Bowdoin College

Bowdoin Digital Commons

Honors Projects | Student Scholarship and Creative Work

2023

The Current Support Theorem in Context

Ethan Winters
Bowdoin College

Follow this and additional works at: https://digitalcommons.bowdoin.edu/honorsprojects

Part of the Geometry and Topology Commons

Recommended Citation

Winters, Ethan, "The Current Support Theorem in Context" (2023). *Honors Projects*. 456.
https://digitalcommons.bowdoin.edu/honorsprojects/456

This Open Access Thesis is brought to you for free and open access by the Student Scholarship and Creative Work at Bowdoin Digital Commons. It has been accepted for inclusion in Honors Projects by an authorized administrator of Bowdoin Digital Commons. For more information, please contact mdoyle@bowdoin.edu, a.sauer@bowdoin.edu.

The Current Support Theorem in Context

An Honors Paper for the Department of Mathematics

By Ethan Winters

Bowdoin College, 2023

©2023 Ethan Winters

Abstract

This work builds up the theory surrounding a recent result of Erlandsson, Leininger, and Sadanand: the Current Support Theorem. This theorem states precisely when a hyperbolic cone metric on a surface is determined by the support of its Liouville current. To provide background for this theorem, we will cover hyperbolic geometry and hyperbolic surfaces more generally, cone surfaces, covering spaces of surfaces, the notion of an orbifold, and geodesic currents. A corollary to this theorem found in the original paper is discussed which asserts that a surface with more than $32(g-1)$ cone points must be rigid. We extend this result to the case that there are more than $3(g-1)$ cone points. An infinite family of cone surfaces which are not rigid and which have precisely $3(g-1)$ cone points is also provided, hence demonstrating tightness.

Acknowledgements

The present thesis is not merely the culmination of my own work and of the great minds that have preceded and built the edifice of knowledge soon to be presented. It is also the product of many individuals who have directly encouraged and guided me throughout this project. Here, I would like to take the opportunity to highlight a few in particular.

First and foremost is my project advisor Dr. Chandrika Sadanand. She has displayed uncountably infinite patience for me while I have struggled to understand and employ the concepts of her field of research. Her constant encouragement has also been an invaluable uplifting force for me when I have found myself in times of doubt. I thank her sincerely for the many hours she has spent with me helping me hone my mathematical abilities, and instilling in me a fresh love for geometry and topology in particular. I could not have hoped to have a better mentor to help me along on in this mathematical journey of mine.

I have also been offered incredible support from my major advisor, prof. Jennifer Taback. From offering me invaluable recommendations and advice about graduate school, to meeting with me and providing motivation when I was experiencing hardship, to even guiding me through my first time flying for a grad school visit! I have her to thank for being able to get as far as I have, for never giving up on me, and never letting me give up on myself.

I would also like to thank Katherine Chui and Dr. Jacob Russell-Madonia of Rice University for providing a very enlightening conversation near the start of this project. I have them to thank for inspiring some of the methods I used for my main result.

Finally, I would of course like to thank my family for their love and support throughout the years. To my parents, for always having nothing but pride and encouragement for me, and providing me with a roof over my head and food to eat for all my life, including the summer's between semesters. To my grandparents, who live a stone's throw away and were always happy to offer me a ride when I needed it, or a visit for a little while which always lifted my spirits, or their time and attention when I was presenting my research. And to my siblings, for having had to deal with my nonsense as a child and yet still staying by my side to this day. To my family, I love you all.

Contents

Chapter 1

Introduction

1.1 Inverse Problems in Geometry

Inverse problems, in the case of geometry, ask whether one can reconstruct the geometry or 'metric' of some object, given a certain geometric data set. Among the originating examples of such problems is Kac's famous question, "Can you hear the shape of a drum?" In that case, the geometric data set under consideration was the spectrum of the Laplace operator on a plane region, and one wished to use this to reconstruct the geometry or 'shape' of the region itself.

Another example, more closely tied to the subject of this thesis, is Otal's ground-breaking work on marked length spectra. If one has a surface with a hyperbolic metric, it turns out that every free homotopy class of loops on this surface has a unique geodesic representative. The length of this closed geodesic can then be taken as the length of the homotopy class itself, and so we obtain a function $\ell : \pi_1(S) \to \mathbb{R}^+$. Otal (1990) proved that in fact this function is sufficient to determine the metric on S (up to isotopy)[1].

A corollary to this fact is that the Liouville current, a certain measure on the space of geodesics of S, also determines the hyperbolic metric on S up to isotopy. This is because the Liouville current is itself determined by the lengths of the homotopy classes of closed loops via a geometric construct known as the "geometric intersection number." However, if S possesses a certain kind of singular point known as a "cone point," then this Liouville current no longer has full support. That is, there are some open sets of geodesics that will have measure 0 when measured by the Liouville current. One can then ask whether knowing this support is sufficient to determine S up to isotopy.

A recent result of Erlandsson, Leininger, and Sadanand [2] answers this in the affirmative for all hyperbolic cone surfaces, minus a measure zero subset of exceptions. Moreover, those surfaces for which the metric is not determined by the support of the Liouville current are fully characterized. Those metrics which are not determined by the support of their Liouville current they refer to as 'flexible,' while those which are they refer to as 'rigid.' In particular, their theorem states that

Theorem 1.1.1 (Current Support Theorem). *If φ_1 and φ_2 are hyperbolic cone metrics on a closed, orientable surface S with supp(L_{φ_1}) = supp(L_{φ_2}), then either $\varphi_1 = \varphi_2$, or there exist locally isometric orbifold branch coverings $p_1 : (S, \varphi_1) \to \mathcal{O}$ and $p_2 : (S, \varphi_2) \to \mathcal{O}$ such that each cone point of S is mapped via p_1 and p_2 to an orbifold point with even order. Moreover, if such an orbifold covering exists for (S, φ), then the set of all hyperbolic cone metrics on S which share the same support of their Liouville currents with φ is parameterized by the Teichmüller space of $\mathcal{O}$.*

With this theorem in hand, Erlandsson et al. were able to apply it to a problem in symbolic dynamics, which asked if one gives the edges of a hyperbolic polygon a cyclic labeling, whether the set of all possible bi-infinite sequences of edge labels that can be obtained from a billiard ball bouncing from edge to edge (the "Bounce Spectrum") was sufficient to determine the polygon up to isometry. Mirroring the result of their Current Support Theorem (CST), they were able to prove that all hyperbolic polygons (excluding a measure zero subset) can be determined up to isometry by their bounce spectra, and moreover fully characterize all exceptional polygons via a single geometric description.

