeq g$ in $\mathbb{R}^2 - \{0\}$. $\square$

Proof of Theorem 1.8, and therefore FTA. Seeking a contradiction, we may suppose that $p(z) = z^d + \sum_{k=0}^{d-1} a_k z^k$ is a complex polynomial with no roots. Let

$$R = 1 + |a_0| + |a_1| + \cdots + |a_{d-1}|$$

and consider the following loops in $\mathbb{R}^2 - \{0\}$:

$$f(t) = e^{2\pi i d t}, \qquad g(t) = R^d e^{2\pi i d t}, \qquad h(t) = p(Re^{2\pi i t}), \qquad k(t) = a_0, \qquad \ell(t) = 1.$$

It can be shown (see homework) that $f \simeq g \simeq h \simeq k \simeq \ell$ in $\mathbb{R}^2 - \{0\}$, and therefore

$$d = \mathrm{w}(f, 0) = \mathrm{w}(g, 0) = \mathrm{w}(h, 0) = \mathrm{w}(k, 0) = \mathrm{w}(\ell, 0) = 0.$$

This is a contradiction, so the assumption that $p(z)$ has no roots must be false. $\square$

Remark. The only use of the assumption on p is to construct the homotopy $h \simeq k$ in $\mathbb{R}^2 - \{0\}$. That is where the contradiction happens. The existence of a root breaks the red "=" sign and avoids the contradiction. (The assumption also rules out the trivial case $a_0 = 0$.)

A powerful black box indeed. Let us now justify its use.

Proof of Theorem 1.14. Here is our strategy.

- (Step 1) Subdivide the interval $[0,1]$ and define $\mathrm{w}(f,0)$ to be the discrete winding number of the polygon obtained from f using this subdivision.
- (Step 2) Show that this definition is independent of the choice of subdivision.
- (Step 3) Show that two homotopic loops have the same winding number.
- (Step 4) Verify that $\mathrm{w}(e^{2\pi i d t},0)=d$.

There is some technical input from analysis (e.g. Math 131). We use the facts that a real-valued continuous function on a compact space is bounded and attains its bounds, and that a continuous function from a compact metric space to another metric space is uniformly continuous. The compact space will be the interval $[0,1]$ for the first two steps, and the unit square $[0,1]\times[0,1]$ for the third step.

We proceed with the proof. Let $f \in \mathrm{Loops}(\mathbb{R}^2 - \{0\})$.

Step 1. Since the interval $[0,1]$ is compact (i.e. closed and bounded):

- Let $m = \min_t(|f(t)|)$, the minimum distance between the loop and 0.
- Let $\delta > 0$ be such that $|t_1 - t_2| < \delta$ implies $|f(t_1) - f(t_2)| < m$.

We know that m exists and is strictly positive, since the minimum distance is attained and f never reaches 0. Then δ exists since f is uniformly continuous.

Let T be a subdivision of the interval $[0,1]$. That is, let

$$T = (t_0, t_1, \ldots, t_n)$$

where

$$0 = t_0 < t_1 < \cdots < t_n = 1.$$

For any such T, we define a 1-chain

$$f_T = \sum_{k=1}^{n} [f(t_{k-1}), f(t_k)].$$

We wish to define $\mathrm{w}(f,0) = \mathrm{w}(f_T,0)$. We must confirm that the edges of f_T do not cross 0. It is useful to define the **mesh-size**

$$\mathrm{mesh}(T) = \max_k(|t_k - t_{k-1}|).$$

Proposition 1.15. *If* $\mathrm{mesh}(T) < \delta$*, then* f_T *is a 1-chain in* $\mathbb{R}^2 - \{0\}$*.*

Proof. For each edge $[f(t_{k-1}), f(t_k)]$ we can argue as follows. Let $r_k = |f(t_k) - f(t_{k-1})|$. Then the entire edge lies inside a disk of radius r_k with center $f(t_k)$. But $|f(t_k)| \geq m$, and $|t_k - t_{k-1}| < \delta$ implies that $r_k < m$. It follows that this disk does not meet 0, and therefore the edge does not meet 0. This is true for every edge, so $|f_T| \subset \mathbb{R}^2 - \{0\}$ as required. $\square$

We can now provisionally define $\mathrm{w}(f,0) = \mathrm{w}(f_T,0)$, where T is any subdivision of $[0,1]$ satisfying $\mathrm{mesh}(T) < \delta$. This completes Step 1.

Step 2. We show that if $\mathrm{mesh}(S), \mathrm{mesh}(T) < \delta$ then $\mathrm{w}(f_S,0) = \mathrm{w}(f_T,0)$.

Suppose first that S is obtained from T by adding one extra point somewhere:

$$\begin{aligned} T &= (t_0, \dots, t_{k-1}, t_k, \dots, t_n), \\ S &= (t_0, \dots, t_{k-1}, s, t_k, \dots, t_n). \end{aligned}$$

We can compare the winding numbers directly: the difference in the formulas involves three edges, all other terms being the same. Let us write $a = f(t_{k-1})$, $b = f(s)$, $c = f(t_k)$. Then

$$\begin{aligned} \mathrm{w}(f_S,0) - \mathrm{w}(f_T,0) &= \tfrac{1}{2\pi}\left(\theta([a,b],0) + \theta([b,c],0) - \theta([a,c],0)\right) \\ &= \tfrac{1}{2\pi}\left(\theta([a,b],0) + \theta([b,c],0) + \theta([c,a],0)\right) \\ &= \mathrm{w}([a,b] + [b,c] + [c,a], 0). \end{aligned}$$

To show that this is zero, notice that $|a| \geq m$ and $|a-b|, |a-c| < m$. It follows that a, b, c, and therefore the three edges between them, are contained in a circular disk that is disjoint from the origin. By selecting a ray at the origin which points away from the center of this disk, we deduce that this winding number is zero and hence $\mathrm{w}(f_S,0) = \mathrm{w}(f_T,0)$.

For the general case, let S, T be two arbitrary subdivisions of mesh-size less than δ and let U be the subdivision obtained as the union of the points in S and T. By adding points to S one-by-one it follows that $\mathrm{w}(f_S,0) = \mathrm{w}(f_U,0)$, and by adding points to T one-by-one it follows that $\mathrm{w}(f_T,0) = \mathrm{w}(f_U,0)$. Thus $\mathrm{w}(f_S,0) = \mathrm{w}(f_T,0)$. This completes Step 2.

Step 3. Suppose f, g are loops that are homotopic in $\mathbb{R}^2 - \{0\}$ through a map H. We will express the difference $\mathrm{w}(f,0) - \mathrm{w}(g,0)$ in a clever way, and show that it is zero.

The domain of H is the square $[0,1] \times [0,1]$. Since this is compact:

- Let $m = \min_t(|H(t)|) > 0$, the minimum distance between the image of H and 0.
- Let $\delta > 0$ be such that $|s_1 - s_2| < \delta$ and $|t_1 - t_2| < \delta$ together imply $|H(s_1,t_1) - H(s_2,t_2)| < m$.

Now let $N > \frac{1}{\delta}$, and consider the subdivision

$$T = \left(0, \tfrac{1}{N}, \tfrac{2}{N}, \dots, \tfrac{N-1}{N}, 1\right)$$

of the unit interval. Then

$$\mathrm{w}(f,0) = \mathrm{w}(f_T,0), \qquad \mathrm{w}(g,0) = \mathrm{w}(f_T,0).$$

This is because δ as chosen above is small enough to function as the δ defined in Step 1, for both f and g.

The next step is to subdivide $[0,1] \times [0,1]$ into N^2 equal squares of side-length $\frac{1}{N}$.

Suppose the vertices of the k-th square are labelled a_k, b_k, c_k, d_k in counterclockwise order; define a polygon

$$\gamma_k = [H(a_k), H(b_k)] + [H(b_k), H(c_k)] + [H(c_k), H(d_k)] + [H(d_k), H(a_k)].$$

By choice of δ, we know that $|H(a_k)| \geq m$ and that each of $H(b_k)$, $H(c_k)$, $H(d_k)$ lies at distance less than m from $H(a_k)$. Thus the entire polygon γ_k is contained in a disk that does not meet the origin. We deduce

$$\mathrm{w}(\gamma_k, 0) = 0$$

by selecting a ray at the origin that does not meet the disk. It follows that

$$\sum_{k=1}^{N^2} \mathrm{w}(\gamma_k, 0) = \sum_{k=1}^{N^2} 0 = 0.$$

On the other hand, we also have

$$\sum_{k=1}^{N^2} \mathrm{w}(\gamma_k, 0) = \mathrm{w}(f_T, 0) - \mathrm{w}(g_T, 0).$$

Indeed, splitting each term on the LHS into its constitutent four terms, we find that the contributions of the inner edges and the left-hand and right-hand edges cancel. What is left are the edges at the bottom (directed rightward) and the edges at the top (directed leftward). This is precisely the RHS.

It follows that $\mathrm{w}(f_T, 0) - \mathrm{w}(g_T, 0) = 0$. This completes Step 3.

Step 4. To calculate $\mathrm{w}(e^{2\pi i d t}, 0)$, let $N > 2d$ and use the subdivision

$$T = \left(0, \tfrac{1}{N}, \tfrac{2}{N}, \ldots, \tfrac{N-1}{N}, 1\right)$$

to get

$$\mathrm{w}(e^{2\pi i d t}, 0) = \mathrm{w}\left(\left[e^{2\pi i d t}\right]_T, 0\right) = \tfrac{1}{2\pi} \sum_{k=1}^{N} \tfrac{2\pi d}{N} = d$$

as required. This completes Step 4.

The proof of the winding number theorem is complete. ∎

The next lemma summarizes an argument used in both Step 2 and Step 3 above.

Lemma 1.16 (Distant Cycle Lemma). *Let $a_1, \ldots, a_n \in \mathbb{R}^2$ satisfy $|a_1 - a_i| < |a_1|$ for all i, and let γ be any 1-cycle constructed using only edges of the form $[a_i, a_j]$. Then $\mathrm{w}(\gamma, 0)$ is defined and equal to 0.*

Proof. The inequalities imply that the a_i are contained in a circular disk disjoint from the origin. All possible edges of γ are contained in the same disk. Thus the winding number is defined. It is zero because there exist rays at the origin which do not meet the disk. ∎

In both steps, we showed that two winding numbers were equal by expressing their difference as a sum of winding numbers (of small triangles or quadrilaterals) that are zero by the Distant Cycle Lemma.

Corollary 1.17. *The winding number is invariant to orientation-preserving reparametrizations of the loop.*

Proof. Here is an 'inside-the-black-box' argument. Let f, g be parametrizations of the same loop in in the same direction, and let $x \in \mathbb{R}^2 - \mathrm{im}(f) = \mathbb{R}^2 - \mathrm{im}(g)$. Then $\mathrm{w}(f, x)$, $\mathrm{w}(g, x)$ can be computed as discrete winding numbers of sufficiently fine polygonal approximations of f, g. We can choose approximations that give the same polygon in $\mathbb{R}^2 - \{x\}$. □

Remark. A more general argument, which stays outside the black box, runs like this. Given a loop $f : [0, 1] \to D$, consider its continuous 'lift'

$$\bar{f} : \mathbb{R} \to D; \quad t + n \mapsto f(t) \quad \text{for all } t \in [0, 1] \text{ and } n \in \mathbb{N}.$$

If $\phi : [0, 1] \to \mathbb{R}$ is a continuous map satisfying $\phi(1) = 1 + \phi(0)$, then $\bar{f} \circ \phi \in \mathrm{Loops}(D)$. This general form of reparametrization allows some temporary doubling back, as well as allowing the basepoint of the loop to change. Then we have

$$\mathrm{w}(f, x) = \mathrm{w}(\bar{f} \circ \phi, x)$$

for any $x \in \mathbb{R}^2 - \mathrm{im}(f)$. Indeed, if we define $\Phi(s, t) = st + (1 - s)\phi(t)$ then $\bar{f} \circ \Phi$ is a homotopy between f and $\bar{f} \circ \phi$. The important point is that Φ satisfies the identity $\Phi(s, t) = 1 + \Phi(s, t + 1)$, since this guarantees that each intermediate stage in the homotopy is a loop.

Lecture 3

1.9. The Brouwer fixed-point theorem.

Theorem 1.18. *Let D be the closed unit disk in $\mathbb{R}^2$, and let $f : D \to D$ be continuous. Then f has a fixed point, i.e. there exists $x \in D$ such that $f(x) = x$.*

Here are some comments on the theorem.

- We can replace D with any space topologically equivalent to D. The theorem for D immediately implies this more general assertion.
- The disk must be closed. For the open disk, let b be any point on the boundary. Then the function $f(x) = \frac{1}{2}(x + b)$ maps the open disk to itself, but has no fixed point.
- The theorem fails for the annulus. Let f be the function 'rotate by θ' (where $\theta \neq 2n\pi$).
- Given a map of Claremont located in Claremont, there exists a point on the map which lies directly over the point in Claremont that it represents. This is a consequence of Brouwer, but it is also a consequence of the (easier) contraction mapping theorem.

- Given two copies of a rectangular piece of paper (such as a map of Claremont), fold or crumple one of them and place it completely over the other. Then there is a point in the folded map which lies directly over the corresponding point in the other map. This follows from Brouwer, but not from the contraction mapping principle.
- The theorem works more generally in $\mathbb{R}^n$. It is usual to prove the general theorem using more sophisticated technology such as homology theory. There is also an elementary combinatorial argument using 'Sperner's Lemma'.
- The theorem in $\mathbb{R}^1$ can be proved using the intermediate value theorem, Indeed, suppose $f : [-1, 1] \to [-1, 1]$ is continuous. Then $g(x) = x - f(x)$ is continuous, with $g(-1) \leq 0$ and $g(1) \geq 0$. Therefore there exists $x \in [-1, 1]$ such that $g(x) = 0$, and hence $f(x) = x$.

The 2-dimensional tool analogous to the IVT is, of course, the winding number.

Proof of the Brouwer fixed-point theorem. Suppose $f : D^2 \to D^2$ is continuous and has no fixed points. Then $g(x) = x - f(x)$ is continuous and maps $D^2 \to \mathbb{R}^2 - \{0\}$. One shows that

$$\begin{aligned} & \mathrm{w}(g(e^{2\pi it}), 0) = \mathrm{w}(g(0), 0) = 0 && \text{homework } \mathbf{11}\text{(i)} \\ \text{and} \quad & \mathrm{w}(g(e^{2\pi it}), 0) = \mathrm{w}(e^{2\pi it}, 0) = 1 && \text{homework } \mathbf{11}\text{(ii)} \end{aligned}$$

which is a contradiction. □

1.10. **Differentiable curves.** Let $f : [0, 1] \to \mathbb{R}^2 - \{0\}$ be differentiable (or piecewise differentiable). We can reinterpret

$$\mathrm{w}(f, 0) = \mathrm{w}(f_T, 0) = \sum_{k=1}^{N} \theta([g(t_{k-1}), g(t_k)], 0)$$

as an integral by taking the limit as $\mathrm{mesh}(T) \to 0$. The idea is to interpret θ as $\Delta \arg$, the change in argument as we move along an edge. Even though arg is not well-defined globally, its differental $d \arg$ makes sense everywhere and we can write

$$\mathrm{w}(g, 0) = \lim_{\mathrm{mesh}(T)\to 0} \mathrm{w}(\gamma_T, 0) = \lim_{\mathrm{mesh}(T)\to 0} \left[\frac{1}{2\pi} \sum_{k=1}^{N} \Delta \arg_{[g(t_{k-1}),g(t_t)]} \right] = \frac{1}{2\pi} \oint_g d \arg .$$

Now we must calculate $d \arg$. For short edges in the half-plane $\{x > 0\}$, we can use

$$\overline{\arg}(x, y) = \arctan(y/x)$$

which leads to the equation

$$d \arg = d\,\overline{\arg} = \frac{1}{1 + (y/x)^2} d(y/x) = \frac{1}{1 + (y/x)^2} \frac{x\,dy - y\,dx}{x^2} = \frac{x\,dy - y\,dx}{x^2 + y^2}.$$

In the other three coordinate half-plane, the alternative lifts

$$\overline{\arg}(x, y) = \arctan(y/x) + \pi \quad \text{or} \quad -\arctan(x/y) + \tfrac{\pi}{2} \quad \text{or} \quad -\arctan(x/y) - \tfrac{\pi}{2}$$

lead to the same expression. We reach the desired formula:

$$\boxed{\mathrm{w}(f, 0) = \frac{1}{2\pi} \oint_f \frac{x\,dy - y\,dx}{x^2 + y^2}}$$

This expression is independent of the parametrization of the loop by the variable t. We can evaluate the integral as

$$\mathrm{w}(f,0) = \frac{1}{2\pi}\int_0^1 \frac{x\,\frac{dy}{dt} - y\,\frac{dx}{dt}}{x^2+y^2}\,dt$$

for any specific parametrization $f(t) = (x(t), y(t))$.

We now relate this to complex analysis. Writing $z = x + iy$, we have

$$\frac{dz}{z} = \frac{dx + i\,dy}{x+iy} = \frac{(dx+i\,dy)(x-iy)}{(x+iy)(x-iy)} = \frac{(x\,dx + y\,dy) + i(x\,dy - y\,dx)}{x^2+y^2}.$$

For any closed loop f we get

$$\oint_f \frac{dz}{z} = \oint_f \frac{x\,dx + y\,dy}{x^2+y^2} + i\oint_f \frac{x\,dy - y\,dx}{x^2+y^2} = \oint_f d\left[\log(\sqrt{x^2+y^2})\right] + 2\pi i\,\mathrm{w}(f,0).$$

Since the function $\log\sqrt{x^2+y^2}$ is globally defined (away from 0), the integral of its differential is automatically zero around a closed loop. Thus we obtain a famous formula from complex analysis:

$$\boxed{\mathrm{w}(f,0) = \frac{1}{2\pi i}\oint_f \frac{dz}{z}}$$

The more general form

$$\boxed{\mathrm{w}(f,a) = \frac{1}{2\pi i}\oint_f \frac{dz}{z-a}}$$

follows immediately by translation.

2. Homology

We now recast our ideas in the language of linear algebra. Soon this will lead to the notion of homology.

2.0. **Introduction.** Our proof of FTA included, and depended on, a proof of the following fact:

Proposition 2.1. *Let $p : \mathbb{C} \to \mathbb{C}$ be a continuous function. If $\mathrm{w}(p(Re^{2\pi it}), 0)$ is (defined and) nonzero then p has a root inside the circle of radius R.* $\square$

If p is a polynomial, or more generally a complex analytic function, then there is a stronger theorem that asserts that $\mathrm{w}(p(Re^{2\pi it}))$ (when defined) counts the roots inside the circle of radius R. Repeated roots are counted with multiplicity. Moreover, the curve does not have to be a circle.

Theorem 2.2. *Let $p : \mathbb{C} \to \mathbb{C}$ be a non-constant complex analytic function. Let $\Lambda \subset \mathbb{C}$ denote its set of roots, and let m_λ denote the multiplicity of $\lambda \in \Lambda$. Then for any continuous loop $f \in \mathrm{Loops}(\mathbb{C} - \Lambda)$ we have*

$$\mathrm{w}(pf, 0) = \sum_{\lambda \in \Lambda} \mathrm{w}(f, \lambda) m_\lambda.$$

Implicit in this description is the fact that the set of roots of a complex analytic function is discrete, which implies that only finitely many of them are enclosed (with nonzero winding number) by f; so the sum on the right-hand side is finite. The multiplicity of a root λ is the exponent of the smallest nonzero term in the Taylor series for $q(z) := p(\lambda + z)$.

The proof works like this:

- Show that the result is true whenever f is a sufficiently small circle around a root. In other words, show that $\mathrm{w}(p(\lambda + re^{2\pi it}), 0) = m_\lambda$ for sufficiently small r. This is done exactly as in homework question **9**(iv).
- Show that a general f is 'equivalent' to a 'sum' of small loops around roots. The number of small loops needed around λ is given by $\mathrm{w}(f, \lambda)$. The equivalence 'respects' winding numbers, so we can calculate $\mathrm{w}(pf, 0)$ as a linear combination of the terms $\mathrm{w}(p(\lambda + re^{2\pi it}), 0) = m_\lambda$.

A good way to remove all the scare quotes ‘’ and make the proof rigorous is to invent homology theory. This is our next task.

2.1. **Chains, cycles, boundaries.** Let $U \subseteq \mathbb{R}^2$. Often U is an open set, but not always. We wish to study the topological properties of U in terms of linear algebra. To this end, we will define three vector spaces $\mathrm{C}_0(U)$, $\mathrm{C}_1(U)$, $\mathrm{C}_2(U)$ and two linear maps ∂_0 and ∂_1. These are organised as follows:

$$\mathrm{C}_0(U) \xleftarrow{\partial_0} \mathrm{C}_1(U) \xleftarrow{\partial_1} \mathrm{C}_2(U)$$

We frequently write ∂ as shorthand for ∂_0 or ∂_1.

Let $\mathbb{F}$ be any field.

Definition 2.3. Let $\mathrm{C}_0 = \mathrm{C}_0(U) = \mathrm{C}_0(U; \mathbb{F})$ to be the vector space over $\mathbb{F}$ with

- a generator $[a]$ for each $a \in U$;
- no linear relations between the generators.

A typical vector looks like

$$\alpha = \lambda_1[a_1] + \cdots + \lambda_n[a_n]$$

and is called a 0-**chain** in U.

Definition 2.4. Let $\mathrm{C}_1 = \mathrm{C}_1(U) = \mathrm{C}_1(U; \mathbb{F})$ to be the vector space over $\mathbb{F}$ with

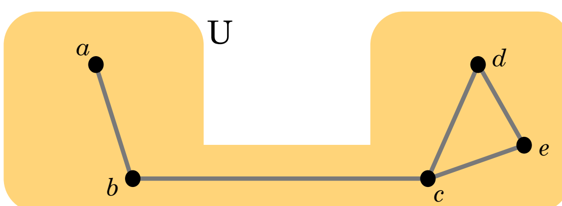
- a generator $[a, b]$ for every a, b distinct with $|[a, b]| \subseteq U$;
- the relation $[a, b] = -[b, a]$ for all such a, b.

A typical vector looks like

$$\alpha = \lambda_1[a_1, b_1] + \cdots + \lambda_n[a_n, b_n]$$

and is called a 1-**chain** in U.

Example 2.5. In the following domain



the 1-chains

$$\alpha = [a, b] + [b, c] + [c, d]$$
$$\beta = [c, e] + [e, d] + [d, c]$$

are defined over any field. The edge $[a, d]$ does not belong to $\mathrm{C}_1(U)$. We can add:

$$\begin{aligned} \alpha + \beta &= ([a, b] + [b, c] + [c, d]) + ([c, e] + [e, d] + [d, c]) \\ &= [a, b] + [b, c] + [c, e] + [e, d]. \end{aligned}$$

If we are working over $\mathbb{R}$, then

$$\gamma = 7[c, e] - \pi[b, a]$$

is a valid 1-chain.

The next step is to define the linear map ∂_0.

