Part IV — Topics in Geometric Group Theory

Based on lectures by H. Wilton

Notes taken by Dexter Chua

Michaelmas 2017

The subject of geometric group theory is founded on the observation that the algebraic and algorithmic properties of a discrete group are closely related to the geometric features of the spaces on which the group acts. This graduate course will provide an introduction to the basic ideas of the subject.

Suppose Γ is a discrete group of isometries of a metric space X. We focus on the theorems we can prove about Γ by imposing geometric conditions on X. These conditions are motivated by curvature conditions in differential geometry, but apply to general metric spaces and are much easier to state. First we study the case when X is *Gromov-hyperbolic*, which corresponds to negative curvature. Then we study the case when X is *CAT(0)*, which corresponds to non-positive curvature. In order for this theory to be useful, we need a rich supply of negatively and non-positively curved spaces. We develop the theory of *non-positively curved cube complexes*, which provide many examples of CAT(0) spaces and have been the source of some dramatic developments in low-dimensional topology over the last twenty years.

Part 1. We will introduce the basic notions of geometric group theory: Cayley graphs, quasiisometries, the Schwarz–Milnor Lemma, and the connection with algebraic topology via presentation complexes. We will discuss the word problem, which is quantified using the Dehn functions of a group.

Part 2. We will cover the basic theory of word-hyperbolic groups, including the Morse lemma, local characterization of quasigeodesics, linear isoperimetric inequality, finitely presentedness, quasiconvex subgroups etc.

Part 3. We will cover the basic theory of CAT(0) spaces, working up to the Cartan–Hadamard theorem and Gromov's Link Condition. These two results together enable us to check whether the universal cover of a complex admits a CAT(1) metric.

Part 4. We will introduce cube complexes, in which Gromov's link condition becomes purely combinatorial. If there is time, we will discuss Haglund–Wise's *special* cube complexes, which combine the good geometric properties of CAT(0) spaces with some strong algebraic and topological properties.

Pre-requisites

Part IB Geometry and Part II Algebraic topology are required.

Contents

1 Cayley graphs and the word metric

1.1 The word metric

There is this unfortunate tendency for people to think of groups as being part of algebra. Which is, perhaps, reasonable. There are elements and there are operations on them. These operations satisfy some algebraic laws. But there is also geometry involved. For example, one of the simplest non-abelian groups we know is D_6, the group of symmetries of a triangle. This is fundamentally a geometric idea, and often this geometric interpretation can let us prove a lot of things about D_6.

In general, the way we find groups in nature is that we have some object, and we look at its symmetry. In geometric group theory, what we do is that we want to remember the object the group acts on, and often these objects have some geometric structure. In fact, we shall see that all groups are symmetries of some geometric object.

Let Γ be a group, and S a generating set for Γ. For the purposes of this course, S will be finite. So in this course, we are mostly going to think about finitely-generated groups. Of course, there are non-finitely-generated groups out there in the wild, but we will have to put in some more effort to make our ideas work.

The simplest geometric idea we can think of is that of distance, i.e. a metric. So we want to use our generating set S to make Γ into a metric space.

Definition (Word length)**.** Let Γ be a group and S a finite generating set. If $\gamma \in \Gamma$, the *word length* of γ is

$$\ell_S(\gamma) = \min\{n : \gamma = s_1^{\pm 1} \cdots s_n^{\pm 1} \text{ for some } s_i \in S\}.$$

Definition (Word metric)**.** Let Γ be a group and S a finite generating set. The *word metric* on Γ is given by

$$d_S(\gamma_1, \gamma_2) = \ell_s(\gamma_1^{-1}\gamma_2).$$

If we stare at the definition, we see that the word metric is left-invariant, i.e. for any $g, \gamma_1, \gamma_2 \in \Gamma$, we have

$$d_S(g\gamma_1, g\gamma_2) = d_S(\gamma_1, \gamma_2).$$

However, it is usually not the case that the metric is right invariant, i.e. $d_S(\gamma_1 g, \gamma_2 g)$ is not equal to $d_S(\gamma_1, \gamma_2)$. This is one of the prices we have to pay for wanting to do geometry.

If we tried to draw our metric space out, it looks pretty unhelpful — it's just a bunch of dots. It is not path connected. We can't really think about it well. The solution is provided by Cayley graphs.

Definition (Cayley graph)**.** The *Cayley graph* $\mathrm{Cay}_S(\Gamma)$ is defined as follows:

- $V(\mathrm{Cay}_S(\Gamma)) = \Gamma$
- For each $\gamma \in \Gamma$ and $s \in S$, we draw an edge from γ to γs.

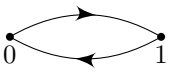
If we want it to be a directed and labelled graph, we can label the edge by the generator s.

How does this relate to the word metric? There is an obvious left action of Γ on $\mathrm{Cay}_S(\Gamma)$ which extends the left action of Γ on itself. This is since the edge is defined by multiplication of the generator on the *right*.

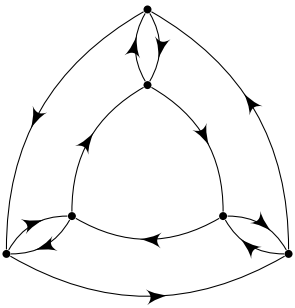
Moreover the word metric on Γ extends to a metric on $\mathrm{Cay}_S(\Gamma)$ in which every edge is isometric to $[0, 1]$.

Finally, note that $\mathrm{Cay}_S(\Gamma)$ is always regular, where the degree of each vertex is $2|S|$.

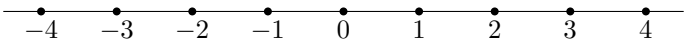
Example. Take $\Gamma = C_2 = \mathbb{Z}/2\mathbb{Z}$, and take $S = \{1\}$. Then the Cayley graph looks like



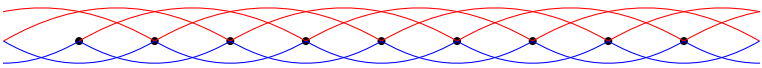
Example. Take $\Gamma = S_3$ and $S = \{(1\ 2), (1\ 2\ 3)\}$. The Cayley graph is



Example. Take $\Gamma = \mathbb{Z}$, and $S = \{1\}$. Then the Cayley graph looks like

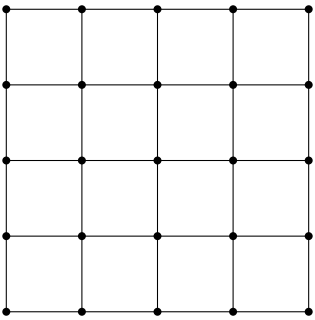


Example. Take $\Gamma = \mathbb{Z}$ and $S = \{2, 3\}$. Then the Cayley graph looks like

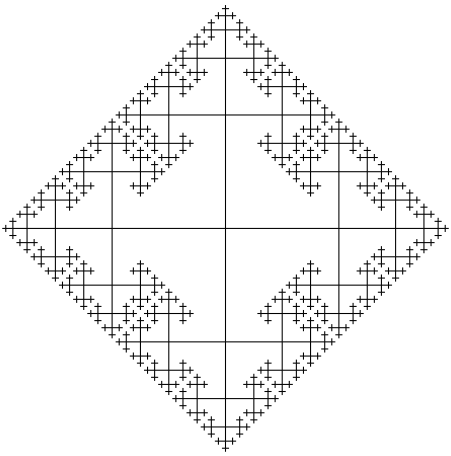


Now this seems quite complicated. It's still the same group $\Gamma = \mathbb{Z}$, but with a different choice of generating set, it looks very different. It seems like what we do depends heavily on the choice of the generating set.

Example. If $\Gamma = \mathbb{Z}^2$ and $S = \{(1, 0), (0, 1)\}$, then the Cayley graph is a grid



Example. If $\Gamma = F_2 = \langle a, b\rangle$, $S = \{a, b\}$, then the Cayley graph looks like



If one has done algebraic topology, then one might recognize this as the universal cover of a space. This is not a coincidence, and we will talk about this later.

We now return to the problem we noticed previously. The Cayley graph depends a lot on the generating set chosen, but we want to study the graph without picking a generating set.

What we observe is that while the two Cayley graphs of $\mathbb{Z}$ we drew seem quite different, if we looked at them from 100 meters away, they look quite similar — they both look like a long, thin line. This is the idea we are trying to formalize.

Definition (Quasi-isometry)**.** Let $\lambda \geqslant 1$ and $c \geqslant 0$. A function between metric spaces $f : X \to Y$ is a (λ, c)-*quasi-isometric embedding* if for all $x_1, x_2 \in X$

$$\frac{1}{\lambda} d_X(x_1, x_2) - c \leqslant d_Y(f(x_1), f(x_2)) \leqslant \lambda d_X(x_{,1}, x_2) + c$$

If, in addition, there is a C such that for all $y \in Y$, there exists $x \in X$ such that $d_Y(y, f(x)) \leqslant C$, we say f is a *quasi-isometry*, and X is *quasi-isometric* to Y. We write $X \underset{qi}{\simeq} Y$.

We can think of the first condition as saying we are *quasi-injective*, and the second as saying we are *quasi-surjective*.

The right way to think about the definition is that the c says we don't care about what happens at scales less than c, and the λ is saying we allow ourselves to stretch distances. Note that if we take $c = 0$, then this is just the notion of a bi-Lipschitz map. But in general, we don't even require f to be continuous, since continuity is a rather fine-grained property.

Exercise. Check that $X \underset{qi}{\simeq} Y$ is an equivalence relation.

This is not immediately obvious, because there is no inverse to f. In fact, we need the axiom of choice to prove this result.

Example. Any bounded metric spaces is quasi-isometric to a point. In particular, if Γ is finite, then $(\Gamma, d_S) \underset{qi}{\simeq} 1$.

For this reason, this is not a great point of view for studying finite groups. On the other hand, for infinite groups, this is really useful.

Example. For any Γ and S, the inclusion $(\Gamma, \Gamma_S) \hookrightarrow (\mathrm{Cay}_S(\Gamma), d_S)$ is a quasi-isometry (take $\lambda = 1, c = 0, C = \frac{1}{2}$).

Of course, the important result is that any two word metrics on the same group are quasi-isomorphic.

Theorem. For any two finite generating sets S, S' of a group Γ, the identity map $(\Gamma, d_S) \to (\Gamma, d_{S'})$ is a quasi-isometry.

Proof. Pick

$$\lambda = \max_{s\in S} \ell_{S'}(s), \quad \lambda' \max_{s\in S'} \ell_S(s),$$

We then see

$$\ell_S(\gamma) \leqslant \lambda' \ell_{S'}(\gamma), \quad \ell_{S'}(\gamma) \leqslant \lambda \ell_s(\gamma).$$

for all $\gamma \in \Gamma$. Then the claim follows. □

So as long as we are willing to work up to quasi-isometry, we have a canonically defined geometric object associated to each finitely-generated group.

Our next objective is to be able to state and prove the Schwarz–Milnor lemma. This is an important theorem in geometric group theory, saying that if a group Γ acts "nicely" on a metric space X, then Γ must be finitely generated, and in fact (Γ, d_s) is quasi-isomorphic to X. This allows us to produce some rather concrete geometric realizations of (Γ, d_s), as we will see in examples.

In order to do that, we must write down a bunch of definitions.

Definition (Geodesic)**.** Let X be a metric space. A *geodesic* in X is an isometric embedding of a closed interval $\gamma : [a, b] \to X$.

This is not exactly the same as the notion in differential geometry. For example, if we have a sphere and two non-antipodal points on the sphere, then there are many geodesics connecting them in the differential geometry sense, but only the shortest one is a geodesic in our sense. To recover the differential geometry notion, we need to insert the words "locally" somewhere, but we shall not care about that.

Definition (Geodesic metric space)**.** A metric space X is called *geodesic* if every pair of points $x, y \in X$ is joined by a geodesic denoted by $[x, y]$.

Note that the notation $[x, y]$ is not entirely honest, since there may be many geodesics joining two points.

Definition (Proper metric space)**.** A metric space is *proper* if closed balls in X are compact.

Example. If $\Gamma = \langle S \rangle$, then $\mathrm{Cay}_S(\Gamma)$ is geodesic. If S is finite, then $\mathrm{Cay}_S(\Gamma)$ is proper.

Example. Let M be a connected Riemannian manifold. Then there is a metric on M defined by

$$d(x, y) = \inf_{\alpha: x\to y} \mathrm{length}(\alpha),$$

where we take the infimum over all smooth paths from x to y.

The *Hopf–Rinow theorem* says if M is complete (as a metric space), then the metric is proper and geodesic. This is great, because completeness is in some sense a local property, but "proper" and "geodesic" are global properties.

We need two more definitions, before we can state the Schwarz–Milnor lemma.

Definition (Proper discontinuous action)**.** An action Γ on a topological space X is *proper discontinuous* if for every compact set K,

$$\{g \in \Gamma : gK \cap K \neq \varnothing\}$$

is finite.

Definition (Cocompact action)**.** An action Γ on a topological space X is *cocompact* if the quotient $\Gamma\backslash X$ is compact.

Lemma (Schwarz–Milnor lemma)**.** Let X be a proper geodesic metric space, and let Γ act properly discontinuously, cocompactly on X by isometries. Then

(i) Γ is finitely-generated.

(ii) For any $x_0 \in X$, the orbit map

$$\begin{aligned}\Gamma &\to X\\ \gamma &\mapsto \gamma x_0\end{aligned}$$

is a quasi-isometry $(\Gamma, d_s) \underset{qi}{\simeq} (X, d)$.

An easy application is that Γ acting on its Caylay graph satisfies these conditions. So this reproduces our previous observation that a group is quasi-isometric to its Cayley graph. More interesting examples involve manifolds.

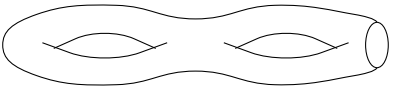
Example. Let M be a closed (i.e. compact without boundary), connected Riemannian manifold. Then the universal cover $\tilde{M}$ is also a complete, connected, Riemannian manifold. By the Hopf–Rinow theorem, this is proper and geodesic.

Since the metric of $\tilde{M}$ is pulled back from M, we know $\pi_1(M)$ acts on $\tilde{M}$ by isometries. Therefore by the Schwarz–Milnor lemma, we know

$$\pi_1(M) \underset{qi}{\simeq} \tilde{M}.$$

Example. The universal cover of the torus $S^1 \times S^1$ is $\mathbb{R}^2$, and the fundamental group is $\mathbb{Z}^2$. So we know $\mathbb{Z}^2 \underset{qi}{\simeq} \mathbb{R}^2$, which is not surprising.

Example. Let $M = \Sigma_2$, the surface of genus 2.



We then have

$$\pi_1\Sigma_2 \cong \langle a_1, b_1, a_2, b_2 \mid [a_1, b_1][a_2, b_2]\rangle.$$

On the other hand, the universal cover can be thought of as the hyperbolic plane $\mathbb{H}^2$, as we saw in IB Geometry.

So it follows that

$$\pi_1\Sigma_2 \underset{qi}{\simeq} \mathbb{H}^2.$$

This gives us some concrete hold on what the group $\pi_1\Sigma_2$ is like, since people have been thinking about the hyperbolic plane for centuries.

After all those applications, we can start to write down a proof.

Proof of Schwarz–Milnor lemma. Let $\bar{B} = \bar{B}(x, R)$ be such that $\Gamma\bar{B} = X$. This is possible since the quotient is compact.

Let $S = \{\gamma \in \Gamma : \gamma\bar{B} \cap \bar{B} \neq \varnothing\}$. By proper discontinuity, this set is finite.

We let

$$r = \inf_{\gamma' \notin S} d(\bar{B}, \gamma'\bar{B}).$$

If we think about it, we see that in fact r is the *minimum* of this set, and in particular $r > 0$.

Finally, let

$$\lambda = \max_{s \in S} d(x_0, sx_0).$$

We will show that S generates Γ, and use the word metric given by S to show that Γ is quasi-isometric to X.

We first show that $\Gamma = \langle S\rangle$. We let $\gamma \in \Gamma$ be arbitrary.

Let $[x_0, \gamma x_0]$ be a geodesic from x_0 to γx_0. Let ℓ be such that

$$(\ell - 1)r \leqslant d(x_0, \gamma x_0) < \ell r.$$

Then we can divide the geodesic into ℓ pieces of length about r. We can choose $x_1, \dots x_{\ell-1}, x_\ell = \gamma x_0$ such that $d(x_{i-1}, x_i) < r$.