In addition, they developed a corollary which makes the theorem somewhat easier to apply. This corollary asserts that if a hyperbolic cone surface has more than $32(g-1)$ cone points, where g is the genus of the surface, then the metric on S must be rigid. Thus in such cases, one does not need to directly prove no requisite orbifold covering exists in order to prove that the metric is rigid. This result is what will be improved upon in this thesis. In particular, we will prove that if the number of cone points exceeds $3(g-1)$, then the metric must be rigid. Moreover, we will supply an infinite family of cone surfaces which are flexible, and which have precisely $3(g-1)$ cone points.

1.2 An Outline of the Sequel

Before coming to the main result of this thesis, we will endeavor to develop the theory surrounding the CST. In particular, in what follows we will first develop an understanding of hyperbolic geometry in general. We will look at two models of it, define the metrics on these models, and categorize the isometries of hyperbolic space. There will also be a discussion of the boundary at infinity of the hyperbolic plane, and see how the isometries of the plane act on it. Finally, we will develop the definition of hyperbolic area, which will play an important role in the proof of our result.

Another piece of understanding we will need is that of covering spaces. We will give examples and a definition for this concept, as well as describe how, under ideal circumstances, the covering spaces of a manifold are in one-to-one correspondence with conjugacy classes of subgroups of its fundamental group. We will next endeavor to detail an explicit construction of the universal cover of a manifold, which corresponds to the trivial subgroup of the fundamental group. Finally, we will finish this chapter by describing a more powerful classification of covering spaces of a given manifold, found in the 'permutation representation.' This classification is more powerful as it will allow us to describe disconnected coverings as well as connected coverings.

Combining these two sections together, we next explore hyperbolic surfaces. Properties of hyperbolic surfaces blend information from both geometry and topology in interesting ways, and we will see how this interaction provides a nice parameterization of the space of all hyperbolic metrics on a surface, i.e. its "Teichmüller space." We than expand on this discussion and generalize to surfaces with cone points, as well as generalize the previous discussion of hyperbolic area to the 'Gauss-Bonnet Theorem,' which gives an easy way to compute the area of a hyperbolic surface.

In chapter 5, we give an outline of the theory of orbifolds, offering examples and elaborating on how they generalize concepts from ordinary manifolds such as coverings, fundamental groups, and Euler characteristics.

Chapter 6 is the last chapter to develop the background theory, and it will focus on Geodesic currents. Beginning with a notion of a space of geodesics on a surface S which is purely topological and divorced from any specific metric. It is on this purely topological construction that we can then define certain measures, and from there build up to the Liouville current.

Chapters 7 and 8 then recap the theory developed in the previous chapters and use them to both explore the CST in more detail, providing examples and applications, as well apply these ideas to the proof that a metric with more than $3(g-1)$ cone points must be rigid.

Chapter 2

Hyperbolic Geometry

2.1 Hyperbolic Geometry and its Models

Hyoerbolic geometry is one of 3 varieties of nice geometry, distinguished by their constant curvature at every point. The first of these is spherical geometry, which possesses a positive curvature at every point. The curvature in this case is inversely proportional to the radius of the sphere. The next is flat geometry, modeled by the Euclidean plane $\mathbb{R}^2$ and having zero curvature at every point. Then finally there is hyperbolic geometry, which has a constant negative curvature at every point. Unlike the sphere and the plane, the hyperbolic plane unfortunately possesses no isometric embedding inside of our space $\mathbb{R}^3$. Thus we have to turn to inaccurate but sufficient models to try and understand this geometry. Luckily there are a variety of these, and we discuss 3 here and how they are related.

The first we will discuss is the hyperboloid model. One advantage of this model is that the way its metric is defined bares the most resemblance to familiar euclidean geometry. In $\mathbb{R}^3$, the distance between any two vectors $\vec{v}$ and $\vec{u}$ is defined to be $\sqrt{(\vec{v}-\vec{u})\cdot(\vec{v}-\vec{u})}$. The measurement of distance is defined in terms of the dot product, which is an example of a quadratic form on $\mathbb{R}^n$. In general, any quadratic form can be written in terms of a matrix A as

$$Q_A(\vec{u},\vec{v}) = \vec{u}^T A\vec{v} = \vec{u}\cdot A\vec{v}.$$

Thus this is a generalization of the dot product. For standard Euclidean geometry, the matrix A is simply the identity matrix. In order to define a model for hyperbolic geometry, we will simply change one of the 1's in the identity matrix to a -1, giving us the quadratic form

[3]

$$Q(\vec{u},\vec{v}) = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = u_1v_1 + u_2v_2 - u_3v_3.$$

From this formula, we can observe that it is actually impossible to form a metric on $\mathbb{R}^3$ in the same way we did in the euclidean plane, i.e by defining $d(\vec{u},\vec{v}) = \sqrt{Q(\vec{u}-\vec{v},\vec{u}-\vec{v})}$, since the value of this quadratic form is sometimes negative depending on the inputs. What we will do instead is restrict attention to the locus of vectors $\vec{v}$ such that $Q(\vec{v},\vec{v}) = -1$. We can use the formula for this quadratic form and solve this equation to get the relation

$$v_3 = \pm\sqrt{1+v_1^2+v_2^2},$$

which is the defining equation for a 2-sheeted hyperboloid in $\mathbb{R}^3$. in order to get a connected subspace, we will restrict attention further to the positive sheet, i.e. where $v_3 > 0$. This restricted space has the important property that every tangent vector $\vec{v}$ to this surface satisfies $Q(\vec{v},\vec{v}) > 0$ (see appendix, section 1). What this means is that we can take any curve on this surface with parameterization $\gamma(t)$, $t \in [0,1]$, and define its length analogously to the way we normally do for Euclidean curves, namely

$$\operatorname{len}(\gamma) = \int_0^1 \sqrt{Q(\gamma'(t),\gamma'(t))}dt.$$

We can then define the distance between two points on this surface, $\vec{u}$ and $\vec{v}$, to be the infimum over the lengths of all possible curves along the surface which join $\vec{u}$ to $\vec{v}$. When this is done, the formula for the distance turns out to be [3]

$$\cosh^{-1}(-Q(\vec{u},\vec{v})).$$

It turns out that the infimum of the lengths of these curves is always achievable by a unique geodesic, and this geodesic turns out to be the curve joining $\vec{u}$ to $\vec{v}$ which is contained in the plane spanned by $\vec{u}$ and $\vec{v}$. To get an equation for this line, we pick a vector $\vec{w}$ in this plane such that $Q(\vec{w},\vec{w}) = 1$ (so it will be off the hyperboloid) and such that it is orthogonal to $\vec{v}$ with respect to this quadratic form, i.e. $Q(\vec{w},\vec{v}) = 0$. Then one can check, by substituting into the formula for the hyperboloid, that the equation of the curve contained in this plane and passing through $\vec{u}$ and $\vec{v}$ is

$$\gamma(t) = \sinh(t)\vec{w} + \cosh(t)\vec{v}.$$

The hyperboloid model is nice in that the way it is defined is most closely aligned with what we would expect given the constructions of spherical and Euclidean geometry. However, as a tool for visualization or analyzing isometries, it is less ideal. We therefore next turn to the Poincaré disk model. This model can be obtained from the hyperboloid model by joining a point on the hyperboloid to the point $(0, 0, -1)$ with a straight line, and mapping it to the point where this line intersects the xy-plane. Doing this for every point on the hyperboloid will fill out the unit disk centered at the origin in the xy-plane:

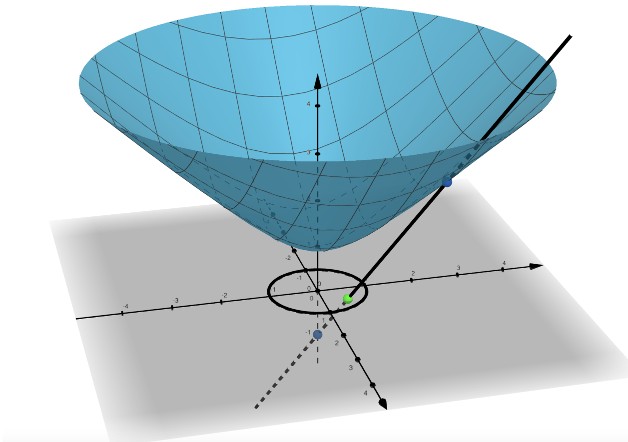
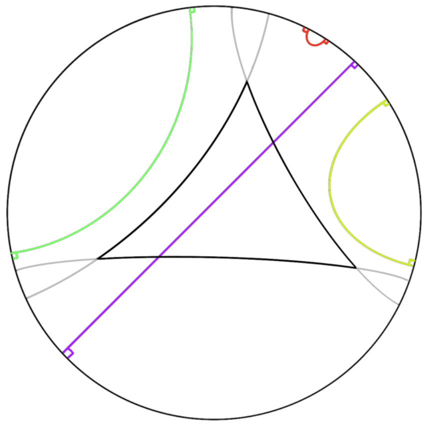


Fig. 2.1.1: Projection onto Poincaré Disk (Image created in GeoGebra)

This projection map is a nice choice because it is conformal, i.e. it preserves angles between intersecting geodesics [4]. It also maps hyperbolic circles in the hyperboloid to Euclidean circles in the plane, though the center of the image circle will likely not coincide with the image of the original hyperbolic circle. A geodesic on the hyperboloid will project to either a circular arc which meets the boundary of the disk at right angles, or to a diameter of the disk if the geodesic lies in a plane containing $(0, 0, -1)$. Thus these will be the geodesics or straight lines of the Poincaré disk. We can illustrate a few of these below. Observe that the black geodesics shown intersect to form a hyperbolic triangle. We will demonstrate this in more detail later, but one can already see that the angles at each vertex appear smaller compared to what we'd expect in the Euclidean plane. Indeed, it will be the case that the sum of the angles of any triangle in the hyperbolic plane is strictly less than π radians. This is one of the important qualitative differences between hyperbolic and Euclidean geometry.



Another important qualitative difference is to do with Parallel lines, or lines which never meet one another. In the Euclidean plane, parallelism defines an equivalence relation on lines. If line A is parallel to line B, and B is parallel to C, then line A must also be parallel to C. This is not so in the hyperbolic plane. For instance, in the picture above the right edge of the triangle and the purple geodesic are both parallel to the green geodesic, yet these two lines intersect and so are not parallel to one another. In fact, one can imagine taking one of these geodesics and swiveling it about the point of intersection away from the green geodesic, and so obtain a whole 1-parameter family of geodesics which are parallel to the green geodesic yet not parallel with each other.

The primary advantage of the Poincaré disk model is in its ease of visualization. One can see the whole plane at once, since it is being represented on a bounded subset of the Euclidean plane. For this reason it will be the model of the hyperbolic plane we will use for the remainder of this work, following this chapter. The fact that it is represented in the Euclidean plane also means we can think of it, in particular, as a subset of the complex plane $\mathbb{C}$. This gives it both an algebraic structure and a coordinate scheme which we can take advantage of. For one thing, any point in the Poincaré disk can be defined by its coordinates (x, y) in the plane. To obtain a metric on this disk, let p and q be points on it. We can first undo the projection and map these points back to the hyperboloid,then measure the distance between them there, and then assign this value to be the distance between p and q in the poincaré disk. In other words, we are defining our metric on the disk such that the projection map from the hyperboloid to the disk is an isometry. Doing this yields the metric [5]

$$d(p, q) = \cosh^{-1}\left(1 + 2\frac{|p - q|^2}{(1 - |p|^2)(1 - |q|^2)}\right),$$

where $|x|$ refers to the Euclidean distance between the point x and the center of the disk, and so $|p - q|$ is the Euclidean distance between points p and q. There is also a simpler formula for the differential of arclength in this model. In the hyperboloid model, this was $ds^2 = dx^2 + dy^2 - dz^2$, given by the quadratic form. One can then integrate a curve with respect to ds to get its length, and by change of coordinate this is equivalent to the integral definition of arclength in the hyperboloid model given above. In the Poincaré disk model, the differential of arclength and the associated arclength integral for a parameterized curve $\gamma(t) = (x(t), y(t))$, $t \in [0, 1]$, are instead [4]

$$ds^2 = 4\frac{dx^2 + dy^2}{(1 - x^2 - y^2)^2}, \quad \mathrm{len}(\gamma) = \int_0^1 \frac{2\sqrt{x'(t)^2 + y'(t)^2}}{1 - (x(t)^2 + y(t)^2)} dt.$$

This formula is a bit simpler than the one given above for the actual metric function. Because the distance between any two points is defined to be the infimum of the arclength of the curves which connect those two points, if an isometry preserves the arclength of curves it must also preserve the distance between points and so be an isometry. Since the formula for the differential of arclength tends to be simpler, we will take advantage of this method in the next section to show that certain transformations of the hyperbolic plane are isometries. Before this though, we will look at one more model of the hyperbolic plane.

The upper half-plane model of the hyperbolic plane consists of the points in the plane which lie strictly above the x-axis, or viewing it as a subset of the complex plane, it is the set of points with positive imaginary part. We can map from the Poincaré disk model to the upper-half plane model via a Möbius transformation. This is a function on the complex plane of the form $f(z) = \dfrac{az + b}{cz + d}$ where a, b, c, d are complex numbers such that $ad - bc \neq 0$, in order to avoid it being a constant map. These transformations are nice since they are conformal, thus we can be sure that angles are accurately represented in the upper half-plane model as well. Moreover, a Möbius transformation is determined by where it sends 3 points. In our case, we can observe that we need to map the boundary of the Poincaré disk to the boundary of the upper-half plane, which is the real axis. Convenient choices of points which will guarantee this are $f(1) = 1$, $f(-1) = -1$, and $f(-i) = 0$. The first two equations can be used to deducae that $a + b = c + d$ and $b - a = c - d$. Adding and subtracting these equations respectively yields $b = c$ and $a = d$. Moreover, for the final equation to be true we need $-ai + b = 0$, or $b = ai$. Putting all this together, we conclude that

$$f(z) = \frac{az + ai}{aiz + a} = \frac{z + i}{iz + 1}.$$

We can then use this function to map the Poincaré disk into the upper half-plane. Just as before, the act of asserting this map to be an isometry between the models forces us into a certain metric. Here we will only give the differential of arc length for this metric and the associated arclength integral, since that is all that we will need for the next section [4]:

2.2 Isometries of the Hyperbolic Plane

In this section, we will see what the isometries of the hyperbolic plane look like algebraically, and also discuss their classification based on the number of points they fix. However, we will focus our attention on orientation-preserving isometries, which do not reflect through a line and so excludes reflections and glide reflections. This is simply because they will not be relevant to what we do later on.