Definition 2.6. The **boundary map** is the map

$$\partial = \partial_0 : \mathrm{C}_1(U) \to \mathrm{C}_0(U)$$

defined on each generator by

$$\partial[a,b] = [b] - [a]$$

and extended by linearity to the whole vector space. This is consistent with the relations $[a,b] = -[b,a]$ because

$$\begin{aligned} \partial\left([a,b]\right) &&= [b]-[a], \\ \partial\left(-[b,a]\right) = -\partial\left([b,a]\right) &= -\left([a]-[b]\right) &= [b]-[a]. \end{aligned}$$

Example 2.7. (Picture of a path *abcde* in a region.)

$$\begin{aligned} \partial\left([a,b]+[b,c]+[c,d]+[d,e]\right) &= [b]-[a]+[c]-[b]+[d]-[c]+[e]-[d] \\ &= [b]-[a]+[c]-[b]+[d]-[c]+[e]-[d] \\ &= [e]-[a] \end{aligned}$$

Example 2.8. (Picture of a pentagram.)

$$\partial\left([a,c]+[c,e]+[e,b]+[b,d]+[d,a]\right) = 0.$$

In linear algebra we learned that there are important vector spaces associated to a linear map $T : V \to W$. The most important are the **kernel**

$$\ker(T) = \{v \in V \mid T(v) = 0\}$$

which is a subspace of V, and the **image**

$$\mathrm{im}(T) = \{w \in W \mid \exists v \in V,\, T(v) = w\}$$

which is a subspace of W. What do the image and kernel of δ tell us about the domain U?

2.2. **The kernel of ∂_0.** The following proposition depends on a property of a field known as its 'characteristic'. Any field that contains a full copy of the integers is said to have characteristic 0. Examples are $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$. Otherwise there is some smallest positive integer p which is equal to zero in the field. This is always a prime number, and we say that the field has characteristic p. For example, the field of two elements $\mathbb{F}_2$ has characteristic 2, since $2 = 1 + 1 = 0$.

Proposition 2.9. *Let $\gamma = \sum_{k=1}^{n} [a_k, b_k]$ be a sum of edges.*

- *if* $\mathrm{char}(\mathbb{F}) = 0$*, then* $\partial\gamma = 0$ *if and only if* γ *is an oriented 1-cycle.*
- *if* $\mathrm{char}(\mathbb{F}) = 2$*, then* $\partial\gamma = 0$ *if and only if* $\tilde{\gamma} = \sum_{k=1}^{n} \{a_k, b_k\}$ *is an unoriented 1-cycle.*

Note. A general 1-chain is a linear combination of edges $\gamma = \sum_k \lambda_k [a_k, b_k]$. This proposition considers the special case where all the coefficients are 1.

Proof. Let $p_1, p_2, \ldots, p_m$ be a list of the distinct points that occur among the a_k and b_k. Then we can write

$$\partial\gamma = \lambda_1[p_1] + \lambda_2[p_2] + \cdots + \lambda_m[p_m]$$

where

$$\begin{aligned}\lambda_j = {} & (\text{the number of times } p_j \text{ occurs among the } b_k) \\ & - (\text{the number of times } p_j \text{ occurs among the } a_k)\,.\end{aligned}$$

Then $\partial\gamma = 0$ if and only if each coefficient $\lambda_j = 0$.

- If $\operatorname{char}(\mathbb{F}) = 0$, this happens precisely when each p_j occurs equally often in the (a_k) as in (b_k); in other words, when γ is an oriented 1-cycle.
- If $\operatorname{char}(\mathbb{F}) = 2$, this happens precisely when each p_j occurs an even number of times in the $(a_k), (b_k)$ combined; in other words, when $\tilde{\gamma}$ is an unoriented 1-cycle.

This completes the proof. □

We now give a very general definition of a cycle.

Definition. Let $U \subseteq \mathbb{R}^2$ and let $\mathbb{F}$ be a field. We define the space of **1-cycles**

$$\mathrm{Z}_1(U;\mathbb{F}) = \ker\left(\partial_0 : \mathrm{C}_1(U;\mathbb{F}) \to \mathrm{C}_0(U;\mathbb{F})\right),$$

a subspace of $\mathrm{C}_1(U;\mathbb{F})$. An element of $\mathrm{Z}_1(U;\mathbb{F})$ is called a 1-cycle (in U, over $\mathbb{F}$).

Remark. When we are working over $\mathbb{R}$, the space of 1-cycles includes much more than just the 'oriented 1-cycles' that we defined earlier. This is because the edges have arbitrary real coefficients.

Remark. In contrast, when we are working over the field of two elements the space of 1-cycles $\mathrm{Z}_1(U;\mathbb{F}_2)$ is precisely equivalent to the set of unoriented 1-cycles.

2.3. **The image of ∂_0.** Can we find a 1-chain γ solving the equation

$$\alpha = \partial\gamma \tag{*}$$

when $\alpha = \sum_k \lambda_k[p_k]$ is a given 0-chain?

Proposition 2.10. *A necessary condition that a solution to (*) exists is that* $\sum_k \lambda_k = 0$.

Definition. Let us name the function that takes a 0-chain α and returns the sum of its coefficients: the **mass function** is the unique linear map

$$\mu : \mathrm{C}_0(U;\mathbb{F}) \to \mathbb{F}$$

that satisfies $\mu([p]) = 1$ for each $p \in U$. (This works because the vectors $[p]$ form a basis.)

Proof of Proposition 2.10. For every generator $[a,b] \in \mathrm{C}_1$ we have

$$\mu\partial([a,b]) = \mu([b]-[a]) = \mu([b]) - \mu([a]) = 1 - 1 = 0.$$

Therefore $\mu\partial$ is the zero map; and $\alpha = \partial\beta$ implies $\mu(\alpha) = 0$. □

We now consider the simplest case of (*) that satisfies this necessary condition.

Theorem 2.11. *Let $U \subseteq \mathbb{R}^2$ and let $a, b \in U$. The equation*

$$(\dagger) \qquad \partial\gamma = [b] - [a]$$

has a solution γ if and only if there exists a polygonal path in U from a to b.

Proof. ($\Leftarrow$) If $P(a_0, a_1, \ldots, a_{n-1}, a_n)$ is a polygonal path in U with $a_0 = a$ and $a_n = b$, then

$$\gamma = [a_0, a_1] + \cdots + [a_{n-1}, a_n]$$

belongs to $\mathrm{C}_1(U)$ and $\partial\gamma = [b] - [a]$.

($\Rightarrow$) Suppose $[b] - [a] = \partial\gamma$ for some γ. We must show that there is a path from a to b. To this end consider the function on U defined by

$$f(p) = \begin{cases} 1 & \text{if there exists a polygonal path in } U \text{ from } a \text{ to } p \\ 0 & \text{otherwise} \end{cases}$$

Note that $f(a) = 0$, by considering the trivial path $P(a)$ of length zero. From f we construct a linear map $\phi : \mathrm{C}_0(U; \mathbb{F}) \to \mathbb{F}$ by defining

$$\phi([p]) = f(p)$$

for every $p \in U$, and extending linearly to the whole vector space.

Claim. The composite map $\phi\partial : \mathrm{C}_1(U; \mathbb{F}) \to \mathbb{F}$ is zero.

Proof. Let $[p, q]$ denote any edge in $\mathrm{C}_1(U; \mathbb{F})$. Then $f(p) = f(q)$, since the edge $[p, q]$ can be used to extend a path from a to p to a path from a to q, and vice versa. Thus

$$\phi\partial([p, q]) = \phi([q] - [p]) = \phi([q]) - \phi([p]) = f(q) - f(p) = 0.$$

Since C_1 is generated (as a vector space) by its edges $[p, q]$, it follows that $\phi\partial = 0$. □

We therefore have

$$0 = \phi\partial\gamma = \phi([b] - [a]) = f(b) - f(a)$$

so $f(b) = f(a) = 1$ which means that there is a polygonal path from a to b. □

Lecture 4

Definition. Consider the equivalence relation on $U \subseteq \mathbb{R}^2$

$$a \sim b \;\Leftrightarrow\; \exists \text{ polygonal path from } a \text{ to } b \text{ in } U$$

An equivalence class of this relation is called a **polygonal-path component** or **PPC**. If U has exactly one PPC, then it is **PP-connected**.

Remark. Each PPC is connected in the traditional sense, so PP-connected implies connected. Conversely, if $U \subseteq \mathbb{R}^2$ is an open set then each PPC is also open (because each point is connected by a segment to each point in a small neighbourhood). Thus the PPCs comprise a partition of U into nonempty open connected sets. As such, they are precisely the connected components of U. Thus, connected implies PP-connected for open sets.

Example. If U is an arc of a circle, then it has uncountably many PPCs.

We now give a full description of $\mathrm{im}(\partial_0)$. Let $\alpha = \sum_k \lambda_k [p_k]$ be a 0-chain in an open set U. In order to solve $\alpha = \partial\gamma$, we have seen that it is necessary for the 'mass' to be zero:

$$\mu(\alpha) = \sum_k \lambda_k = 0$$

On the other hand, we have seen that $\partial\gamma$ cannot be used to 'transfer' mass between different connected components.

Following this idea, let $\{U_\kappa \mid \kappa \in K\}$ be the set of PPCs of U. For each U_κ we define a map $\mu_\kappa : \mathrm{C}_0(U; \mathbb{F}) \to \mathbb{F}$ by setting

$$\mu_\kappa([p]) = \begin{cases} 1 & \text{if } p \in U_\kappa \\ 0 & \text{otherwise} \end{cases}$$

and extending linearly to the whole vector space. This gives the 'mass on U_κ' of a 0-chain.

Theorem 2.12. *Let $U \subseteq \mathbb{R}^2$ be open. A 0-chain $\alpha \in \mathrm{C}_0(U)$ belongs to $\mathrm{im}(\partial_0)$ if and only if $\mu_\kappa(\alpha) = 0$ for every polygonal-path component U_κ of U.*

Proof. (Homework **17**.) □

Definition. Write

$$\mathrm{B}_0(U; \mathbb{F}) = \mathrm{im}\left(\partial_0 : \mathrm{C}_1(U; \mathbb{F}) \to \mathrm{C}_0(U; \mathbb{F})\right)$$

Then $\mathrm{B}_0(U; \mathbb{F}) \leq \mathrm{C}_0(U; \mathbb{F})$, and is known as the space of **0-boundaries** (in U, over $\mathbb{F}$).

2.4. **2-chains and their boundaries.** We now define the space of 2-chains and the boundary map from 2-chains to 1-chains.

An **oriented triangle** is a triple $[a, b, c]$ where $a, b, c \in \mathbb{R}^2$ are distinct. If we wish to think of this as a subset of the plane, we write:

$$\begin{aligned} |[a, b, c]| &= \text{the set of points in the filled-in triangle} \\ &= \{\lambda a + \mu b + \nu c \mid \lambda + \mu + \nu = 1,\ \lambda \geq 0,\ \mu \geq 0,\ \nu \geq 0\} \\ &= \{\text{convex linear combinations of } a, b, c\} \end{aligned}$$

In pictures, we usually draw the triangle with a curly arrow to indicate the cyclic order (a, b, c) of the vertices.

Remark. We allow a, b, c to be collinear, but they must be distinct.

Definition. Let $\mathrm{C}_2 = \mathrm{C}_2(U) = \mathrm{C}_2(U; \mathbb{F})$ to be the vector space over $\mathbb{F}$ with

- a generator $[a, b, c]$ for every a, b, c distinct with $|[a, b, c]| \subseteq U$;
- a relation $[a, b, c] = [b, c, a]$ for all such a, b, c;
- a relation $[a, b, c] = -[b, a, c]$ for all such a, b, c.

A typical vector looks like

$$\sigma = \lambda_1[a_1, b_1, c_1] + \cdots + \lambda_n[a_n, b_n, c_n]$$

and is called a **2-chain** in U.

Remark. If we apply the two relations repeatedly, we find that the six permutations of a triangle $[a, b, c]$ are related as follows

$$[a, b, c] = [b, c, a] = [c, a, b] = -[a, c, b] = -[b, a, c] = -[c, b, a]$$

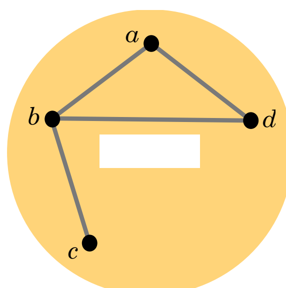
and therefore belong to the same 1-dimensional subspace.

It is useful to define the **support** of $\sigma = \sum_k \lambda_k[a_k, b_k, c_k]$ to be the set

$$|\sigma| = \bigcup_k |[a_k, b_k, c_k]|$$

of all points contained in the triangles of σ.

Example 2.13. In the following domain $[a, b, d] \in \mathrm{C}_2(U)$ whereas $[a, b, c] \notin \mathrm{C}_2(U)$:



We now define the boundary map $\partial_1 : \mathrm{C}_2 \to \mathrm{C}_1$. There are three steps.

- Define ∂ on the generators of C_2, by the formula:

$$\partial[a, b, c] = [b, c] - [a, c] + [a, b]$$

- Extend linearly to the whole vector space C_2:

$$\partial\left(\sum_k \lambda_k[a_k, b_k, c_k]\right) = \sum_k \lambda_k \partial[a_k, b_k, c_k]$$

- Verify that the definition is consistent with the prescribed relations:

$$\begin{aligned}\partial[a, b, c] = [b, c] - [a, c] + [a, b] &= \quad [c, a] - [b, a] + [b, c] = \quad \partial[b, c, a] \\ &= -[a, c] + [b, c] - [b, a] = -\partial[b, a, c]\end{aligned}$$

This completes the definition of $\partial = \partial_1$ from 2-chains to 1-chains. We now give names to the image and kernel of this boundary map:

$$\begin{aligned}\mathrm{B}_1(U) &= \mathrm{im}\,(\partial_1 : \mathrm{C}_2(U) \to \mathrm{C}_1(U)) \;\; = \text{“the space of 1-boundaries”} \leq \mathrm{C}_1(U) \\ \mathrm{Z}_2(U) &= \ker\,(\partial_1 : \mathrm{C}_2(U) \to \mathrm{C}_1(U)) = \text{“the space of 2-cycles”} \qquad \leq \mathrm{C}_2(U)\end{aligned}$$

The following relationship is of central importance in homology theory:

Proposition 2.14. $\mathrm{B}_1(U) \leq \mathrm{Z}_1(U)$.

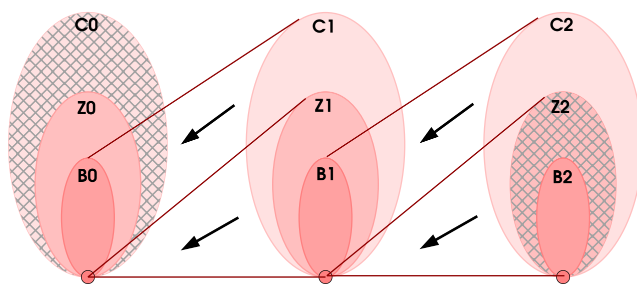
Proof. The assertion can be expressed in several ways, easily seen to be equivalent:

$$\begin{aligned}
\mathrm{B}_1(U) \leq \mathrm{Z}_1(U) &\Leftrightarrow \text{ every 1-boundary is a 1-cycle} \\
&\Leftrightarrow \text{ if } \gamma \in \mathrm{im}(\partial_1) \text{ then } \gamma \in \ker(\partial_0) \\
&\Leftrightarrow \text{ if } \gamma = \partial_1\sigma \text{ for some } \sigma \in \mathrm{C}_2(U), \text{ then } \partial_0\gamma = 0. \\
&\Leftrightarrow \text{ if } \sigma \in \mathrm{C}_2(U), \text{ then } \partial_0\partial_1\sigma = 0. \\
&\Leftrightarrow \partial_0\partial_1 = 0
\end{aligned}$$

To prove that $\partial_0\partial_1 = 0$ it is enough to show that this linear map sends every generator of C_2 to zero. This can be checked immediately:

$$\partial_0\partial_1[a,b,c] = \partial_0\left([b,c] - [a,c] + [a,b]\right) = [c] - [b] - [c] + [a] + [b] - [a] = 0 \qquad \square$$

We end up with the following configuration of vector spaces, with the arrows indicating the respective boundary maps:



We have not yet defined Z_0 or B_2; these are respectively the kernel of a map ∂_{-1} and the image of a map ∂_2. It turns out that $\mathrm{Z}_0 = \mathrm{C}_0$ and $\mathrm{B}_2 = \mathrm{Z}_2$, as suggested by the cross-hatching.

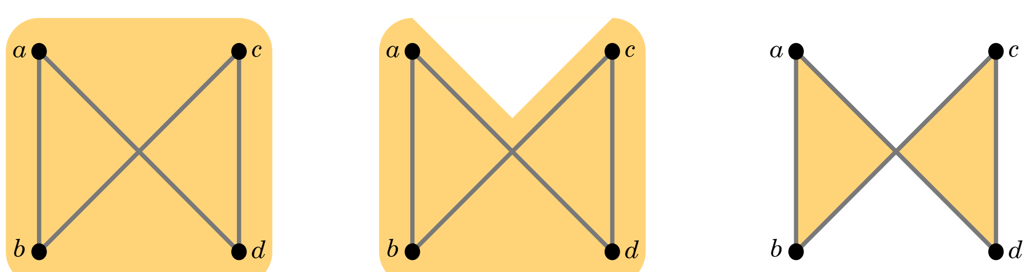
2.5. **The image of ∂_1.** We consider the equation:

$$\gamma = \partial_1\sigma$$

When this holds we say that γ **bounds** σ, or that σ **spans** γ.

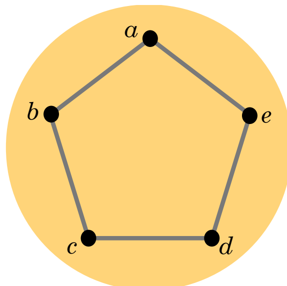
Question. For a given $\gamma \in \mathrm{Z}_1(U)$ does there exist $\sigma \in \mathrm{C}_2(U)$ which spans it?

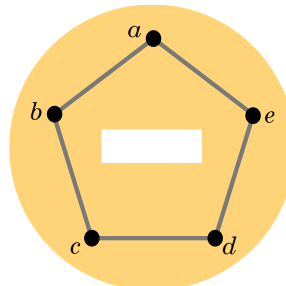
Example 2.15. In the domain on the left, the 1-cycle $\gamma = [a,b] + [b,c] + [c,d] + [d,a]$, is spanned by the 2-chain $\sigma = [a,b,c] + [a,c,d]$.



The domain in the middle doesn't admit σ as a 2-chain, but γ is still a boundary: it is spanned by $[a, b, d] + [b, c, d]$ for instance. It turns out that the smallest domain in which γ bounds a 2-chain is the domain on the right. See homework **17** and Theorem 2.18.

Example 2.16. Let $\gamma = [a, b] + [b, c] + [c, d] + [d, e] + [e, a]$ in the two domains shown here.





Then γ bounds a 2-chain in the domain on the left, for instance $\sigma = [a, b, c]+[a, c, d]+[a, d, e]$. For the domain on the right, there is no obvious 2-chain that spans γ. Indeed, it seems that no such 2-chain should exist. But how can we prove it?

The trick: In order to show that $\gamma = \partial_1\sigma$ has no solution, we find a linear map T such that $T(\gamma) \neq 0$ but where $T\partial_1$ is the zero map. This prevents the existence of σ.

The linear map in this case is the winding number about some point inside the white rectangle. Then γ has winding number 1 about that point, but the boundaries of the individual triangles of σ have winding number 0. So when we add them up, we get the contradiction $1 = 0$.

Let's do this a little more carefully. We have defined the winding number only for 1-cycles of the form $\sum_k [a_k, b_k]$ rather than the more general form $\sum_k \lambda_k [a_k, b_k]$. So let's pause to broaden the definition of the winding number.

Definition. Let $U \subsetneq \mathbb{R}^2$, and let $x \in \mathbb{R}^2 - U$. Define a linear map $\mathrm{w}_x : \mathrm{C}_1(U; \mathbb{R}) \to \mathbb{R}$ by setting

$$\mathrm{w}_x([a, b]) = \tfrac{1}{2\pi}\,\theta([a, b], x)$$

for all generators $[a, b]$, and extending linearly to the whole space $\mathrm{C}_1(U; \mathbb{R})$.

It is immediate that if a sum of edges $\gamma = \sum_k [a_k, b_k]$ is a 1-cycle, then $\mathrm{w}_x(\gamma) = \mathrm{w}(\gamma, x)$. So we recover the winding number for those 1-cycles and, better still, w_x is defined on all 1-chains.

Remark. We can no longer say that w_x of a 1-cycle is necessarily an integer: scalar multiplication means that we can get any real number (unless w_x is identically zero). It is only guaranteed to be an integer when γ is a linear combination of edges with integer coefficients.

Proposition 2.17. *Let $U \subsetneq \mathbb{R}^2$ and let $x \in \mathbb{R}^2 - U$. Then $\mathrm{w}_x\,\partial_1$ is zero on $\mathrm{C}_2(U; \mathbb{R})$.*

Proof. Consider a generator $[a, b, c] \in \mathrm{C}_2(U; \mathbb{R})$. Thus $x \notin |[a, b, c]|$. We have

$$w_x\partial_1([a, b, c]) = \mathrm{w}_x([b, c] - [a, c] + [a, b]) = \mathrm{w}([b, c] - [a, c] + [a, b]) = 0$$

by considering a ray from x pointing away from the triangle. A linear map which is zero for every generator must be the zero map, so $\mathrm{w}_x \partial_1 = 0$. □

Recall that $|\sigma|$ denotes the support of σ, that is, the set of all points contained in the triangles of σ.

Theorem 2.18 (coverage theorem)**.** *Suppose $\gamma = \partial_1 \sigma$, where $\gamma \in \mathrm{Z}_1(\mathbb{R}^2)$ and $\sigma \in \mathrm{C}_2(\mathbb{R}^2)$. Then $|\sigma|$ contains all points of $|\gamma|$ as well as all points $x \in \mathbb{R}^2 - |\gamma|$ for which $\mathrm{w}_x(\gamma) \neq 0$.*

For instance, in Example 2.15, any 2-chain spanning γ must cover the two small triangles.

Proof. Let $U = |\sigma|$. Then $\gamma = \partial_1 \sigma$ is a 1-cycle in U, so $|\gamma| \subseteq |\sigma|$. By Proposition 2.17, $x \notin |\sigma|$ implies that $\mathrm{w}_x(\gamma) = \mathrm{w}_x \partial_1(\sigma) = 0$. Thus $\mathrm{w}_x(\gamma) \neq 0$ implies $x \in |\sigma|$. □

2.6. **Sensor network coverage.** Consider a 2-dimensional domain containing a large number of robotic sensors. The sensors are to broadcast or receive information in a small radius around them. Their capacities are limited. Fix $R > 0$. We assume:

- each sensor has a unique identifier;
- each sensor can identify all sensors which lie within distance R of it;
- this information can be relayed back to a coordinating computer.

We do not assume that the sensors can identify their locations precisely. There is no GPS. We may sometimes be able to deliberately position some sensors in known locations.

Radial coverage. A point $x \in U$ is **r-covered** if it is contained in a disk of radius $R/\sqrt{3}$ centered on one of the sensors.

Vietoris–Rips coverage. A point $x \in U$ is **VR-covered** if it is contained in a triangle spanned by three sensors that lie pairwise within distance R of each other.

It is a theorem that VR-coverage implies r-coverage. This follows from the trigonometric result that if the three sides of a triangle have length at most R then the three disks of radius $R/\sqrt{3}$ centered at the vertices cover the entire triangle. (The ratio $1/\sqrt{3}$ cannot be made smaller: consider an equilateral triangle.)