By assumption, we can pick $\gamma_i \in \Gamma$ such that $x_i \in \gamma_i\bar{B}$, and further we pick $\gamma_\ell = \gamma$, $\gamma_0 = e$. Now for each i, we know

$$d(\bar{B}, \gamma_{i-1}^{-1}\gamma_i\bar{B}) = d(\gamma_{i-1}\bar{B}, \gamma_i\bar{B}) \leqslant d(x_{i-1}, x_i) < r.$$

So it follows that $\gamma_{i-1}^{-1}\gamma_i \in S$. So we have

$$\gamma = \gamma_\ell = (\gamma_0^{-1}\gamma_1)(\gamma_1^{-1}\gamma_2)\cdots(\gamma_{\ell-1}^{-1}\gamma_\ell) \in \langle S\rangle.$$

This proves $\Gamma = \langle S\rangle$.

To prove the second part, we simply note that

$$r\ell - r \leqslant d(x_0\gamma x_0).$$

We also saw that ℓ is an upper bound for the word length of γ under S. So we have

$$r\ell_s(\gamma) - r \leqslant d(x_0, \gamma x_0).$$

On the other hand, by definition of λ, we have

$$d(x_0, \gamma x_0) \leqslant \lambda\ell_s(\gamma).$$

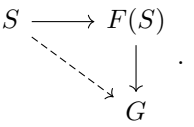
So the orbit map is an orbit-embedding, and quasi-surjectivity follows from cocompactness. □

If we want to understand a group, then a reasonable strategy now would be to come up with a well-understood (proper geodesic) space for Γ to act (proper discontinuously, cocompactly) on, and see what we can tell from this quasi-isometry. There is no general algorithm for doing so, but we shall see that for certain groups, there are some "obvious" good choices to make.

1.2 Free groups

Let's begin by understanding the free group. In some sense, they are the "simplest" groups, but they also contain some rich structure. Roughly speaking, a free group on a set S is a group that is "freely" generated by S. This freeness is made precise by the following universal property.

Definition (Free group)**.** Let S be a (usually finite) set. A group $F(S)$ with a map $S \to F(S)$ is called the *free group on* S if it satisfies the following universal property: for any set map $S \to G$, there is a unique group homomorphism $F(S) \to G$ such that the following diagram commutes:



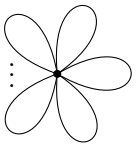
Usually, if $|S| = r$, we just write $F(S) = F_r$.

In fancy category-theoretic language, this says the functor $F : \mathbf{Sets} \to \mathbf{Grps}$ is left adjoint to the forgetful functor $U : \mathbf{Grps} \to \mathbf{Sets}$.

This definition is good for some things, but not good for others. One thing it does well is it makes it clear that $F(S)$ is unique up to isomorphism if it exists. However, it is not immediately clear that $F(S)$ exists! (unless one applies the adjoint functor theorem)

Thus, we must concretely construct a group satisfying this property, and this construction is often useful if we actually want to work with the set.

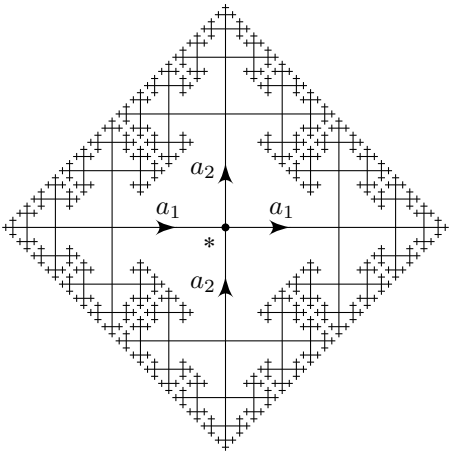
For concreteness, suppose $S = \{a_1, \dots, a_r\}$ is a finite set. We consider a graph

$X_r =$ 

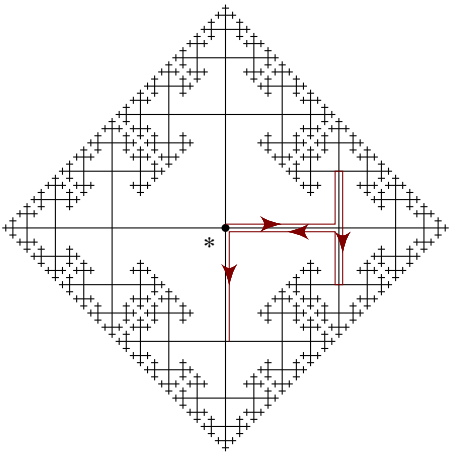
where there are r "petals". This is known as the *rose with r petals*.

We claim that its fundamental group is $F(S)$. To see this, we have to understand the universal cover $\tilde{X}_r$ of X_r. We know $\tilde{X}_r$ is a simply connected graph, so it is a tree. Moreover, since it is a covering space of X_r, it is regular with degree $2r$ on each vertex.

If $r = 2$, then this is our good old picture



We can use this to understand the fundamental group $\pi_1(X_r)$. The elements on $\pi_1(X_r)$ are homotopy classes $[\gamma]$ of based loops in X_r. Using the homotopy lifting lemma, this map can be lifted to a path $\tilde{\gamma}$ in the universal cover starting at $*$.



Since this is a loop, we know it ends at another vertex. Because $\tilde{X}_r$ is a tree, we may homotope $\tilde{\gamma}$ rel endpoints to the unique embedded path from $*$ to the endpoint of $\tilde{\gamma}$. We can push this homotopy down to X_r to get a homotopy of γ. We can see that γ is of the following form:

$$\gamma = a_{i_1}^{\pm 1} a_{i_2}^{\pm 1} a_{i_3}^{\pm 1} \cdots a_{i_n}^{\pm 1}.$$

Moreover, since $\tilde{\gamma}$ is now embedded, we know γ is reduced, i.e. we never see anything looking like $a_i^{\pm 1} a_i^{\mp 1}$. Importantly, this form is unique. There is only one way to write γ in this form. This gives us a *normal form theorem* for elements in $\pi_1(X_r)$ — every element in $\pi_1(X_r)$ is represented by a unique reduced word in the generators. So if we have two elements in $\pi_1(X_r)$ expressed as products of generators and their inverses, then to determine if they are equal, we can individually reduce the two words in any way we like, and check if the result is the same.

Example. Take the case $\pi_1 X_2 = \langle a, b\rangle$. We might be given a word

$$a^3a^{-2}b^{-2}b^3a^5a^{-3}a^{-3}b^{-1}.$$

We can reduce this to

$$aba^{-1}b^{-1}.$$

In particular, since this is not the identity, we know the original path was not null-homotopic.

As a consequence of this, we see that

Corollary. $\pi_1(X_r)$ has the universal property of $F(S)$, with $S = \{a_1, \dots, a_r\}$. So $\pi_1(X_r) \cong F_r$.

Proof. Given any map $f : S \to G$, define $\tilde{f} : F(S) \to G$ by

$$\tilde{f}(a_{i_1}^{\pm 1} \cdots a_{i_n}^{\pm 1}) = \tilde{f}(a_i)^{\pm 1} \cdots \tilde{f}(a_{i_n})^{\pm 1}$$

for any reduced word $a_{i_1}^{\pm 1} \cdots a_{i_n}^{\pm 1}$. This is easily seen to be well-defined and is the unique map making the diagram commute. □

1.3 Finitely-presented groups

Now if a group Γ is generated by S, then we have a surjective homomorphism $F(S) \to \Gamma$. Let $K = \ker \eta$. Then

$$\Gamma \cong \frac{F(S)}{K}.$$

Since we understand $F(S)$ quite explicitly, it would be nice if we have a solid gasp on K as well. If R normally generates K, so that $K = \langle\langle R\rangle\rangle$, then we say that $\langle S \mid R\rangle$ is a *presentation* for Γ. We will often write that

$$\Gamma \cong \langle S \mid R\rangle.$$

Example.

$$\mathbb{Z}^2 = \langle a, b \mid aba^{-1}b^{-1}\rangle$$

Definition (Finitely-presented group)**.** A *finitely-presentable group* is a group Γ such that there are finite S and R such that $\Gamma \cong \langle S \mid R\rangle$.

A *finitely-presented group* is a group Γ equipped S and R such that $\Gamma \cong \langle S \mid R\rangle$.

Presentations give us ways to geometrically understand a group. Given a presentation $\mathcal{P} = \langle S \mid R\rangle$, we can construct space $X_\mathcal{P}$ such that $\pi_1 X_\mathcal{P} \cong \langle S \mid R\rangle$

To do so, we first construct a rose with $|S|$ many petals, each labeled by an element of S. For each $r \in R$, glue a disk onto the rose along the path specified by r. The *Seifert–van Kampen theorem* then tells us $\pi_1 X_\mathcal{P} \cong \Gamma$.

Example. We take the presentation $\mathbb{Z}^2 \cong \langle a, b \mid aba^{-1}b^{-1}\rangle$. If we think hard enough (or a bit), we see this construction gives the torus.

Conversely, if we are given a connected cell complex X, we can obtain a presentation of the fundamental group. This is easy if the cell complex has a single 0-cell. If it has multiple 0-cells, then we choose a maximal tree in $X^{(1)}$, and the edges $S = \{a_i\}$ not in the maximal tree define a generating set for $\pi_1 X^{(1)}$. The attaching maps of the 2-cells in X define elements $R = \{r_j\}$ of $\pi_1 X^{(1)}$, and these data define a presentation $\mathcal{P}_X = \langle S \mid R\rangle$ for $\pi_1 X$.

This is not canonical, since we have to pick a maximal tree, but let's not worry too much about that. The point of maximal trees is to get around the problem that we might have more than one vertex.

Exercise. If X has one vertex, then $\mathrm{Cay}_S \pi_1 X = \tilde{X}^{(1)}$.

1.4 The word problem

Before we try to understand the word problem, let's give some historical perspectives. Long time ago, Poincaré was interested in characterizing the 3-sphere. His first conjecture was that if a closed 3-manifold M had $H_*(M) \cong H_*(S^3)$, then $M \cong S^3$. Poincaré thought very hard about the problem, and came up with a counterexample. The counterexample is known as the *Poincaré homology sphere*. This is a manifold P^3 with

$$H_*(P) \cong H_*(S^3),$$

but it turns out there is a surjection $\pi_1(P^3) \to A_5$. In particular, P^3 is not homeomorphic to a sphere.

So he made a second conjecture, that if M is a compact manifold with $\pi_1 M \cong 1$, then $M \cong S^3$. Note that the condition already implies $H_*(M) \cong H_*(S^3)$ by Hurewicz theorem and Poincaré duality. This is known as the Poincaré conjecture, which was proved in 2002 by Perelman.

But there is more to say about Poincaré's initial conjecture. Some time later, Max Dehn constructed an *infinite family* of 3d homology spheres. In order to prove that he genuinely had infinitely many distinct homology spheres, he had to show that his homology spheres had different fundamental groups. He did manage to write down the presentations of the fundamental groups, and so what was left to do is to distinguish whether the presentations actually presented the same group.

To make sense of this statement, we must have some way to distinguish between the homology spheres. To do so, he managed to write down the presentation of the fundamental group of his homology spheres, and he had to figure out if those groups are genuinely the same.

For our purposes, perhaps we should be slightly less ambitious. Suppose we are given a presentation $\langle S \mid R\rangle$ of a finitely-presented group Γ. Define the set of alphabets

$$S^{\pm} = S \amalg S^{-1} = \{s, s^{-1} : s \in S\},$$

and let S^* be the set of all finite words in $S^{\pm}$. Then the fundamental question is, given a word $w \in S^*$, when does it represent the identity element $1 \in \Gamma$. Ideally, we would want an *algorithm* that determines whether this is the case.

Example. Recall that in the free group $F(S) = \langle S \mid \varnothing\rangle$, we had a normal form theorem, namely that every element in Γ can be written *uniquely* as a product

of generator (and inverses) such that there aren't any occurrences of ss^{-1} or $s^{-1}s$. This gives a way of determining whether a given word $w \in S^*$ represents the identity. We perform *elementary reductions*, removing every occurrence of ss^{-1} or $s^{-1}s$. Since each such procedure reduces the length of the word, this eventually stops, and we would end up with a *reduced word*. If this reduced word is empty, then w reduces the identity. If it is non-empty, w does not.

Thus, we have a complete solution to the word problem for free groups, and moreover the algorithm is something we can practically implement.

For a general finitely-presented group, this is more difficult. We can first reformulate our problem. The presentation $\Gamma = \langle S \mid R \rangle$ gives us a map $F(S) \to \Gamma$, sending s to s. Each word gives us an element in $F(S)$, and thus we can reformulate our problem as identifying the elements in the kernel of this map.

Lemma. Let $\Gamma = \langle S \mid R \rangle$. Then the elements of $\ker\{F(S) \to \Gamma\}$ are precisely those of the form

$$\prod_{i=1}^{d} g_i r_i^{\pm 1} g_i^{-1},$$

where $g_i \in F(S)$ and $r_i \in R$.

Proof. We know that $\ker\{F(S) \to \Gamma\} = \langle\langle R \rangle\rangle$, and the set described above is exactly $\langle\langle R \rangle\rangle$, noting that $gxyg^{-1} = (gxg^{-1})(gyg^{-1})$. □

This tells us the set of elements in S^* that represent the identity is *recursively enumerable*, i.e. there is an algorithm that lists out all words that represent the identity. However, this is not very helpful when we want to identify whether a word represents the identity. If it does, then our algorithm will eventually tell us it is (maybe after 3 trillion years), but if it doesn't, the program will just run forever, and we will never know that it doesn't represent the identity.

It turns out the answer to this question has some geometric interpretations. Let's first look at the example.

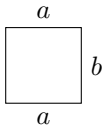
Example. Let $\Gamma = \mathbb{Z}^2 \cong \langle a, b \mid [a, b] \rangle$. Consider the word

$$w = a^n b^n a^{-n} b^{-n}$$

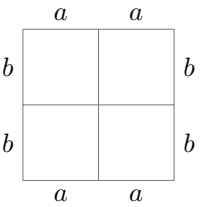
We see that this represents our identity element. But how long would it take for our algorithm to figure this fact out? Equivalently, what d do we need to pick so that w is of the form

$$\prod_{i=1}^{d} g_i r_i^{\pm 1} g_i^{-1}?$$

We can draw a picture. Our element $[a, b]$ looks like



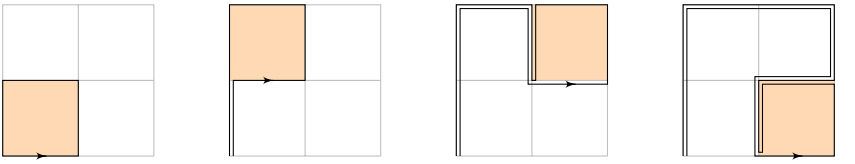
If, say, $n = 2$, then w is given by



If we think about it hard, then we see that what we need to do is to find out how to fill in this big square with the small squares, and we need 4 squares. Indeed, we can read off the sequence

$$w = [a, b] \cdot (b[a, b]b^{-1}) \cdot (b^2ab^{-1}[a, b]ba^{-1}b^{-2}) \cdot (b^2a^2b^{-1}a^{-1}b^{-1}[a, b]baba^{-2}b^{-2}),$$

which corresponds to filling in the square one by one as follows:



Of course, this is fine if we know very well what the Cayley graph of the group looks like, but in general it is quite hard. Indeed, solving the word problem is part of what you do if you want to draw the Cayley graph, since you need to know when two words give the same element.

So how do we solve the word problem? Our previous partial algorithm would make a good, full solution if we knew how far we had to search. If we know that we will only need at most $d = 10^{10^{100}}$, then if we searched for all expressions of the form $\prod_{i=1}^{d} g_i r_i^{\pm} g_i^{-1}$ for $d < 10^{10^{100}}$, and didn't find w, then we know w does not represent the identity element (we will later argue that we don't have to worry about there being infinitely many g_i's to look through).

Definition (Null-homotopic)**.** We say $w \in S^*$ is *null-homotopic* if $w = 1$ in Γ.

Definition (Algebraic area)**.** Let $w \in S^*$ be mull-homotopic. Its *algebraic area* is

$$\mathrm{Area}_{a,\mathcal{P}}(w) = \min\left\{ d : w = \prod_{i=1}^{d} g_i r_i^{\pm 1} g_i^{-1} \right\}.$$

We write the subscript a to emphasize this is the algebraic area. We will later define the geometric area, and we will write it as Area_g. Afterwards, we will show that they are the same, and we will stop writing the subscript.