To motivate the description of the orientation-preserving isometries of the hyperbolic plane, we will first make some observations. Firstly, because the upper half-plane model was obtained from the hyperboloid model by a composition of conformal maps, angles are preserved in the upper-half plane model. For a map to be an isometry, we therefore need it to be a conformal map of the complex plane as well so that it does not change angles. Moreover, Euclidean circles in the Poincaré disk model were also hyperbolic circles. And since Möbius transformations like that used to map from the disk to the upper half-plane preserve Euclidean circles (so long as they do not pass through the pre-image of infinity), we conclude that Euclidean circles in the upper half-plane model are also hyperbolic circles since no circle inside the disk can pass through the preimage of infinity, which is i. For an isometry to preserve hyperbolic circles, it must therefore also preserve Euclidean circles. The only orientation-preserving maps from the complex plane to itself which are both conformal and circle-preserving are the Möbius transformations. These Möbius transformations in our case must also preserve the boundary of the plane which is the real line, and so the parameters a, b, c, and d must be real themselves.

Finally, we can also compute the imaginary part of the output of one of these Möbius transformations fairly easily:

$$\begin{aligned}
\mathrm{Im}\left(\frac{az+b}{cz+d}\right) &= \mathrm{Im}\left(\frac{(az+b)(c\overline{z}+d}{||cz+d||^2}\right) \\
&= \frac{1}{||cz+d||^2}\mathrm{Im}\left(ac||z||^2+adz+bc\overline{z}+bd\right) \\
&= \frac{1}{||cz+d||^2}(ad\mathrm{Im}\,(z)+bc\mathrm{Im}(z)) \\
&= \frac{ad-bc}{||cz+d||^2}\mathrm{Im}(z).
\end{aligned}$$

If a map is to be an isometry of the upper half-plane, it most certainly send the upper half-plane to itself. Therefore we need this imaginary part to be positive. We know by assumption that $\mathrm{Im}(z) > 0$, and also of course $||cz + d||^2 > 0$. Thus for any isometry of the upper half-plane, we must require that $ad - bc > 0$. Having wittled down the candidates for isometries so much, we may wonder if what we have now are in fact all isometries of the plane. In fact they are. In particular, we have the proposition

Proposition 2.2.1. *If $f(z) = \dfrac{az + b}{cz + d}$ with $a, b, c, d \in \mathbb{R}$ and $ad - bc > 0$, then f is an isometry of the upper half plane model.*

A nice property of Möbius transformations is that they correspond to matrix multiplication. That is, the map ψ which sends the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ to $\dfrac{az + b}{cz + d}$ is a homomorphism from the group of real linear transformations of $\mathbb{R}^2$ under matrix multiplication to the group of Möbius transformations under function composition. The kernel of this homomorphism is the matrices of the form λI, for $\lambda \in \mathbb{R} - \{0\}$. Thus the first isomorphism theorem from group theory tells us that, if this proposition is correct, the group of isometries of the hyperbolic plane is isomorphic to $PSL(2, \mathbb{R})$. Let us now prove the above proposition.

Proof. Let $f(z) = \dfrac{az + b}{cz + d}$ be as stated in the proposition. Using the formula for the arclength of a curve in the upper half-plane model we saw in the previous section, we can calculate that for any path $(x(t), y(t) = \gamma(t)$, for $t \in [0, 1]$, we have

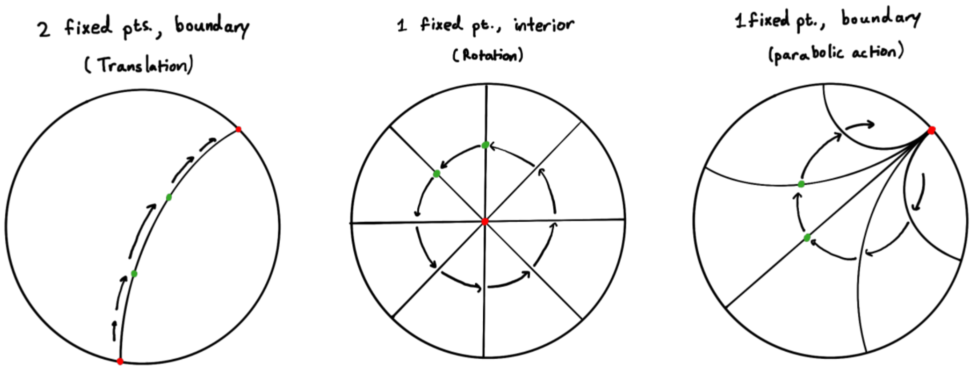
$$\begin{aligned}
\mathrm{len}(f(\gamma)) &= \int_0^1 \frac{||\frac{d}{dt}\frac{a\gamma(t)+b}{c\gamma(t)+d}||}{\mathrm{Im}\left(\frac{a\gamma(t)+b}{c\gamma(t)+d}\right)} dt \\
&= \int_0^1 \frac{||\frac{a\gamma'(t)(c\gamma(t)+d)-(a\gamma(t)+b)\cdot c\gamma'(t)}{(cz+d)^2}||}{\frac{ad-bc}{||c\gamma(t)+d||^2}\mathrm{Im}(\gamma(t))} dt \\
&= \int_0^1 \frac{||ad\gamma'(t) - bc\gamma'(t)||}{(ad - bc)\mathrm{Im}(\gamma(t))} dt \\
&= \int_0^1 \frac{||\gamma'(t)||}{\mathrm{Im}(\gamma(t))} dt = \mathrm{len}(\gamma).
\end{aligned}$$

Thus the arclength of any curve in the upper half-plane is preserved by such a transformation, from which it follows that the distance between any two distinct points must be preserved as well, as this distance is defined as the infimum of the arclengths of the curves joining the two points. ☐

Having a better understanding of hyperbolic isometries now, we can begin to classify them further. In particular, recall that a Möbius transformation is determined uniquely by where it sends three points. As the identity map is a Möbius transformation, we can conclude that if a Möbius transformation fixes 3 points, it must be the identity map. Any hyperbolic isometry which is not the identity then must fix either 1 or 2 points of the hyperbolic plane or of the boundary. It cannot fix zero points of the plane or boundary, since we know that the corresponding isometry of the Poincaré disk model maps the disk to itself, and so by the Brouwer fixed point theorem must have a fixed point somewhere inside the disk.