Controlled boundary coverage test. Let U be a domain bounded by a simple closed polygon. Place sensors $s_1, s_2, \ldots, s_n = s_0$ at all the corners of the polygon and along the edges so that each edge $[s_{k-1}, s_k]$ has length at most R. Other sensors are distributed within the domain. Consider the finite dimensional vector spaces

$$\hat{C}_1 = \text{subspace of } C_1(\mathbb{R}^2) \text{ generated by edges between sensors separated by at most } R,$$

$$\hat{C}_2 = \text{subspace of } C_2(\mathbb{R}^2) \text{ generated by triangles between sensors separated by at most } R.$$

and the linear map $\hat{\partial}_1 : \hat{C}_2 \to \hat{C}_1$ defined as the restriction of ∂_1. If the 'fence' cycle

$$\gamma = \sum_{k=1}^{n} [s_{k-1}, s_k]$$

lies in the image of $\hat{\partial}_1$ then return YES otherwise return NO.

Theorem 2.18 implies that YES guarantees that all of U is VR-covered (and therefore also r-covered). It is also true that NO implies that some point of U is not VR-covered; but the only proof I know is rather complicated.

2.7. **Homology and Betti numbers.** Let $U \subseteq \mathbb{R}^2$ be open. We will define 2 vector spaces

$$\mathrm{H}_0(U),\ \mathrm{H}_1(U)$$

called the **homology** of U, and two integers

$$\mathrm{b}_0(U) = \dim(\mathrm{H}_0(U))$$
$$\mathrm{b}_1(U) = \dim(\mathrm{H}_1(U))$$

called the **Betti numbers** of U. These will tell us about the topology of U. In fact, we eventually see that

$$\mathrm{b}_0(U) = \#\text{ connected components of } U$$

and

$$\begin{aligned} \mathrm{b}_1(U) &= \#\text{ 'holes' in } U \\ &= \#\text{ bounded components of } \mathbb{R}^2 - U \end{aligned}$$

(at least, in certain special cases).

Definition. Consider the sequence of vector spaces and linear maps

$$0 \xleftarrow{\partial_{-1}} \mathrm{C}_0(U) \xleftarrow{\partial_0} \mathrm{C}_1(U) \xleftarrow{\partial_1} \mathrm{C}_2(U)$$

(with an extra boundary map, zero, appended at the left). Then

$$\begin{aligned} \mathrm{H}_0(U) &= \frac{\ker(\partial_{-1})}{\mathrm{im}(\partial_0)} = \frac{\mathrm{C}_0(U)}{\mathrm{B}_0(U)} \\ \mathrm{H}_1(U) &= \frac{\ker(\partial_0)}{\mathrm{im}(\partial_1)} = \frac{\mathrm{Z}_1(U)}{\mathrm{B}_1(U)} \end{aligned}$$

The definition is a special case of a very general construction. A **chain complex** is a sequence of vector spaces and linear maps arranged as follows:

$$0 \xleftarrow{\partial_{-1}} \mathrm{C}_0(U) \xleftarrow{\partial_0} \mathrm{C}_1(U) \xleftarrow{\partial_1} \mathrm{C}_2(U) \xleftarrow{\partial_2} \mathrm{C}_2(U) \xleftarrow{\partial_3} \dots$$

such that $\partial_{k-1}\partial_k = 0$ for every k. This last condition implies that $B_k = \mathrm{im}(\partial_k)$ is a subspace of $Z_k = \ker(\partial_{k-1})$. We define the k-th homology of the chain complex to be $H_k = Z_k/B_k$.

Quotient vector spaces. Let V be a vector space. Any subspace $U \leq V$ gives rise to an equivalence relation

$$v_1 \sim v_2 \iff v_2 - v_1 \in U$$

on V. The set of equivalence classes is itself a vector space, on the understanding that

$$0 = [0], \qquad \lambda[v_1] = [\lambda v_1], \qquad [v_1] + [v_2] = [v_1 + v_2].$$

This is called the quotient vector space (of V by U) and is written V/U. The linear map $V \to V/U$ defined by $v \mapsto [v]$ is called the **canonical projection**.

Example. Let C^∞ denote the vector space of smooth real-valued functions of a real variable. The differentiation operator

$$Df = \frac{df}{dx}$$

is a linear map $D : C^\infty \to C^\infty$, but indefinite integration

$$If = \int f(x)\,dx$$

cannot be described that way, because of the ambiguity of the '$+C$'. Consider the subspace $\mathbb{R} \leq C^\infty$ of constant functions. Since any two antiderivatives of f differ by a constant, we can instead think of indefinite integration as a linear map $I : C^\infty \to C^\infty/\mathbb{R}$.

In homology theory, it is useful to express statements about V/U in terms of V and U directly. For example, let $v_1, \ldots, v_n \in V$ and let $x \in V$.

- The statement

$$[x] \in \operatorname{span}([v_1], \ldots, [v_n])$$

is equivalent to

$$x = \lambda_1 v_1 + \cdots + \lambda_n v_n + u \quad \text{for some } \lambda_1, \ldots, \lambda_n \in \mathbb{F} \text{ and } u \in U.$$

- The statement

$$[v_1], \ldots, [v_n] \text{ are linearly independent}$$

is equivalent to

$$\text{if } \lambda_1 v_1 + \cdots + \lambda_n v_n \in U \text{ then every } \lambda_k = 0.$$

Lecture 5

Definition. We say that two k-cycles are **homologous**
$\iff$ they represent the same element of $\mathrm{H}_k = \mathrm{Z}_k/\mathrm{B}_k$
$\iff$ their difference is a boundary.

Repeated vertices. Henceforth, we will allow repeated vertices in the edges and triangles of our chain spaces. Mostly this makes no difference to the theory. Indeed

$$[a, a] = -[a, a] \quad \Rightarrow \quad 2[a, a] = 0 \quad \Rightarrow \quad [a, a] = 0$$

and similarly

$$[a, a, b] = -[a, a, b] \quad \Rightarrow \quad 2[a, a, b] = 0 \quad \Rightarrow \quad [a, a, b] = 0$$

as long as the characteristic of the field is not equal to 2. If we are working over $\mathbb{F}_2$ or some other field that contains $\mathbb{F}_2$, then $[a, a] \neq 0$. We can agree to live with that. It turns out not to affect the homology of our spaces. Meanwhile, we no longer have to carefully avoid these degenerate edges and triangles in our arguments.

We are now in a position to calculate the homology of the plane.

Theorem 2.19. *The homology of the plane is given by*

$$\begin{aligned} \mathrm{H}_0(\mathbb{R}^2) &\cong \mathbb{F}, \qquad \textit{in other words} \qquad \mathrm{b}_0(\mathbb{R}^2) = 1; \\ \mathrm{H}_1(\mathbb{R}^2) &= 0, \qquad \textit{in other words} \qquad \mathrm{b}_1(\mathbb{R}^2) = 0. \end{aligned}$$

Remark. It is clear that $\mathrm{b}_0 \geq 1$, because not every 0-chain is equal to a boundary. For instance, $[0]$ is not a boundary because $\mu([0]) = 1$ and $\mu\partial_0 = 0$, so we cannot write $[0] = \partial_0\gamma$.

Proof. The two statements can be interpreted as follows.

- $\mathrm{b}_0 \leq 1 \;\Leftrightarrow$ Every 0-chain α is homologous to some $\lambda[0]$, where $\lambda \in \mathbb{F}$ is a scalar.
- $\mathrm{b}_1 = 0 \;\Leftrightarrow$ Every 1-cycle γ is bf null-homologous, i.e. $\gamma = \partial\sigma$, where $\sigma \in \mathrm{C}_2$.

For the first statement, given $\alpha = \sum \lambda_k[a_k]$ we set $\gamma = \sum \lambda_k[0, a_k]$. Then

$$\partial\gamma = \alpha - \mu(\alpha)[0]$$

so α is homologous to $\lambda[0]$ with $\lambda = \mu(\alpha)$. For the second statement, given a 1-cycle $\gamma = \sum \lambda_k[a_k, b_k]$ we set $\sigma = [0, a_k, b_k]$. Then

$$\partial\sigma = \gamma$$

because $\partial\gamma = 0$ implies that the terms arising from $[0, a_k]$, $[0, b_k]$ in $\partial\sigma$ all cancel out. Specifically, the coefficient of a particular $[0, p]$ in $\partial\sigma$ is given by computing

$$(\text{sum of the } \lambda_k \text{ for which } a_k = p) - (\text{sum of the } \lambda_k \text{ for which } b_k = p)$$

and this is zero for every p because $\partial\gamma = 0$. $\square$

What properties of the plane did we use? We made use of $[0]$. For every $[a]$ occurring in a 0-chain, we made use of $[0, a]$. And for every $[a, b]$ occurring in a 1-cycle, we made use of $[0, a, b]$. This gives an immediate generalisation.

Definition. We say that $U \subseteq \mathbb{R}^2$ is **star-shaped** if there exists a point $x \in U$ such that if $a \in U$ then $|[x, a]| \subseteq U$. To specify x, we say that U is a **star on** x.

Theorem 2.20. *Let $U \subseteq \mathbb{R}^2$ be star-shaped. Then* $\mathrm{b}_0(U) = 1$ *and* $\mathrm{b}_1(U) = 0$. $\square$

2.8. **Chain homotopy.** Let $U \subseteq \mathbb{R}^2$ be star-shaped about a point x. The proof that $\mathrm{b}_0(U) = 1$ and $\mathrm{b}_1(U) = 0$ is based on the idea that all chains can be moved to x. We now give a more streamlined description of this proof.

Consider the following diagram of vector spaces and linear maps, where the vertical arrows represent identity maps:

$$\begin{tikzcd}
\mathbb{F} \arrow[d] \arrow[dr, "K"] & \mathrm{C}_0(U) \arrow[l, "\mu"'] \arrow[d] \arrow[dr, "K"] & \mathrm{C}_1(U) \arrow[l, "\partial"'] \arrow[d] \arrow[dr, "K"] & \mathrm{C}_2(U) \arrow[l, "\partial"'] \arrow[d] \\
\mathbb{F} & \mathrm{C}_0(U) \arrow[l, "\mu"'] & \mathrm{C}_1(U) \arrow[l, "\partial"'] & \mathrm{C}_2(U) \arrow[l, "\partial"']
\end{tikzcd}$$

The maps K are defined in the usual way, by specifying their values on generators:

$$\begin{aligned}
&K : \mathbb{F} \to \mathrm{C}_0 && \text{is defined by} \quad K(1) &&= [x] \\
&K : \mathrm{C}_0 \to \mathrm{C}_1 && \text{is defined by} \quad K([a]) &&= [x, a] \\
&K : \mathrm{C}_1 \to \mathrm{C}_2 && \text{is defined by} \quad K([a, b]) &&= [x, a, b]
\end{aligned}$$

The reason for doing this is the next result.

Proposition 2.21. *We have the identities:*

$$\begin{aligned}
K\mu + \partial K &= 1 \qquad \text{on } \mathrm{C}_0(U) \\
K\partial + \partial K &= 1 \qquad \text{on } \mathrm{C}_1(U)
\end{aligned}$$

The first identity implies that $\mathrm{b}_0(U) \leq 1$*, and the second identity implies that* $\mathrm{b}_1(U) = 0$*.*

Proof. It is enough to verify the first identity on generators $[a] \in \mathrm{C}_0(U)$. We have

$$\begin{aligned}
K\mu[a] &= K(1) \quad = [x] \\
\partial K[a] &= \partial[x, a] = [a] - [x]
\end{aligned}$$

so $(K\mu + \partial K)[a] = [a]$ as required. It follows that any 0-chain α is homologous to a multiple of $[x]$, because

$$\alpha = K\mu\alpha + \partial K\alpha = (\mu\alpha)[x] + \text{boundary}.$$

Therefore H_0 is at most 1-dimensional; that is to say $\mathrm{b}_0 \leq 1$.

(To complete the proof that $\mathrm{b}_0 = 1$, recall that $[x]$ is not a boundary since $\mu\partial = 0$ and $\mu[x] \neq 0$.)

It is enough to verify the second identity on generators $[a, b] \in \mathrm{C}_1(U)$. We have

$$\begin{aligned}
K\partial[a, b] &= K([b] - [a]) = [x, b] - [x, a] \\
\partial K[a, b] &= \partial[x, a, b] \qquad = [a, b] - [x, b] + [x, a]
\end{aligned}$$

so $(K\mu + \partial K)[a, b] = [a, b]$ as required. It follows that any 1-cycle γ is a 1-boundary: Indeed

$$\gamma = K\partial\gamma + \partial K\gamma = 0 + \partial K\gamma = \text{boundary}$$

using $\partial\gamma = 0$. Therefore $\mathrm{Z}_1 = \mathrm{B}_1$ and so $\mathrm{b}_1 = 0$. ∎

Remark. The collection of linear maps K plays the role in algebra that a homotopy plays in topology. Such a collection is known as a **chain homotopy**.

2.9. **Several PP-components.** We generalise $\mathrm{b}_0(\mathbb{R}^2) = 1$ as follows.

Theorem 2.22. *Let $U \subseteq \mathbb{R}^2$ be a subset of the plane with N polygonal-path components $U_1, \ldots, U_N$. Then $\mathrm{b}_0(U) = N$.*

Proof. We will construct a linear map from $\mathrm{H}_0(U) \to \mathbb{F}^N$ and show that it is both surjective and injective, and therefore an isomorphism.

Step 1: constructing the map. For each connected component U_k, let $\mu_k : \mathrm{C}_0 \to \mathbb{F}$ denote the mass function for that component. Since $\mu_k \partial_0 = 0$, it follows that μ_k is zero on the space of boundaries B_0. Therefore we can interpret μ_k as a linear map on the quotient space $\mathrm{H}_0 = \mathrm{C}_0/\mathrm{B}_0$.

$$\begin{tikzcd} \mathrm{C}_0(U) \arrow[r, "\mu_k"] \arrow[d] & \mathbb{F} \\ \mathrm{H}_0(U) \arrow[ur, dotted, "\mu_k"'] & \end{tikzcd}$$

Specifically, if $\alpha, \alpha' \in \mathrm{C}_0$ are homologous then

$$\mu_k(\alpha) = \mu_k(\alpha' + \partial\gamma) = \mu_k(\alpha') + \mu_k\partial\gamma = \mu_k(\alpha')$$

so the value of μ_k depends only on the equivalence class $[\alpha] = [\alpha']$ rather than on the specific 0-chain α, α' within the equivalence class.

Let $\bar{\mu}$ denote the vector:

$$\bar{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}$$

Then $\bar{\mu}$ is a linear map $\mathrm{H}_0(U) \to \mathbb{F}^N$.

Step 2: surjectivity. Let $a_k \in U_k$. Then $\bar{\mu}([a_k])$ is equal to the standard basis vector $\mathbf{e}_k$ in $\mathbb{F}^N$. Since $\mathrm{im}(\bar{\mu})$ contains every standard basis vector, it follows that $\bar{\mu}$ is onto.

Step 3: injectivity. This is the assertion that $\bar{\mu}([\alpha]) = 0$ implies that $[\alpha] = 0$. In other words, if the mass of a 0-chain α is zero on every component, then α is a boundary. This is homework question **17**.

It follows that $\bar{\mu}$ is an isomorphism between $\mathrm{H}_0(U)$ and $\mathbb{F}^N$. □

2.10. **The punctured plane.** We finish this chapter by calculating the homology of a punctured plane.

Theorem 2.23. *Let $p_1, \ldots, p_M \in \mathbb{R}^2$ be distinct and let $U = \mathbb{R}^2 - \{p_1, \ldots, p_M\}$. Then*

$$\mathrm{H}_1(U) \cong \mathbb{F}^M.$$

(We already know that $\mathrm{H}_0(U) \cong \mathbb{F}$, because U is PP-connected.)

Remark. In some sense the result is overdetermined: there are many ways to prove it. It's important not to ascribe too much importance to the details of any specific proof. Here we will use a local-to-global argument: the result is verified locally (Lemma 2.26) and extended to a global theorem.

Some notation will be helpful. Let $D_{p,r}$ denote the closed disk with center p and radius r, and let $D^*_{p,r} = D_{p,r} - \{p\}$ be the punctured closed disk. We write $D = D_{0,1}$ and $D^* = D^*_{0,1}$.

Within the unit disk we have the triangle 1-cycle $\Delta = \Delta_{0,1} = [1, \omega] + [\omega, \omega^2] + [\omega^2, 1]$ where $\omega = e^{2\pi i/3}$. This is called $E(3,1)$ in various homework questions. More generally we set

$$\Delta_{p,r} = [p + r, p + r\omega] + [p + r\omega, p + r\omega^2] + [p + r\omega^2, p + r].$$

This 1-cycle will serve as our 'standard reference' generator for $\mathrm{H}_1(D^*_{p,r})$.

Lemma 2.24. *There is a linear map* $\mathrm{w}_{\mathbb{F}}(-, 0) : \mathrm{H}_1(\mathbb{R}^2 - \{0\}) \to \mathbb{F}$ *with* $\mathrm{w}_{\mathbb{F}}(\Delta, 0) = 1$.

Proof. See homework **22**. We already know this for $\mathbb{F} = \mathbb{R}$ or $\mathbb{F}_2$. □

Corollary 2.25. *By translation we define* $\mathrm{w}_{\mathbb{F}}(-, p)$ *about any point* $p \in \mathbb{R}^2$*; then* $\mathrm{w}_{\mathbb{F}}(\Delta_{p,r}, p) = \mathrm{w}_{\mathbb{F}}(\Delta_{0,r}, 0) = \mathrm{w}_{\mathbb{F}}(\Delta, 0) = 1$*, since* $\Delta_{0,r}$ *is homologous to* Δ *in* $\mathbb{R}^2 - \{0\}$.

Lemma 2.26. *Every 1-cycle in the punctured unit disk* D^* *(over* $\mathbb{F}$*) is homologous to a scalar multiple of the triangle 1-cycle* Δ.

(This is scalar multiplication in the vector space $\mathrm{Z}_1(D^*)$, not geometric rescaling.)

Proof. See homework **23**. Every cycle can be pushed to the boundary of the disk so that its edges are chords, and the chords can be assumed to connect adjacent vertices around the circle. Each such chord must occur with the same coefficient, so the cycle is a multiple of a polygon winding once around the boundary. Any two such polygons are homologous, and in particular every such polygon is homologous to $E(3,1)$. □

Corollary 2.27. *Any* 1*-cycle in the punctured closed disk* $D^*_{p,r}$ *is homologous to* $\Delta_{p,r}$. □

Proof of Theorem 2.23. We mimic the proof of Theorem 2.22. Let $(r_k \mid 1 \le k \le M)$ be positive real numbers chosen so that the closed disks D_{p_k,r_k} are disjoint.

Step 1: constructing the map. Let $\mathrm{w}_k = \mathrm{w}_{\mathbb{F}}(-, p)$ denote the winding number about p_k, interpreted as a linear map $\mathrm{H}_1(U) \to \mathbb{F}$. If we write

$$\bar{\mathrm{w}} = \begin{bmatrix} \mathrm{w}_1 \\ \vdots \\ \mathrm{w}_M \end{bmatrix}$$

then $\bar{\mathrm{w}}$ is a map $\mathrm{H}_1(U) \to \mathbb{F}^M$.

Step 2: surjectivity. For $0 < \epsilon \le \min(r_k)$, we have $\bar{\mathrm{w}}(\Delta_{p,\epsilon}) = \mathbf{e}_k$, the standard basis vector in $\mathbb{F}^M$. Since $\mathrm{im}(\bar{\mathrm{w}})$ contains every standard basis vector it follows that $\bar{\mathrm{w}}$ is onto.

Step 3: injectivity. We must show that if γ is a 1-cycle in U whose winding number is zero about each p_k, then γ is a boundary in U. In fact, we will show that for any $\gamma \in \mathrm{Z}_1(U)$ we have

$$\gamma = \sum_k \lambda_k \Delta_{p_k,r_k} + (\text{boundary in } U)$$

in U. By applying $\bar{\mathrm{w}}$ to both sides it follows that $\lambda_k = \mathrm{w}_k(\gamma)$ so the desired result follows as a special case.

Let $\gamma \in \mathrm{Z}_1(U)$. Since $\mathrm{H}_1(\mathbb{R}^2) = 0$ we have that $\gamma = \partial\sigma$ for some $\sigma \in \mathrm{C}_2(\mathbb{R}^2)$. What we must do is modify σ to avoid the points p_k. First we make the edges and triangles small. Define linear maps

$$\begin{aligned} S &: \mathrm{C}_1(U) \to \mathrm{C}_1(U) \quad \text{by} \quad [a,b] \mapsto [a, \tfrac{1}{2}(a+b)] + [\tfrac{1}{2}(a+b), b], \\ K &: \mathrm{C}_1(U) \to \mathrm{C}_2(U) \quad \text{by} \quad [a,b] \mapsto [a, \tfrac{1}{2}(a+b), b]. \end{aligned}$$

Then $\partial K = S - 1$ on $\mathrm{C}_1(U)$, so we have

$$S\gamma = \gamma + (\text{boundary in } U).$$

Furthermore, define

$$S : \mathrm{C}_2(\mathbb{R}^2) \to \mathrm{C}_2(\mathbb{R}^2) \quad \text{by} \quad [a,b,c] \mapsto [a,n,m] + [b,l,n] + [c,m,l] + [l,m,n]$$

where $l = \frac{1}{2}(b+c)$, $m = \frac{1}{2}(c+a)$, $n = \frac{1}{2}(a+b)$. Then $\partial S = S\partial$ on $\mathrm{C}_2(\mathbb{R}^2)$, so

$$S\gamma = S\partial\sigma = \partial S\sigma.$$

Iterating n times, we get a 1-cycle

$$S^n\gamma = \gamma + (\text{boundary in } U)$$

with $S^n\gamma = \partial S^n\sigma$ in $\mathbb{R}^2$.

Remark. We can think of the maps S as a 'splitting' or 'subdivision' operation on chains. This particular operation doesn't generalize well to higher-dimensional chains.

By construction, each of the four triangles in $S[a,b,c]$ has half the diameter of $[a,b,c]$. Let R be the maximum diameter of the triangles in γ. By choosing $n \ge \log_2(R/\min(r_k))$, it follows that each triangle in $S^n\sigma$ has diameter at most $\min(r_k)$.

Let τ be the 2-chain obtained from $S^n\sigma$ by removing terms coming from triangles that meet any of the p_k. Specifically we can write

$$S^n\sigma = \rho_1 + \rho_2 + \cdots + \rho_M + \tau$$

where none of the triangles of τ meet any p_k, and all of the triangles of ρ_k meet p_k. Since these latter triangles have diameter at most $\min(r_k)$ it follows that $\rho_k \in \mathrm{C}_2(\Delta_{p_k,r_k})$ for all k.

Lemma 2.28. *Each $\partial\rho_k$ is a 1-cycle in its respective punctured disk $D^*_{p_k,r_k}$.*

Proof. The 1-cycle in question

$$\partial\rho_k = \partial S^n\sigma - \partial\tau - \sum_{j\neq k}\partial\rho_j = S^n\gamma - \partial\tau - \sum_{j\neq k}\partial\rho_j$$

is contained in $\mathbb{R}^2 - \{p_k\}$ (by considering the right-hand side) and also in D_{p_k,r_k} (by considering the left-hand side). Thus every edge is contained in $D^*_{p_k,r_k}$. □

Note. Despite appearances, it does not follow that $\partial\rho_k \in \mathrm{B}_1(D^*_{p_k,r_k})$. For that, it would need to be the boundary of a 2-chain in the punctured disk; whereas ρ_k meets the puncture and is not such a 2-chain (unless it happens to be zero).