Let us use $|\cdot|_S$ to denote word length in $F(S)$, while ℓ_S continues to denote the word length in Γ.

Definition (Dehn function)**.** Then *Dehn function* is the function $\delta_{\mathcal{P}} : \mathbb{N} \to \mathbb{N}$ mapping

$$n \mapsto \max\{\mathrm{Area}_{a,\mathcal{P}}(w) \mid |w|_S \leqslant n, w \text{ is null-homotopic}\}.$$

This Dehn function measures the difficulty of the word problem in $\mathcal{P}$.

Proposition. The word problem for $\mathcal{P}$ is solvable iff $\delta_{\mathcal{P}}$ is a computable function.

We will postpone the proof. The hard part of the proof is that we don't have to worry about the infinitely many possible g_i that may be used to conjugate.

It would be really nice if the Dehn function can be viewed as a property of the group, and not the presentation. This requires us to come up with a notion of equivalence relation on functions $\mathbb{N} \to [0, \infty)$ (or $[0, \infty) \to [0, \infty)$).

Definition ($\preccurlyeq$)**.** We write $f \preccurlyeq g$ iff for some $C > 0$, we have

$$f(x) \leqslant Cg(Cx + C) + Cx + C.$$

for all x.

We write $f \approx g$ if $f \preccurlyeq g$ and $g \preccurlyeq f$.

This captures the (super-linear) asymptotic behaviour of f.

Example. For $\alpha, \beta \geqslant 1$, $n^\alpha \preccurlyeq n^\beta$ iff $\alpha \leqslant \beta$.

Proposition. If P and Q are two finite presentations for Γ, then $\delta_\mathcal{P} \approx \delta_\mathcal{Q}$.

We start with two special cases, and then show that we can reduce everything else to these two special cases.

Lemma. If $R' \subseteq \langle\langle R \rangle\rangle$ is a finite set, and

$$\mathcal{P} = \langle S \mid R \rangle, \quad \mathcal{Q} = \langle S \mid R \cup R' \rangle,$$

then $\delta_\mathcal{P} \simeq \delta_\mathcal{Q}$.

Proof. Clearly, $\delta_\mathcal{Q} \leqslant \delta_\mathcal{P}$. Let

$$m = \max_{r' \in R'} \mathrm{Area}_\mathcal{P}(r').$$

It is then easy to see that

$$\mathrm{Area}_\mathcal{P}(w) \leqslant m \, \mathrm{Area}_\mathcal{Q}(w).$$

□

Lemma. Let $\mathcal{P} = \langle S \mid R \rangle$, and let

$$\mathcal{Q} = \langle S \coprod T \mid R \coprod R' \rangle,$$

where

$$R' = \{t w_t^{-1} : t \in T, w_t \in F(S)\}.$$

Then $\delta_\mathcal{P} \approx \delta_\mathcal{Q}$.

Proof. We first show $\delta_\mathcal{P} \preccurlyeq \delta_\mathcal{Q}$. Define

$$\begin{aligned} \rho : F(S \coprod T) &\to F(s) \\ s &\mapsto s \\ t &\mapsto w_t. \end{aligned}$$

In particular, $\rho(r) = r$ for all $r \in R$ and $\rho(r') = 1$ for all $r' \in R'$.

Given $w \in F(S)$, we need to show that

$$\text{Area}_{\mathcal{P}}(w) \leqslant \text{Area}_{\mathcal{Q}}(w).$$

Let $d = \text{Area}_{\mathcal{Q}}(w)$. Then

$$w = \prod_{i=1}^{d} g_i r_i^{\pm 1} g_i^{-1},$$

where $g_i \in F(S \coprod T), r_i \in R \cup R'$. We now apply ρ. Since ρ is a retraction, $\rho(w) = w$. Thus,

$$w = \prod_{i=1}^{d} \rho(g_i)\rho(r_i)^{\pm 1}\rho(g_i)^{-1}.$$

Now $\rho(r_i)$ is either r_i or 1. In the first case, nothing happens, and in the second case, we can just forget the ith term. So we get

$$w = \prod_{i=1, r_i \in R}^{d} \rho(g_i) r_i^{\pm 1} \rho(g_i)^{-1}.$$

Since this is a valid proof in $\mathcal{P}$ that $w = 1$, we know that

$$\text{Area}_{\mathcal{P}}(w) \leqslant d = \text{Area}_{\mathcal{Q}}(w).$$

We next prove that $\delta_{\mathcal{Q}} \preccurlyeq \delta_{\mathcal{P}}$. It is unsurprising that some constants will appear this time, instead of just inequality on the nose. We let

$$C = \max_{t \in T} |w_t|_S.$$

Consider a null-homotopic word $w \in F(S \amalg T)$. This word looks like

$$w = s_{i_1}^{\pm} s_{i_2}^{\pm} \cdots t_{j_1} \cdots s \cdots t \cdots \in F(S \amalg T).$$

We want to turn these t's into s's, and we need to use the relators to do so.

We apply relations from R' to write this as

$$w' = s_{i_1}^{\pm} s_{i_2}^{\pm} \cdots w_{t_{j_1}} \cdots s \cdots w_t \cdots \in F(S).$$

We certainly have $|w'|_S \leqslant C|w|_{S \amalg T}$. With a bit of thought, we see that

$$\begin{aligned} \text{Area}_{\mathcal{Q}}(w) &\leqslant \text{Area}_{\mathcal{P}}(w') + |w|_{S \amalg T} \\ &\leqslant \delta_{\mathcal{P}}(C|w|_{S \amalg T}) + |w|_{S \amalg T}. \end{aligned}$$

So it follows that

$$\delta_{\mathcal{Q}}(n) \leqslant \delta_{\mathcal{P}}(Cn) + n. \qquad \square$$

Proof of proposition. Let $\mathcal{P} = \langle S \mid R \rangle$ and $\mathcal{Q} = \langle S' \mid R' \rangle$. If $\mathcal{P}, \mathcal{Q}$ present isomorphic groups, then we can write

$$s = u_s \in F(S') \text{ for all } s \in S$$

Similarly, we have

$$s' = v_{s'} \in F(S) \text{ for all } s' \in S'$$

We let

$$T = \{su_s^{-1} \mid s \in S\}$$
$$T' = \{s'v_{s'}^{-1} \mid s' \in S'\}$$

We then let

$$\mathcal{M} = \langle S \amalg S' \mid R \cup R' \cup T \cup T' \rangle.$$

We can then apply our lemmas several times to show that $\delta_{\mathcal{P}} \approx \delta_{\mathcal{M}} \approx \delta_{\mathcal{Q}}$. □

Now we can talk about δ_Γ, the Dehn function of Γ, as long as we know we are only talking up to $\approx$. In fact, it is true that

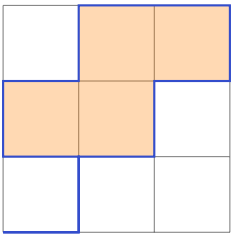
Fact. If $\Gamma_1 \underset{qi}{\simeq} \Gamma_2$, then $\delta_{\Gamma_1} \approx \delta_{\Gamma_2}$.

The proof is rather delicate, and we shall not attempt to reproduce it.

2 Van Kampen diagrams

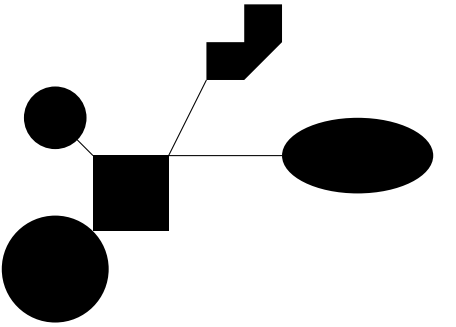
Recall we previously tried to understand the Dehn function of $\mathbb{Z}^2$ in terms of filling in squares on a grid. In this chapter, we pursue this idea further, and come up with an actual theorem.

To make this more precise, recall that in the case of $\mathbb{Z}^2$, we could embed a null-homotopic word into the plane of the form



We saw, at least heuristically, the algebraic area of w is just the geometric area of this figure. Of course, sometimes it is slightly more complicated. For example, if we have the word $w = [a, b]^2$, then we have to use the cell $[a, b]$ twice.

Definition (Singular disc diagram)**.** A *(singular) disc diagram* is a compact, contractible subcomplex of $\mathbb{R}^2$, e.g.



We usually endow the disc diagram D with a base point p, and define

$$\text{Area}_g(D) = \text{number of 2-cells of } D$$

and

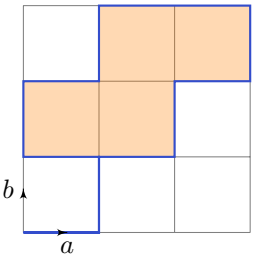
$$\text{Diam}_p(D) = \text{length of the longest embedded path in } D^{(1)}, \text{ starting at } p.$$

If we orient $\mathbb{R}^2$, then D has a well-defined boundary cycle. A disc diagram is *labelled* if each (oriented) edge is labelled by an element $s \in S^{\pm}$. The set of *face labels* is the set of boundary cycles of the 2-cells.

If all the face labels are elements of $R^{\pm}$, then we say D is a *diagram* over $\langle S \mid R \rangle$.

Definition (van Kampen diagram)**.** If $w \in \langle\langle R \rangle\rangle$, and w is the boundary cycle of a singular disc diagram D over the presentation $\langle S \mid R \rangle$, we say that D is a *van Kampen diagram* for w.

Example. The diagram



is a van Kampen diagram for $abababa^{-2}b^{-1}a^{-1}b^{-1}ab^{-1}a^{-1}$.

Note that in this case, the map $S^1 \to X_{\mathcal{P}}$ that represents w factors through a map $D \to X_{\mathcal{P}}$.

Lemma (van Kampen's lemma)**.** Let $\mathcal{P} = \langle S \mid R \rangle$ be a presentation and $w \in S^*$. Then the following are equivalent:

(i) $w = 1$ in Γ presented by $\mathcal{P}$ (i.e. w is null-homotopic)

(ii) There is a van Kampen diagram for w given $\mathcal{P}$.

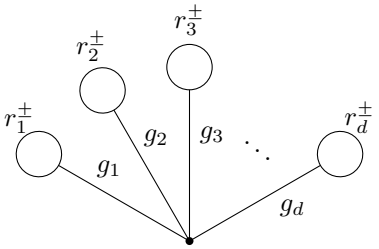
If so, then

$$\mathrm{Area}_a(w) = \min\{\mathrm{Area}_g(D) : D \text{ is a van Kampen diagram for } w \text{ over } \mathcal{P}\}.$$

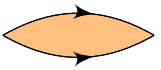
Proof. In one direction, given

$$w = \prod_{i=1}^{n} g_i r_i^{\pm} g_i^{-1} \in F(S)$$

such that $w = 1 \in \Gamma$, we start by writing down a "lollipop diagram"



This defines a diagram for the word $\prod_{i=1}^{n} g_i r_i^{\pm} g_i^{-1}$ which is equal in $F(S)$ to w, but is not exactly w. However, note that performing elementary reductions (or their inverses) correspond to operations on the van Kampen diagram that does not increase the area. We will be careful and not say that the area doesn't change, since we may have to collapse paths that look like



In the other direction, given a diagram D for w, we need to produce an expression

$$w = \prod_{i=1}^{d} g_i r_i^{\pm} g_i^{-1}$$

such that $d \leqslant \text{Area}(D)$.

We let e be the first 2-cell that the boundary curve arrives at, reading anticlockwise around ∂D. Let g be the path from p to e, and let r be the relator read anti-clockwise around e. Let

$$D' = D - e,$$

and let $w' = \partial D'$. Note that

$$w = gr^{\pm 1}g^{-1}w'.$$

Since $\text{Area}(D') = \text{Area}(D) - 1$, we see by induction that

$$w = \prod_{i=1}^{d} g_i r_i^{\pm 1} g_i^{-1}$$

with $d = \text{Area}(D)$. □

We observe that in our algorithm about, the g_i's produced have length $\leqslant \text{Diam}(D)$. So we see that

Corollary. If w is null-homotopic, then we can write w in the form

$$w = \prod_{i=1}^{d} g_i r_i^{\pm} g_i^{-1},$$

where

$$|g_i|_S \leqslant \text{Diam}\, D$$

with D a minimal van Kampen diagram for w. We can further bound this by

$$(\max |r_i|_S)\,\text{Area}(D) + |w|_S \leqslant \text{constant} \cdot \delta_{\mathcal{P}}(|w|_S) + |w|_S.$$

So in particular, if we know $\delta_{\mathcal{P}}(|w|_S)$, then we can bound the maximum length of g_i needed.

It is now easy to prove that

Proposition. The word problem for a presentation $\mathcal{P}$ is solvable iff $\delta_{\mathcal{P}}$ is computable.

Proof.

($\Leftarrow$) By the corollary, the maximum length of a conjugator g_i that we need to consider is computable. Therefore we know how long the partial algorithm needs to run for.

($\Rightarrow$) To compute $\delta_{\mathcal{P}}(n)$, we use the word problem solving ability to find all null-homotopic words in $F(S)$ of length $\leqslant n$. Then for each d, go out and look for expressions

$$w = \prod_{i=1}^{d} g_i r_i^{\pm 1} g_i^{-1}.$$

A naive search would find the smallest area expression, and this gives us the Dehn function.

□

It is a hard theorem that

Theorem (Novikov–Boone theorem)**.** There exists a finitely-presented group with an unsolvable word problem.

Corollary. $\delta_{\mathcal{P}}$ is sometimes non-computable.

Let's look at some applications to geometry. In geometry, people wanted to classify manifolds. Classifying (orientable) 2-dimensional manifolds is easy. They are completely labeled by the genus. This classification can in fact be performed by a computer. If we manage to triangulate our manifold, then we can feed the information of the triangulation into a computer, and then the computer can compute the Euler characteristic.

Is it possible to do this in higher dimensions? It turns out the word problem gives a severe hindrance to doing so.

Theorem. Let $n \geqslant 4$ and $\Gamma = \langle S \mid R\rangle$ be a finitely-presented group. Then we can construct a closed, smooth, orientable manifold M^n such that $\pi_1 M \cong \Gamma$.

This is a nice, little surgery argument.

Proof. Let $S = \{a_1, \dots, a_m\}$ and $R = \{r_1, \dots, r_n\}$. We start with

$$M_0 = \#_{i=0}^m (S^1 \times S^{n-1}).$$

Note that when we perform this construction, as $n \geqslant 3$, we have

$$\pi_1 M_0 \cong F_m$$

by Seifert van-Kampen theorem. We now construct M_k from M_{k-1} such that

$$\pi_1 M_k \cong \langle a_1, \dots, a_m \mid r_1, \dots, r_k\rangle.$$

We realize r_k as a loop in M_{k-1}. Because $n \geqslant 3$, we may assume (after a small homotopy) that this is represented by a smooth embedded map $r_k : S^1 \to M_{k-1}$.

We take N_k to be a smooth tubular neighbourhood of r_k. Then $N_k \cong S^1 \times D^{n-1} \subseteq M_{k-1}$. Note that $\partial N_k \cong S^1 \times S^{n-2}$.

Let $U_k = D^2 \times S^{n-2}$. Notice that $\partial U_k \cong \partial N_k$. Since $n \geqslant 4$, we know U_k is simply connected. So we let

$$M'_{k-1} = M_k \backslash \mathring{N}_k,$$

a manifold with boundary $S^1 \times S^{n-2}$. Choose an orientation-reversing diffeomorphism $\varphi_k : \partial U_k \to \partial M'_{k-1}$. Let

$$M_k = M'_{k-1} \cup_{\varphi_k} U_k.$$

Then by applying Seifert van Kampen repeatedly, we see that

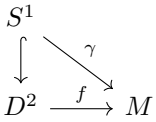
$$\pi_1 M_k = \pi_1 M_{k-1} / \langle\langle r_k \rangle\rangle,$$

as desired. □

Thus, if we have an algorithm that can distinguish between the fundamental groups of manifolds, then that would solve (some variant of) the word problem for us, which is impossible.

Finally, we provide some (even more) geometric interpretation of the Dehn function.

Definition (Filling disc)**.** Let (M, g) be a closed Riemannian manifold. Let $\gamma : S^1 \to M$ be a smooth null-homotopic loop. A filling disk for γ in M is a smooth map $f : D^2 \to M$ such that the diagram



Since there is a metric g on M, we can pull it back to a (possibly degenerate) metric f^*g on D^2, and hence measure quantities like the length $\ell(\gamma)$ of γ and the area $\mathrm{Area}(f)$ of D^2.