It seems there are therefore 5 camps of isometries: 2 fixed points on the boundary, 1 fixed point on the boundary and 1 in the interior, 2 fixed points in the interior, 1 fixed point on the boundary, and 1 fixed point in the interior. However, we can rule out the second and third possibilities, since an isometry which which fixes two points must also fix the geodesic joining them (between a point in the plane and a point on the boundary there is also a unique geodesic, as we will see in the next section). Distances between points along this geodesic and one of the fixed points must also be preserved by the isometry, and so these will be fixed points as well. Thus this isometry would in fact be the identity as it has more than 2 fixed points.

The remaining cases supply us with 3 classes of isometries. There are translations, which possess 2 fixed points on the boundary which lie are the endpoints of the geodesic we translate along. There are rotations which have just one fixed point in the interior. And there are so-called "parabolic" isometries which have 1 fixed point on the boundary. These are unique to hyperbolic geometry. We can picture these isometries below:



2.3 The Boundary at Infinity

The Poincaré disk model and upperhalf-plane model we saw in section 1 of this chapter have a natural boundary associated to them, namely the boundary of the disk itself in the case of the Poincaré disk and the real axis in the case of the upper half-plane. These points are not part of the plane itself, as they lie at an infinite distance away from any point in the plane.

These boundary points instead represent sequences of points in the plane which head off to infinity in some precise sense.

Of course, not just any divergent sequence of points will be valid, for instance a sequence of points that spirals outward toward infinite would not seem to approach any particular boundary point. One way to specify which divergent sequences of points are valid is simply to take a sequence of points in the Poincaré disk model of the hyperbolic plane, re-interpret them as points inside the Euclidean complex plane, and check if their limit in that more familiar metric space is a point on the unit circle.

If we wish to define these points strictly in terms of the hyperbolic metric though, we can instead consider geodesic rays that stay bounded distance apart. That is, if we have two hyperbolic rays and supply them with unit-speed parameterizations $\gamma_1(t)$ and $\gamma_2(t)$ for $t \in [0, \infty)$, then these two rays will share an endpoint at infinity if and only if there exists a constant $C > 0$ for which $d(\gamma_1(t), \gamma_2(t)) < C$ for all t. If we relate two hyperbolic geodesics if they're unit speed parameterizations stay bounded distance apart, then this in fact defines an equivalence relation on the set of hyperbolic rays. It is obviously reflexive and symmetric by the uniqueness of unit-speed parameterizations and the symmetry of the metric, and moreover the triangle inequality gives us transitivity. We can then define the points on the boundary of the hyperbolic plane as equivalence classes of geodesic rays that stay a bounded distance apart.

This space of boundary points is typically denoted $\partial\mathbb{H}^2$. It is homeomorphic to the circle S^1, as we intuitively expect from looking at the Poincaré disk model. Moreover, we can also see that between any two distinct boundary point there is exactly one geodesic arc. Reasoning within the Poincaré disk model, we know that any geodesic is either a diameter or the arc of another circle which meets the boundary of the disk at right angles. Given two points on the boundary, we can draw the tangent lines to the boundary at these points. If these tangent lines meet at a point, then reflectional symmetry through the midpoint of the chord joining our two boundary points tells us that this intersection point must be the same distance from both boundary points, and so they lie on a circular arc centered at that intersection. Since the tangents meet the boundary at right angles, so will the circular arc joining the two boundary points which is perpendicular to the tangents. This settles existence. For uniqueness, simply observe that intersection points of lines must be unique. And otherwise if the tangents do not intersect, then they must be parallel and so bound a diameter of the circle which is a geodesic in this model. Thus every pair of distinct boundary points has a unique geodesic joining them. This fact will be useful later when we want to define the space of geodesics of the hyperbolic plane and of hyperbolic surfaces.

To finish off this section, we will note an important property of the isometries of the hyperbolic plane, which is that they act transitively on the set of triples of distinct points on the boundary. That is, given three distinct points (x, y, z) as inputs and 3 distinct points (x', y', z') as outputs, there exists a unique isometry which maps the corresponding inputs to the corresponding outputs. Note that we can talk meaningfully about hyperbolic isometries acting on the boundary despite it not being apart of the hyperbolic plane proper, since these isometries are Möbius transformations of the whole complex plane and so extend to a continuous action on the unit circle in particular. This result will be useful in the next section where we derive certain properties of hyperbolic triangles.

2.4 Hyperbolic Areas and Triangles

At this point, we have developed the notions of hyperbolic lengths, isometries, and angles. The final geometric notion we will need from hyperbolic geometry is that of hyperbolic area. In order to find a good definition of hyperbolic area, we will employ an axiomatic approach. A few reasonable conditions that any useful notion of area should satisfy are

(i) The area of any region $A \subseteq \mathbb{H}^2$ is non-negative.

(ii) If $A, B \subseteq \mathbb{H}^2$ are disjoint regions, then $\mathrm{Area}(A \cup B) = \mathrm{Area}(A) + \mathrm{Area}(B)$.

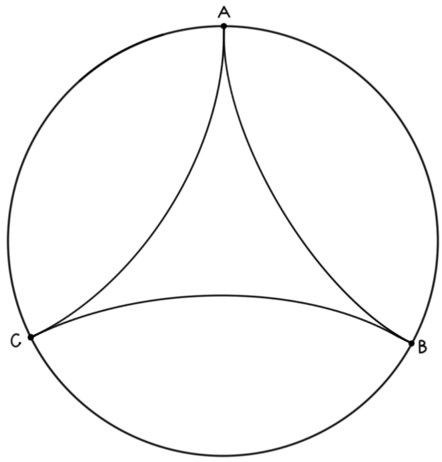
(iii) If γ is an isometry of $\mathbb{H}^2$ and $A \subseteq \mathbb{H}^2$, then $\mathrm{Area}(\gamma \cdot A) = \mathrm{Area}(A)$.

(iv) The area of any point or line is zero.