Finally, by Lemma 2.26, each $\partial\rho_k$ is homologous to some scalar multiple of Δ_{p_k,r_k} in its punctured disk and therefore in U. We have

$$\begin{aligned}
\gamma &= S^n\gamma && + (\text{boundary in } U)\\
&= \partial S^n\sigma && + (\text{boundary in } U)\\
&= \partial\rho_1 + \cdots + \partial\rho_M + \partial\tau && + (\text{boundary in } U)\\
&= \partial\rho_1 + \cdots + \partial\rho_M && + (\text{boundary in } U)\\
&= \lambda_1\Delta_{p_1,r_1} + \cdots + \lambda_M\Delta_{p_M,r_M} && + (\text{boundary in } U)
\end{aligned}$$

as required. □

Corollary. Using Theorem 2.23 and the explicit description of the isomorphism map $\bar{\mathrm{w}}$, it is now not too difficult (homework **24**) to establish the formula

$$\mathrm{w}(pf, 0) = \sum_{\lambda\in\Lambda} \mathrm{w}(f,\lambda)m_\lambda$$

for a complex polynomial p and a continuous loop f that avoids its roots (Theorem 2.2).

Lecture 6

3. Categories

3.1. **Definition.** A category $\mathbf{C}$ is specified by the following data:

- a class $\mathbf{C}_0 = \mathrm{Obj}(\mathbf{C})$, the *objects* of the category;
- a class $\mathbf{C}_1 = \mathrm{Arr}(\mathbf{C}) = \mathrm{Mor}(\mathbf{C})$, the *arrows* or *morphisms* of the category.

Each arrow has a source (domain) and a target (codomain) in $\mathbf{C}_0$. If $x = \mathrm{source}(f)$ and $y = \mathrm{target}(f)$ then we may draw a picture such as this:

$$y \xleftarrow{f} x$$

We write $\mathbf{C}(x, y)$ or $\mathrm{Mor}_{\mathbf{C}}(x, y)$ or $\mathrm{Mor}(x, y)$ to denote the set of arrows $x \to y$.

Note. While $\mathbf{C}_0$ is permitted to be a proper class, each $\mathbf{C}_0(x, y)$ is required to be a set.

- There is a *composition operation* $\circ : \mathbf{C}(x, y) \times \mathbf{C}(y, z) \to \mathbf{C}(x, z)$.

$$\begin{tikzcd} z & y \arrow[l, "g"'] & x \arrow[l, "f"'] \arrow[ll, bend left, "gf"] \end{tikzcd}$$

- Composition is *associative*: $(hg)f = h(gf)$ when either side is defined.

$$\begin{tikzcd} & \bullet \arrow[dl, "h"'] & & \bullet \arrow[ll, "gf"'] \arrow[dl, "f"] \\ \bullet & & \bullet \arrow[ul, "g"'] \arrow[ll, "hg"] & \end{tikzcd}$$

- Every object $x \in \mathbf{C}_0$ has an element $1_x \in \mathbf{C}(x, x)$ which is an *identity* in the sense that $f = f1_x$ whenever $\mathrm{source}(f) = x$, and $g = 1_x g$ whenever $\mathrm{target}(f) = x$.

$$\begin{tikzcd} \bullet & x \arrow[l, "f"'] & x \arrow[l, "1_x"'] \arrow[ll, bend left, "f"] \end{tikzcd} \qquad \begin{tikzcd} x & x \arrow[l, "1_x"'] & \bullet \arrow[l, "g"'] \arrow[ll, bend left, "g"] \end{tikzcd}$$

3.2. **Concrete categories.** A concrete category takes the following form:

- object = set with additional structure
- arrow = function 'compatible' with the structure
- composition = composition of functions
- identity = identity function

Here are several examples.

category	object	arrow
Set	set	function
Vect	vector space	linear map
Poset	partially-ordered set	monotone function
Group	group	homomorphism
Top	topological space	continuous map

Some of these categories have *full subcategories* that are in common use. These are formed by restricting to a smaller class of objects while keeping the arrows between those objects unchanged:

- **Ab** (abelian groups) is a full subcategory of **Group**.
- **Haus** (Hausdorff spaces) is a full subcategory of **Top**.
- **Cpct** (compact Hausdorff spaces) is a full subcategory of **Top**.

3.3. **Other examples.** There are many important examples of categories whose arrows do not correspond to functions.

Example 3.1. A partially-ordered set $(P, \leq)$ may be interpreted as a category $\mathbf{P}$ as follows.

- The objects are the elements of P.
- If $x \leq y$ then there is a unique arrow $x \to y$; otherwise there is no arrow from x to y.
- Transitivity guarantees that $x \to y$ and $y \to z$ can be composed to $x \to z$.
- Reflexivity guarantees the existence of identity arrows $x \to x$.

Remark. We do not make use of the anti-symmetry axiom ($x \leq y$ and $y \leq x$ implies $x = y$). A set with a binary relation $\leq$ that is reflexive and transitive is called a **pre-ordered set**, and can be interpreted as a category exactly as above. Pre-ordered sets correspond exactly to **thin categories**, which are categories where each $\mathrm{Mor}(x, y)$ contains at most one arrow.

Example 3.2. A group G may be interpreted as a category $\mathbf{G}$ as follows.

- There is a single object $*$.
- There is an arrow for each group element $g \in G$.
- Composition is defined by group multiplication.
- The identity 1_* is the arrow corresponding to the group identity.

Remark. We do not make use of the existence of inverses. A set with an associative binary operation and an identity element is called a **semigroup with identity**, and can be interpreted as a category exactly as above. Semigroups with identity correspond exactly to **monoids**, which are categories with one object.

The two examples above represent two extremes: (i) plenty of objects, at most one arrow with given source and target; (ii) one object, plenty of arrows.

Example 3.3. A directed multigraph D determines a category $\mathbf{D}$ as follows.

- The objects are the vertices of D.
- There is an arrow $x \to y$ for each path from x to y along directed edges.
- Composition is concatenation of paths.
- The identity arrow 1_x is the empty path at x.

We call $\mathbf{D}$ the *path category* of D.

For instance, the graph $\bullet_x \xrightarrow{e} \bullet_y$ gives rise to the category

$$\begin{tikzcd} x \arrow[loop above, "1_x"] \arrow[rr, "e"] & & y \arrow[loop above, "1_y"] \end{tikzcd}$$

with 2 objects and 3 arrows; and the graph $\begin{tikzcd} \bullet_x \arrow[r, "e"] & \bullet_y \arrow[r, bend left, "f"] \arrow[r, bend right, "g"'] & \bullet_z \end{tikzcd}$ gives rise to the category

$$\begin{tikzcd} x \arrow[loop above, "1_x"] \arrow[r, "e"] \arrow[rr, bend left=50, "fe"] \arrow[rr, bend right=50, "ge"'] & y \arrow[loop above, "1_y"] \arrow[r, bend left, "f"] \arrow[r, bend right, "g"'] & z \arrow[loop above, "1_z"] \end{tikzcd}$$

with 3 objects and 8 morphisms. Notice that the directed path $\xrightarrow{e}$ followed by $\xrightarrow{f}$ is written $\xrightarrow{fe}$ rather than $\xrightarrow{ef}$, in accordance with the syntax of composition of arrows.

One important example of this type arises from the one-loop graph:

$$\begin{tikzcd} \bullet \arrow[loop right, "e"] \end{tikzcd}$$

This results in a category with one object and morphisms $\{1, e, e^2, e^3, \dots\}$. In terms of Example 3.2, this is the monoid corresponding to the semigroup $\mathbb{N} = \{0, 1, 2, 3, \dots\}$. This category lies at the heart of the theory of discrete dynamical systems.

Example 3.4. Every category $\mathbf{C}$ has an *opposite category* $\mathbf{C}^{\mathrm{op}}$. The objects of $\mathbf{C}^{\mathrm{op}}$ are the same as the objects of $\mathbf{C}$, but the morphisms are reversed: $\mathbf{C}^{\mathrm{op}}(x, y) = \mathbf{C}(y, x)$. Thus each arrow $x \xrightarrow{f} y$ in $\mathbf{C}$ becomes an arrow $x \xleftarrow{f^{\mathrm{op}}} y$ in $\mathbf{C}^{\mathrm{op}}$. We define composition in the opposite category by the only possible rule:

$$f^{\mathrm{op}} g^{\mathrm{op}} = (gf)^{\mathrm{op}}.$$

Clearly $(\mathbf{C}^{\mathrm{op}})^{\mathrm{op}} = \mathbf{C}$.

3.4. **Iso- and other morphisms.** An arrow $x \xrightarrow{f} y$ is an **isomorphism** if there exists an arrow $x \xleftarrow{g} y$ such that $gf = 1_x$ and $fg = 1_y$.

Remark. The map g is unique if it exists, since $g = g1_y = g(f\hat{g}) = (gf)\hat{g} = 1_x\hat{g} = \hat{g}$ for any two such maps $g, \hat{g}$. Thus we can write $f^{-1} := g$ without ambiguity.

Example 3.5. Here are several examples.

- Every identity map 1_x is an isomorphism.
- The isomorphisms in **Set** are the bijections.

In a concrete category an isomorphism must be a bjiection. The converse may fail.

- In **Vect**, **Group** and **Ab**, the isomorphisms are precisely the maps which are bijections. This is because the inverse of such a map is itself structure-preserving. For instance, the inverse of a linear bijection is itself linear.
- The isomorphisms in **Top** are the homeomorphisms. A continuous bijection is not necessarily a homeomorphism; the inverse is required to be continuous. In the full subcategory **Cpct** of compact Hausdorff spaces, however, a continuous bijection *is* always a homeomorphism.
- In the category **P** arising from a poset $(P, \leq)$, the only isomorphisms are the identity arrows. Indeed, if there are arrows $x \to y$ and $y \to x$ then $x = y$, by antisymmetry.
- In the category **G** arising from a group G, every arrow is an isomorphism.
- In the path category **D** of a directed multigraph D, the only isomorphisms are the identity arrows 1_x. This is because concatenation is strictly additive on path-length; there is no cancellation.

In the category **Set**, an isomorphism is a function which is both **1–1** and **onto**. These two separate notions can be generalized to an arbitrary category as follows.

Definition 3.6. Let $x \xrightarrow{f} y$ be an arrow in a category **C**.

- We say that f is a **monomorphism** (or f is **monic**) if $fg = fh$ implies $g = h$ for any pair of arrows g, h with target x and a common source.

$$\begin{tikzcd} \bullet \arrow[r, shift left, bend left, "g"] \arrow[r, shift right, bend right, "h"'] & x \arrow[r, "f"] & y \end{tikzcd}$$

In other words, f is a monomorphism if it is algebraically left-cancellable.

- We say that f is an **epimorphism** (or f is **epic**) if $gf = hf$ implies $g = h$ for any pair of arrows g, h with source y and a common target.

$$\begin{tikzcd} x \arrow[r, "f"] & y \arrow[r, bend left, "g"] \arrow[r, bend right, "h"'] & \bullet \end{tikzcd}$$

In other words, f is an epimorphism if it is algebraically right-cancellable.

Proposition 3.7. *An isomorphism is both a monomorphism and an epimorphism.*

Proof. We can (left|right)-cancel f by (left|right)-composing with f^{-1}. ☐

We immediately note that the converse isn't true. An arrow that is both a monomorphism and an epimorphism need not be an isomorphism. In some categories (including **Set**, **Vect**, **Group**) it can be proven that it is, but in general it need not be. See homework **25**, **26**.

The lack of a converse should not surprise us: for f to be an isomorphism we need an arrow $y \to x$, but the definitions of monic and epic say nothing about the existence of arrows. In fact, there may be no arrows $y \to x$ at all. In the 2-object 3-arrow category

$$\begin{tikzcd} x \arrow[loop above, "1_x"] \arrow[rr, "f"] & & y \arrow[loop above, "1_y"] \end{tikzcd}$$

the arrow f is both monic and epic, for trivial reasons, but it is certainly not an isomorphism.

Proposition 3.8. *In the category* **Set** *(i) a function is a monomorphism if and only if it is* 1–1*; and (ii) a function is an epimorphism if and only if it is onto.*

Proof. Let $f : A \to B$.

(i) ($\Leftarrow$) Suppose f is 1–1, and suppose $fg = fh$ for functions $g, h : X \to A$. For all $x \in X$, we have $fg(x) = fh(x)$ and therefore $g(x) = h(x)$ since f is 1–1. Thus $g = h$.

(i) ($\Rightarrow$) Suppose f is a monomorphism. We show that f is 1–1. Suppose $a_1, a_2 \in A$ satisfy $f(a_1) = f(a_2)$. Let $X = \{*\}$ and define $g, h : X \to A$ by $g(*) = a_1$ and $h(*) = a_2$. Then $fg = fh$, and therefore $g = h$ since f is monic, and therefore $a_1 = a_2$.

(ii) ($\Leftarrow$) Suppose f is onto, and suppose $gf = hf$ for functions $g, h : B \to X$. For every $b \in B$, there exists $a \in A$ such that $f(a) = b$; so $g(b) = gf(a) = hf(a) = h(b)$. Thus $g = h$.

(ii) ($\Rightarrow$) Suppose f is an epimorphism. We show that f is onto. Let $X = \{0, 1\}$ and define maps $g, h : B \to X$ as follows:

$$g(b) = 0 \quad \text{for all } b \in B \qquad \text{and} \qquad h(b) = \begin{cases} 0 & \text{for all } b \in \operatorname{im}(f) \\ 1 & \text{for all } b \notin \operatorname{im}(f) \end{cases}$$

Then $gf = hf$, and therefore $g = h$ since f is epic, and therefore $\operatorname{im}(f) = B$. ☐

3.5. **Universal objects.** Initial objects and terminal objects play a special role in a category. Let $\mathbf{C}$ be a category and $x \in \mathbf{C}_0$.

- x is **initial** $\Leftrightarrow$ for every $y \in \mathbf{C}_0$ there is exactly one arrow $x \to y$.
- x is **terminal** $\Leftrightarrow$ for every $w \in \mathbf{C}_0$ there is exactly one arrow $w \to x$.

Not every category has an initial or a terminal object. Some categories have both.

Example 3.9. The empty set $\emptyset$ is the unique initial object in **Set**. Every singleton set $\{x\}$ is terminal in **Set**.

Example 3.10. Every 0-dimensional vector space is both initial and terminal in **Vect**.

Example 3.11. A poset category has an initial (respectively, terminal) object $\Leftrightarrow$ the poset has a least (respectively, greatest) element.

Initial and terminal objects are essentially unique, and in that sense 'universal':

Proposition 3.12. *Any two initial (respectively, terminal) objects in a category are isomorphic, by a uniquely determined isomorphism.*

Proof. Let x, y be initial (respectively, terminal). Then there are unique arrows

$$\begin{tikzcd} x \arrow[r, bend left, "f"] & y \arrow[l, bend left, "g"] \end{tikzcd}$$

between them. In fact, these comprise an isomorphism: since there is a unique arrow $x \to x$, it follows that $gf = 1_x$; and since there is a unique arrow $y \to y$ it follows that $fg = 1_y$. $\square$

Many familiar constructions in mathematics can be thought of as universal constructions. We give an example of this. Let $U \xrightarrow{\alpha} V$ be an arrow in **Vect**. We will construct its cokernel $V/\operatorname{im}(\alpha)$ in a crafty way.

Step 1. Define a category $\mathbf{N}$ of 'nullifiers' of α:

- Objects are linear maps β with source V such that $\beta\alpha = 0$.

$$\begin{tikzcd} U \arrow[r, "\alpha"] \arrow[rr, bend right, "0"'] & V \arrow[r, "\beta"] & W \end{tikzcd}$$

- Morphisms are linear maps $\phi : W \to W'$ such that $\phi\beta = \beta'$:

$$\begin{tikzcd} & W \arrow[d, "\phi"] \\ V \arrow[ur, "\beta"] \arrow[dr, "\beta'"'] & \\ & W' \end{tikzcd}$$

Step 2. Define $\operatorname{Coker}\alpha$ to be any initial object in this category. If one exists, then it is well defined in the sense that any two choices are uniquely isomorphic to each other.

Step 3. Exhibit a specific instance of $\operatorname{Coker}\alpha$:

$$\pi = \text{quotient map } V \to V/\operatorname{im}(\alpha), \text{ defined by } v \mapsto [v]_{\operatorname{mod}\operatorname{im}(\alpha)}$$

This certainly nullifies α. Moreover, given any other nullifier β we seek to construct $\bar{\beta}$ in the following diagram:

$$\begin{tikzcd} V \arrow[r, "\pi"] \arrow[dr, "\beta"'] & V/\operatorname{im}(\alpha) \arrow[d, dotted, "\bar{\beta}"] \\ & W \end{tikzcd}$$

To make the diagram commute, we are forced to define $\bar{\beta}[v] = \beta v$ for all $v \in V$. This is well-defined because $[v] = [v']$ means that $v' = v + \alpha u$ for some $u \in U$, and then

$$\beta v' = \beta(v + \alpha u) = \beta v + \beta\alpha u = \beta v$$

since $\beta\alpha = 0$. Thus $\bar{\beta}$ is uniquely defined, and π is an initial object in $\mathbf{N}$, as claimed.

3.6. **Functors.** Let $\mathbf{C}, \mathbf{D}$ be categories. A *functor* $F : \mathbf{C} \to \mathbf{D}$ is specified by the following data:

- Every object $x \in \mathbf{C}_0$ is assigned an object $F(x) \in \mathbf{D}_0$.
- Every arrow $\alpha \in \operatorname{Mor}_{\mathbf{C}}(x, y)$ is assigned an arrow $F[\alpha] \in \operatorname{Mor}_{\mathbf{D}}(F(x), F(y))$.
- The functor respects composition: $F[\beta\alpha] = F[\beta]F[\alpha]$.
- The functor respects identities: $F[1_x] = 1_{F(x)}$.

Example 3.13. There is a *forgetful functor* $\mathbf{Top} \to \mathbf{Set}$ which takes a topological space to its underlying set, and which takes each continuous function to itself.

Example 3.14. There is a *forgetful functor* $\mathbf{Vect} \to \mathbf{Set}$ which takes a vector space to its underlying set, and which takes each linear map to itself.

Example 3.15. Let $\mathbf{P}$ be the category of subsets of the plane, with a unique arrow $U \to V$ whenever $U \subseteq V$ and no arrow otherwise. Then H_1 is a functor $\mathbf{P} \to \mathbf{Vect}$. Indeed, for every U we construct

$$\mathrm{C}_0(U) \xleftarrow{\partial_0} \mathrm{C}_1(U) \xleftarrow{\partial_1} \mathrm{C}_2(U)$$

and define $\mathrm{H}_1(U) = \ker\partial_0/\operatorname{im}\partial_1$. Whenever $U \subseteq V$, we can draw a commutative diagram

$$\begin{tikzcd} \mathrm{C}_0(U) \arrow[d] & \mathrm{C}_1(U) \arrow[l, "\partial_0"'] \arrow[d] & \mathrm{C}_2(U) \arrow[l, "\partial_1"'] \arrow[d] \\ \mathrm{C}_0(V) & \mathrm{C}_1(V) \arrow[l, "\partial_0"'] & \mathrm{C}_2(V) \arrow[l, "\partial_1"'] \end{tikzcd}$$

where the vertical maps are inclusions of vector spaces. Then there is a map

$$\mathrm{H}_1[U \subseteq V] : \mathrm{H}_1(U) \to \mathrm{H}_1(V); \quad [\gamma]_{\operatorname{mod}\mathrm{B}_1(U)} \mapsto [\gamma]_{\operatorname{mod}\mathrm{B}_1(V)}$$

which is well-defined because $\mathrm{B}_0(U) \subseteq \mathrm{B}_0(V)$, and trivially linear. It is immediate that $\mathrm{H}_1[-]$ respects composition and identities.

Example 3.16. The *free functor* $F : \mathbf{Set} \to \mathbf{Vect}_{\mathbb{F}}$ is defined as follows:

$$F(A) = \{\text{functions } v : A \to \mathbb{F} \text{ such that } v(a) = 0 \text{ for all but finitely many } a\}$$

This is a vector space (addition and scalar multiplication carried out pointwise) and it has a basis consisting of elements $(e_a \mid a \in A)$ defined as follows:

$$\begin{aligned} e_a(a) &= 1; \\ e_a(x) &= 0, \quad \text{if } x \neq a. \end{aligned}$$

Given a function $\alpha : A \to B$ we define $F[\alpha] : F(A) \to F(B)$ by setting

$$\big[F[\alpha](v)\big] : b \mapsto \sum_{a \in \alpha^{-1}(b)} v(a)$$

for any function $v \in F(A)$. This is the unique linear map $F(A) \to F(B)$ with

$$F[\alpha](e_a) = e_{\alpha(a)}$$

on each basis vector e_a.

Lecture 7

Example 3.17. Let G, H be groups and let $\mathbf{G}, \mathbf{H}$ be the associated categories (with one object, and a morphism for each group element). A functor $\Phi : \mathbf{G} \to \mathbf{H}$ is determined by the map $g \mapsto \Phi[g]$, and this map is required to respect composition and the identity. Thus functors $\mathbf{G} \to \mathbf{H}$ are the same thing as group homomorphisms $G \to H$.

Example 3.18. Let P, Q be posets and $\mathbf{P}, \mathbf{Q}$ be the associated categories (with an object for every poset element and an arrow for every relation $x \leq y$). Then functors $\mathbf{P} \to \mathbf{Q}$ are the same thing as maps $P \to Q$ that are order-preserving.

3.7. **Contravariant Functors.** A **contravariant** functor is a 'functor which reverses arrows'. The standard formulation is to define a contravariant functor to be a functor $\mathbf{C} \to \mathbf{D}^{\mathrm{op}}$, where $\mathbf{D}^{\mathrm{op}}$ is the opposite category to $\mathbf{D}$. In practice, this means the following:

- Every object $x \in \mathbf{C}_0$ is assigned an object $F(x) \in \mathbf{D}_0$.
- Every arrow $\alpha \in \mathrm{Mor}_{\mathbf{C}}(x, y)$ is assigned an arrow $F[\alpha] \in \mathrm{Mor}_{\mathbf{D}}(F(y), F(x))$.
- The functor respects composition: $F[\beta\alpha] = F[\alpha]F[\beta]$.
- The functor respects identities: $F[1_x] = 1_{F(x)}$.

The reversal of the composition law is the only definition that makes sense in this context:

$$\begin{tikzcd} z & y \arrow[l, "\beta"'] & x \arrow[l, "\alpha"'] \arrow[ll, bend left, "\beta\alpha"] \end{tikzcd} \qquad \text{becomes} \qquad \begin{tikzcd} F(z) \arrow[r, "F{[\beta]}"] \arrow[rr, bend right, "F{[\alpha]}F{[\beta]}"'] & F(y) \arrow[r, "F{[\alpha]}"] & F(x) \end{tikzcd}$$

Remark. Functors in the usual sense may be called **covariant** functors, when distinguishing them from contravariant functors. I will usually simply speak of functors $\mathbf{C} \to \mathbf{D}$ and functors $\mathbf{C} \to \mathbf{D}^{\mathrm{op}}$. I may sometimes, illogically, refer to a 'contravariant functor $\mathbf{C} \to \mathbf{D}^{\mathrm{op}}$' when I simply mean a functor $\mathbf{C} \to \mathbf{D}^{\mathrm{op}}$. In those instances, the word 'contravariant' is intended to emphasize the $^{\mathrm{op}}$ rather than cancel it out.