A classic result is

Theorem (Douglas, Radú, Murray)**.** If γ is embedded, then there is a least-area filling disc.

So we can define

Definition (FArea)**.**

$$\mathrm{FArea}(\gamma) = \inf\{\mathrm{Area}(f) \mid f : D^2 \to M \text{ is a filling disc for } \gamma\}.$$

Definition (Isoperimetric function)**.** The isoperimetric function of (M, g) is

$$\begin{aligned} FiU^M &: [0, \infty) \to [0, \infty) \\ \ell &\mapsto \sup\{\mathrm{FArea}(\gamma) : \gamma : S^1 \to M \text{ is a smooth null-homotopic loop, } \ell(\gamma) \leqslant \ell\} \end{aligned}$$

Theorem (Filling theorem)**.** let M be a closed Riemannian manifold. Then $FiU^M \simeq \delta_{\pi_1 M}$.

By our previous construction, we know this applies to every finitely-presented group.

3 New groups from old: A crash course in Bass–Serre theory

The idea is that we want to be able to do things similar to how we built new manifolds by gluing together old manifolds.

3.1 Graphs of spaces

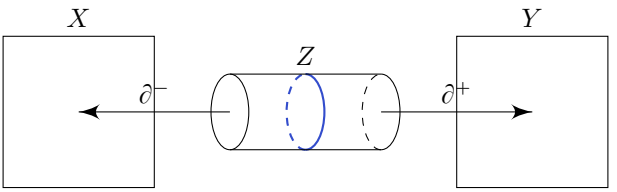
Disclaimer: All spaces are cell complexes.

We want to build new spaces by gluing well-understood spaces together via homotopy pushouts.

Definition (Homotopy pushout)**.** Let X, Y, Z be spaces, and $\partial^- : Z \to X$ and $\partial^+ : Z \to Y$ be maps. We define

$$X \coprod_Z Y = (X \coprod Y \coprod Z \times [-1, 1])/\sim,$$

where we identify $\partial^\pm(z) \sim (z, \pm 1)$ for all $z \in Z$.



By Seifert van-Kampen, we know $\pi_1(X \coprod_Z Y)$ is the pushout

$$\begin{array}{ccc} \pi_1 Z & \xrightarrow{\partial^+_*} & X] \\ \downarrow{\scriptstyle \partial^-_*} & & \downarrow \\ \pi_1 Y & \longrightarrow & \pi_1(X \coprod_Z Y) \end{array}$$

If $\pi^\pm_*$ are injective, then we have

$$\pi_1(X \coprod_Z Y) \cong \pi_1 X \underset{\pi_1 Z}{*} \pi_1 Y.$$

We want to generalize this.

Definition (Graph of spaces)**.** A *graph of spaces* $\mathcal{X}$ consists of the following data

- A connected graph Ξ.
- For each vertex $v \in V(\Xi)$, a path-connected space X_v.
- For each edge $e \in E(\Xi)$, a path-connected space X_e.
- For each edge $e \in E(\Xi)$ attached to $v^\pm \in V(\Xi)$, we have π_1-injective maps $\partial^\pm_e : X_e \to X_{v^\pm}$.

The *realization* of $\mathcal{X}$ is

$$|\mathcal{X}| = X = \coprod_{v\in V(\Xi)} X_v \amalg \coprod_{e\in E(\Xi)} (X_e \times [-1,]1)/(\forall e \in E(\Xi), \forall x \in X_e, (x,\pm 1) \sim \partial_e^{\pm}(x)).$$

These conditions are not too restrictive. If our vertex or edge space were not path-connected, then we can just treat each path component as a separate vertex/edge. If our maps are not π_1 injective, as long as we are careful enough, we can attach 2-cells to kill the relevant loops.

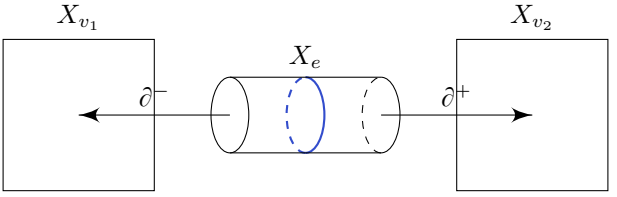
Example. Let X be an orientable surface and Y be a multi-curve, i.e. a disjoint union of essentially simple closed curves on X. Take a tubular neighbourhood of each curve in Y, and let $N(Y)$ be the union of all these tubular neighbourhoods, which is a union of closed annuli. Let

$$\coprod X_i = X - \mathring{N}(Y)$$

be the decomposition of $X - \mathring{N}(Y)$ into connected components.

We can then take $\coprod X_e = Y$. Then X is just the realization of this graph, when we pick the appropriate attaching maps.

Example. In the case of



Recall that Seifert–van Kampen tells us

$$\pi_1 X \cong \pi_1 X_{v_1} \underset{\pi_1 X_e}{*} \pi_1 X_{v_2}.$$

Example. If we instead have

then if we think carefully, we find that we have

$$\pi_1 X = \frac{\pi_1 X_v * \langle t \rangle}{\langle\langle (\partial_e^+)_*(g) = t(\partial_e^-)_*(g)t^{-1} \forall g \in \pi_1 X_e \rangle\rangle}.$$

Here t represents a loop that starts at X_v, goes around X_e, and returns to the original point in X_v.

This is known as an *HNN extension*. The way to think about this is as follows — we have a group $\pi_1 X_v$, and we have two subgroups that are isomorphic to each other. Then the HNN extension is the "free-est" way to modify the group so that these two subgroups are conjugate.

How about for a general graph of spaces? If Ξ is a graph of spaces, then its fundamental group has the structure of a *graph of groups* $\mathcal{G}$.

Definition (Graph of groups)**.** A *graph of groups* $\mathcal{G}$ consists of

– A graph Γ

- Groups G_v for all $v \in V(\Gamma)$
- Groups G_e for all $e \in E(\Gamma)$
- For each $e \mapsto v^{\pm}(e)$, injective group homomorphisms

$$\partial_e^{\pm} : G_e \to G_{v^{\pm}(e)}.$$

In the case of a graph of spaces, it was easy to define a realization. One way to do so is that we have already see how to do so in the two above simple cases, and we can build the general case up inductively from the simple cases, but that is not so canonical. However, this has a whole lot of choices involved. Instead, we are just going to do is to associate to a graph of groups $\mathcal{G}$ a graph of spaces $\mathcal{X}$ which "inverts" the natural map from graphs of spaces to graphs of groups, given by taking π_1 of everything.

This can be done by taking Eilenberg–Maclane spaces.

Definition (Aspherical space)**.** A space X is *aspherical* if $\tilde{X}$ is contractible. By Whitehead's theorem and the lifting criterion, this is true iff $\pi_n(X) = 0$ for all $n \geqslant 2$.

Proposition. For all groups G there exists an aspherical space $BG = K(G, 1)$ such that $\pi_1(K(G, 1)) \cong G$. Moreover, for any two choices of $K(G, 1)$ and $K(H, 1)$, and for every homomorphism $f : G \to H$, there is a unique map (up to homotopy) $\bar{f} : K(G, 1) \to K(H, 1)$ that induces this homomorphism on π_1. In particular, $K(G, 1)$ is well-defined up to homotopy equivalence.

Moreover, we can choose $K(G, 1)$ functorially, namely there are choices of $K(G, 1)$ for each G and choices of $\bar{f}$ such that $\overline{f_1 \circ f_2} = \bar{f}_1 \circ \bar{f}_2$ and $\overline{\mathrm{id}_G} = \mathrm{id}_{K(G,1)}$ for all f, G, H.

When we talked about presentations, we saw that this is true if we don't have the word "aspherical". But the aspherical requirement makes the space unique.

These $K(G, 1)$ are known as *Eilenberg–MacLane spaces*

Proof. See Hatcher (or Segal (or May)). □

Using Eilenberg–MacLane spaces, given any graph of groups $\mathcal{G}$, we can construct a graph of spaces Ξ such that when we apply π_1 to all the spaces in $\mathcal{X}$, we recover $\mathcal{G}$.

We can now set

$$\pi_1\mathcal{G} = \pi_1|\mathcal{X}|.$$

For more details, read *Trees* by Serre, or *Topological methods in group theory* by Scott and Wall.

Note that if Γ is finite, and all the G_v's are finitely-generated, then $\pi_1\mathcal{G}$ is also finitely-generated, which one can see by looking into the construction of $K(G, 1)$. If Γ is finite, all G_v's are finitely-presented and all G_e's are finitely-generated, then $\pi_1\mathcal{G}$ is finitely-presented.

3.2 The Bass–Serre tree

Given a graph of groups $\mathcal{G}$, we want to analyze $\pi_1\mathcal{G} = \pi_1 X$, and we do this by the natural action of G on $\tilde{X}$ by deck transformations.

Recall that to understand the free group, we looked at the universal cover of the rose with r petals. Since the rose was a graph, the universal cover is also a graph, and because it is simply connected, it must be a tree, and this gives us a normal form theorem. The key result was that the covering space of a graph is a graph.

Lemma. If $\mathcal{X}$ is a graph of spaces and $\hat{X} \to X$ is a covering map, then $\hat{X}$ naturally has the structure of a graph of spaces $\hat{\mathcal{X}}$, and p respects that structure.

Note that the underlying graph of $\hat{\mathcal{X}}$ is not necessarily a covering space of the underlying graph of $\mathcal{X}$.

Proof sketch. Consider

$$\bigcup_{v \in V(\Xi)} X_V \subseteq X.$$

Let

$$p^{-1}\left(\bigcup_{v \in V(\Xi)} X_V\right) = \coprod_{\hat{v} \in V(\hat{\Xi})} \hat{W}_{\hat{v}}.$$

This defines the vertices of $\hat{\Xi}$, the underlying graph of $\hat{\mathcal{X}}$. The path components $\hat{X}_{\hat{v}}$ are going to be the vertex spaces of $\hat{\mathcal{X}}$. Note that for each $\hat{v}$, there exists a unique $v \in V(\Xi)$ such that $p : \hat{X}_{\hat{v}} \to X_v$ is a covering map.

Likewise, the path components of

$$p^{-1}\left(\bigcup_{e \in E(\Xi)} X_e \times \{0\}\right)$$

form the edge spaces $\coprod_{e \in E(\Xi)} \hat{X}_{\hat{e}}$ of $\hat{\mathcal{X}}$, which again are covering spaces of the edge space of X.

Now let's define the edge maps $\partial^{\pm}_{\hat{e}}$ for each $\hat{e} \in E(\hat{\Xi}) \mapsto e \in E(\Xi)$. To do so, we consider the diagram

$$\begin{array}{ccccc}
\hat{X}_{\hat{e}} & \xhookrightarrow{\sim} & \hat{X}_{\hat{e}} \times [-1,1] & \dashrightarrow & \hat{X} \\
\downarrow & & \downarrow & & \downarrow p \\
X_e & \xhookrightarrow{\sim} & X_e \times [-1,1] & \longrightarrow & X
\end{array}$$

By the lifting criterion, for the dashed map to exist, there is a necessary and sufficient condition on $(\hat{X}_{\hat{e}} \times [-1,1] \to X_e \times [-1,1] \to X)_*$. But since this condition is homotopy invariant, we can check it on the composition $(\hat{X}_{\hat{e}} \to X_e \to X)_*$ instead, and we know it must be satisfied because a lift exists in this case.

The attaching maps $\partial^{\pm}_{\hat{e}} : \hat{X}_e \to \hat{X}$ are precisely the restriction to $\hat{X}_e \times \{\pm 1\} \to \hat{X}$.

Finally, check using covering space theory that the maps $\hat{X}_{\hat{e}} \times [-1, 1] \to \hat{X}$ can injective on the interior of the cylinder, and verify that the appropriate maps are π_1-injective. □

Now let's apply this to the universal cover $\tilde{X} \to X$. We see that $\tilde{X}$ has a natural action of $G = \pi_1 X$, which preserves the graph of spaces structure.

Note that for any graph of spaces X, there are maps

$$\iota : \Xi \to X$$
$$\rho : X \to \Xi$$

such that $\rho \circ \iota \simeq \mathrm{id}_\Xi$. In particular, this implies ρ_* is surjective (and of course ρ itself is also surjective).

In the case of the universal cover $\tilde{X}$, we see that the underlying graph $\tilde{\Xi} = T$ is connected and simply connected, because $\pi_1 \tilde{\Xi} = \rho_*(\pi_1 \tilde{X}) = 1$. So it is a tree!

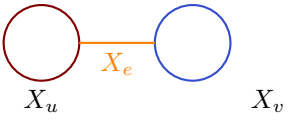
The action of G on $\tilde{X}$ descends to an action of G on $\tilde{\Xi}$. So whenever we have a graph of spaces, or a graph of groups $\mathcal{G}$, we have an action of the fundamental group on a tree. This tree is called the *Bass–Serre tree* of $\mathcal{G}$.

Just like the case of the free group, careful analysis of the Bass–Serre tree leads to theorems relating $\pi_1(\mathcal{G})$ to the vertex groups G_v and edge groups G_e.

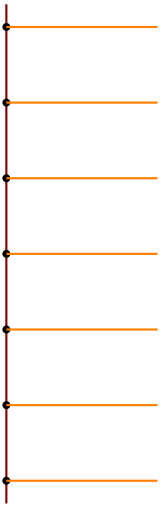
Example. Let

$$G = F_2 = \langle a \rangle * \langle b \rangle = \mathbb{Z} * \mathbb{Z}.$$

In this case, we take X to be



Here we view this as a graph where the two vertex spaces are circles, and there is a single edge connecting them. The covering space $\tilde{X}$ then looks like



This is *not* the Bass–Serre tree. The Bass–Serre tree looks like

Thus, this is an infinite-valent bipartite tree.

The simply-connected of the Bass–Serre tree translates into a normal form for the elements of G.

This point of view gives two important results that relate elements of G to the vertex groups G_{v_i} and the edge maps $\partial^{\pm}_{e_j} : G_{e_j} \to G_{v_j^{\pm}}$.

Lemma (Britton's lemma)**.** For any vertex Ξ, the natural map $G_v \to G$ is injective.

Unsurprisingly, this really requires that the edge maps are injective. It is an exercise to find examples to show that this fails if the boundary maps are not injective.

Proof sketch. Observe that the universal cover $\tilde{X}$ can be produce by first building universal covers of the vertex space, which are then glued together in a way that doesn't kill the fundamental groups. □

The next theorem is a normal form theorem. Pick a base vertex $v \in V(\Xi)$. We can then represent elements of G via loops in $\Xi = \Gamma$, the underlying graph based at v and labelled by elements of vertex groups. These loops can be written formally as

$$\gamma = g_0 e_1^{\pm 1} g_1 e_2^{\pm 1} \cdots e_n^{\pm 1} g_n$$

We say a *pinch* is a sub-word of the form

$$e^{\pm 1} \partial_e^{\pm 1}(g) e^{\mp 1},$$

which can be replaced by $\partial_e^{\mp 1}(g)$.

We say a loop is *reduced* if it contains no pinches.

Theorem (Normal form theorem)**.** Every element can be represented by a reduced loop, and the only reduced loop representing the identity is the trivial loop.

This is good enough, since if we can recognize is something is the identity, then we see if the product of one with the inverse of the other is the identity. An exact normal form for all words would be tricky.

Proof idea. It all boils down to the fact that the Bass–Serre tree is a tree. Connectedness gives the existence, and the simply-connectedness gives the "uniqueness". □

4 Hyperbolic groups

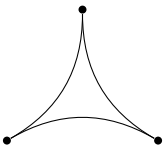
So far, we obtained a lot of information about groups by seeing how they act on trees. Philosophically, the reason for this is that trees are very "negatively curved" spaces. We shall see in this chapter than in general, whenever a group acts on a negatively curved space, we can learn a lot about the group.

4.1 Definitions and examples

We now want to define a negatively-curved space in great generality. Let X be a geodesic metric space. Given $x, y \in X$, we will write $[x, y]$ for a choice of geodesic between x and y.

Definition (Geodesic triangle)**.** A *geodesic triangle* Δ is a choice of three points x, y, z and geodesics $[x, y], [y, z], [z, x]$.

Geodesic triangles look like this:



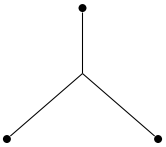
Note that in general, the geodesics may intersect.

Definition (δ-slim triangle)**.** We say Δ is δ-slim if every side of Δ is contained in the union of the δ-neighbourhoods of the other two sides.