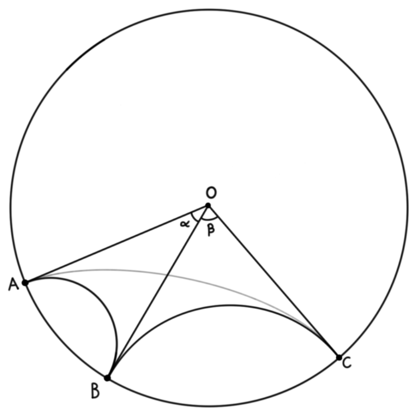
For our purposes in this document, a "region" will more precisely be any (not necessarily bounded) polygonal region. And as every finite-sided polygon has a finite triangulation, condition (ii) allows us to restrict even further to just finding areas of triangular regions. It is then a somewhat remarkable fact that these axioms are sufficient to determine the area function for triangles up to a scalar multiple. In particular, we will prove the following theorem:

Proposition 2.4.1. *If $T \subseteq \mathbb{H}^2$ is a triangle with vertex angles α, β, and γ, then we have $\mathrm{Area}(T) = k\,(\pi - (\alpha + \beta + \gamma))$, for some constant k which is independent of T.*

Proof. Let A, B, and C be points on the boundary of the hyperbolic plane, $\partial\mathbb{H}^2$. If we connect these points at infinity pair-wise with geodesics, we form a region of the plane known as the ideal triangle:



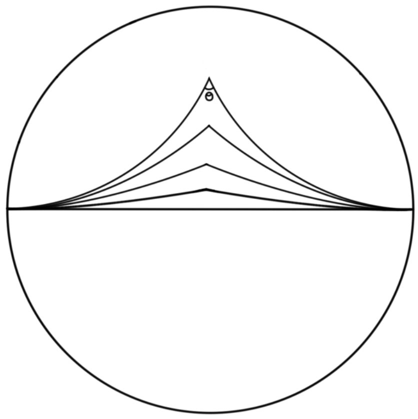
We learned in section 3 of this chapter that the isometries of $\mathbb{H}^2$ act transitively on the set of triples of points on the boundary, meaning we can take any ideal triangle to any other. So it is justified to refer to this shape as “the" ideal triangle of $\mathbb{H}^2$. Let us simply define the area of this triangle to be M. We will now show how the area of any other triangle is completely determined after this choice. To begin, let us first consider the next simplest case of a triangle which has two vertices at infinity and one point in the interior of $\mathbb{H}^2$. If we have two such triangles, one with a vertex angle of α and another with a vertex angle of β, we can translate one of the triangles so that the two are joined at a side:



Let the area of such a triangle with vertex angle θ be $f(\theta)$. We can observe that in the picture above the triangle OAC is a triangle with exactly one non-infinite vertex having a vertex angle of $\alpha+\beta$, so its area will be $f(\alpha+\beta)$. On the other hand, we can also obtain its area by adding the two triangle OAB and OBC that cover it, and subtracting of the area of the ideal triangle ABC. Thus we have

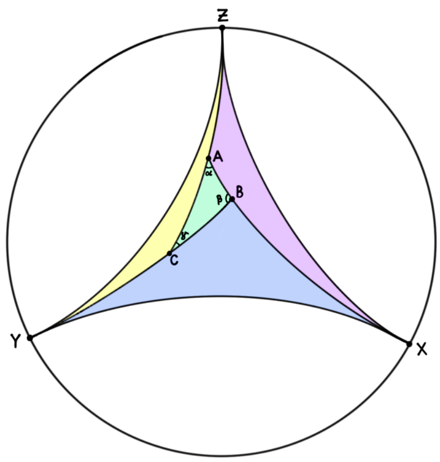
$$f(\alpha+\beta) = f(\alpha) + f(\beta) - M.$$

In other words, if we are willing to believe the function is continuous, then this functional equation implies $f(\theta)$ is an affine function, $f(\theta) = c\theta + d$. It is easy to see that d must be equal to M, given the above functional equation. To determine c we note that as we let θ approach π our triangle approaches a geodesic:



Thus we should have $f(\pi) = c\pi + M = 0$. In other words, $c = -\dfrac{M}{\pi}$.

Now let us consider a general triangle in $\mathbb{H}^2$, say ABC. We can form the rays $\overrightarrow{AB}$, $\overrightarrow{BC}$ and $\overrightarrow{CA}$ and denote their respective endpoints at infinity as X, Y, and Z. By connecting these endpoints at infinity with geodesics, we create the following picture:



Note that the vertex angles at A, B, and C are α, β, and γ respectively. So we see that the triangle AXZ has a vertex angle of $\pi - \alpha$, meaning its area is $-\dfrac{M}{\pi}(\pi - \alpha) + M = \dfrac{M}{\pi}\alpha$. We can similarly conclude that triangle BXY has area $\dfrac{M}{\pi}\beta$ and CZY has area $\dfrac{M}{\pi}\gamma$. Triangle XYZ by definition has area M, so it follows that

$$\begin{aligned}\text{Area}(ABC) &= \text{Area}(XYZ) - (\text{Area}(AXZ) + \text{Area}(BXY) + \text{Area}(CZY)) \\ &= M - \left(\frac{M}{\pi}\alpha + \frac{M}{\pi}\beta + \frac{M}{\pi}\gamma\right).\end{aligned}$$

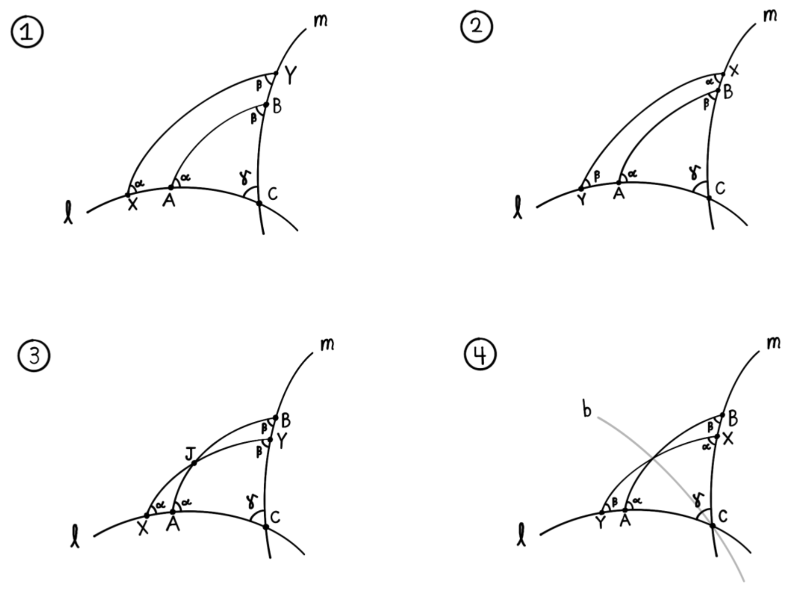
If we then rename M as $k\pi$, we have

$$\text{Area}(ABC) = k\pi - (k\alpha + k\beta + k\gamma) = k(\pi - (\alpha + \beta + \gamma)),$$

which proves the proposition. □

It follows from this proposition that any hyperbolic triangle must have angles which sum to strictly less than π radians, so that its area will be positive. This observation will allow us to conclude something rather strong about hyperbolic triangles: the vertex angles of a triangle determine the triangle up to isometry. A similar statement is true in spherical geometry as well in fact, and really it is Euclidean geometry which is special in that it allows for similar triangles. But since any dilation map would need to change the curvature of our hyperbolic space, and we are fixing the curvature of our space at -1, it is not so surprising that we do not have similar triangles and so a triangle is determined by its angles. Let us see why this is rigorously.