Example 3.19. Let $\mathbf{P}, \mathbf{Q}$ be categories associated to posets P, Q. Functors $\mathbf{P} \to \mathbf{Q}^{\mathrm{op}}$ correspond to order-reversing maps $P \to Q$.

Example 3.20. A special case of the previous example is the Galois correspondence for a Galois extension $\mathbb{E}/\mathbb{F}$ of fields: the poset of intermediate fields is reverse-isomorphic to the poset of subgroups of the group $\mathrm{Gal}(\mathbb{E}/\mathbb{F})$ of automorphisms of $\mathbb{E}$ that fix $\mathbb{F}$.

Example 3.21. Vector-space duality defines a functor $(-)^* : \mathbf{Vect} \to \mathbf{Vect}^{\mathrm{op}}$:

- Given a vector space V, we define $V^* = \mathrm{Hom}_{\mathbb{F}}(V, \mathbb{F})$.
- Given a linear map $T : V \to W$, we define $T^* : W^* \to V^*$ to be the map $(\alpha \mapsto \alpha T)$.

$$\begin{tikzcd} V \arrow[r, "T"] \arrow[dr, "\alpha T"'] & W \arrow[d, "\alpha"] \\ & \mathbb{F} \end{tikzcd}$$

- The composition law $(TS)^* = S^* T^*$ follows by contemplating the diagram

$$\begin{tikzcd} U \arrow[r, "S"] \arrow[rr, bend left, "TS"] \arrow[drr, "\alpha TS"'] & V \arrow[r, "T"] \arrow[dr, "\alpha T"] & W \arrow[d, "\alpha"] \\ & & \mathbb{F} \end{tikzcd}$$

  or by writing $S^* T^*(\alpha) = S^*(\alpha T) = \alpha TS = (TS)^* \alpha$.
- The identity law $(1_V)^* = 1_{V^*}$ is obvious from the definition.

Remark. Vector space double-duality is the composite of two contravariant functors, and therefore a covariant functor $(-)^{**} : \mathbf{Vect} \to \mathbf{Vect}$. A finite-dimensional vector space is 'naturally isomorphic' to its double dual, and in general every vector space is 'naturally' embedded as a subspace of its double dual. The concept of a natural transformation between functors is what gives meaning to the word 'natural' here.

3.8. **Diagram categories and natural transformations.** The collection of functors between two categories is something that can be studied in its own right. Let $\mathbf{C}, \mathbf{D}$ be categories, and define

$$\mathbf{C}^{\mathbf{D}} = \{\text{functors } \mathbf{D} \to \mathbf{C}\}.$$

There are two different flavours here.

- If $\mathbf{D}$ is a small category (meaning that its collection of objects is a set rather than a proper class), then we typically study the collection of all such functors. In many cases we think of $\mathbf{C}^{\mathbf{D}}$ as a category of 'diagrams in $\mathbf{C}$ with shape $\mathbf{D}$'. Any two such diagrams can be compared by defining morphisms between them in a sensible way.
- if $\mathbf{D}$ is a large category, then we are typically interested in studying specific functors from $\mathbf{D}$ to $\mathbf{C}$. A well-chosen functor allows us to exploit what we know about objects and arrows in $\mathbf{C}$ to answer questions about objects and arrows in $\mathbf{D}$. For example, algebraic topology makes extensive use of functors from topological categories to algebraic categories. In doing so, it becomes important to compare functors.

The mechanism for comparing two functors or diagrams is the *natural transformation*. It has the same definition in both situations. We will mostly consider the case where $\mathbf{D}$ is a small category, so that functors are thought of as diagrams and the entire collection $\mathbf{C}^{\mathbf{D}}$ is of interest.

Technical Comment. If $\mathbf{D}$ is a large category then the right-hand side in the definition of $\mathbf{C}^{\mathbf{D}}$ is a collection of classes rather than a collection of sets. This is not something we usually have a name for in the standard set-theoretic foundations. There are ways of working around this. The simplest is to agree never to 'collect' all possible functors at any one time; just the few that are needed for any given purpose. One can continue to use the $\mathbf{C}^{\mathbf{D}} = \{\dots\}$ notation as a convenient pretence to keep the high-level discussion tidy.

When $\mathbf{D}$ is small, it follows that $\mathbf{C}^{\mathbf{D}}$ is itself a category by taking natural transformations to be the arrows. And when $\mathbf{D}$ is large, $\mathbf{C}^{\mathbf{D}}$ is a somewhat illegal object which nonetheless behaves like a category.

Example 3.22. (i) Consider the following directed graph:

$$\bullet \longrightarrow \bullet \longrightarrow \bullet$$

Consider the category of vector space diagrams shaped like the graph. An object is a diagram of vector spaces and linear maps

$$V_0 \xrightarrow{v_1^0} V_1 \xrightarrow{v_2^1} V_2$$

and a morphism between two such objects is a diagram

$$\begin{tikzcd}
V_0 \arrow[r, "v_1^0"] \arrow[d, "\phi_0"] & V_1 \arrow[r, "v_2^1"] \arrow[d, "\phi_1"] & V_2 \arrow[d, "\phi_2"] \\
W_0 \arrow[r, "w_1^0"'] & W_1 \arrow[r, "w_2^1"'] & W_2
\end{tikzcd}$$

where the two squares commute: $\phi_1 v_1^0 = w_1^0 \phi_0$ and $\phi_2 v_2^1 = w_2^1 \phi_1$.

Note. The outer rectangle then commutes automatically: $\phi_2(v_2^1 v_1^0) = w_2^1 \phi_1 v_1^0 = (w_2^1 w_1^0)\phi_0$.

(ii) Vector-space diagrams shaped like the directed graph in (i) are the same thing as functors to **Vect** on the path category

$$\begin{tikzcd} 0 \arrow[loop above] \arrow[r] \arrow[rr, bend right] & 1 \arrow[loop above] \arrow[r] & 2 \arrow[loop above] \end{tikzcd}$$

associated to the directed graph. Indeed, such a functor F is constructed by specifying data as follows.

- Select three vector spaces V_0, V_1, V_2 and set $F(k) = V_k$ for $k = 0, 1, 2$.
- Select a linear map $v_1^0 : V_0 \to V_1$ and set $F[0 \to 1] = v_1^0$.
- Select a linear map $v_2^1 : V_1 \to V_2$ and set $F[1 \to 2] = v_2^1$.

This is exactly the same data that define the diagrams of (i). To complete the construction of the functor, we must decide what to do with the identity arrows and the composite arrow $0 \to 1 \to 2$. But these choices are forced by the defining properties of a functor: we must have $F[1_k] = 1_{V_k}$ for $k = 0, 1, 2$ and $F[0 \to 1 \to 2] = F[1 \to 2] \circ F[0 \to 1] = v_2^1 v_1^0$.

Remark. The same principle applies to any directed graph D and its path category $\mathbf{D}$: diagrams in a category $\mathbf{C}$ shaped like D are the same thing as functors $\mathbf{D} \to \mathbf{C}$.

Definition 3.23. Let $\mathbf{C}, \mathbf{D}$ be categories and let $F, G : \mathbf{D} \to \mathbf{C}$ be functors. A **natural transformation** η from F to G, written

$$\eta : F \Rightarrow G \qquad \text{or} \qquad F \overset{\eta}{\Longrightarrow} G \qquad \text{or} \qquad \begin{tikzcd} \mathbf{C} \arrow[r, bend left, "F"] \arrow[r, bend right, "G"'] \arrow[r, phantom, "\Downarrow\eta"] & \mathbf{D} \end{tikzcd},$$

is defined as follows.

- To each object $x \in \mathbf{D}$ we assign an arrow $\eta_x : F(x) \to G(x)$ of $\mathbf{C}$.
- For each arrow $x \xrightarrow{\alpha} y$ of $\mathbf{C}$ we require that the diagram

  $$\begin{tikzcd} F(x) \arrow[r, "F[\alpha]"] \arrow[d, "\eta_x"'] & F(y) \arrow[d, "\eta_y"] \\ G(x) \arrow[r, "G[\alpha]"] & G(y) \end{tikzcd}$$

  commutes.

The morphisms defined in Example 3.22 (and homework **30**) are natural transformations.

For each F we define the identity natural transformation $F \overset{1_F}{\Longrightarrow} F$ by $(1_F)_x = 1_{F(x)}$. And given two natural transformations

$$F \overset{\eta}{\Longrightarrow} G \quad \text{and} \quad G \overset{\zeta}{\Longrightarrow} H$$

we define their composite

$$F \overset{\zeta\eta}{\Longrightarrow} H$$

by $(\zeta\eta)_x = \zeta_x \eta_x$. In this way, $\mathbf{C}^{\mathbf{D}}$ becomes a category.

Remark. There are other ways to compose natural transformations. The version that we have given, which makes $\mathbf{C}^{\mathbf{D}}$ a category, is sometimes known as 'vertical' composition. There is also a 'horizontal' composition which combines two natural transformations

$$\begin{tikzcd} \mathbf{C} \arrow[r, bend left, "F"] \arrow[r, bend right, "G"'] \arrow[r, phantom, "\Downarrow\eta"] & \mathbf{D} \end{tikzcd} \quad \text{and} \quad \begin{tikzcd} \mathbf{D} \arrow[r, bend left, "H"] \arrow[r, bend right, "K"'] \arrow[r, phantom, "\Downarrow\zeta"] & \mathbf{E} \end{tikzcd}$$

to obtain a third natural transformation

$$\begin{tikzcd} \mathbf{C} \arrow[r, bend left, "HF"] \arrow[r, bend right, "KG"'] \arrow[r, phantom, "\Downarrow"] & \mathbf{E} \end{tikzcd}$$

written $\eta * \zeta : HF \Rightarrow KG$.

3.9. **Limits and colimits.** Many important and familiar constructions in mathematics can be expressed as the *limit* or *colimit* of a functor. Perhaps the most prevalent instances are *products* and *coproducts* (homework **28** and **29**). Limits and colimits are formed by constructing a suitable category and looking for terminal or initial objects in that category. They are 'universal' constructions and that explains their importance.

Remark. We saw in the previous section that diagrams shaped like a directed graph D are the same thing as functors from the path category $\mathbf{D}$. Therefore we can talk about the limit or colimit of a diagram. Most of our examples are of this type.

Limits. (i) Let $F : \mathbf{D} \to \mathbf{C}$ be a functor. A **cone** on F is defined by the following data:

- An object X of $\mathbf{C}$.
- An arrow $X \xrightarrow{p_d} F(d)$ for each $d \in \mathbf{D}_0$. (These are arrows in $\mathbf{C}$.)
- For each arrow $d_1 \xrightarrow{\alpha} d_2$ of $\mathbf{D}$, the diagram

$$\begin{tikzcd} & F(d_1) \arrow[dd, "F[\alpha]"] \\ X \arrow[ur, "p_{d_1}"] \arrow[dr, "p_{d_2}"'] & \\ & F(d_2) \end{tikzcd}$$

  is required to commute.

Thus a cone is an object of $\mathbf{C}$ and a selection of 'compatible' arrows to each $F(d)$.

(ii) A morphism between cones $(X, (p_d))$, $(Y, (q_d))$ is defined to be an arrow $X \xrightarrow{\xi} Y$ such that the diagram

$$\begin{tikzcd}
X \arrow[d, "\xi"'] \arrow[dr, "p_d"] & \\
 & F(d) \\
Y \arrow[ur, "q_d"'] &
\end{tikzcd}$$

commutes for all $d \in D$. Thus we get a category $\mathbf{Cone}(F)$.

(iii) If $\mathbf{Cone}(F)$ has a terminal object, it is called a **limit** of F. Since terminal objects are unique up to a unique isomorphism, it is customary to speak of 'the' limit of F.

Colimits. (i) Let $F : \mathbf{D} \to \mathbf{C}$ be a functor. A **cocone** on F is defined by the following data:

- An object X of $\mathbf{C}$.
- An arrow $F[d] \xrightarrow{i_d} X$ for each $d \in \mathbf{D}_0$. (These are arrows in $\mathbf{C}$.)
- For each arrow $d_1 \xrightarrow{\alpha} d_2$ of $\mathbf{D}$, the diagram

$$\begin{tikzcd}
F(d_1) \arrow[d, "F[\alpha]"'] \arrow[dr, "i_{d_1}"] & \\
 & X \\
F(d_2) \arrow[ur, "i_{d_2}"'] &
\end{tikzcd}$$

  is required to commute.

Thus a cone is an object of $\mathbf{C}$ and a selection of 'compatible' arrows from each $F(d)$.

(ii) A morphism between cocones $(X, (i_d))$, $(Y, (j_d))$ is defined to be an arrow $X \xrightarrow{\xi} Y$ such that the diagram

$$\begin{tikzcd}
 & X \arrow[dd, "\xi"] \\
F(d) \arrow[ur, "i_d"] \arrow[dr, "j_d"'] & \\
 & Y
\end{tikzcd}$$

commutes for all $d \in D$. Thus we get a category $\mathbf{coCone}(F)$.

(iii) If $\mathbf{coCone}(F)$ has an initial object, it is called a **colimit** of F. Since initial objects are unique up to a unique isomorphism, it is customary to speak of 'the' colimit of F.

$\boxed{\text{A limit is a terminal cone. A colimit is an initial cocone.}}$

Example 3.24 (products and coproducts)**.** Consider diagrams in a category $\mathbf{C}$ shaped like the graph

$$\bullet_1 \qquad \bullet_2$$

with two vertices and no edges. Such a diagram is simply a choice of two objects of $\mathbf{C}$:

$$A_1 \qquad A_2$$

The **product** of A_1 and A_2 is defined to be the limit of this diagram, if it exists. It is terminal in the category whose objects look like this:

$$A_1 \xleftarrow{p_1} X \xrightarrow{p_2} A_2$$

See homework **29** for an example.

Dually, the **coproduct** of A_1 and A_2 is the colimit of the diagram if it exists. It is initial in the category whose objects look like this:

$$A_1 \xrightarrow{j_1} X \xleftarrow{j_2} A_2$$

See homework **28** for an example.

Here are some standard examples of products and coproducts:

category	product	coproduct
Set	Cartesian product $A_1 \times A_2$	disjoint union $A_1 \sqcup A_2$
Top	product space $X_1 \times X_2$	disjoint union $X_1 \sqcup X_2$
Vect	direct sum $V_1 \oplus V_2$	direct sum $V_1 \oplus V_2$
Group	Cartesian product $G_1 \times G_2$	free product $G_1 * G_2$

The maps p_i, p_2 or j_1, j_2 are part of the description of each product or coproduct. In each of the table entries it should be clear what those maps are.

Note. Products and coproducts of three or more objects can be defined using empty graphs with the appropriate number of vertices (finite or infinite). If the number of objects is finite, then it can be proved that the same result is obtained by iterating the two-object construction enough times.

Example 3.25. The limit of a diagram of the form

$$\begin{tikzcd} & B \arrow[d, "g"] \\ A \arrow[r, "f"'] & C \end{tikzcd} \qquad \text{or equivalently} \qquad A \xrightarrow{f} C \xleftarrow{g} B$$

is called a **pullback**. A cone on this diagram is a space X and maps p_A, p_B, p_C making the left diagram commute:

$$\begin{tikzcd} X \arrow[r, "p_B"] \arrow[d, "p_A"'] \arrow[dr, "p_C"] & B \arrow[d, "g"] \\ A \arrow[r, "f"'] & C \end{tikzcd} \qquad\qquad \begin{tikzcd} X \arrow[r, "p_B"] \arrow[d, "p_A"'] & B \arrow[d, "g"] \\ A \arrow[r, "f"'] & C \end{tikzcd}$$

However, we can omit p_C and simply look for p_A, p_B making the right diagram commute. This is because p_C is necessarily equal to $fp_A = gp_B$ so it is redundant to specify it.

Note. The colimit of the diagram is uninteresting: it is C with the maps $f, g, 1_C$.

Proposition 3.26. *In* **Set** *the pullback of* $A \xrightarrow{f} C \xleftarrow{g} B$ *is the set*

$$N = \{(a,b) \mid a \in A,\, b \in B,\, f(a) = g(b)\} \subseteq A \times B$$

together with the projections n_A, n_B onto A and B.

Proof. First we verify that (N, n_A, n_B) is a cone. For any $(a,b) \in N$ we have

$$fn_A(a,b) = f(a) = g(b) = gn_B(a,b),$$

so $fn_A = gn_B$.

Now consider an arbitrary cone (X, p_A, p_B). We seek a map α making the following diagram commute:

$$\begin{tikzcd} X \arrow[dr, dotted, "\alpha"] \arrow[drr, bend left, "p_B"] \arrow[ddr, bend right, "p_A"'] & & \\ & N \arrow[r, "n_B"] \arrow[d, "n_A"'] & B \arrow[d, "g"] \\ & A \arrow[r, "f"'] & C \end{tikzcd}$$

Since $n_A\alpha = p_A$ and $n_B\alpha = p_B$, we must have $\alpha(x) = (p_A(x), p_B(x))$ for all $x \in X$. And if we take this as the definition of α, the diagram does commute. Thus the morphism α exists and is unique. ☐

Example 3.27. The colimit of a diagram of the form

$$\begin{tikzcd} A & \\ C \arrow[u, "f"'] \arrow[r, "g"'] & B \end{tikzcd} \qquad \text{or equivalently} \qquad A \xleftarrow{f} C \xrightarrow{g} B$$

is called a **pushout**. A cone on this diagram is a space X and maps j_A, j_B, j_C making the left diagram commute:

$$\begin{tikzcd} A \arrow[r, "j_A"] & X \\ C \arrow[u, "f"] \arrow[r, "g"'] \arrow[ur, "j_C"'] & B \arrow[u, "j_B"'] \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "j_A"] & X \\ C \arrow[u, "f"] \arrow[r, "g"'] & B \arrow[u, "j_B"'] \end{tikzcd}$$

However, we can omit j_C and simply request j_A, j_B making the right diagram commute. This is because j_C is necessarily equal to $j_Af = j_Bg$ so it is redundant to specify it.

Remark. The colimit of the diagram is uninteresting: it is C with the maps $f, g, 1_C$.

Proposition 3.28. *In* **Set** *the pushout of* $A \xleftarrow{f} C \xrightarrow{g} B$ *is the set*

$$U = \frac{\{(a,0) \mid a \in A\} \cup \{(b,1) \mid b \in B\}}{\sim} = \frac{A \sqcup B}{\sim}$$

where $\sim$ *is the smallest equivalence relation which contains the relations* $(f(c),0) \sim (g(c),1)$ *for all* $c \in C$*; along with the maps*

$$u_A : A \to U;\ a \mapsto [(a,0)], \qquad \textit{and} \qquad u_B : B \to U;\ b \mapsto [(b,1)].$$

Proof. First we verify that (U, u_A, u_B) is a cocone. Given $c \in C$ we have

$$u_A(f(c)) = [(f(c),0)] = [(g(c),1)] = u_B(g(c))$$

so $u_A f = u_B g$.

Now consider an arbitrary cocone (X, i_A, i_B). We seek a map α making the following diagram commute:

$$\begin{tikzcd} & & X \\ A \arrow[r, "u_A"] \arrow[urr, bend left, "i_A"] & U \arrow[ur, dotted, "\alpha"] & \\ C \arrow[u, "f"] \arrow[r, "g"'] & B \arrow[u, "u_B"'] \arrow[uur, bend right, "i_B"'] & \end{tikzcd}$$

Since $\alpha u_A = i_A$ and $\alpha u_B = i_B$, we must have $\alpha([(a,0)]) = i_A(a)$, and $\alpha([(b,1)]) = i_B(b)$. To show that α exists we must show that these formulas define a consistent value for the elements of each equivalence class. It is enough to test this for the generating relations. And indeed

$$\alpha([(f(c),0)]) = i_A(f(c)) = i_B(g(c)) = \alpha([(g(c),1)])$$

as required. Thus the morphism α exists and is unique. $\square$

The situation for **Set** has the following form: a pullback is a subset of the product of A, B and a pushout is a quotient set of the coproduct of A, B. Notice also that the pullback of

$$A \longrightarrow \{*\} \longleftarrow B$$

is precisely the product $A \times B$, using the fact that $\{*\}$ is terminal; and the pushout of

$$A \longleftarrow \emptyset \longrightarrow B$$

is precisely the coproduct $A \sqcup B$, using the fact that $\emptyset$ is initial.

3.10. **The homotopy category.** Algebraic topologists spend much of their time working in the *homotopy category* **hTop**. This is a modification of **Top** defined as follows:

- The objects of **hTop** are topological spaces.
- Given objects X, Y the set of arrows $\mathbf{hTop}(X,Y)$ is defined to be the set of equivalence classes of continuous functions $X \to Y$ with respect to the homotopy relation.

We have already, covertly, worked in the homotopy category: the continuous winding number on loops is a function $\mathbf{hTop}(S^1, \mathbb{R}^2 - \{0\}) \to \mathbb{Z}$. Let us review the general definition.

Two continuous maps $f, g : X \to Y$ are **homotopic** if there exists a homotopy between them; that is, a continuous function

$$F : X \times [0,1] \to Y$$

such that

$$F(x,0) = f(x) \quad \text{and} \quad F(x,1) = g(x) \quad \text{for all } x \in X.$$

We write $f \simeq g$ or $f \simeq_F g$. The relation is reflexive, symmetric and transitive, by considering the homotopies

$$R(x,t) = f(x), \quad S(x,t) = F(x, 1-t), \quad T(x,t) = \begin{cases} G(x,2t) & (0 \le t \le \frac{1}{2}) \\ H(x, 2t-1) & (\frac{1}{2} \le t \le 1) \end{cases}$$

which give

$$f \simeq_R f, \quad g \simeq_S f \text{ when } f \simeq_F g, \quad f \simeq_T h \text{ when } f \simeq_G g \text{ and } g \simeq_H h$$

respectively.

To confirm that **hTop** is a category, we need to check one thing:

Proposition 3.29. *Composition $[f][g] = [fg]$ is well-defined in* **hTop**.

Proof. Let $f_1, f_2 : X \to Y$ and $g_1, g_2 : Y \to Z$ be continuous maps with $f_1 \simeq_F f_2$ and $g_1 \simeq_G g_2$. Then $g_1 f_1 \simeq_H g_2 f_2$ where $H(x,t) = G(F(x,t),t)$. □

One major consequence of working in **hTop** is that there are many more isomorphisms. Two spaces X, Y are isomorphic in **hTop** if there are continuous maps $f : X \to Y$, and $g : Y \to X$ such that

$$[g][f] = [1_X] \quad \text{and} \quad [f][g] = [1_Y];$$

in other words, such that

$$gf \simeq 1_X \quad \text{and} \quad fg \simeq 1_Y.$$

In that case we say that the spaces X, Y are **homotopy equivalent** and we write $X \simeq Y$. The map f is a **homotopy equivalence** (as is the map g).