Definition (Hyperbolic space)**.** A metric space is *(Gromov) hyperbolic* if there exists $\delta \geqslant 0$ such that every geodesic triangle in X is δ-slim. In this case, we say it *is δ-hyperbolic.*

Example. $\mathbb{R}^2$ is not Gromov hyperbolic.

Example. If X is a tree, then X is 0-hyperbolic! Indeed, each triangle looks like



We call this a *tripod*.

Unfortunately, none of these examples really justify why we call these things hyperbolic. Let's look at the actual motivating example.

Example. Let $X = \mathbb{H}^2$, the *hyperbolic plane*. Let $\Delta \subseteq \mathbb{H}^2$ be a triangle. Then Δ is δ-slim, where the maximum radius of an inscribed semi-circle D in Δ with the center on one of the edges.

But we know that the radius of D is bounded by some increasing function of the area of D, and the area of D is bounded above by the area of Δ. On the other hand, by hyperbolic geometry, we know the area of *any* triangle is bounded by π.

So $\mathbb{H}^2$ is δ-hyperbolic for some δ.

For the arguments we are going to do, we don't really care about what δ is.

Example. Let X be any bounded metric space, e.g. S^2. Then X is Gromov hyperbolic, since we can just take δ to be the diameter of the metric space.

This is rather silly, but it makes sense if we take the "coarse point of view", and we have to ignore bounded things.

What we would like to do is to say is that a group Γ if for every finite generating set S, the Cayley graph $\mathrm{Cay}_S(\Gamma)$ equipped with the word metric is δ-hyperbolic for some δ.

However, this is not very helpful, since we have to check it for all finite generating sets. So we want to say that being hyperbolic is quasi-isometry invariant, in some sense.

This is slightly difficult, because we lose control of how the geodesic behaves if we only look at things up to isometry. To do so, we have to talk about quasi-geodesics.

4.2 Quasi-geodesics and hyperbolicity

Definition (quasi-geodesic)**.** A (λ, ε)-quasi-geodesic for $\lambda \geqslant 1, \varepsilon \geqslant 0$ is a (λ, ε)-quasi-isometric embedding $I \to X$, where $I \subseteq \mathbb{R}$ is a closed interval.

Note that our definition allows for I to be unbounded, i.e. I may be of the forms $[a, b]$, $[0, \infty)$ or $\mathbb{R}$ respectively. We call these *quasi-geodesic intervals*, *quasi-geodesic rays* and *quasi-geodesic lines* respectively.

Example. The map $[0, \infty) \to \mathbb{R}^2$ given in polar coordinates by

$$t \mapsto (t, \log(1 + t))$$

is a quasigeodesic ray.

This should be alarming. The quasi-geodesics in the Euclidean plane can look unlike any genuine geodesic.

Theorem (Morse lemma)**.** For all $\delta \geqslant 0, \lambda \geqslant 1$ there is $R(\delta, \lambda, \varepsilon)$ such that the following holds:

If X is a δ-hyperbolic metric space, and $c : [a, b] \to X$ is a (λ, ε)-quasigeodesic from p to q, and $[p, q]$ is a choice of geodesic from p to q, then

$$d_{Haus}([p, q], \mathrm{im}(c)) \leqslant R(\delta, \lambda, \varepsilon),$$

where

$$d_{Haus}(A, B) = \inf\{\varepsilon > 0 \mid A \subseteq N_\varepsilon(B) \text{ and } B \subseteq N_\varepsilon(A)\}$$

is the *Hausdorff distance*.

This has nothing to do with the Morse lemma in differential topology.

Corollary. There is an $M(\delta, \lambda, \varepsilon)$ such that a geodesic metric space X is δ-hyperbolic iff any (λ, ε)-quasigeodesic triangle is M-slim.

Corollary. Suppose X, X' are geodesic metric spaces, and $f : X \to X''$ is a quasi-isometric embedding. If X' is hyperbolic, then so is X.

In particular, hyperbolicity is a quasi-isometrically invariant property, when restricted to geodesic metric spaces.

Thus, we can make the following definition:

Definition (Hyperbolic group)**.** A group Γ is *hyperbolic* if it acts properly discontinuously and cocompactly by isometries on a proper, geodesic hyperbolic metric space. Equivalently, if it is finitely-generated and with a hyperbolic Cayley graph.

Let's prove the Morse lemma. To do so, we need the following definition:

Definition (Length of path)**.** Let $c : [a, b] \to X$ be a continuous path. Let $a = t_0 < t_1 < \cdots < t_n = b$ be a *dissection* $\mathcal{D}$ of $[a, b]$. Define

$$\ell(c) = \sup_{\mathcal{D}} \sum_{i=1}^{n} d(c(t_{i-1}), c(t_i)).$$

In general, this number is extremely hard to compute, and may even be finite.

Definition (Rectifiable path)**.** We say a path c is *rectifiable* if $\ell(c) < \infty$.

Example. Piecewise geodesic paths are rectifiable.

Lemma. Let X be a geodesic space. For any (λ, ε)-quasigeodesic $c : [a, b] \to X$, there exists a continuous, rectifiable (λ, ε')-quasigeodesic $c'[a, b] \to X$ with $\varepsilon' = 2(\lambda + \varepsilon)$ such that

(i) $c'(a) = c(a)$, $c'(b) = c(b)$.

(ii) For all $a \leqslant t < t' \leqslant b$, we have

$$\ell(c'|_{[t,t']}) \leqslant k_1 d(c'(t), c'(t')) + k_2$$

where $k_1 = \lambda(\lambda + \varepsilon)$ and $k_2 = (\lambda\varepsilon' + 3)(\lambda + 3)$.

(iii) $d_{Haus}(\operatorname{im} c, \operatorname{im} c') \leqslant \lambda + \varepsilon$.

Proof sketch. Let $\Sigma = \{a, b\} \cup ((a, b) \cap \mathbb{Z})$. For $t \in \Sigma$, we let $c'(t) = c(t)$, and define c' to be geodesic between the points of Σ. Then claims (i) and (iii) are clear, and to prove quasigeodesicity and (ii), let $\sigma : [a, b] \to \Sigma$ be a choice of closest point in Σ, and then estimate $d(c'(t), c'(t'))$ in terms of $d(c(\sigma(t)), c(\sigma(t')))$. □

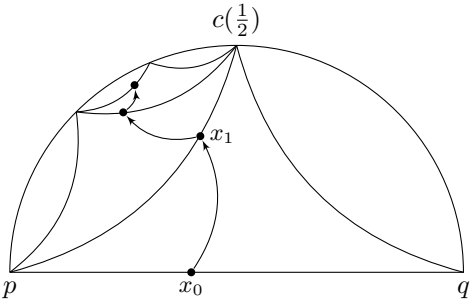
Lemma. let X be δ-hyperbolic, and $c : [a, b] \to X$ a continuous, rectifiable path in X joining p to q. For $[p, q]$ a geodesic, for any $x \in [p, q]$, we have

$$d(x, \operatorname{im} c) \leqslant \delta |\log_2 \ell(c)| + 1.$$

Proof. We may assume $c : [0, 1] \to X$ is parametrized proportional to arc length. Suppose

$$\frac{\ell(c)}{2^N} < 1 \leqslant \frac{\ell(c)}{2^{N-1}}.$$

Let $x_0 = x$. Pick a geodesic triangle between $p, q, c(\frac{1}{2})$. By δ-hyperbolicity, there exists a point x_1 lying on the other two edges such that $d(x_0, x_1) \leqslant \delta$. We wlog assume $x_1 \in [p, c(\frac{1}{2})]$. We can repeat the argument with $c|_{[0,\frac{1}{2}]}$.



Formally, we proceed by induction on N. If $N = 0$ so that $\ell(c) < 1$, then we are done by taking desired point on $\operatorname{im} c$ to be p (or q). Otherwise, there is some $x_1 \in [p, c(\frac{1}{2})]$ such that $d(x_0, x_1) \leqslant \delta$. Then

$$\frac{\ell(c|_{[0,\frac{1}{2}]})}{2^{N-1}} < 1 \leqslant \frac{\ell(c|_{[0,\frac{1}{2}]})}{2^{N-2}}.$$

So by the induction hypothesis,

$$\begin{aligned} d(x_1, \operatorname{im} c) &\leqslant \delta |\log_2 \ell(c|_{[0,\frac{1}{2}]}| + 1 \\ &= \delta \left|\frac{1}{2}\log_2 \ell(c)\right| + 1 \\ &= \delta(|\log_2 \ell(c)| - 1) + 1. \end{aligned}$$

Note that we used the fact that $\ell(c) > 1$, so that $\log_2 \ell(c) > 0$.

Then we are done since

$$d(x, \operatorname{im} c) \leqslant d(x, x_1) + d(x_1, \operatorname{im} c).$$

□

Proof of Morse lemma. By the first lemma, we may assume that c is continuous and rectifiable, and satisfies the properties as in the lemma.

First, we show that $[p, q]$ is contained in a bounded neighbourhood of $\operatorname{im} c$. Let

$$D = \sup\{d(x, \operatorname{im} c) \mid x \in [p, q]\}.$$

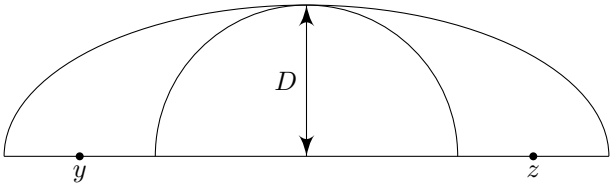
Since $[p, q]$ and $\operatorname{im} c$ are compact, we can let x_0 be a point where the supremum is obtained.

Let p, q be the end points of c, and $[p, q]$ a geodesic. First we show that $[p, q]$ is contained in a bounded neighbourhood of $\operatorname{im} c$. Let

$$D = \sup_{x \in [p,q]} d(x, \operatorname{im} c).$$

By compactness of the interval, let $x_0 \in [p, q]$ where the supremum is attained. Then by assumption, $\operatorname{im} c$ lies outside $\mathring{B}(x_0, D)$. Choose $y, z \in [p, q]$ be such that $d(x_0, y) = d(x_0, z) = 2D$ and y, x_0, z appear on the geodesic in this order (take $y = p$, or $z = q$ if that is not possible).

Let $y' = c(s) \in \operatorname{im} c$ be such that $d(y', y) \leqslant D$, and similarly let $z' = c(t) \in \operatorname{im} c$ be such that $d(z, z') \leqslant D$.



Let $\gamma = [\gamma, \gamma'] \cdot c|_{[s,t]} \cdot [z', z]$. Then

$$\ell(\gamma) = d(y, y') + d(z, z') + \ell(c|_{[s,t]}) \leqslant D + D + k_1 d(y', z') + k_2,$$

by assumption. Also, we know that $d(y', z') \leqslant 6D$. So we have

$$\ell(\gamma) \leqslant 6k_1 D + 2D + k_2.$$

But we know that

$$d(x_0, \operatorname{im} \gamma) = D.$$

So the second lemma tells us

$$D \leqslant \delta |\log_2(6k_1 D + 2K + k_2)| + 1.$$

The left hand side is linear in D, and the right hand side is a logarithm in D. So it must be the case that D is bounded. Hence $[p, q] \subseteq N_{D_0}(\operatorname{im} c)$, where D_0 is some constant.

It remains to find a bound M such that $\operatorname{im} c \subseteq N_M([p, q])$. Let $[a', b']$ be a maximal subinterval of $[a, b]$ such that $c[a', b']$ lies entirely outside $\mathring{N}_{D_0}([p, q])$. Since $\bar{N}_D(c[a, a'])$ and $\bar{N}_D(c[b', b])$ are both closed, and they collectively cover the connected set $[p, q]$, there exists

$$w \in [p, q] \cap \bar{N}_{D_0}(c[a, a']) \cap \bar{N}_{D_0}(c[b', b]).$$

Therefore there exists $t \in [a, a']$ and $t' \in [b', b]$ such that $d(w, c(t)) \leqslant D_0$ and $d(w, c(t')) \leqslant D_0$. In particular, $d(c(t), c(t')) \leqslant 2D_0$.

By the first lemma, we know

$$\ell(c|_{[t,t']}) \leqslant 2k_1 D_0 + k_2.$$

So we know that for $s \in [a', b']$, we have

$$d(c(s), [p, q]) \leqslant d(c(s), w) \leqslant d(c(s), c(t)) + d(c(t), w) \leqslant \ell(c|_{[t,t']}) + D_0 \leqslant D_0 + 2k_1 D_0 +$$

So we are done. □

With the Morse lemma proven, we can sensibly talk about hyperbolic groups. We have already seen many examples of hyperbolic spaces, such as trees and the hyperbolic plane.

Example. Any finite group is hyperbolic since they have bounded Cayley graphs.

Example. Finitely-generated free groups are hyperbolic, since their Cayley graphs are trees.

Example. Surface groups Σ_g for $g \geqslant 2$ acts on the hyperbolic plane properly discontinuously and cocompactly by isometries (since the universal cover of the genus g surface is the hyperbolic plane by tessellating the hyperbolic plane with $4g$-gons, c.f. IB Geometry).

Definition (virtually-P)**.** Let P be a property of groups. Then we say G is virtually-P if there is a finite index subgroup $G_0 \leqslant G$ such that G_0 is P.

Example. Finite groups are virtually trivial!

Note that $G_0 \leqslant G$ is finite-index, then G_0 acts on $\mathrm{Cay}_S(G)$ in a way that satisfies the properties of the Schwarz–Milnor lemma (the only property we may worry about is cocompactness, which is what we need finite-index for) and hence $G_0 \underset{qi}{\simeq} G$. In particular,

Example. A virtually hyperbolic group is hyperbolic. For example, virtually free groups such that $(\mathbb{Z}/2) * (\mathbb{Z}/3) = \mathrm{PSL}_2\mathbb{Z}$ are hyperbolic.

These are some nice classes of examples, but they will be dwarfed by our next class of examples.

A "random group" is hyperbolic. More precisely, fix $m \geqslant 2$ and $n \geqslant 1$. We temporarily fix $\ell \geqslant 0$. Consider groups of the form

$$\Gamma = \langle a_1, \dots, a_m \mid r_1, \dots, r_n \rangle.$$

such that r_i are all *cyclicly reduced* words of length ℓ. Put the uniform probability distribution on the set of all such groups. This defines a group-valued random variable Γ_ℓ. For a property P, we say that "a random group is P" if

$$\lim_{\ell \to \infty} \mathbb{P}(\Gamma_\ell \text{ is hyperbolic}) = 1$$

for all n, m.

Theorem (Gromov)**.** A random group is infinite and hyperbolic.

There are, of course, other ways to define a "random group". As long as we control the number of relations well so that we don't end up with finite groups all the time, the "random group" should still be hyperbolic.

Recall that in differential geometry, geodesics are defined locally. However, we defined our geodesics to be embedded interval, which is necessarily a global notion. We want an analogous local version. However, if we want to work up to quasi-isomorphism, then we cannot go completely local, because locally, you are allowed to be anything.

Definition (k-local geodesic)**.** Let X be a geodesic metric space, and $k > 0$. A path $c : [a, b] \to X$ is a *k-local geodesic* if

$$d(c(t), c(t')) = |t - t'|$$

for all $t, t' \in [a, b]$ with $|t - t'| \leqslant k$.

Theorem. Let X be δ-hyperbolic and $c : [a, b] \to X$ be a k-local geodesic where $k > 8\delta$. Then c is a (λ, ε)-quasigeodesic for $\lambda = \lambda(\delta, k)$ and $\varepsilon = \varepsilon(\delta, k)$.

First, we prove

Lemma. Let X be δ-hyperbolic and $k > 8\delta$. If $c : [a, b] \to X$ is a k-local geodesic, then $\operatorname{im} c$ is contained in the 2δ-neighbourhood $[c(a), c(b)]$.

Observe that by iterating the definition of hyperbolicity, on a δ-hyperbolic space, any point on an n-gon is at most $(n-2)\delta$ away from a point on another side.

Proof. Let $x = c(t)$ maximize $d(x, [c(a), c(b)])$. Let

$$y = c\left(t - \frac{k}{2}\right), \quad z = c\left(t - \frac{k}{2}\right).$$

If $t - \frac{k}{2} < a$, we set $y = c(a)$ instead, and similarly for z.

Let $y' \in [c(a), c(b)]$ minimize $d(y, y')$, and likewise let $z' \in [c(a), c(b)]$ minimize $d(z, z')$.