Suppose triangles ABC and XYZ both have vertex angles α, β, and γ, but are not isometric. We can translate Z to C and rotate the triangles so that the edges adjacent to the angle γ in each triangle are coincident. After doing this, there are four possibilities for what our triangles could look like:



In the first case, we notice that the quadrilateral $ABYX$ has vertex angles of α, β, $\pi-\alpha$, and $\pi-\beta$, which sum to 2π. Thus, by triangulating this quadrilateral and applying our formula for the area, we see that this quadrilateral should have area 0, which implies X and Y must lie on top of A and B respectively. This is also true in the second case, since the vertex angles are the same but just in a different order.

The third case is similar and more direct, since if we label the intersection point of $\overline{AB}$ and $\overline{XY}$ as J, then triangle AJX has vertex angles of α and $\pi-\alpha$ which sum to π and imply this triangle has non-positive area. Similarly for triangle BJY. So we will again require that X lies on top of A and Y on top of B.

For the fourth and final case, consider what would happen to the line segment XY if we reflect over the angle bisector of γ. This reflection will necessarily swap the lines ℓ and m which form the angle γ, and so will map Y somewhere onto ℓ and X somewhere onto m. Since the reflection will preserve the angles at these points as well, we see that we will always end up back in either case 1 or case 3, where again conclude that X must lie on top of A and Y on top of B.

Thus, in every case we found and isometry that maps the vertices of XYZ onto the vertices of ABC, which means that these two triangles differ only up to isometry, as desired.

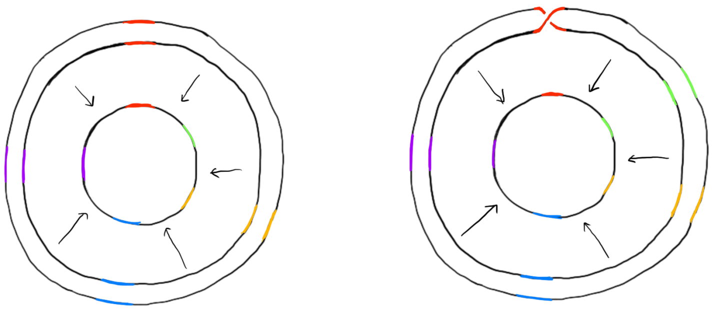
Chapter 3

Covering Spaces

3.1 Concept and Preliminary Examples

Before broadening our study of hyperbolic geometry from the hyperbolic plane to more general hyperbolic surfaces, we will take a quick detour to consider basic covering space theory. This will be necessary both for seeing how hyperbolic surfaces can be constructed from the hyperbolic plane, and to derive properties of hyperbolic surfaces from known properties of the hyperbolic plane.

The notion of a *covering space* is of central importance to topology. When studying topological spaces, it is often useful to form larger spaces by taking copies of your space and "stitching them together" by some means. These larger spaces share certain topological features with the smaller spaces they are formed from, but can also be simpler in other ways such as having trivial fundamental groups in some cases. In general, this stitching process can be thought of in terms of taking an open cover of your space, making n copies of each open set for some chosen value of n and identifying these copies together in an appropriate fashion to form some larger space. The process is illustrated below for a couple examples of coverings of the circle, one by two copies of the circle, and another by a single copy of the circle which "wraps around itself" twice:



Another, more explicit example comes from the projective plane $\mathbb{R}P^2$. It is well-known that this space is non-orientable, and cannot be represented in 3D euclidean space. Hence, we must devise other means of forming pictures of it. One such way this can be done is by recalling the fact that $\mathbb{R}P^2$ is a quotient of the 2-sphere S^2 by the $\mathbb{Z}_2$-action which swaps antipodal points (equivalently, multiplies each point on the sphere by -1). Therefore, a point of the projective plane can be visualized as a pair of antipodal points on the sphere.

Around any two antipodal points of the sphere, we can find open sets containing each which quotient to the same open set of the projective plane. Moreover, this quotient map is a homeomorphism when restricted to either open set. Choosing such open sets for each pair of antipodal points on the sphere produces an open cover of the sphere for which each open set is one of a pair of open sets copied from $\mathbb{R}P^2$, and which of course "stitch together" to form the sphere S^2. So this example does indeed appear to satisfy the informal idea of covering space which we are developing.

Now we must make this idea of a covering space rigorous by providing an unambiguous definition (cf. [6]):

Definition 3.1.1 (Covering)**.** *A covering of a topological space B is a pair (E, p), where E is a topological space and $p : E \to B$ is an n-to-1, continuous function such that for any $b \in B$ there exists an open neighborhood U of b for which $p^{-1}(U)$ consists of n disjoint open sets which map homeomorphically via p to U.*

Some additional terminology can be introduced in light of this definition. We call B the "**base space**" of the covering, and E the "**covering space**" of B. Meanwhile p is the "**covering map**". The "**degree**" of this covering map is the value n referenced in the definition, and this degree can be infinite. The components of $p^{-1}(b)$ which map homeomorphically to U are referred to as "**sheets**" of the covering over b, and we say that b is "**evenly covered**." We often describe a covering as an "n**-sheeted covering**" if it has degree n. Note that a covering is automatically surjective by this definition since each point $b \in B$ needs to be mapped to by $n \geq 1$ sheets.

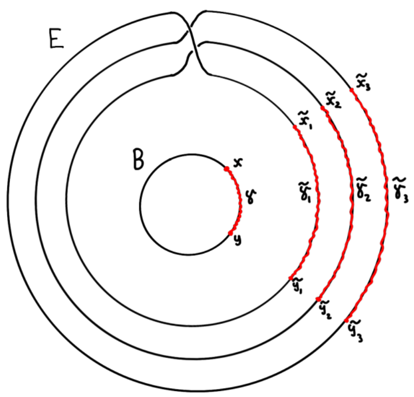
This chapter will adapt most of its results and proofs from [6].