Proposition 3.30. *The 1-point space $\{*\}$ is homotopy equivalent to any star-shaped subset $U \subseteq \mathbb{R}^n$. In particular, $\{*\} \simeq \mathbb{R}^n$ for all n.*

Proof. Let $u_0 \in U$ be a star-center, so that $|[u_0, u]| \subseteq U$ for every $u \in U$. Define maps in both directions:

$$j : \{*\} \to U; \quad * \mapsto u_0$$
$$p : U \to \{*\}; \quad u \mapsto *$$

Then $pj = 1_{\{*\}}$ trivially, and $jp \simeq 1_U$ by the homotopy $H(u,t) = u_0 + t(u - u_0)$. □

In general, we say that a space X is **contractible** if it is isomorphic to $\{*\}$ in the homotopy category. It is not hard to see that X is contractible if and only if 1_X is homotopic to a constant function $X \to X$ (indeed, any constant function).

Proposition 3.31. *The inclusion map $j : S^{n-1} \to \mathbb{R}^n - \{0\}$ is a homotopy equivalence.*

Proof. We define a homotopy inverse

$$p : \mathbb{R}^n - \{0\} \to S^{n-1}; \quad x \mapsto x/|x|.$$

Then $pj = 1_{S^{n-1}}$ and $jp \simeq 1_{\mathbb{R}^n - \{0\}}$ by the homotopy $H(x, t) = x|x|^{t-1}$. $\square$

One of the reasons for working in the homotopy category is that it enables proof strategies like the one in the following theorem.

Theorem 3.32. *Suppose there exists a functor $\mathrm{H} : \mathbf{hTop} \to \mathbf{Vect}$ such that $\mathrm{H}(\{*\}) = 0$ and $\mathrm{H}(S^{n-1}) \neq 0$. Then the Brouwer fixed-point theorem is true in n-dimensions.*

Repeated for emphasis: we can prove the Brouwer fixed-point theorem by finding a functor.

Proof. Suppose there exists a fixed-point free map $g : D^n \to D^n$. Then the map

$$f : S^{n-1} \to \mathbb{R}^n - \{0\}; \quad u \mapsto u - g(u)$$

is homotopic in $\mathbb{R}^n - \{0\}$ to the inclusion map $j(u) = u$ by the homotopy

$$J(u, t) = u - tg(u).$$

At the same time, f is homotopic in $\mathbb{R}^n - \{0\}$ to the constant map $k(u) = -g(0)$ by the homotopy

$$K(u, t) = tu - g(tu).$$

Thus $\mathrm{H}[j] = \mathrm{H}[f] = \mathrm{H}[k]$.

Now we show that $\mathrm{H}[j] \neq 0$. Let $p : \mathbb{R}^n - \{0\} \to S^{n-1}$ be the normalization map $x \mapsto x/|x|$, so that the diagram

$$\begin{tikzcd} S^{n-1} \arrow[r, "j"] \arrow[dr, "1_{S^{n-1}}"'] & R^n - \{0\} \arrow[d, "p"] \\ & S^{n-1} \end{tikzcd}$$

commutes. Applying the functor H, we get a commutative diagram

$$\begin{tikzcd} \mathrm{H}(S^{n-1}) \arrow[r, "\mathrm{H}[j]"] \arrow[dr, "1_{\mathrm{H}(S^{n-1})}"'] & \mathrm{H}(R^n - \{0\}) \arrow[d, "\mathrm{H}[p]"] \\ & \mathrm{H}(S^{n-1}) \end{tikzcd}$$

of vector spaces. The identity map on the diagonal is not zero because $\mathrm{H}(S^{n-1}) \neq 0$, and it follows that $\mathrm{H}[j] \neq 0$.

Now we show that $\mathrm{H}[k] = 0$. Since k is constant, it can be factored through the one-point space $\{*\}$ so that the diagram

$$\begin{tikzcd} S^{n-1} \arrow[r, "q"] \arrow[dr, "k"'] & \{*\} \arrow[d, "r"] \\ & S^{n-1} \end{tikzcd}$$

commutes. The horizontal map q is the unique map to $*$, and the vertical map r takes $*$ to the constant value taken by k. Applying the functor H, we get a commutative diagram

$$\begin{tikzcd} \mathrm{H}(S^{n-1}) \arrow[r, "\mathrm{H}[q]"] \arrow[dr, "\mathrm{H}[k]"'] & 0 \arrow[d, "\mathrm{H}[r]"] \\ & \mathrm{H}(S^{n-1}) \end{tikzcd}$$

which gives a factorization of $\mathrm{H}[k]$ through the zero vector space. It follows that $\mathrm{H}[k] = 0$.

Thus $\mathrm{H}[j] \neq \mathrm{H}[k]$, which contradicts our earlier calculation that $\mathrm{H}[j] = \mathrm{H}[k]$. □

Lecture 8

3.11. **Chain complexes.** We now describe two algebraic categories which are essential to our study of algebraic topology. We work with vector spaces over a field $\mathbb{F}$; there are analogous categories for abelian groups and for modules over a commutative ring.

A **chain complex** is a collection $V_* = (V_k, \partial_k)$ of vector spaces and linear maps organised in a sequence

$$V_0 \xleftarrow{\partial_0} V_1 \xleftarrow{\partial_1} V_2 \xleftarrow{\partial_2} V_3 \xleftarrow{\partial_3} \dots$$

where successive maps compose to zero; in other words $\partial^2 = 0$.

A **chain map** $T : V_* \to W_*$ is a collection of linear maps $T = (T_k : V_k \to W_k)$ for which the diagram

$$\begin{tikzcd} V_0 \arrow[d, "T_0"'] & V_1 \arrow[l, "\partial_0"'] \arrow[d, "T_1"'] & V_2 \arrow[l, "\partial_1"'] \arrow[d, "T_2"'] & V_3 \arrow[l, "\partial_2"'] \arrow[d, "T_3"'] & \dots \arrow[l, "\partial_3"'] \\ W_0 & W_1 \arrow[l, "\partial_0"] & W_2 \arrow[l, "\partial_1"] & W_3 \arrow[l, "\partial_2"] & \dots \arrow[l, "\partial_3"] \end{tikzcd}$$

commutes; in other words $T\partial = \partial T$.

A **chain homotopy** $K : S \Rightarrow T$ between two chain maps $S, T : V_* \to W_*$ is a collection of linear maps $(K_k : V_k \to W_{k+1})$

$$\begin{tikzcd} V_0 \arrow[d, shift right] \arrow[d, shift left] & V_1 \arrow[l] \arrow[d, shift right] \arrow[d, shift left] \arrow[dl, "K_0"'] & V_2 \arrow[l] \arrow[d, shift right] \arrow[d, shift left] & V_3 \arrow[l] \arrow[d, shift right] \arrow[d, shift left] & \dots \arrow[l] \\ W_0 & W_1 \arrow[l] & W_2 \arrow[l] & W_3 \arrow[l] & \dots \arrow[l] \end{tikzcd}$$

such that $\partial K + K\partial = T - S$.
(This means $\partial_k K_k + K_{k-1}\partial_{k-1} = T_k - S_k$ when k is positive, and also $\partial_0 K_0 = T_0 - S_0$.)

If there exists a chain homotopy $S \Rightarrow T$ we write $S \simeq T$ (or $S \simeq_K T$ if we wish to name the chain homotopy) and say that S, T are chain-homotopic. It is easy to verify that this is an equivalence relation. From this, we can define two categories:

- **CC** is the category whose objects are chain complexes and whose arrows are chain maps.
- **hCC** is the category whose objects are chain complexes and whose arrows are chain-homotopy equivalence classes of chain maps.

We need to show that composition is well-defined in **hCC**. Indeed if we have maps

$$\begin{tikzcd} V_* \arrow[r, shift left, "S"] \arrow[r, shift right, "S'"'] & W_* \arrow[r, "T"] & X_* \end{tikzcd}$$

with $S \simeq_K S'$ then $TS \simeq_{TK} TS'$, and if we have maps

$$\begin{tikzcd} V_* \arrow[r, "S"] & W_* \arrow[r, shift left, "T"] \arrow[r, shift right, "T'"'] & X_* \end{tikzcd}$$

with $T \simeq_L T'$ then $TS \simeq_{LS} T'S$.

Clearly there is a functor $\mathbf{CC} \to \mathbf{hCC}$ which keeps objects unchanged and which takes each chain map to its equivalence class. It is similar in spirit to a forgetful functor, in that it 'forgets' the distinction between chain maps that are chain-homotopic to each other.

Definition 3.33. For $k \geq 0$, let $\mathrm{C}_k, \mathrm{Z}_k, \mathrm{B}_k$ be the functors $\mathbf{CC} \to \mathbf{Vect}$ defined by

$$\begin{aligned} &\mathrm{C}_k(V_*) = V_k, && \mathrm{Z}_k(V_*) = \ker(\partial_{k-1}), && \mathrm{B}_k(V_*) = \mathrm{im}(\partial_k), \\ &\mathrm{C}_k[T] \;= T_k, && \mathrm{Z}_k[T] \;= T_k|_{\ker \partial_{k-1}}, && \mathrm{B}_k[T] \;= T_k|_{\mathrm{im}\, \partial_k}. \end{aligned}$$

It is easily shown (homework **30**) that these are well-defined functors.

Definition 3.34. For $k \geq 0$, let H_k be the functor $\mathbf{CC} \to \mathbf{Vect}$ defined by

$$\mathrm{H}_k(V_*) = \frac{\ker(\partial_{k-1})}{\mathrm{im}(\partial_k)}, \qquad \mathrm{H}_k[T] \colon [\gamma] \longmapsto [T_k \gamma].$$

It is easily shown (homework **35**) that this is a well-defined functor, and moreover it descends to a functor $\mathrm{H}_k : \mathbf{hCC} \longrightarrow \mathbf{Vect}$.

Homology theory in algebraic topology takes advantage of the following diagram:

$$\begin{tikzcd} \mathbf{Top} \arrow[r] \arrow[d] & \mathbf{CC} \arrow[d] \arrow[dr] & \\ \mathbf{hTop} \arrow[r] & \mathbf{hCC} \arrow[r, "\mathrm{H}_k"] & \mathbf{Vect} \end{tikzcd}$$

The trick is to construct the horizontal functors $\mathbf{Top} \longrightarrow \mathbf{CC}$ and $\mathbf{hTop} \longrightarrow \mathbf{hCC}$ and then let H_k work its magic. Everything that happens after the chain complex is constructed is called *homological algebra*.

We finish this section with a short discussion of **exact sequences**. A three-term sequence of vector spaces and linear maps

$$U \xrightarrow{f} V \xrightarrow{g} W$$

is **exact** if $\ker g = \operatorname{im} f$. We may say that it is 'exact at V' to emphasize the vector space whose two subspaces are being compared. A longer sequence such as

$$U_1 \xrightarrow{f_1} U_2 \xrightarrow{f_2} U_3 \xrightarrow{f_3} \dots \xrightarrow{f_{n-1}} U_n$$

may be exact at any of the individual terms $U_2, U_3, \dots, U_{n-1}$, or at all of them, in which case it is simply said to be exact.

Example 3.35. A chain complex V_* is exact at V_k iff its homology $\mathrm{H}_k(V_*)$ is zero.

Example 3.36. The sequence $U \xrightarrow{f} V \longrightarrow 0$ is exact iff f is surjective.

Example 3.37. The sequence $0 \longrightarrow U \xrightarrow{f} V$ is exact iff f is injective.

Example 3.38. The sequence $0 \longrightarrow U \xrightarrow{f} V \longrightarrow 0$ is exact iff f is an isomorphism.

We finish with a calculation that will prove useful. Suppose $U \xrightarrow{f} V \xrightarrow{g} W$ is exact. Then, starting with the rank–nullity formula, we get

$$\begin{aligned} \dim(V) &= \dim(\ker g) + \dim(\operatorname{im} g) \\ &= \dim(\operatorname{im} f) + \dim(\operatorname{im} g) = \operatorname{rank}(f) + \operatorname{rank}(g). \end{aligned}$$

It is worth recalling the proof of the rank–nullity formula: take a basis $v_1, \dots, v_k$ for $\ker(g)$; extend it to a basis $v_1, \dots, v_k, v_{k+1}, \dots, v_n$ for V; show that the vectors $g(v_{k+1}), \dots, g(v_n)$ are linearly independent in W and therefore constitute a basis of $\operatorname{im}(g)$.

In homework **44** we consider what happens when $gf = 0$ and the sequence is not necessarily exact.

4. Simplicial Complexes

Simplicial complexes are discrete objects that may be used to study continuous spaces. We begin by studying simplicial complexes in their own right, and then we establish the connection with continuous topology.

4.1. Abstract simplicial complexes.

4.1. Abstract simplicial complexes. A simplicial complex is a nonempty set X of finite sets which is closed under taking subsets:

$$\sigma \in X \text{ and } \tau \subseteq \sigma \text{ implies } \tau \in X$$

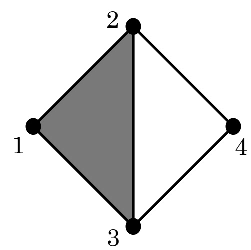
Example. The set

$$M = \{\{2,4\}, \{3,4\}, \{1,2,3\}\}$$

is not a simplicial complex, but its closure under taking subsets

$$X = \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{3,4\}, \{1,2,3\}\}$$

is a simplicial complex. This can be drawn as follows:



Remark. Here we can recover M as the set of **maximal** elements of X: those elements that are not subsets of some larger element of X.

There is plenty of terminology.

- If the elements of X are subsets of a set V, we say that X is a simplicial complex on V.
- The smallest such V is given by considering the singleton elements of X. In the example above, the singletons are $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$ so we can take $V = \{1,2,3,4\}$. This is the set of **vertices** of X, which we may write $V = V(X)$.
- Each $\sigma \in X$ is called a **simplex** (plural: **simplices**) of X.
- We call σ a k**-simplex** (or k**-cell**) and write $\dim(\sigma) = k$, if its cardinality is $k+1$.
- The **dimension** of the complex is $\dim(X) = \max(\dim(\sigma) \mid \sigma \in X)$.

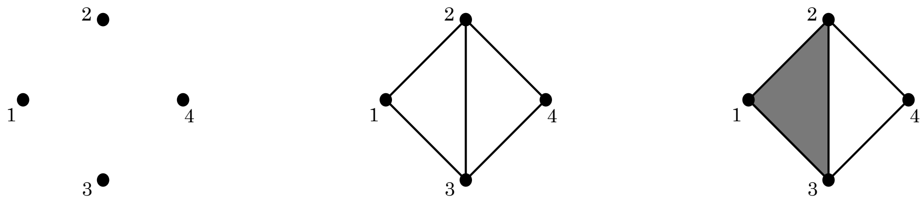
In the example above, we have

$$\begin{array}{ll} \text{0-cells:} & \{1\}, \{2\}, \{3\}, \{4\} \\ \text{1-cells:} & \{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{3,4\} \\ \text{2-cells:} & \{1,2,3\} \end{array}$$

and we can think of $\emptyset$ as a (-1)-cell. The complex X is 2-dimensional.

- The **k-skeleton** $X^{(k)}$ consists of all cells of X of dimension at most k.

Here are the skeleta $X^{(0)}$, $X^{(1)}$, and $X^{(2)} = X$ for the example above.



- If $\tau \subseteq \sigma$ we say that τ is a **face** of σ, or σ is a **coface** of τ. We write $\tau \leq \sigma$.
- If $\tau \leq \sigma$ and $\dim(\tau) = \dim(\sigma) - 1$, we say that τ is a **facet** of σ.

Example. The simplex $\{1, 2, 3\}$ has 8 faces (6 'proper' faces, as well as $\varnothing$ and the cell itself). Three of these are facets: $\{2, 3\}$, $\{1, 3\}$, $\{1, 2\}$.

Example. An n-simplex $\sigma = \{a_0, a_1, \dots, a_n\}$ has $n + 1$ facets:

$$\sigma_k = \{a_0, \dots, \hat{a}_k, \dots, a_n\} = \{a_0, \dots, a_{k-1}, a_{k+1}, \dots, a_n\}$$

(Here the caret $\hat{}$ indicates that a_k is to be omitted from the list.)

4.2. **Examples and constructions.** Here are some important constructions.

The empty complex. This is the complex

$$E = \{\varnothing\}$$

whose only simplex is the (-1)-cell $\varnothing$.

The standard n-simplex.

$$B_n = \{\text{all subsets of } \{0, 1, \dots, n\}\}$$

This is the simplicial equivalent of an n-dimensional ball.

The boundary of the standard n-simplex.

$$\begin{aligned} \partial B_n &= B_n^{(n-1)} \\ &= \{\text{all subsets of } \{0, 1, \dots, n\} \text{ of cardinality at most } n\} \end{aligned}$$

This is the simplicial equivalent of a sphere of dimension $n - 1$.

Cones. A complex C is a **cone** if there exists a vertex $*$ such that

$$\sigma \in C \text{ implies } \sigma \cup \{*\} \in C.$$

The vertex $*$ is called a **cone point** for C.

Example. The complex

$$\{\emptyset,\ \{1\}, \{2\},\ \{3\},\ \{1,2\},\ \{2,3\}\}$$

is a cone with respect to the vertex 2, whereas

$$\{\emptyset,\ \{1\}, \{2\},\ \{3\},\ \{1,2\},\ \{1,3\},\ \{2,3\}\}$$

is not a cone with respect to any vertex.

Remark. Cones are the simplicial equivalent of star-shaped domains.

The cone on a complex. Every simplicial complex X is contained in a cone CX, constructed as follows: introduce a new vertex $*$ not already in $V = V(X)$, and include a new cell $\sigma \cup \{*\}$ for every $\sigma \in X$. Formally:

$$CX = \{\sigma,\ \sigma \cup \{*\} \mid \sigma \in X\} = \{\sigma \subseteq V \cup \{*\} \mid \sigma \cap V \in X\}$$

Notice that CX contains exactly twice as many simplices as X (counting the empty cell).

Example. If

$$X = \{\emptyset, \{1\},\ \{2\},\ \{3\},\ \{1,2\}\}$$

then

$$CX = \{\emptyset,\ \{1\},\ \{2\},\ \{3\},\ \{1,2\},\ \{*\},\ \{*,1\},\ \{*,2\},\ \{*,3\},\ \{*,1,2\}\}.$$

Remark. If we iterate the cone construction, for instance to construct CCX, we need a new symbol for each new cone point.

The relative cone on a subcomplex. Sometimes we 'cone off' just part of a simplicial complex. Let X, Y be complexes with $Y \subseteq X$. Then

$$C(X,Y) = X \cup CY$$

is the **relative cone** of the inclusion map $Y \to X$.

Example. If

$$X = \{\emptyset,\ \{1\},\ \{2\},\ \{1,2\}\}, \qquad Y = \{\emptyset,\ \{1\},\ \{2\}\}$$

then

$$C(X,Y) = \{\emptyset,\ \{1\},\ \{2\},\ \{1,2\},\ \{*\},\ \{*,1\},\ \{*,2\}\}.$$

The suspension of a complex. This is the union of two cones on X with respect to points n and s (which are assumed not to belong to $V = V(X)$). Formally:

$$\Sigma X = \{\sigma,\, \sigma \cup \{n\},\, \sigma \cup \{s\} \mid \sigma \in X\}$$

Notice that ΣX has thrice as many simplices as X. We may think of X as being 'suspended' as the equator between a north pole n and a south pole s.

Example. If

$$X = \{\emptyset, \{1\}, \{2\}\}$$

then

$$\Sigma X = \{\emptyset,\, \{1\},\, \{2\},\, \{n\},\, \{n, 1\},\, \{n, 2\},\, \{s\},\, \{s, 1\},\, \{s, 2\}\}.$$

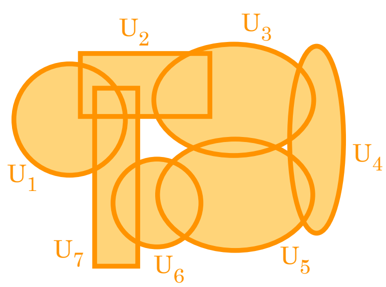
An alternative definition is $\Sigma X = C(CX, X)$.

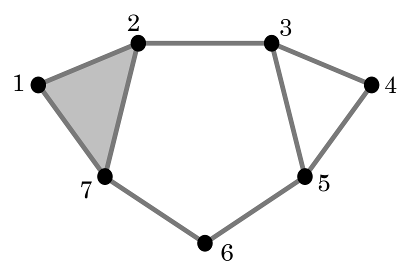
The nerve of a covering. Let $U_1, \dots, U_n$ be sets, and write $\mathcal{U} = (U_1, \dots, U_n)$. The **nerve** of $\mathcal{U}$ is a simplicial complex on the vertex set $\{1, 2, \dots, n\}$ defined by the following condition:

$$(\cap) \qquad \sigma \in \mathrm{Nerve}(\mathcal{U}) \;\Leftrightarrow\; \bigcap_{k \in \sigma} U_k \neq \emptyset$$

Remark. Quite commonly, the (U_k) occur as subsets covering a larger set $U = U_1 \cup \dots \cup U_n$.

Example. A collection of sets in the plane, and the nerve of this collection:





Remark. The construction is equally valid for an infinite family of sets $\mathcal{U} = (U_k \mid k \in K)$. $\mathrm{Nerve}(\mathcal{U})$ consists of the finite subsets $\sigma \subset K$ that satisfy condition $(\cap)$.

The Vietoris–Rips complex. Let X be a metric space and let $R \geq 0$. The **Vietoris–Rips complex** with diameter R is a simplicial complex on X defined by the condition:

$$\sigma = \{x_0, x_1, \dots, x_n\} \in \mathrm{Rips}(X, R) \;\Leftrightarrow\; d(x_i, x_j) \leq R \text{ for all } 0 \leq i, j \leq n$$

Example. Let X be a finite collection of robotic sensors located in the plane. The metric is Euclidean distance. The coverage criterion (see Theorem 2.18) makes use of the 2-skeleton $\mathrm{Rips}(X, R)^{(2)}$ of the Vietoris–Rips complex. The 1-cells are used to define the fence cycle; the 2-cells are used to determine if the fence cycle is a boundary; if the test is successful then unused 0-cells can be switched off to conserve power.

Chains and antichains. Let P be a partially-ordered set (poset). There are two natural complexes on P, determined by the partial-order structure.

The complex of chains:[2]

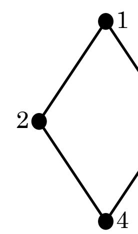
$$\begin{aligned}\sigma \in \mathrm{Ch}(P) &\Leftrightarrow \sigma \text{ is a finite } \textbf{chain} \text{ in } P\\ &\Leftrightarrow \sigma \text{ is a totally ordered finite subset of } P\end{aligned}$$

The complex of antichains:

$$\begin{aligned}\sigma \in \mathrm{Anti}(P) &\Leftrightarrow \sigma \text{ is a finite } \textbf{antichain} \text{ in } P\\ &\Leftrightarrow \sigma \text{ is a totally unordered finite subset of } P\end{aligned}$$

If σ is a chain (resp. antichain) and $\tau \subseteq \sigma$, then τ is also a chain (resp. antichain), so these definitions yield simplicial complexes.

Example. Let P be the poset with the following Hasse diagram:



Thus 1 is the largest element, 4 is the smallest, and $2, 3$ are incomparable with each other.