Fix geodesics $[y, y']$ and $[z, z']$. Then we have a geodesic rectangle with vertices y, y', z, z'. By δ-hyperbolicity, there exists w on the rectangle *not* on $\operatorname{im} c$, such that $d(x, w) = 2\delta$.

If $w \in [y', z']$, then we win. Otherwise, we may wlog assume $w \in [y, y']$. Note that in the case $y = c(a)$, we must have $y = y'$, and so this would imply $w = y \in [c(a), c(b)]$. So we are only worried about the case $y = c\left(t - \frac{k}{2}\right)$. So $d(y, x) > 4\delta$. But then by the triangle inequality, we must have $d(y, w) > 2\delta$.

However,

$$d(x, y') \leqslant d(x, w) + d(w, y') < d(y, w) + d(w, y') = d(y, y').$$

So it follows that

$$d(x, [c(a), c(b)]) < d(y, y') = d(y, [c(a), c(b)]).$$

This contradicts our choice of c. □

Proof of theorem. Let $c : [a, b] \to X$ be a k-local geodesic, and $t \leqslant t' \in [a, b]$. Choose $t_0 = t < t_1 < \cdots < t_n < t'$ such that $t_i = t_{i-1} + k$ for all i and $t' - t_n < k$.

Then by definition, we have

$$d(c(t_{i-1}), c(t_i)) = k, \quad d(c(t_n), c(t')) = |t_n - t'|.$$

for all i. So by the triangle inequality, we have

$$d(c(t), c(t')) \leqslant \sum_{i=1}^{n} d(c(t_{i-1}), c(t_i)) + d(t_n, t') = |t - t'|.$$

We now have to establish a coarse *lower* bound on $d(c(t), c(t'))$.

We may wlog assume $t = a$ and $t' = b$. We need to show that

$$d(c(a), c(b)) \geqslant \frac{1}{\lambda}|b - a| - \varepsilon.$$

We divide c up into regular subintervals $[x_i, x_{i+1}]$, and choose x_i' close to x_i. The goal is then to prove that the x_i' appear in order along $[c(a), c(b)]$.

Let

$$k' = \frac{k}{2} + 2\delta > 6\delta.$$

Let $b - a = Mk' + \eta$ for $0 \leqslant \eta < k'$ and $M \in \mathbb{N}$. Put $x_i = c(ik')$ for $i = 1, \dots, M$, and let x_i' be a closest point on $[c(a), c(b)]$ to x_i. By the lemma, we know $d(x_i, x_i') \leqslant 2\delta$.

Claim. $x_1', \dots, x_m'$ appear in the correct order along $[c(a), c(b)]$.

Let's finish the proof assuming the claim. If this holds, then note that

$$d(x_i', x_{i+1}') \geqslant k' - 4\delta > 2\delta$$

because we know $d(x_i, x_{i+1}) = 6\delta$, and also $d(x_m, c(b)) \geqslant \eta - 2\delta$. Therefore

$$d(c(a), c(b)) = \sum_{i=1}^{M} d(x_i, x_{i-1}) + d(x_m, c(b)) \geqslant 2\delta M + \eta - 2\delta \geqslant 2\delta(M-1).$$

On the other hand, we have

$$M = \frac{|b-a| - \eta}{k'} \geqslant \frac{|b-a|}{k'} - 1.$$

Thus, we find that

$$d(c(a), c(b)) \geqslant \frac{2\delta}{k'}|b-a| - 4\delta.$$

To prove the claim, let $x_i = c(t_i)$ for all i. We let

$$y = c(t_{i-1} + 2\delta)$$
$$z = c(t_{i+1} - 2\delta).$$

Define

$$\Delta_- = \Delta(x_{i-1}, x_{i-1}', y)$$
$$\Delta_+ = \Delta(x_{i+1}, x_{i+1}', z).$$

Both Δ_- and Δ_+ are disjoint from $B(x_i, 3\delta)$. Indeed, if $w \in \Delta_-$ with $d(x_i, w) \leqslant 3\delta$, then by δ-slimness of Δ_-, we know $d(w, x_{i-1}) \leqslant 3\delta$, and so $d(x_i, x_{i-1}) \leqslant 6\delta$, which is not possible.

Therefore, since the rectangle y, z, x_{i+1}', x_{i-1}' is 2δ-slim, and x_i is more than 2δ away from the sides yx_{i-1}' and zx_{i+1}'. So there must be some $x_i'' \in [x_{i-1}', x_{i+1}']$ with $d(x_i, x_i'') \leqslant 2\delta$.

Now consider $\Delta = \Delta(x_i, x_i', x_i'')$. We know $x_i x_i'$ and $x_i x_i''$ are both of length $\leqslant 2\delta$. Note that every point in this triangle is within 3δ of x_i by δ-slimness. So $\Delta \subseteq B(x_i, 3\delta)$, and this implies Δ is disjoint from $B(x_{i-1}, 3\delta)$ and $B(x_{i+1}, 3\delta)$ as before.

But $x_{i-1}' \in B(x_{i-1}, 3\delta)$ and $x_{i+1}' \in B(x_{i+1}, 3\delta)$. Moreover, Δ contains the segment of $[c(a), c(b)]$ joining x_i' and x_i''. Therefore, it must be the case that $x_i' \in [x_{i-1}', x_{i+1}']$. □

4.3 Dehn functions of hyperbolic groups

We now use our new understanding of quasi-geodesics in hyperbolic spaces to try to understand the word problem in hyperbolic groups. Note that by the Schwarz–Milnor lemma, hyperbolic groups are finitely-generated and their Cayley graphs are hyperbolic.

Corollary. Let X be δ-hyperbolic. Then there exists a constant $C = C(\delta)$ such that any non-constant loop in X is *not* C-locally geodesic.

Proof. Take $k = 8\delta + 1$, and let

$$C = \max\{\lambda\varepsilon, k\}$$

where λ, ε are as in the theorem.

Let $\gamma : [a, b] \to X$ be a closed loop. If Γ were C-locally geodesic, then it would be (λ, ε)-quasigeodesic. So

$$0 = d(\gamma(a), \gamma(b)) \geqslant \frac{|b - a|}{\lambda} - \varepsilon.$$

So

$$|b - a| \leqslant \lambda\varepsilon < C.$$

But γ is a C-local geodesic. This implies γ is a constant loop. □

Definition (Dehn presentation)**.** A finite presentation $\langle S \mid R\rangle$ for a group Γ is called *Dehn* if for every null-homotopic reduced word $w \in S^*$, there is (a cyclic conjugate of) a relator $r \in R$ such that $r = u^{-1}v$ with $\ell_S(u) < \ell_S(v)$, and $w = w_1 v w_2$ (without cancellation).

The point about this is that if we have a null-homotopic word, then there is some part in the word that can be replaced with a shorter word using a single relator.

If a presentation is Dehn, then the naive way of solving the word problem just works. In fact,

Lemma. If Γ has a Dehn presentation, then δ_Γ is linear.

Proof. Exercise. □

Theorem. Every hyperbolic group Γ is finitely-presented and admits a Dehn presentation.

In particular, the Dehn function is linear, and the word problem is solvable.

So while an arbitrary group can be very difficult, the generic group is easy.

Proof. Let S be a finite generating set for Γ, and δ a constant of hyperbolicity for $\mathrm{Cay}_S(\Gamma)$.

Let $C = C(\delta)$ be such that every non-trivial loop is *not* C-locally geodesic.

Take $\{u_i\}$ to be the set of all words in $F(S)$ representing geodesics $[1, u_i]$ in $\mathrm{Cay}_S(\Gamma)$ with $|u_i| < C$. Let $\{v_j\} \subseteq F(S)$ be the set of all *non*-geodesic words of length $\leqslant C$ in $\mathrm{Cay}_S(\Gamma)$.

Let $R = \{u_i^{-1} v_j \in F(S) : u_i =_\Gamma v_j\}$.

We now just observe that this gives the desired Dehn presentation! Indeed, any non-trivial loop must contain one of the v_j's, since $\mathrm{Cay}_S(\Gamma)$ is not C-locally geodesic, and re can replace it with u_i! □

This argument was developed by Dehn to prove results about the fundamental group of surface groups in 1912. In 1980, Gromov noticed that Dehn's argument works for an arbitrary hyperbolic group!

One can keep on proving new things about hyperbolic groups if we wished to, but there are better uses of our time. So for the remaining of the chapter, we shall just write down random facts about hyperbolic groups without proof.

So hyperbolic groups have linear Dehn functions. In fact,

Theorem (Gromov, Bowditch, etc)**.** If Γ is a finitely-presented group and $\delta_\Gamma \precnsim n^2$, then Γ is hyperbolic.

Thus, there is a "gap" in the isoperimetric spectrum. We can collect our results as

Theorem. If Γ is finitely-generated, then the following are equivalent:

(i) Γ is hyperbolic.

(ii) Γ has a Dehn presentation.

(iii) Γ satisfies a linear isoperimetric inequality.

(iv) Γ has a subquadratic isoperimetric inequality.

In general, we can ask the question — for which $\alpha \in \mathbb{R}$ is $n^\alpha \simeq$ a Dehn function of a finitely-presented group? As we saw, α cannot lie in $(1, 2)$, and it is a theorem that the set of such α is dense in $[2, \infty)$. In fact, it includes all rationals in the interval.

Subgroup structure

When considering subgroups of a hyperbolic group Γ, it is natural to consider "geometrically nice" subgroups, i.e. finitely-generated subgroups $H \subseteq \Gamma$ such that the inclusion is a quasi-isometric embedding. Such subgroups are called *quasi-convex*, and they are always hyperbolic.

What sort of such subgroups can we find? There are zillions of free quasi-convex subgroups!

Lemma (Ping-pong lemma)**.** Let Γ be hyperbolic and torsion-free (for convenience of statement). If $\gamma_1, \gamma_2 \in \Gamma$ do not commute, then for large enough n, the subgroup $\langle \gamma_1^n, \gamma_2^n \rangle \cong F_2$ and is quasi-convex.

How about non-free subgroups? Can we find surface groups? Of course, we cannot always guarantee the existence of such surface groups, since all subgroups of free groups are free.

Question. Let Γ be hyperbolic and torsion-free, and not itself free. Must Γ contain a quasi-convex subgroup isomorphic to $\pi_1\Sigma$ for some closed hyperbolic surface Σ?

We have no idea if it is true or false.

Another open problem we can ask is the following:

Question. If Γ is hyperbolic and not the trivial group, must Γ have a proper subgroup of finite index?

Proposition. Let Γ be hyperbolic, and $\gamma \in \Gamma$. Then $C(\gamma)$ is quasiconvex. In particular, it is hyperbolic.

Corollary. Γ does not contain a copy of $\mathbb{Z}^2$.

The boundary

Recall that if Σ is a compact surface of genus $g \geqslant 2$, then $\pi_1\Sigma \underset{qi}{\simeq} \mathbb{H}^2$. If we try to draw the hyperbolic plane in the disc model, then we would probably draw a circle and fill it in. One might think the drawing of the circle is just an artifact of the choice of the model, but it's not! It's genuinely there.

Definition (Geodesic ray)**.** Let X be a δ-hyperbolic geodesic metric space. A *geodesic ray* is an isometric embedding $r : [0, \infty) \to X$.

We say $r_1 \sim r_2$ if there exists M such that $d(r_1(t), r_2(t)) \leqslant M$ for all t. In the disc model of $\mathbb{H}^2$, this is the scenario where two geodesic rays get very close together as $t \to \infty$ (in the upper half plane model, this contains vertical, parallel lines).

We define $\partial_\infty X = \{\text{geodesic rays}\}/\sim$. This can be topologized in a sensible way, and in this case $X \cup \partial_\infty X$ is compact. By the Morse lemma, for hyperbolic spaces, this is quasi-isometry invariant.

Example. If $\Gamma = \pi_1\Sigma$, with Σ closed hyperbolic surface, then $\partial_\infty\Gamma = S^1$ and the union $X \cup \partial_\infty X$ gives us the closed unit disc.

Theorem (Casson–Jungreis, Gabai)**.** If Γ is hyperbolic and $\partial_\infty\Gamma \cong S^1$, then Γ is virtually $\pi_1\Sigma$ for some closed hyperbolic Σ.

Example. If Γ is free, then $\partial_\infty\Gamma$ is the Cantor set.

Conjecture (Cannon)**.** If Γ is hyperbolic and $\partial_\infty\Gamma \cong S^2$, then Γ is virtually $\pi_1 M$ for M a closed hyperbolic 3-manifold.

5 CAT(0) spaces and groups

From now on, instead of thinking of geodesics as being isometric embeddings, we reparametrize them linearly so that the domain is always $[0, 1]$.

5.1 Some basic motivations

Given a discrete group Γ, there are two basic problems you might want to solve.

Question. Can we solve the word problem in Γ?

Question. Can we compute the (co)homology of Γ?

Definition (Group (co)homology)**.** The *(co)homology* of a group Γ is the (co)homology of $K(\Gamma, 1)$.

We can define this in terms of the group itself, but would require knowing some extra homological algebra. A very closely related question is

Question. Can we find an explicit X such that $\Gamma = \pi_1 X$ and $\tilde{X}$ is contractible?

We know that these problems are not solvable in general:

Theorem (Novikov–Boone theorem)**.** There exists a finitely-presented group with an unsolvable word problem.

Theorem (Gordan)**.** There exists a sequence of finitely generated groups Γ_n such that $H_2(\Gamma_n)$ is not computable.

As before, we might expect that we can solve these problems if our groups come with some nice geometry.

Let M be a compact manifold of non-positive sectional curvature. It is a classical fact that such a manifold satisfies a quadratic isoperimetric inequality. This is not too surprising, since the "worst case" we can get is a space with constant zero curvature, which implies $\tilde{M} \cong \mathbb{R}^n$.

If we know this, then by the Filling theorem, we know the Dehn function of the fundamental group is at worst quadratic, and in particular it is computable. This solves the first question.

What about the second question?

Theorem (Cartan–Hadamard theorem)**.** Let M be a non-positively curved compact manifold. Then $\tilde{M}$ is diffeomorphic to $\mathbb{R}^n$. In particular, it is contractible. Thus, $M = K(\pi_1 M, 1)$.

For example, this applies to the torus, which is not hyperbolic.

So non-positively curved manifolds are good. However, there aren't enough of them. Why? In general, the homology of a group can be very complicated, and in particular can be infinite dimensional. However, manifolds always have finite-dimensional homology groups. Moreover, they satisfy Poincaré duality.

Theorem (Poincaré duality)**.** Let M be an orientable compact n-manifold. Then

$$H_k(M; \mathbb{R}) \cong H_{n-k}(M; \mathbb{R}).$$

This is a very big constraint, and comes very close to characterizing manifolds. In general, it is difficult to write down a group whose homology satisfies Poincaré duality, unless we started off with a manifold whose universal cover is contractible, and then took its fundamental group.

Thus, we cannot hope to realize lots of groups as the π_1 of a non-positively curved manifold. The idea of CAT(0) spaces is to mimic the properties of non-positively curved manifolds in a much more general setting.

5.2 CAT(κ) spaces

Let $\kappa = -1, 0$ or 1, and let M_κ be the unique, connected, complete 2-dimensional Riemannian manifold of curvature κ. Thus,

$$M_1 = S^2, \quad M_0 = \mathbb{R}^2, \quad M_{-1} = \mathbb{H}^2.$$

We can also talk about M_κ for other κ, but we can just obtain those by scaling $M_{\pm 1}$.

Instead of working with Riemannian manifolds, we shall just think of these as *complete geodesic metric spaces*. We shall now try to write down a "CAT(κ)" condition, that says the curvature is bounded by κ in some space.

Definition (Triangle)**.** A *triangle* with vertices $\{p, q, r\} \subseteq X$ is a choice

$$\Delta(p, q, r) = [p, q] \cup [q, r] \cup [r, p].$$

If we want to talk about triangles on a sphere, then we have to be a bit more careful since the sides cannot be too long. Let $D_\kappa = \operatorname{diam} M_\kappa$, i.e. $D_\kappa = \infty$ for $\kappa = 0, -1$ and $D_\kappa = \pi$ for $\kappa = +1$.

Suppose $\Delta = \Delta(x_1, x_2, x_3)$ is a triangle of perimeter $\leqslant 2D_\kappa$ in some complete geodesic metric space (X, d). Then there is, up to isometry, a unique *comparison triangle* $\bar{\Delta} = \Delta(\bar{x}_1, \bar{x}_2, \bar{x}_3) \subseteq M_\kappa$ such that

$$d_{M_k}(\bar{x}_i, \bar{x}_j) = d(x_i, x_j).$$

This is just the fact we know from high school that a triangle is determined by the lengths of its side. The natural map $\bar{\Delta} \to \Delta$ is called the *comparison triangle*.