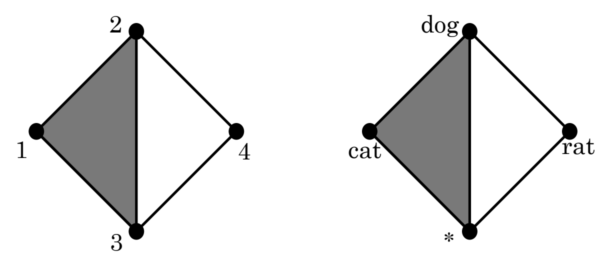
The maximal chains of P are $\{1,2,4\}$ and $\{1,3,4\}$ so the complex of chains consists of all subsets of these two sets:

$$\mathrm{Ch}(P) = \{\varnothing,\ \{1\},\ \{2\},\ \{3\},\ \{4\},\ \{1,2\},\ \{1,3\},\ \{1,4\},\ \{2,4\},\ \{3,4\},\ \{1,2,4\},\ \{1,3,4\}\}$$

The maximal antichains are $\{1\}$, $\{4\}$, and $\{2,3\}$ so the complex of antichains is

$$\mathrm{Anti}(P) = \{\varnothing,\ \{1\},\ \{2\},\ \{3\},\ \{4\},\ \{2,3\}\}$$

4.3. **Elementary equivalence.** There is a clear notion of isomorphism between simplicial complexes. For instance, the complexes



are isomorphic under the bijection

$$1 \leftrightarrow \text{cat},\ 2 \leftrightarrow \text{dog},\ 3 \leftrightarrow *,\ 4 \leftrightarrow \text{rat}$$

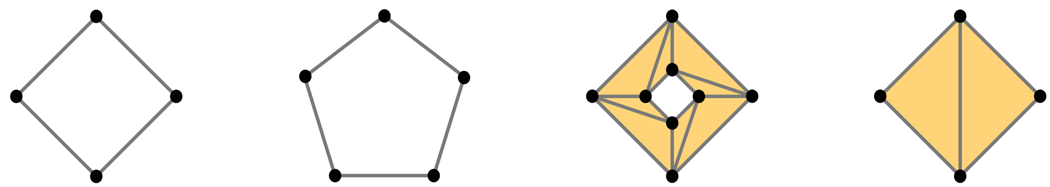
[2]It is an accident that this sounds like 'chain complex'.

between their vertex sets; and also under the bijection

$$1 \leftrightarrow \text{cat},\ 2 \leftrightarrow *,\ 3 \leftrightarrow \text{dog},\ 4 \leftrightarrow \text{rat}.$$

We write $X \cong Y$ to indicate that the complexes X, Y are isomorphic.

Isomorphism is rather stringent, and has nothing to say about our intuition that the first three of the following complexes have a certain similarity not shared by the fourth:



According to the isomorphism relation, all four complexes are different. We wish to define a weaker relation, under which the first three of these complexes are equivalent to each other and are not obviously equivalent to the fourth.

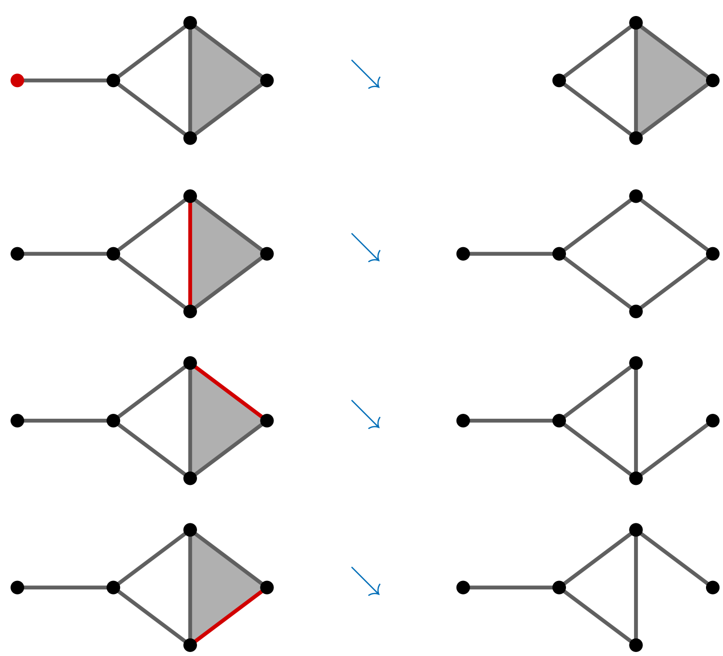
Let X be a simplicial complex and suppose $\tau \in X$ is a non-empty simplex with exactly one proper coface $\sigma \in X$ ('proper' meaning 'not equal to τ itself'). Then,

$$Y = X - \{\tau, \sigma\}$$

is a simplicial complex. We call Y an **elementary collapse** of X, and write $X \searrow Y$. Reciprocally, we may call X an **elementary expansion** of Y, and write $Y \nearrow X$.

The idea is that elementary collapses are topology-preserving, in some sense to be made precise later.

Example. The simplicial complex on the left has four possible elementary collapses:

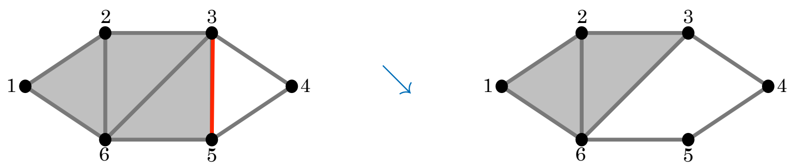


Notice, for example, that the number of holes (one) remains unchanged in each collapse.

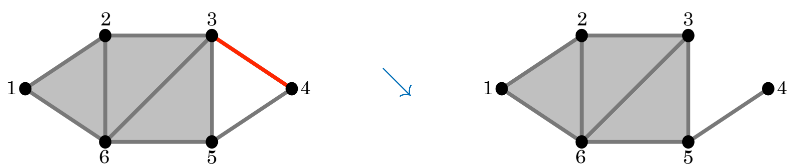
Example. What is special about having exactly one proper coface? In each of the following situations, we remove a 1-simplex and all of its cofaces.

The edge $\{3, 6\}$ has two proper cofaces:

The edge $\{3, 5\}$ has one proper coface:



The edge $\{3, 4\}$ has no proper cofaces:



The second example is the only one where the essential topological structure of the complex is not changed by the operation. In the first example a new hole is created, whereas in the third example an existing loop is broken.

Definition 4.1. Two simplicial complexes X, Y are **Whitehead equivalent** if there is a sequence $X_0, X_1, \ldots, X_n$ of simplicial complexes such that

- $X \cong X_0$
- $Y \cong X_n$
- For each $k = 1, \ldots, n$, either $X_{k-1} \nearrow X_k$ or $X_{k-1} \searrow X_k$.

We write $X \simeq Y$ to indicate that X, Y are Whitehead equivalent. It is easy to see that this is an equivalence relation. In fact, it is the smallest (i.e. most discriminating) equivalence relation such that isomorphic complexes are equivalent and where $X \searrow Y$ implies $X \simeq Y$.

The usual way to show that two complexes are Whitehead equivalent is to find an explicit sequence of collapses and expansions relating them.

Theorem 4.2. *For $n \geq 3$, let C_n denote the cycle of length n; that is, the simplicial complex on $\{1, 2, \ldots, n\}$ whose maximal simplices are*

$$\{1,2\},\ \{2,3\},\ \ldots,\ \{n-1,n\},\ \{n,1\}.$$

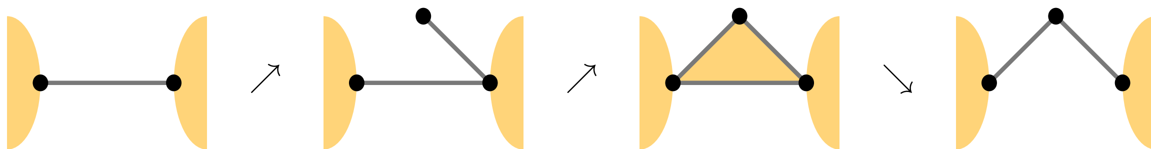
Then $C_m \simeq C_n$ for every $m, n \geq 3$.

Lemma 4.3 (edge splitting)**.** *Let X be a simplicial complex, and suppose that $\{a,b\} \in X$ has no proper cofaces. Let Y be the complex obtained by splitting the edge in two along a new vertex $m \notin V(X)$:*

$$Y = X \cup \{\{m\},\ \{a,m\},\ \{m,b\}\} - \{\{a,b\}\}$$

Then $Y \simeq X$.

Proof. In pictures, this is



Formally we define

$$\begin{aligned} X_0 &= X \\ X_1 &= X_0 \cup \{\{m\},\ \{a,m\}\} \\ X_2 &= X_1 \cup \{\{b,m\},\ \{a,b,m\}\} \\ X_3 &= X_2 - \{\{a,b\},\ \{a,b,m\}\} \end{aligned}$$

and then $X = X_0$, $Y = X_3$, and $X_0 \nearrow X_1 \nearrow X_2 \searrow X_3$ as required. □

Proof of Theorem 4.2. Since C_{n+1} is obtained from C_n by splitting the edge $\{n,1\}$ in two along a new vertex $n+1$, the result follows from the lemma. □

A simplicial complex is **collapsible** if it is Whitehead equivalent to the 1-point complex B_0.

Theorem 4.4. *Every finite cone is collapsible.*

Proof. Let CX be a cone with cone point $*$. We use induction on the number of cells of X. If X is the empty complex then we are done, because

$$CE = \{\emptyset,\ \{*\}\}$$

is isomorphic to B_0. Otherwise let τ be a maximal simplex of X, so that $X - \tau$ is also a simplicial complex. Then τ has exactly one coface in CX, namely $\sigma = \tau \cup \{*\}$. Then

$$CX \searrow CX - \{\sigma, \tau\} = C(X - \tau)$$

which is collapsible by the inductive hypothesis. □

There are two natural tasks that arise when considering any equivalence relation:

(1) Show that $X \simeq Y$. (2) Show that $X \not\simeq Y$.

For Whitehead equivalence, we can answer (1) by exhibiting a suitable sequence of collapses and expansions. It is not so easy to answer (2) by brute force, because we have to rule out all possible sequences. There is no bound on how long such a sequence might be, and the complexes X_k that occur on the way may well be enormous. We need invariants.

Proposition 4.5 (connected components)**.** *Let* $\mathrm{b}_0(X)$ *denote the number of connected components of* X*. This is defined as the number of equivalence classes of vertices of* X *under the relation*

$$v \sim w \;\Leftrightarrow\; \textit{there is a path of edges from } v \textit{ to } w.$$

Then b_0 *is Whitehead invariant.*

(Later we will define b_0 homologically.)

Proof. We show that b_0 is unchanged by an elementary collapse. Consider $Y = X - \{\sigma, \tau\}$ where σ is the only proper coface of τ.

- If $\dim(\tau) = 0$, we may write $\tau = \{a\}$, $\sigma = \{a, b\}$. Then X has the same connected components as Y, except that the component containing b has an extra vertex a.
- If $\dim(\tau) = 1$, we may write $\tau = \{a, b\}$, $\sigma = \{a, b, c\}$. Then X has the same vertices as Y but one fewer edge $\{a, b\}$. The equivalence relation is the same, because the 'lost' relation $a \sim b$ is recovered in Y by combining $a \sim c$ and $c \sim b$.
- If $\dim(\tau) \geq 2$, then the set of vertices and the equivalence relation are unchanged.

In all three cases, $\mathrm{b}_0(X) = \mathrm{b}_0(Y)$. □

Example 4.6. The m-point space and the n-point space

$$\{\varnothing, \{1\}, \{2\}, \ldots, \{m\}\} \quad \text{and} \quad \{\varnothing, \{1\}, \{2\}, \ldots, \{n\}\}$$

are not Whitehead equivalent if $m \neq n$.

Proposition 4.7. *The parity of the total number of cells*

$$\mathrm{par}(X) = \sum_{k=0}^{\dim(X)} \#\{k\text{-cells of } X\} \pmod 2$$

is Whitehead invariant.

(We choose not to count the (-1)-cell $\varnothing$.)

Proof. Exactly two cells are removed at each elementary collapse. □

Example 4.8. The hollow triangle and the point

$$\partial B_2 = \{\varnothing,\ \{0\},\ \{1\},\ \{2\},\ \{0,1\},\ \{0,2\},\ \{1,2\}\} \quad \text{and} \quad B_0 = \{\varnothing,\ \{0\}\}$$

are not Whitehead equivalent because $\mathrm{par}(\partial B_2) = 0$ whereas $\mathrm{par}(B_0) = 1$.

Proposition 4.9. *The Euler characteristic*

$$\chi(X) = \sum_{k=0}^{\dim(X)} (-1)^k \#\{k\textit{-cells of } X\}$$

is Whitehead invariant.

Proof. Exactly two cells of adjacent dimensions are removed at each elementary collapse. Their contributions to the Euler characteristic cancel. □

Example 4.10. The hollow triangle ∂B_2 and the hollow tetrahedron ∂B_2 are not Whitehead equivalent because $\chi(\partial B_2) = 3 - 3 = 0$ whereas $\chi(\partial B_3) = 4 - 6 + 4 = 2$.

Remark. Notice that the Euler characteristic is an integer lift of the parity invariant and therefore a more discriminating invariant. We will shortly see that the Euler characteristic itself can be 'lifted' to an invariant sequence of nonnegative integers $\mathrm{b} = (\mathrm{b}_0, \mathrm{b}_1, \mathrm{b}_2, \dots)$ that is more discriminating still.

Lecture 9

4.4. **Simplicial homology.** Let X be a simplicial complex. In this section we define the simplicial homology of X. As usual we will work with a field $\mathbb{F}$. We will construct a sequence of vector spaces $\mathrm{H}_k(X) = \mathrm{H}_k(X; \mathbb{F})$.

Note. In general we can work over any commutative ring, and obtain modules over that ring.

Definition 4.11. We define $\mathrm{C}_k(X)$, the space of k-chains in X, as follows:

- There is a generator $[a_0, a_1, \dots, a_k]$ for any $(k+1)$-tuple of vertices for which the set $\{a_0, a_1, \dots, a_k\}$ belongs to X.
- There is a relation $[a_0, a_1, \dots, a_k] = 0$ whenever two of the a_i are equal. Briefly, we express this relation as
$$[\dots, a, \dots, a, \dots] = 0.$$
- There is a relation $[a_0, a_1, \dots, a_k] = -[a'_0, a'_1, \dots, a'_k]$ whenever the lists (a_i) and (a'_i) differ by a single transposition. Briefly, we express this relation as
$$[\dots, a, \dots, b, \dots] = -[\dots, b, \dots, a, \dots],$$
the ellipses representing sequences that are the same on both sides of the equation.

Note. The 'repeated term' relation follows from the 'transposition' relation except when $\mathrm{char}(\mathbb{F}) = 2$. We include it to cover that case also. It turns out to be more convenient to allow repeated vertices and set those generators to be zero, than to ban repeated vertices and have to work carefully to avoid getting them by accident. It does mean that we have to check one more thing every time we define a linear map, but this is usually quite easy.

For an arbitrary permutation π of $\{0, 1, \dots, k\}$ we obtain
$$[a_0, a_1, \dots, a_k] = (-1)^\pi [a_{\pi(0)}, a_{\pi(1)}, \dots, a_{\pi(k)}]$$
by expressing π as a product of transpositions. It follows that
$$\dim \mathrm{C}_k(X) = \#\{k\text{-simplices of } X\}.$$

Definition 4.12. The boundary map $\partial_k : \mathrm{C}_{k+1}(X) \to \mathrm{C}_k(X)$ is defined on generators by

$$\partial_k[a_0, \dots, a_{k+1}] = \sum_{i=0}^{k+1} (-1)^i [\dots, \hat{a}_i, \dots]$$

where $[\dots, \hat{a}_i, \dots]$ is an abbreviation for $[a_0, \dots, a_{i-1}, a_{i+1}, \dots, a_{k+1}]$, which is the result of deleting a_i from $[a_0, \dots, a_{k+1}]$.

We confirm that the definition of ∂ respects both relations:

Consider a generator with repeated vertices in positions i, j with $i < j$. All but two terms of its boundary are immediately zero, by the repeated-term relation. We then have

$$\begin{aligned} \partial[\dots, a, \dots, a, \dots] &= (-1)^i[\dots, \dots, a, \dots] + (-1)^j[\dots, a, \dots, \dots] \\ &= (-1)^i[\dots, \dots, a, \dots] + (-1)^j(-1)^{j-i-1}[\dots, \dots, a, \dots] = 0 \end{aligned}$$

since we need $j - i - 1$ transpositions to make the second term look like the first term.

Consider two generators which differ by a transposition of positions i, j with $i < j$. Then

$$\partial[\dots, a, \dots, b, \dots] = (-1)^i[\dots, \dots, b, \dots] + (-1)^j[\dots, a, \dots, \dots] + (k-1 \text{ other terms})$$

$$\partial[\dots, b, \dots, a, \dots] = (-1)^j[\dots, b, \dots, \dots] + (-1)^i[\dots, \dots, a, \dots] + (k-1 \text{ other terms})$$

The two sets of "$k-1$ other terms" are immediately equal and opposite, by the transposition relation. The same is true for the first two pairs of terms by the following calculations:

$$(-1)^j[\dots, b, \dots, \dots] = (-1)^j(-1)^{j-i-1}[\dots, \dots, b, \dots] = -(-1)^i[\dots, \dots, b, \dots]$$

$$(-1)^i[\dots, \dots, a, \dots] = (-1)^i(-1)^{j-i-1}[\dots, a, \dots, \dots] = -(-1)^j[\dots, a, \dots, \dots]$$

Thus

$$\partial[\dots, b, \dots, a, \dots] = -\partial[\dots, a, \dots, b, \dots].$$

These checks are generally quite easy to carry out by inspection, but they can be awkward to write down formally. We will sometimes give the details and sometimes not.

Proposition 4.13. *The boundary maps satisfy $\partial^2 = 0$, that is $\partial_k \partial_{k+1} = 0$ for all $k \geq 0$.*

Proof. For any generator $[a_0, a_1, \dots, a_{k+2}]$ the double boundary is a linear combination of terms of the form $[\dots, \hat{a}_i, \dots, \hat{a}_j, \dots]$. Each such term occurs in exactly two ways

$$[\dots, a_i, \dots, a_j, \dots] \xrightarrow{(-1)^i} [\dots, \hat{a}_i, \dots, a_j, \dots] \xrightarrow{(-1)^{j-1}} [\dots, \hat{a}_i, \dots, \hat{a}_j, \dots]$$

$$[\dots, a_i, \dots, a_j, \dots] \xrightarrow{(-1)^j} [\dots, a_i, \dots, \hat{a}_j, \dots] \xrightarrow{(-1)^i} [\dots, \hat{a}_i, \dots, \hat{a}_j, \dots]$$

which sum to an overall coefficient of $(-1)^i(-1)^{j-1} + (-1)^j(-1)^i = 0$. $\square$

Thus we have a chain complex, the **simplicial chain complex** of X.

Definition 4.14. The **simplicial homology** of X is defined in each dimension $k \geq 0$ to be the vector space

$$\mathrm{H}_k(X) = \mathrm{H}_k(\mathrm{C}_*(X))$$

obtained by applying the homology functor H_k to the simplicial chain complex of X.

Definition 4.15. The **Betti vector** of X is the sequence

$$\mathrm{b} = (\mathrm{b}_0, \mathrm{b}_1, \mathrm{b}_2, \dots)$$

of **Betti numbers**

$$\mathrm{b}_k = \mathrm{b}_k(X) = \dim \mathrm{H}_k(X).$$

Note. If $k \geq \dim(X)$ then $\mathrm{C}_k(X) = 0$ and therefore $\mathrm{H}_k(X) = 0$ and $\mathrm{b}_k(X) = 0$.

Example 4.16. The Betti vector of the empty complex is $(0, 0, 0, \dots)$.

Example 4.17. Let E_n be a simplicial complex with n vertices and no higher-dimensional simplices. It's simplicial chain complex is isomorphic to

$$\mathbb{F}^n \longleftarrow 0 \longleftarrow 0 \longleftarrow 0 \longleftarrow \dots$$

so $\mathrm{H}_0(E_n) \cong \mathbb{F}^n$ and $\mathrm{H}_k(E_n) = 0$ for $k > 0$. Thus $\mathrm{b}(E_n) = (n, 0, 0, 0, \dots)$.

A familiar argument (using mass functions and paths) gives the next result:

Proposition 4.18. *The zeroth Betti number* $\mathrm{b}_0(X)$ *is equal to the number of connected components of* X*, as defined in Propostion 4.5.* □

Proposition 4.19. *The cycle* C_n *with vertices* $1, 2, \dots, n$ *and edges*

$$\{1,2\},\ \{2,3\},\ \dots,\ \{n-1,n\},\ \{n,1\}$$

has Betti vector $\mathrm{b}(C_n) = (1, 1, 0, 0, 0, \dots)$.

Proof. There is one connected component, so $\mathrm{b}_0 = 1$; and $\mathrm{b}_k = 0$ for $k \geq 2$ trivially. Finally, each 1-cycle must be a scalar multiple of

$$[1,2] + [2,3] + \dots + [n-1,n] + [n,1]$$

because the cycle condition implies that adjacent edges occur with the same coefficient. Since there are no 2-simplices we have $\mathrm{B}_1 = 0$ and therefore $\mathrm{b}_1 = \dim \mathrm{H}_1 = \dim \mathrm{Z}_1 = 1$. □

Proposition 4.20. *The Betti vector of a cone is* $(1, 0, 0, 0, \dots)$.

(In other words, a cone has the same homology as a point. A cone is a 'homology point'.)

Proof. Every vertex is connected to the cone point $*$, so $\mathrm{b}_0 = 1$. Now consider the chain homotopy K defined by the maps

$$K_k : \mathrm{C}_k \to \mathrm{C}_{k+1};\ [a_0, \dots, a_k] \mapsto [*, a_0, \dots, a_k]$$

It is immediate that this definition is consistent with the repeated-term and transposition relations, so this is a well-defined map for all k. Now let $k \geq 1$. One calculates easily that

$$(\partial K + K\partial)[a_0, \dots, a_k] = [a_0, \dots, a_k]$$

for every generator $[a_0, \dots, a_k]$ of C_k, so if γ is a k-cycle we have

$$\gamma = (\partial K + K\partial)\gamma = \partial K \gamma.$$

Thus $\mathrm{Z}_k = \mathrm{B}_k$ and so $\mathrm{H}_k = 0$. □

We now contrive to make the homology of a point (or a cone) even more simple than it is.

Definition 4.21. The **augmented chain complex** of a simplicial complex X is obtained from the simplicial chain complex by placing a copy of the field $\mathbb{F}$ in dimension -1 and defining $\partial_{-1} = \mu$, the mass function:

$$\mathbb{F} \xleftarrow{\mu} \mathrm{C}_0(X) \xleftarrow{\partial_0} \mathrm{C}_1(X) \xleftarrow{\partial_1} \mathrm{C}_2(X) \xleftarrow{\partial_2} \dots$$

The **reduced homology** of X is defined to be the homology of the augmented chain complex. Thus we have a vector space, written $\tilde{\mathrm{H}}_k(X)$, for every $k \geq -1$. We write $\tilde{\mathrm{b}}_k(X)$ for the corresponding 'reduced' Betti numbers.

The empty simplicial complex E has augmented chain complex

$$\mathbb{F} \longleftarrow 0 \longleftarrow 0 \longleftarrow 0 \longleftarrow \dots$$

and therefore $\tilde{\mathrm{b}}_{-1}(E) = 1$ and $\tilde{\mathrm{b}}_k(E) = 0$ otherwise. Now suppose that X is not the empty complex. Then:

- $\tilde{\mathrm{b}}_{-1}(X) = 0$, since the map μ is surjective.
- $\tilde{\mathrm{b}}_0(X) = \mathrm{b}_0(X) - 1$, because the space of 0-cycles is one dimension smaller in reduced homology, while the space of 0-boundaries remains unchanged.
- $\tilde{\mathrm{b}}_k(X) = \mathrm{b}_k(X)$ otherwise, since k-chains and k-boundaries are unchanged for $k \geq 1$.