Similarly, given a point $p \in [x_i, x_j]$, there is a *comparison point* $\bar{p} \in [\bar{x}_i, \bar{y}_j]$. Note that p might be only multiple edges, so it could have multiple comparison points. However, the comparison point is well-defined as long as we specify the edge as well.

Definition (CAT(κ) space)**.** We say a space (X, d) is CAT(κ) if for any geodesic triangle $\Delta \subseteq X$ of diameter $\leqslant 2D_\kappa$, any $p, q \in \Delta$ and any comparison points $\bar{p}, \bar{q} \in \bar{\Delta}$,

$$d(p, q) \leqslant d_{M_\kappa}(\bar{p}, \bar{q}).$$

If X is locally CAT(κ), then K is said to be of *curvature* at most κ.

In particular, a locally CAT(κ) space is called a *non-positively curved space*.

We are mostly interested in CAT(0) spaces. At some point, we will talk about CAT(1) spaces.

Example. The following are CAT(0):

(i) Any Hilbert space.

(ii) Any simply-connected manifold of non-positive sectional curvature.

(iii) Symmetric spaces.

(iv) Any tree.

(v) If X, Y are CAT(0), then $X \times Y$ with the ℓ_2 metric is CAT(0).

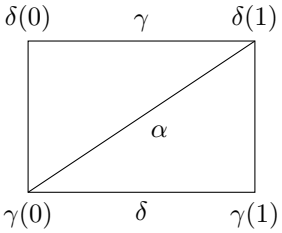
(vi) In particular, a product of trees is CAT(0).

Lemma (Convexity of the metric)**.** Let X be a CAT(0) space, and $\gamma, \delta : [0, 1] \to X$ be geodesics (reparameterized). Then for all $t \in [0, 1]$, we have

$$d(\gamma(t), \delta(t)) \leqslant (1 - t)d(\gamma(0), \delta(0)) + td(\gamma(1), \delta(1)).$$

Note that we have strict equality if this is in Euclidean space. So it makes sense that in CAT(0) spaces, we have an inequality.

Proof. Consider the rectangle



Let $\alpha : [0, 1] \to X$ be a geodesic from $\gamma(0)$ to $\delta(1)$. Applying the CAT(0) estimate to $\Delta(\gamma(0), \gamma(1), \delta(1))$, we get

$$d(\gamma(t), \alpha(t)) \leqslant d(\overline{\gamma(t)}, \overline{\alpha(t)}) = td(\overline{\gamma(1)}, \overline{\alpha(1)}) = td(\gamma(1), \alpha(1)) = td(\gamma(1), \delta(1)),$$

using what we know in plane Euclidean geometry. The same argument shows that

$$d(\delta(t), \alpha(t) \leqslant (1 - t)d(\delta(0), \gamma(0)).$$

So we know that

$$d(\gamma(t), \delta(t)) \leqslant d(\gamma(t), \alpha(t)) + d(\alpha(t), \delta(t)) \leqslant (1 - t)d(\gamma(0), \delta(0)) + td(\gamma(1), \delta(1)).$$

□

Lemma. If X is CAT(0), then X is *uniquely geodesic*, i.e. each pair of points is joined by a unique geodesic.

Proof. Suppose $x_0, x_1 \in X$ and $\gamma(0) = \delta(0) = x_0$ and $\gamma(1) = \delta(1) = x_1$. Then by the convexity of the metric, we have $d(\gamma(t), \delta(t)) \leqslant 0$. So $\gamma(t) = \delta(t)$ for all t. □

This is not surprising, because this is true in the Euclidean plane and the hyperbolic plane, but not in the sphere.

Lemma. Let X be a proper, uniquely geodesic metric space. Then geodesics in X vary continuously with their end points in the compact-open topology (which is the same as the uniform convergence topology).

This is actually true without the word "proper", but the proof is harder

Proof. This is an easy application of the Arzelá–Ascoli theorem. □

Proposition. Any proper CAT(0) space X is contractible.

Proof. Pick a point $x_0 \in X$. Then the map $X \to \mathrm{Maps}([0,1], X)$ sending x to the unique geodesic from x_0 to x is continuous. The adjoint map $X \times [0,1] \to X$ is then a homotopy from the constant map at x_0 to the identity map. □

Definition (CAT(0) group)**.** A group is CAT(0) if it acts properly discontinuously and cocompactly by isometries on a proper CAT(0) space.

Usually, for us, the action will also be free. This is the case, for example, when a fundamental group acts on the covering space.

Example. $\mathbb{Z}^n$ for any n is CAT(0), since it acts on $\mathbb{R}^n$.

Example. More generally, $\pi_1 M$ for any closed manifold M of non-positive curvature is CAT(0).

Example. Uniform lattices in semi-simple Lie groups. Examples include $\mathrm{SL}_n \mathcal{O}_K$ for certain number fields K.

Example. Any free group, or direct product of free groups is CAT(0).

We remark

Proposition. Any CAT(0) group Γ satisfies a quadratic isoperimetric inequality, that is $\delta_\Gamma \simeq n$ or $\sim n^2$.

We will not prove it.

Note that if Γ is in fact CAT(-1), then Γ is hyperbolic, which is not terribly difficult to see, since $\mathbb{H}^2$ is hyperbolic. But if a group is hyperbolic, is it necessarily CAT(-1)? Or even just CAT(0)? This is an open question. The difficulty in answering this question is that hyperbolicity is a "coarse condition", but being CAT(0) is a fine condition. For example, Cayley graphs are not CAT(0) spaces unless they are trees.

5.3 Length metrics

In differential geometry, if we have a covering space $\tilde{X} \to X$, and we have a Riemannian metric on X, then we can lift this to a Riemannian metric to $\tilde{X}$. This is possible since the Riemannian metric is a purely local notion, and hence we can lift it locally. Can we do the same for metric spaces?

Recall that if $\gamma : [a, b] \to X$ is a path, then

$$\ell(\gamma) = \sup_{a=t_0<t_1<\cdots<t_n=b} \sum_{i=1}^{n} d(\gamma(t_{i-1}), \gamma(t_i)).$$

We say γ is *rectifiable* if $\ell(\gamma) < \infty$.

Definition (Length space)**.** A metric space X is called a *length space* if for all $x, y \in X$, we have

$$d(x, y) = \inf_{\gamma: x \to y} \ell(\gamma).$$

Given any metric space (X, d), we can construct a *length pseudometric* $\hat{d} : X \times X \to [0, \infty]$ given by

$$\hat{d}(x, y) = \inf_{\gamma: x \to y} \ell(\gamma).$$

Now given a covering space $p : \tilde{X} \to X$ and (X, d) a metric space, we can define a length pseudometric on $\tilde{X}$ by, for any path $\tilde{\gamma} : [a, b] \to \tilde{X}$,

$$\ell(\tilde{\gamma}) = \ell(p \circ \tilde{\gamma}).$$

This induces an induced pseudometric on $\tilde{X}$.

Exercise. If X is a length space, then so is $\tilde{X}$.

Note that if A, B are length spaces, and $X = A \cup B$ (such that the metrics agree on the intersection), then X has a natural induced length metric. Recall we previously stated the Hopf–Rinow theorem in the context of differential geometry. In fact, it is actually just a statement about length spaces.

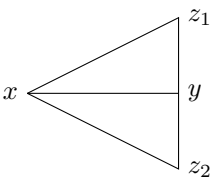
Theorem (Hopf–Rinow theorem)**.** If a length space X is complete and locally compact, then X is proper and geodesic.

This is another application of the Arzelá–Ascoli theorem.

5.4 Alexandrov's lemma

Alexandrov's lemma is a lemma that enables us to glue CAT(0) space together to obtain new examples.

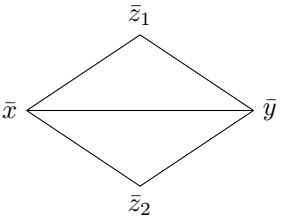
Lemma (Alexandrov's lemma)**.** Suppose the triangles $\Delta_1 = \Delta(x, y, z_1)$ and $\Delta_2 = \Delta(x, y, z_2)$ in a metric space satisfy the CAT(0) condition, and $y \in [z_1, z_2]$.



Then $\Delta = \Delta(x, z_1, z_2)$ also satisfies the CAT(0) condition.

This is the basic result we need if we want to prove "gluing theorems" for CAT(0) spaces.

Proof. Consider $\bar{\Delta}_1$ and $\bar{\Delta}_2$, which together form a Euclidean quadrilateral $\bar{Q}$ with with vertices $\langle x, \bar{z}_1, \bar{z}_2, \bar{y}$. We claim that then the interior angle at $\bar{y}$ is $\geqslant 180^\circ$. Suppose not, and it looked like this:

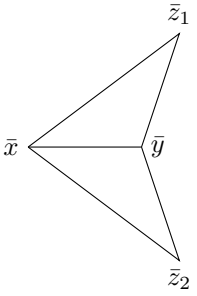


If not, there exists $\bar{p}_i \in [\bar{y}, \bar{z}_i]$ such that $[\bar{p}_1, \bar{p}_2] \cap [\bar{x}, \bar{y}] = \{\bar{q}\}$ and $\bar{q} \neq \bar{y}$. Now

$$\begin{aligned} d(p_1, p_2) &= d(p_2, y) + d(y, p_2) \\ &= d(\bar{p}_1, \bar{y}) + d(\bar{y}, \bar{p}_1) \\ &= d(\bar{p}, \bar{y}) + d(\bar{y}, \bar{p}_2) \\ &> d(\bar{p}_1, \bar{q}) + d(\bar{q}, \bar{p}_2) \\ &\geqslant d(p_1, q) + d(q, p_2) \\ &\geqslant d(p_1, p_2), \end{aligned}$$

which is a contradiction.

Thus, we know the right picture looks like this:



To obtain $\bar{\Delta}$, we have to "push" $\bar{y}$ out so that the edge $\bar{z}_1\bar{z}_2$ is straight, while keeping the lengths fixed. There is a natural map $\pi : \bar{\Delta} \to \bar{Q}$, and the lemma follows by checking that for any $a, b \in \bar{\Delta}$, we have

$$d(\pi(a), \pi(b)) \leqslant d(a, b).$$

This is an easy case analysis (or is obvious). □

A sample application is the following:

Proposition. If X_1, X_2 are both locally compact, complete CAT(0) spaces and Y is isometric to closed, subspaces of both X_1 and X_2. Then $X_1 \cup_Y X_2$, equipped with the induced length metric, is CAT(0).

5.5 Cartan–Hadamard theorem

Theorem (Cartan–Hadamard theorem)**.** If X is a complete, connected length space of non-positive curvature, then the universal cover $\tilde{X}$, equipped with the induced length metric, is CAT(0).

This was proved by Cartan and Hadamard in the differential geometric setting.

Corollary. A (torsion free) group Γ is CAT(0) iff it is the π_1 of a complete, connected space X of non-positive curvature.

We'll indicate some steps in the proof of the theorem.

Lemma. If X is proper, non-positively curved and uniquely geodesic, then X is CAT(0).

Proof idea. The idea is that given a triangle, we cut it up into a lot of small triangles, and since X is locally CAT(0), we can use Alexandrov's lemma to conclude that the large triangle is CAT(0).

Recall that geodesics vary continuously with their endpoints. Consider a triangle $\Delta = \Delta(x, y, z) \subseteq \bar{B} \subseteq X$, where $\bar{B}$ is a compact ball. By compactness, there is an ε such that for every $x \in \bar{B}$, the ball $B_x(\bar{\varepsilon})$ is CAT(0).

We let β_t be the geodesic from x to $\alpha(t)$. Using continuity, we can choose $0 < t_1 < \cdots < t_N = 1$ such that

$$d(\beta_{t_i}(s), \beta_{t_{i+1}}(s)) < \varepsilon$$

for all $s \in [0, 1]$.

Now divide Δ up into a "patchwork" of triangles, each contained in an ε ball, so each satisfies the CAT(0) condition, and apply induction and Alexandrov's lemma to conclude. □

Now to prove the Cartan–Hadamard theorem, we only have to show that the universal cover is uniquely geodesic. Here we must use the simply-connectedness condition.

Theorem. Let X be a proper length space of non-positive curvature, and $p, q \in X$. Then each homotopy class of paths from p to q contains a *unique* (local) geodesic representative.

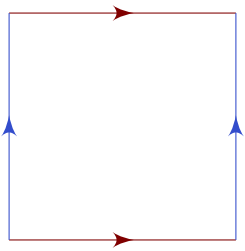
5.6 Gromov's link condition

Gromov's link condition is a criterion that makes it very easy to write down interesting examples of non-positively-curved metric spaces.

Definition (Euclidean cell complex)**.** A locally finite cell complex X is *Euclidean* if every cell is isometric to a convex polyhedron in Euclidean space and the attaching maps to identify faces isometrically with lower-dimensional cells.

Such a complex X has a natural length metric which is proper and geodesic by Hopf–Rinow. What we'd like to do is to come up with a condition that ensures X is CAT(0).

Example. The usual diagram for a torus gives an example of a Euclidean complex.



Example. We can construct a sphere this way, just by taking a cube!

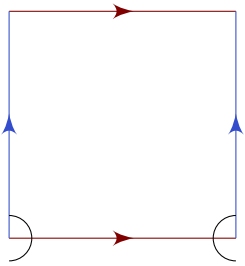
We know that T^2 is non-positively curved (since it is flat), but the cube is not, because by Cartan–Hadamard, it would be CAT(0), hence contractible, but it is not.

Definition (Link)**.** Let X be a Euclidean complex, and let v be a vertex of X, and let $0 < \varepsilon \ll$ shortest 1-cell. Then the *link* of v is

$$Lk(v) = S_v(\varepsilon) = \{x \in X : d(x, v) = \varepsilon\}.$$

Note that $Lk(V)$ is a cell complex: the intersection of $Lk(v)$ with a cell of X of dimension is a cell of dimension $n - 1$.

Example. In the torus, there is only one vertex. The link looks like



So the link is S^1.

Example. If we take the corner of a cube, then $Lk(v)$ is homeomorphic to S^1, but it is a weird one, because it is made up of three right angles, not two.

How can we distinguish between these two S^1's? *Angle* puts a metric on $Lk(v)$. We can do this for general metric spaces, but we only need it for Euclidean complexes, in which case there is not much to do.

Restricted to each cell, the link is just a part of a sphere, and it has a natural spherical metric, which is a length metric. These then glue together to a length metric on $Lk(v)$. Note that this is not the same as the induced metric.

Example. In the torus, the total length of the link is 2π, while that of the cube is $\frac{3\pi}{2}$.

Theorem (Gromov's link criterion)**.** A Euclidean complex X is non-positively-curved iff for every vertex v of X, $Lk(v)$ is CAT(1).

Note that by definition, we only have to check the CAT(1) inequality on triangles of perimeter $< 2\pi$.

Exercise. Check these in the case of the torus and the cube.

Thus, given a group, we can try to construct a space whose π_1 is it, and then put a Euclidean structure on it, then check Gromov's link criterion.

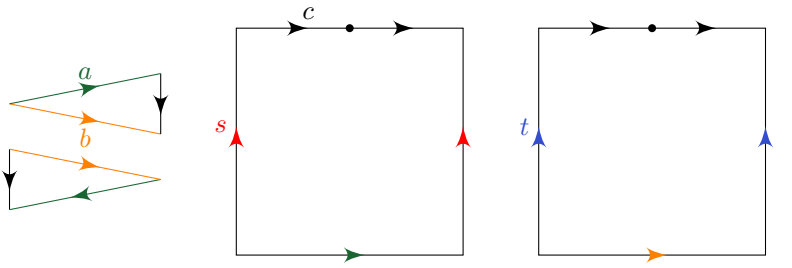
In general, Gromov's link condition might not be too useful, because we still don't know how to check if something is CAT(1)! However, it is very simple in dimension 2.

Corollary. If X is a 2-dimensional Euclidean complex, then for all vertices v, $Lk(v)$ is a metric graph, and X is CAT(0) iff $Lk(v)$ has no loop of length $< 2\pi$ for all v.