Example 4.22. The reduced Betti numbers of a cone are all zero.

Plan. Over the next few sections, one of our goals is to show that simplicial homology is Whitehead invariant. This can be done by directly comparing the chain complexes for X and Y whenever $X \searrow Y$. The two chain complexes can be shown to be chain-homotopy equivalent. This is not too difficult, but we will take a different approach. We will instead spend some time developing some of the classic machinery of homology theory. We will show how simplicial homology is functorial, we will define the relative homology of a pair of simplicial complexes, and we will construct the long exact sequence of such a pair. When this is done, it will be very easy to show that simplicial homology is preserved by elementary collapses.

4.5. **Simplicial maps.** Let X, Y be simplicial complexes on the respective vertex sets $V(X), V(Y)$.

Definition 4.23. A **simplicial map** from X to Y is a function $f : V_X \to V_Y$ which carries simplices to simplices. In other words,

$$\sigma = \{a_0, a_1, \dots, a_k\} \in X \quad \Rightarrow \quad f(\sigma) = \{f(a_0), f(a_1), \dots, f(a_k)\} \in Y.$$

Note that $f(\sigma)$ need not have the same cardinality as σ, because f need not be injective.

Thus we get a category **Simp** whose objects are simplicial complexes and whose arrows are simplicial maps. We generally write $f : X \to Y$ for an arrow in this category, with the understanding that there is an underlying function $f : V(X) \to V(Y)$. Composition and identities are defined in terms of the underlying map.

We can now make simplicial homology *functorial*.

Definition 4.24. Let $X \xrightarrow{f} Y$ be an arrow in **Simp**. We define a chain map

$$\mathrm{C}_*(X) \xrightarrow{\mathrm{C}_*[f]} \mathrm{C}_*(Y)$$

by the formula

$$\mathrm{C}_k[f] : [a_0, a_1, \ldots, a_k] \longmapsto [f(a_0), f(a_1), \ldots, f(a_k)]$$

in each dimension k.

(It is immediate that this definition is consistent with the repeated-term and transposition relations, and that the linear maps $\mathrm{C}_k[f]$ together constitute a chain map.)

Then in each dimension k there is a linear map

$$\mathrm{H}_k(X) \xrightarrow{\mathrm{H}_k[f]} \mathrm{H}_k(Y)$$

induced by the chain map $\mathrm{C}_*[f]$. In this way, each H_k is a functor.

Example 4.25. —do an example—

Plan (continued). The Whitehead invariance of simplicial homology can be expressed in sharper form. Suppose $X \searrow Y$. Then $Y = X - \{\sigma, \tau\}$ is a subcomplex of X. Let $Y \xrightarrow{i} X$ be the inclusion map. The sharp form of Whitehead invariance asserts that the linear map

$$\mathrm{H}_k(Y) \xrightarrow{\mathrm{H}_k[i]} \mathrm{H}_k(X)$$

is an isomorphism. In other words, the isomorphism between $\mathrm{H}_k(Y)$ and $\mathrm{H}_k(X)$ is specified.

4.6. Relative homology.

4.7. The long exact sequence of a pair.

In this section we encounter a particularly famous theorem in homological algebra. There are many theorems of this general flavour.

A **short exact sequence** (SES) of vector spaces is an exact sequence of the following form:

$$0 \longrightarrow A \xrightarrow{p} B \xrightarrow{q} C \longrightarrow 0$$

We can break this down into several separate statements:

- The composite qp is zero.
- Exactness at A means that p is injective. Thus A is isomorphic to $p(A)$.
- Exactness as C means that q is surjective. Thus C is isomorphic to $B/q^{-1}(0)$.
- Exactness at B means that $p(A) = q^{-1}(0)$. $\boxed{\text{Thus } C \cong B/p(A).}$

This last boxed statement is the point of short exact sequences. If U is a subspace of V and V/U is the quotient space, then there is a SES

$$0 \longrightarrow U \xrightarrow{j} V \xrightarrow{q} V/U \longrightarrow 0$$

where j is the inclusion and q is the quotient map. Conversely, the discussion above indicates that every SES is isomorphic to a SES of this type.

A SES of chain complexes is a diagram of chain complexes

$$0 \longrightarrow A_* \xrightarrow{P} B_* \xrightarrow{Q} C_* \longrightarrow 0$$

which restricts to a SES of vector spaces at each index k. We can draw this as a vast commuting diagram

$$\begin{tikzcd}
 & \vdots \arrow[d] & \vdots \arrow[d] & \vdots \arrow[d] & \\
0 \arrow[r] & A_{k+1} \arrow[r, "P_{k+1}"] \arrow[d] & B_{k+1} \arrow[r, "Q_{k+1}"] \arrow[d] & C_{k+1} \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & A_k \arrow[r, "P_k"] \arrow[d] & B_k \arrow[r, "Q_k"] \arrow[d] & C_k \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & A_{k-1} \arrow[r, "P_{k-1}"] \arrow[d] & B_{k-1} \arrow[r, "Q_{k-1}"] \arrow[d] & C_{k-1} \arrow[r] \arrow[d] & 0 \\
 & \vdots & \vdots & \vdots &
\end{tikzcd}$$

where the vertical maps are the boundary maps and each row is a SES.

Theorem 4.26. *A short exact sequence of chain complexes gives rise to a long exact sequence*

$$\ldots \longrightarrow \mathrm{H}_{k+1}(C_*) \xrightarrow{\partial} \mathrm{H}_k(A_*) \xrightarrow{\mathrm{H}_k[P]} \mathrm{H}_k(B_*) \xrightarrow{\mathrm{H}_k[Q]} \mathrm{H}_k(C_*) \xrightarrow{\partial} \mathrm{H}_{k-1}(A_*) \longrightarrow \ldots$$

where the maps ∂, to be defined in the proof, are called **connecting homomorphisms**.

In many cases our chain complexes stop at $k = 0$. This means that the rows of our vast commuting diagram are zero for negative indices and the LES terminates in the following terms:

$$\ldots \longrightarrow \mathrm{H}_1(C_*) \xrightarrow{\partial} \mathrm{H}_0(A_*) \xrightarrow{\mathrm{H}_0[P]} \mathrm{H}_0(B_*) \xrightarrow{\mathrm{H}_0[Q]} \mathrm{H}_0(C_*) \longrightarrow 0$$

Remark. Theorems like this are proved using a technique called 'diagram-chasing'. It is often easier to explain the arguments in real time than to write them down. It is also often easier to reconstruct the arguments oneself than to follow a written version.

"*Proof.* This is a routine diagram-chase. □"

is a common substitute for a written-out proof.

Proof. To establish the theorem one must

(i) construct ∂,

and then verify the relations

(ii) $\mathrm{H}_k[P] \circ \partial = 0$, (ii′) $\ker \mathrm{H}_k[P] \subseteq \operatorname{im} \partial$,

(iii) $\mathrm{H}_k[Q] \circ \mathrm{H}_k[P] = 0$, (iii′) $\ker \mathrm{H}_k[Q] \subseteq \operatorname{im} \mathrm{H}_k[P]$,

(iv) $\partial \circ \mathrm{H}_k[Q] = 0$, (iv′) $\ker \partial \subseteq \operatorname{im} \mathrm{H}_k[Q]$.

We address each item in turn. (Items (iii) and (iii)′ are independent of (i).)

(i) To construct $\partial : \mathrm{H}_{k+1}(C_*) \to \mathrm{H}_k(A_*)$, we work with part of the diagram:

$$
\begin{tikzcd}
 & B_{k+1} \arrow[r, "Q"] \arrow[d, "\partial"'] & C_{k+2} \arrow[d, "\partial"'] \\
A_{k+1} \arrow[r, "P"] \arrow[d, "\partial"'] & B_{k+1} \arrow[r, "Q"] \arrow[d, "\partial"'] & C_{k+1} \arrow[d, "\partial"'] \\
A_k \arrow[r, "P"] \arrow[d, "\partial"'] & B_k \arrow[r, "Q"] \arrow[d, "\partial"'] & C_k \\
A_{k-1} \arrow[r, "P"] & B_{k-1} &
\end{tikzcd}
$$

All subscripts have been stripped from the maps, and the leading and trailing zero spaces have been omitted. We remember that the maps P are injective and the maps Q are surjective.

Consider a cycle $\gamma \in C_{k+1}$. Since Q is surjective we can find a 'lift' $\beta \in B_{k+1}$ with $\gamma = Q\beta$. Then $Q\partial\beta = \partial Q\beta = \partial\gamma = 0$ so by exactness there is a unique $\alpha \in A_k$ such that $\partial\beta = P\alpha$. Moreover, $\partial\alpha = 0$ since P is injective and $P\partial\alpha = \partial P\alpha = \partial\partial\beta = 0$. See Figure 1 (left).

The choice of lift of γ need not be unique. Suppose we pick a different lift β' and obtain α'. Since $Q(\beta' - \beta) = \gamma - \gamma = 0$, it follows from exactness that there exists $\hat{\alpha} \in A_{k+1}$ with $P\hat{\alpha} = \beta' - \beta$. Then $P\partial\hat{\alpha} = \partial P\hat{\alpha} = \partial(\beta' - \beta) = P(\alpha' - \alpha)$, and the injectivity of P implies $\partial\hat{\alpha} = \alpha' - \alpha$. See Figure 1 (middle).

The upshot is that the homology class $[\alpha] \in \mathrm{H}_k(A_*)$ is uniquely defined for any given cycle $\gamma \in \mathrm{Z}_{k+1}(C_*)$ by the procedure above.

The function $\gamma \mapsto [\alpha]$ is linear. Indeed, if $\gamma_1, \gamma_2 \in \mathrm{Z}_{k+1}(C_*)$ have respective lifts β_1, β_2 which yield cycles α_1, α_2, then for $\gamma_1 + \gamma_2$ we can take $\beta_1 + \beta_2$ as its lift and obtain the cycle $\alpha_1 + \alpha_2$. A similar argument works for scalar multiplication.

Finally, we show that $[\alpha] \in \mathrm{H}_k(A_*)$ depends only on the homology class $[\gamma] \in \mathrm{H}_{k+1}(C_*)$. Consider a boundary $\partial\hat{\gamma} \in \mathrm{B}_{k+1}(C_*)$ obtained from some $\hat{\gamma} \in \mathrm{C}_{k+2}(C_*)$. For any lift $\hat{\beta}$ of $\hat{\gamma}$, we have $Q\partial\hat{\beta} = \partial Q\hat{\beta} = \partial\hat{\gamma}$ which means that $\partial\hat{\beta}$ is a lift of $\partial\hat{\gamma}$. But now $\partial\partial\hat{\beta} = 0$ so the resulting α is also zero. See Figure 1 (right).

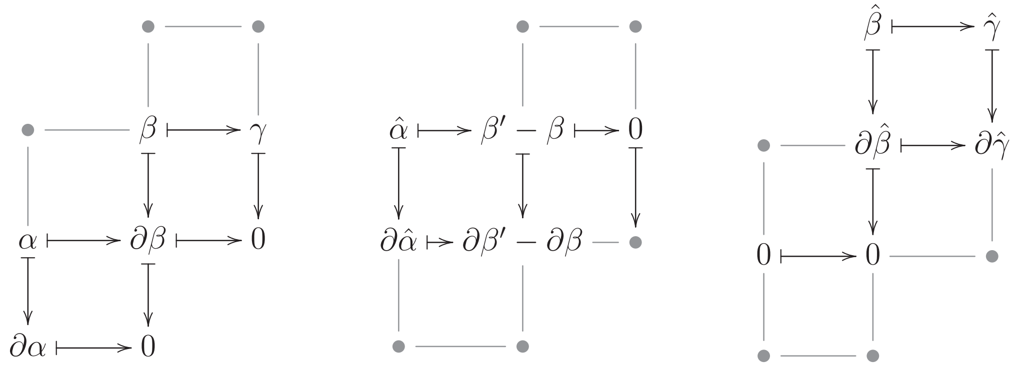


FIGURE 1. The construction of the connecting homomorphism: (left) given a cycle $\gamma \in \mathrm{Z}_{k+1}(C_*)$ and a lift $\beta \in \mathrm{C}_{k+1}(B_*)$ we obtain a cycle $\alpha \in \mathrm{Z}_k(A_*)$; (middle) the homology class $[\alpha] \in \mathrm{H}_k(A_*)$ is independent of the choice of lift β; (right) the homology class $[\alpha] \in \mathrm{H}_k(A_*)$ depends only on the homology class $[\gamma] \in \mathrm{H}_{k+1}(C_*)$.

Putting these statements together, it follows that we have a well defined linear map

$$\mathrm{H}_{k+1}(C_*) \to \mathrm{H}_k(A_*), \quad [\gamma] \mapsto [\alpha]$$

and this is the connecting homomorphism we seek.

(ii) Suppose $[\alpha] = \partial[\gamma]$. By construction, $\mathrm{H}_k[P]$ carries $[\alpha]$ to $[P\alpha] = [\partial\beta] = 0$.

(ii)′ Suppose $\mathrm{H}_k[P]$ carries $[\alpha]$ to 0, so $[P\alpha] = [0]$. This means that $P\alpha = \partial\beta$ for some $\beta \in \mathrm{C}_{k+1}(B_*)$. Then $\gamma = Q\beta$ is a cycle, because $\partial\gamma = \partial Q\beta = Q\partial\beta = QP\alpha = 0$. Now we have reproduced the configuration of Figure 1, and so $[\alpha] = \partial[\gamma]$.

(The next two items, (iii) and (iii)′ are independent of (i), and could have been verified at the outset.)

(iii) Since H_k is a functor, we have $\mathrm{H}_k[Q] \circ \mathrm{H}_k[P] = \mathrm{H}_k[QP] = \mathrm{H}_k[0]$ and this is equal to zero.

(iii)′ Suppose $\mathrm{H}_k[Q]$ carries $[\beta] \in \mathrm{H}_k(B_*)$ to zero, so $Q\beta = \partial\hat\gamma$ for some $\hat\gamma \in \mathrm{C}_{k+1}(C_*)$, and we can take a lift $\hat\beta \in \mathrm{C}_{k+1}$ so that $\hat\gamma = Q\hat\beta$. Then $Q\partial\hat\beta = \partial Q\hat\beta = \partial\hat\gamma = Q\beta$, so that $Q(\beta - \partial\hat\beta) = 0$. Exactness now implies that there exists $\alpha \in \mathrm{C}_k(A_*)$ such that $P\alpha = \beta - \partial\hat\gamma$, and moreover $\partial\alpha = 0$ because P is injective and $P\partial\alpha = \partial P\alpha = \partial(\beta - \partial\hat\beta) = 0$. Thus we have $[\alpha] \in \mathrm{H}_k(A_*)$ which is carried by $\mathrm{H}_k[P]$ to $[P\alpha] = [\beta - \partial\hat\beta] = [\beta]$.

(iv) Let $[Q\beta]$ be the image under $\mathrm{H}_k[Q]$ of some $[\beta] \in \mathrm{H}_k(B_*)$. We can construct $[\alpha] = \partial[Q\beta]$ in the usual way, choosing β as the lift of $Q\beta$. But β is a cycle, so $\partial\beta = 0$ and so $\alpha = 0$.

(iv)′ Suppose $\partial[\gamma] = [\alpha] = 0$ for some $[\gamma] \in \mathrm{H}_{k+1}(C_*)$. We may suppose that $P\alpha = \partial\beta$ for some lift $\beta \in \mathrm{C}_{k+1}(B_*)$ of γ. Now $[\alpha] = 0$ implies that $\alpha = \partial\hat\alpha$ for some $\hat\alpha \in \mathrm{C}_{k+1}(A_*)$. Now $QP\hat\alpha = 0$ and $\partial P\hat\alpha = P\partial\hat\alpha = P\alpha = \partial B$. It follows that $Q(\beta - P\hat\alpha) = \gamma$ and $\partial(\beta - P\hat\alpha) = 0$. Thus we have a homology class $[\beta - P\hat\alpha] \in \mathrm{H}_k(B_*)$ which is carried by $\mathrm{H}_k[Q]$ to $[\gamma]$.

This completes the proof of the theorem. □

Theorem 4.27 (The LES of a pair). *Let X, Y be simplicial complexes with $Y \subseteq X$. Let i, j denote the inclusion maps*

$$Y \xrightarrow{i} X \xrightarrow{j} (X, Y).$$

Then there is a long exact sequence

$$\boxed{\ldots \longrightarrow \mathrm{H}_{k+1}(X,Y) \xrightarrow{\partial} \mathrm{H}_k(Y) \xrightarrow{\mathrm{H}_k[i]} \mathrm{H}_k(X) \xrightarrow{\mathrm{H}_k[j]} \mathrm{H}_k(X,Y) \xrightarrow{\partial} \mathrm{H}_{k-1}(Y) \longrightarrow \ldots}$$

relating the homology of X and Y and the relative homology of (X, Y).

Proof. Apply Theorem 4.26 to the following tautological SES of chain complexes

$$0 \longrightarrow \mathrm{C}_*(Y) \longrightarrow \mathrm{C}_*(X) \longrightarrow \mathrm{C}_*(X)/\mathrm{C}_*(Y) \longrightarrow 0$$

and recall the definition $\mathrm{C}_*(X,Y) = \mathrm{C}_*(X)/\mathrm{C}_*(Y)$. □

Here is the theorem we have been aiming for:

Theorem 4.28 (Whitehead invariance of simplicial homology). *Let X, Y be simplicial complexes with $X \searrow Y$ and let $Y \xrightarrow{i} X$ denote the inclusion map. Then*

$$\mathrm{H}_k(X) \xrightarrow{\mathrm{H}_k[i]} \mathrm{H}_k(Y)$$

is an isomorphism for all k.

Proof. We first show that $\mathrm{H}_k(X,Y) = 0$ for all k. Recall that $Y = X - \{\sigma, \tau\}$ where σ is the unique coface of τ. Thus the chain complex $\mathrm{C}_*(X,Y)$ is isomorphic to

$$0 \longleftarrow \ldots \longleftarrow 0 \longleftarrow \mathbb{F} \xleftarrow{\cong} \mathbb{F} \longleftarrow 0 \longleftarrow \ldots$$

with a single generator in each of the dimensions $\dim(\tau)$ and $\dim(\sigma) = \dim(\tau) + 1$, and a nonzero boundary map between them. Evidently $\mathrm{H}_k(X,Y) = 0$ in all dimensions.

It follows that the LES for (X, Y) contains the following excerpt, for every k:

$$\ldots \longrightarrow 0 \longrightarrow \mathrm{H}_k(Y) \xrightarrow{\mathrm{H}_k[i]} \mathrm{H}_k(X) \longrightarrow 0 \longrightarrow \ldots$$

Exactness now implies that each $\mathrm{H}_k[i]$ is an isomorphism. □

Some calculations run more quickly in reduced homology. Here is the appropriate adaptation of Theorem 4.27. The content of the result is identical except that $\mathrm{H}_0(X), \mathrm{H}_0(Y)$ are replaced by $\tilde{\mathrm{H}}_0(X), \tilde{\mathrm{H}}_0(Y)$.

Theorem 4.29 (The LES of a pair, reduced homology). *Under the same circumstances as Theorem 4.27, there is a long exact sequence*

$$\boxed{\ldots \longrightarrow \mathrm{H}_{k+1}(X,Y) \xrightarrow{\partial} \tilde{\mathrm{H}}_k(Y) \xrightarrow{\tilde{\mathrm{H}}_k[i]} \tilde{\mathrm{H}}_k(X) \xrightarrow{\tilde{\mathrm{H}}_k[j]} \mathrm{H}_k(X,Y) \xrightarrow{\partial} \tilde{\mathrm{H}}_{k-1}(Y) \longrightarrow \ldots}$$

in reduced homology.

Proof. We must verify that the sequence of chain complexes is exact.

$$0 \longrightarrow \tilde{\mathrm{C}}_*(Y) \xrightarrow{\tilde{\mathrm{C}}_*[i]} \tilde{\mathrm{C}}_*(X) \xrightarrow{\tilde{\mathrm{C}}_*[j]} \mathrm{C}_*(X)/\mathrm{C}_*(Y) \longrightarrow 0\ .$$

In fact, the sequence is equal to the SES for the unreduced case in all rows except $k = -1$. The row $k = -1$ is by definition equal to

$$0 \longrightarrow \mathbb{F} \xrightarrow{1} \mathbb{F} \longrightarrow 0 \longrightarrow 0$$

and is therefore also exact. □

Example 4.30. For $n \geq 0$, the reduced homology of a hollow n-simplex is as follows:

$$\tilde{\mathrm{H}}_k(\partial B_n) \cong \begin{cases} \mathbb{F} & \text{for } k = n-1 \\ 0 & \text{for } k \neq n-1 \end{cases}$$

Proof. Let $X = \partial B_n$ and let $Y = X - \sigma$, where σ is a simplex of dimension $n-1$. Notice that Y is isomorphic to a cone $C(\partial B_{n-1})$. Thus $\tilde{\mathrm{H}}_k(Y) = 0$ for all k, and so the snippet

$$\ldots \longrightarrow 0 \longrightarrow \tilde{\mathrm{H}}_k(X) \longrightarrow \mathrm{H}_k(X,Y) \longrightarrow 0 \longrightarrow \ldots$$

from the reduced-homology LES implies that $\tilde{\mathrm{H}}_k(X) \cong \mathrm{H}_k(X,Y)$ for all k. The latter is easy to compute: the chain complex $\mathrm{C}_*(X,Y)$ has one generator in dimension $n-1$ and no other generators, so the homology is 1-dimensional in dimension $n-1$ and zero elsewhere. □

Note. We can immediately deduce the unreduced homology of ∂B_n from this, following the discussion at the end of Section 4.4. The following comparison of reduced and unreduced Betti vectors indicates why reduced homology is marginally more convenient for this calculation.

$$\begin{aligned}
\tilde{\mathrm{b}}(\partial B_0) &= [1; 0,0,0,0,0,\ldots], & \mathrm{b}(\partial B_0) &= [0,0,0,0,0,\ldots]. \\
\tilde{\mathrm{b}}(\partial B_1) &= [0; 1,0,0,0,0,\ldots], & \mathrm{b}(\partial B_1) &= [2,0,0,0,0,\ldots]. \\
\tilde{\mathrm{b}}(\partial B_2) &= [0; 0,1,0,0,0,\ldots], & \mathrm{b}(\partial B_2) &= [1,1,0,0,0,\ldots]. \\
\tilde{\mathrm{b}}(\partial B_3) &= [0; 0,0,1,0,0,\ldots], & \mathrm{b}(\partial B_3) &= [1,0,1,0,0,\ldots]. \\
\tilde{\mathrm{b}}(\partial B_4) &= [0; 0,0,0,1,0,\ldots], & \mathrm{b}(\partial B_4) &= [1,0,0,1,0,\ldots].
\end{aligned}$$

(For the reduced Betti vectors, the term $\tilde{\mathrm{b}}_{-1}$ is listed before the semicolon.)

Lecture 10

4.8. Contiguous maps.

4.9. Vietoris–Rips complexes.