Example (Wise's example)**.** Consider the group

$$W = \langle a, b, s, t \mid [a, b] = 1, s^{-1}as = (ab)^2, t^{-1}bt = (ab)^2 \rangle.$$

Letting $c = ab$, it is easy to see that this is the π_1 of the following Euclidean complex:



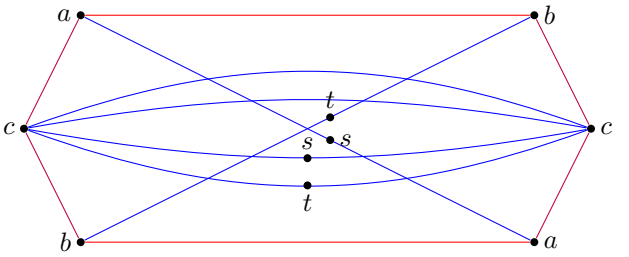
This is metrized in the obvious way, where all edges have length 2 except the black ones of length 1. To understand the links, we set

$$\alpha = \sin^{-1}\frac{1}{4}, \quad \beta = \cos^{-1}\frac{1}{4}.$$

Then the triangles each has angles $2\alpha, \beta, \beta$.

We show that W is non-positively curved, and then use this to show that there is a homomorphism $W \to W$ that is surjective but not injective. We say W is *non-Hopfian*. In particular, this will show that WW is not linear, i.e. it is not a matrix group.

To show that X is non-positively curved, we check the link condition:



To check the link condition, we have to check that there are no loops of length $< 2\pi$ in $Lk(v)$. Note that by trigonometric identities, we know $\alpha + \beta = \frac{\pi}{2}$. So we can just observe that everything is fine.

To see that W is non-Hopfian, we define

$$\begin{aligned} f : W &\to W \\ a &\mapsto a^2 \\ b &\mapsto b^2 \\ s &\mapsto s^2 \\ t &\mapsto t \end{aligned}$$

We check that this is a well-defined homomorphism, and is surjective, since we observe

$$a = sa^2b^2s^{-1}, \quad b = ta^2b^2t^{-1},$$

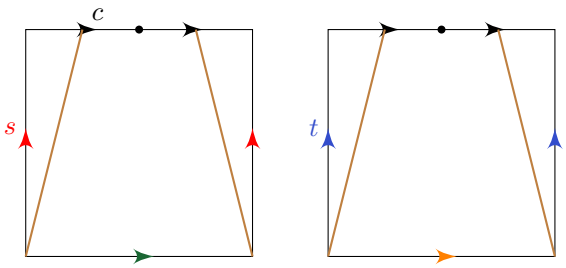
and so $a, b, s, t \in \operatorname{im} t$. The non-trivial claim is that $\ker f \neq 0$. Let

$$g = [scs^{-1}, tct^{-1}].$$

Note that

$$f(g) = [f(scs^{-1}), f(tct^{-1})] = [sc^2s^{-1}, tc^2t^{-1}] = [a, b] = 1.$$

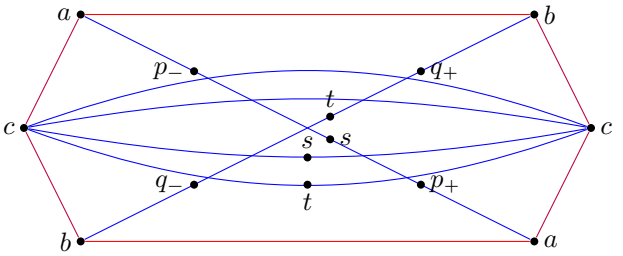
So the crucial claim is that $g \neq 1$. This is where geometry comes to the rescue. For convenience, write $p = scs^{-1}, q = tct^{-1}$. Consider the following local geodesics in the two squares:



The left- and right-hand local geodesics represent p and q respectively. Then

$$g = p \cdot q \cdot \bar{p} \cdot \bar{q}.$$

We claim that this represents g by a local geodesic. The only place that can go wrong is the points where they are joined. By the proof of the Gromov link condition, to check this, we check that the three "turns" at the vertex are all of angle $\geqslant \pi$.



Here p_+ is the point where p ends and p_- is the point where p starts, and similarly for q. Moreover, the distance from $p_\pm, q_\pm$ to the top/bottom left/right vertices is β. So we can just check and see that everything works.

Recall that every homotopy class of paths in X contains a *unique* locally geodesic representative. Since the constant path is locally geodesic, we know $g \neq 1$.

Definition (Residually finite group)**.** A group G is *residually finite* if for every $g \in G\backslash\{1\}$, there is a homomorphism $\varphi : G \to$ finite group such that $\varphi(g) \neq 0$.

Theorem (Mal'cev)**.** Every finitely generated linear subgroup (i.e. a subgroup of $\mathrm{GL}_n(\mathbb{C})$) is resudially finite.

Proof sketch. If the group is in fact a subgroup of $\mathrm{GL}_n(\mathbb{Z})$, then we just reduce mod p for $p \gg 0$. To make it work over $\mathrm{GL}_n(\mathbb{C})$, we need a suitable version of the Nullstellensatz. □

Theorem (Mal'cev)**.** Every finitely generated residually finite group is Hopfian.

Proof. Finding a proof is a fun exercise! □

We know that Wise's example is not Hopfian, hence not residually finite, hence not a linear group.

Contrast this with the amazing theorem of Sela that all hyperbolic groups are Hopfian. However, a major open question is whether all hyperbolic groups are residually finite. On the other hand, it is known that not all hyperbolic groups are not linear.

How are we supposed to think about residually finite groups?

Lemma (Scott's criterion)**.** Let X be a cell complex, and $G = \pi_1 X$. Then G is residually finite if and only if the following holds:

Let $p : \tilde{X} \to X$ be the universal cover. For all compact subcomplexes $K \subseteq \tilde{X}$, there is a finite-sheeted cover $X' \to X$ such that the natural covering map $p' : \tilde{X} \to X'$ is injective on K.

A good (though not technically correct) way to think about this is follows: if we have a map $f : K \to X$ that may be complicated, and in particular is not injective, then we might hope that there is some "finite resolution" $X' \to X$ such that f lifts to X', and the lift is injective. Of course, this is not always possible, and a necessary condition for this to work is that the lift to the universal cover is injective. If this is not true, then obviously no such resolution can exist. And residual finiteness says if there is not obvious reason to fail, then it in fact does not fail.

5.7 Cube complexes

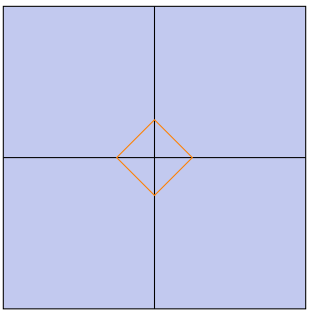
Definition (Cube complex)**.** A Euclidean complex is a *cube complex* if every cell is isometric to a cube (of any dimension).

This is less general than Euclidean complexes, but we can still make high dimension things out of this. Of course, general Euclidean complexes also work in high dimensions, but except in two dimensions, the link condition is rather

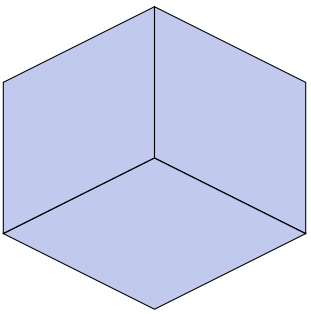
tricky to check in higher dimensional situations. It turns out the link condition is easy to check for cube complexes.

Purely topologically, the link of each vertex is made out of simplices, and, subdividing if necessary, they are simplicial complexes. The metric is such that every edge has length $\frac{\pi}{2}$, and this is called the "all-right simplicial complex".

Recall that



is non-positively curved, while



is not.

Definition (Flag simplicial complex)**.** A simplicial complex is not *flag* if for some $n \geqslant 2$, it contains a 1-dimensional subcomplex isomorphic to the 1-skeleton of an n-simplex, which is not contained in the boundary of an n-simplex.

A flag simplicial complex is one that is not not flag.

So we cannot contain an empty triangle, or hollow tetrahedra.

Note that topologically, flag complexes are not special in any sense. For any simplicial complex K, the first barycentric subdivision is flag.

Theorem (Gromov)**.** A cube complex is non-positively curved iff every link is flag.

Now the property of being flag is purely combinatorial, and easy to check. So this lets us work with cube complexes.

Right-angled Artin groups

Definition (right-angled Artin group)**.** Let N be a simplicial graph, i.e. a graph where the vertices determine the edges, i.e. a graph as a graph theorist would consider. Then

$$A_N = \langle V(N) \mid [u, v] = 1 \text{ for all } (u, v) \in E(N) \rangle$$

is the *right-angled Artin group*, or *graph group*.

Example. If N is the discrete graph on n vertices, then $A_N = F_n$.

Example. If N is the complete graph on n vertices, then $A_N = \mathbb{Z}^n$.

Example. If N is a square, i.e. the complete bipartite graph $K_{2,2}$, then $A_N = F_2 \times F_2$.

Example. When N is the path with 4 vertices, then this is a complicated group that doesn't have a good, alternative description. This is quite an interesting group.

Definition (Salvetti complex)**.** $\mathcal{S}_N$ is a cube complex defined as follows:

- Set $\mathcal{S}_N^{(2)}$ is the presentation complex for A_N.
- For any immersion of the 2-skeleton of a d-dimensional cube, we glue in an d-dimensional cube to $\mathcal{S}_N^{(2)}$.

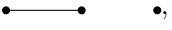
Alternatively, we have a natural inclusion $\mathcal{S}_N^{(2)} \subseteq (S^1)^{|V(N)|}$, and $\mathcal{S}_N$ is the largest subcomplex whose 2-skeleton coincides with $\mathcal{S}_N^{(2)}$.

Example. If N is

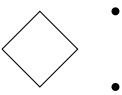
then $A_n = \mathbb{Z} * \mathbb{Z}^2$, and $\mathcal{S}_N$ is a circle glued to a torus.

Definition (Flag complex)**.** The *flag complex* of N, written $\bar{N}$, is the only flag simplicial complex with 1-skeleton N.

Example. If N is



then $\mathcal{S}_N$ is $T^2 \vee S^1$. There is a unique vertex v, and its link looks like



In fact, there is a recipe for getting the link out of N.

Definition (Double)**.** the *double* $D(K)$ of a simplicial complex K is defined as follows:

- The vertices are $\{v_1^+, \dots, v_n^+, v_1^-, \dots, v_n^-\}$, where $\{v_1, \dots, v_n\}$ are the vertices of k.
- The simplices are those of the form $\langle v_{i_0}^{\pm}, \dots, v_{i_k}^{\pm} \rangle$, where $\langle v_{i_0}, \dots, v_{i_k} \rangle \in K$.

Example. The double of N is just $Lk(v)$!

Note that $D(K)$ contains many copies of K, especially K^+, which is spanned by the v_i^+, and K^-, which is spanned by the v_i^-. In fact, K^+ (and also K^-) is a retract of $D(K)$, using the map that sends $v_i^{\pm}$ to v_i. Note also that $D(K)$ is flag iff K is flag.

Lemma. For any (simplicial) graph N, the link of the unique vertex of $\mathcal{S}_N$ is $D(\bar{N})$. In particular, $\mathcal{S}_N$ is non-positively curved.

Thus, right-angled Artin groups and their Salvetti complexes give examples of non-positively curved spaces with very general links. It turns out

Theorem. Right-angled Artin groups embed into $\mathrm{GL}_n\mathbb{Z}$ (where n depends on N).

5.8 Special cube complexes

Let X be a non-positively curved cube complex. We will write down explicit geometric/combinatorial conditions on X so that $\pi_1 X$ embeds into A_N for some N.

Hyperplanes and their pathologies

If $C \cong [-1,1]^n$, then a *midcube* $M \subseteq C$ is the intersection of C with $\{x_i = 0\}$ for some i.

Now if X is a non-positively curved cube complex, and M_1, M_2 are midcubes of cubes in X, we say $M_1 \sim M_2$ if they have a common face, and extend this to an equivalence relation. The equivalence classes are *immersed hyperplanes*. We usually visualize these as the union of all the midcubes in the equivalence class.

An immersed hyperplane can be thought of as a locally isometric map $H \looparrowright X$, where H is a cube complex. Let N_H be the pullback interval bundle over H. That is, N_H Is obtained by gluing together $\{M \times (-1,1) \mid M \text{ is a cube in } H\}$. The point of this is to resolve the self-intersections.

N_H is always a $(-1,1)$-bundle over H, and this corresponds to a homomorphism $\pi_1 H \to \mathrm{Isom}(-1,1) \cong \mathbb{Z}/2$. If this homomorphism is trivial, then we say H is *two-sided*. Otherwise, H is *one-sided*.

We have now met two pathologies of hyperplanes: they may be one-sided (e.g. a Möbius band), or have self-intersections. There are two more. The first is *self osculations*:

Note that it makes sense to distinguish between direct and indirect osculations only if our hyperplane is two-sided.

Finally, we have *inter-osculations*:

Definition (Special cube complex (Haglund–Wise))**.** A cube complex is *special* if its hyperplanes do not exhibit any of the following four pathologies:

- One-sidedness
- Self-intersection
- Direct self-osculation
- Inter-osculation

Example. A cube is a special cube complex.

Example. Traditionally, the way to exhibit a surface as a cube complex is to first tile it by right-angled polygons, so that every vertex has degree 4, and then the dual exhibits the surface as a cube complex. The advantage of this approach is that the hyperplanes are exactly the edges in the original tiling!

From this, one checks that we in fact have a special curve complex.

This is one example, but it is quite nice to have infinitely many. It is true that covers of special things are special. So this already gives us infinitely many special cube complexes. But we want others.

Example. If $X = \mathcal{S}_N$ is a Salvetti complex, then it is a special cube complex, and it is not terribly difficult to check.

Theorem (Haglund–Wise)**.** If X is a compact special cube complex, then there exists a graph N and a local isometry of cube complexes

$$\varphi_X : X \looparrowright \mathcal{S}_N.$$

Corollary. $\pi_1 X \hookrightarrow A_N$.

Proof. If $g \in \pi_1 X$, then g is uniquely represented by a local geodesic $\gamma : I \to X$. Then $\varphi \circ \gamma$ is a local geodesic in $\mathcal{S}_N$. Since homotopy classes of loops are represented by unique local geodesics, this implies $\varphi_X \circ \gamma$ is not null-homotopic. So the map $(\varphi_X)_*$ is injective. □

So if we know some nice group-theoretic facts about right-angled Artin groups, then we can use them to understand $\pi_1 X$. For example,

Corollary. If X is a special cube complex, then $\pi_1 X$ is linear, residually finite, Hopfian, etc.

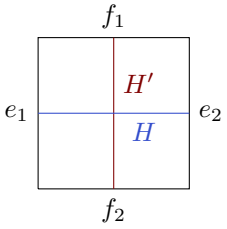
We shall try to give an indication of how we can prove the Haglund–Wise theorem. We first make the following definition.

Definition (Virtually special group)**.** A group Γ is *virtually special* if there exists a finite index subgroup $\Gamma_0 \leqslant \Gamma$ such that $\Gamma_0 \cong \pi_1 X$, where X is a compact special cube complex.

Sketch proof of Haglund–Wise. We have to first come up with an N. We set the vertices of N to be the hyperplanes of X, and we join two vertices iff the hyperplanes cross in X. This gives $\mathcal{S}_N$. We choose a transverse orientation on each hyperplane of X.

Now we define $\varphi_X : X \looparrowright \mathcal{S}_N$ cell by cell.

- Vertices: There is only one vertex in $\mathcal{S}_N$.
- Edges: Let e be an edge of X. Then e crosses a unique hyperplane H. Then H is a vertex of N. This corresponds to a generator in A_N, hence a corresponding edge $e(H)$ of $\mathcal{S}_N$. Send e to $e(H)$. The choice of transverse orientation tells us which way round to do it
- Squares: given a hyperplane



Note that we already mapped e_1, e_2 to $e(H)$, and f_1, f_2 to $e(H')$. Since H and H' cross in X, we know $e(H)$ and $e(H')$ bound a square in $\mathcal{S}_N$. Send this square in X to that square in $\mathcal{S}_N$.

– There is nothing to do for the higher-dimensional cubes, since by definition of $\mathcal{S}_N$, they have all the higher-dimensional cubes we can hope for.

We haven't used a lot of the nice properties of special cube complexes. They are needed to show that the map is a local isometric embedding. What we do is to use the hypothesis to show that the induced map on links is an isometric embedding, which implies φ_X Is a local isometry. □

This really applies to a really wide selection of objects.

Example. The following groups are virtually special groups:

– $\pi_1 M$ for M almost any 3-manifold.

– Random groups

This is pretty amazing. A "random group" is linear!

Index