3.2 Path Lifting Property

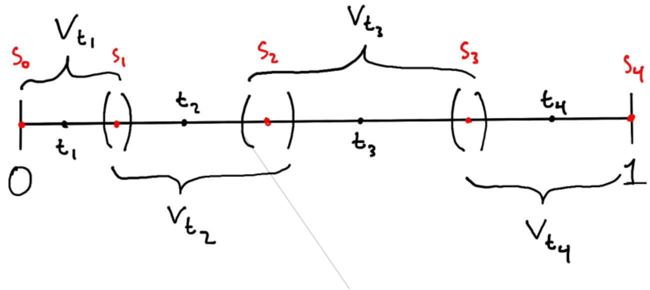
An important property of covering spaces is the "lifting property" of paths:

Proposition 3.2.1 (Path Lifting Property)**.** *Let $\gamma : [0,1] \to B$ be a continuous path in B with $\gamma(0) := b$, $p : E \to B$ an n-sheeted cover, and $\tilde{b}$ a preimage of b, i.e. $\tilde{b} \in p^(b)$. Then there exists a unique path $\tilde{\gamma} : [0,1] \to E$ such that $p \circ \tilde{\gamma} = \gamma$ and $\tilde{\gamma}(0) = \tilde{b}$.*

For instance, if we consider the 3-sheeted cover of the circle by itself, any path in the base space will have 3 "lifts" to the covering space:



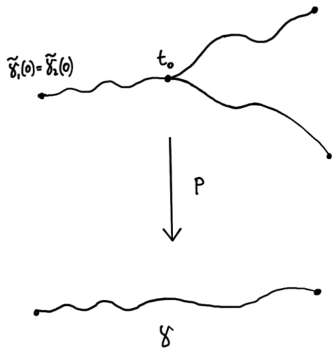
Proof. To prove this we need to demonstrate separately that a path with these properties exists, and also that this path is unique. To show existence, we will leverage the compactness of the interval $[0, 1]$. Recall that a space is compact if any open cover of the space (a distinct notion from a *covering*) has a finite subcover. To construct a useful open cover of $[0, 1]$, we will take each point $t \in [0, 1]$ and an open neighborhood U_t of $\gamma(t)$ such that its preimage under p consists of disjoint sheets that map homeomorphically to it. Since γ is a continuous map, it follows that the preimages $\gamma^{-1}(U_t)$ must each be open in $[0, 1]$ as well. Let V_t be the connected component of U_t containing t. Note that the V_t's form an open cover of $[0, 1]$. The compactness of $[0, 1]$ then gives us a finite subcover, say $V_{t_1}, \ldots, V_{t_k}$. We can then choose a sequence of values $0 = s_0 < s_1 < \ldots < s_{k-1} < s_k = 1$ such that $[s_{i-1}, s_i] \subseteq V_{t_i}$ for all $1 \leq i \leq k$:



To construct $\tilde{\gamma}$, we will note that since we have chosen $\tilde{b} \in p^{-1}(b)$ we know that $\tilde{\gamma}(0)$ must be equal to $\tilde{b}$. Thus we have constructed $\tilde{\gamma}$ on $[0, s_0]$. Now by way of induction, suppose we have constructed $\tilde{\gamma}$ on $[0, s_i]$. We know that $[s_i, s_{i+1}]$ is contained in V_{t_i} by construction, and $p^{-1}(\gamma(V_{t_i}))$ consists of n disjoint sheets homeomorphic to $\gamma(v_{t_i})$. Only one of these sheets can contain $\tilde{\gamma}(s_i)$, say $\tilde{V}_{t_i}$. We can then define $\tilde{\gamma}(t) = (p|_{\tilde{V}_{t_i}})^{-1}(\gamma(t))$ for $t \in [s_i, s_{i+1}]$. This definition agrees with the value of $\tilde{\gamma}(s_i)$ we already have constructed, and so gives us a continuous extension of $\tilde{\gamma}(t)$ to $[0, s_{i+1}]$. We can continue in this fashion until we have define $\tilde{\gamma}$ on $[0, s_k] =$

$[0,1]$. By our construction, it is also clear that $p \circ \tilde{\gamma} = \gamma$, so this is our desired lift.

Finally for uniqueness, suppose that $p \circ \tilde{\gamma}_1 = p \circ \tilde{\gamma}_2 = \gamma$ and that $\tilde{\gamma}_1(0) = \tilde{\gamma}_2(0)$. Then in E these two lifts may look like:



We therefore define $t_0 = \sup\{t \in [0,1] \mid \tilde{\gamma}_1(s) = \tilde{\gamma}_2(s)\ \forall\ s \leq t\}$. This supremum is well-defined since the set contains 0 by our assumption on $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$, and is upper bounded by 1. If $t_0 = 1$ then we are done since $\tilde{\gamma}_1 = \tilde{\gamma}_2$. Otherwise, let U_{t_0} be an open neighborhood of $\gamma(t_0)$ whose preimages under p are homeomorphic sheets. Then the sheet $\tilde{U}_{t_0}$ containing $\tilde{\gamma}_1(t_0) = \tilde{\gamma}_2(t_0)$ must also contain $\tilde{\gamma}_1(t_0 + \varepsilon)$ and $\tilde{\gamma}_2(t_0 + \varepsilon)$ for sufficiently small ε by the continuity of these paths, and moreover we can take a sufficiently small ε such that $\tilde{\gamma}_1(t_0 + \varepsilon) \neq \tilde{\gamma}_2(t_0 + \varepsilon)$ by the fact that t_0 is a supremum. However, since we must have $p(\tilde{\gamma}_1(t_0+\varepsilon)) = \gamma(t_0+\varepsilon) = p(\tilde{\gamma}_2(t_0+\varepsilon))$, this contradicts the fact that p is a homeomorphism (and thus is injective) on $\tilde{U}_{t_0}$. So we cannot have $t_0 < 1$, and the lifts must in fact be the same. $\square$

As a small note, what we have proven here is generalized by the "homotopy lifting property." This states that if $F : Y \times [0,1] \to X$ is a homotopy and $p : \tilde{X} \to X$ is a covering, then there is a unique lift $\tilde{F}$ of F for any specified initialization of $\tilde{F}$ on $Y \times \{0\}$ [6]. In the case of the path lifting property we simply take Y to be the space consisting of just one point. We will only be needing the path lifting property for the rest of this document though, so we will content ourselves with proving just this special case.

3.3 Universal Cover Construction and Deck Transformations

Recall the connected example of a covering space of the circle we saw in the first section. It was a circle which we 'unwound' twice. Or more explicitly, if we identify the circle with the set of complex values e^{it} for $t \in \mathbb{R}$, then we obtain a double cover of the circle from the map $e^{it} \mapsto e^{2it}$, which wraps the circle around itself twice over. We can get an n-sheeted covering from this idea as well, by intsead employing the map $e^{it} \mapsto e^{nit}$. We can imagine unwinding the circle more and more to get larger and larger covering spaces, and we can even unwind completely to obtain the real line $\mathbb{R}$. Indeed, $\mathbb{R}$ is an infinite-sheeted covering space for S^1 using the map $t \mapsto e^{it}$. Visually, one can imagine deforming the real line to form an infinite helical shape in $\mathbb{R}^3$, then projecting it down onto the circle centered on its axis of symmetry.

A covering space which is obtained by "unwinding completely" the base space is what we call the **universal cover**. To see what we mean more precisely by "unwind completely," call the generator for the fundamental group of the covering space $\tilde{S}^1$ say λ, and the generator for the base space S^1 say γ. Then λ is sent by the covering map to the loop γ^n. So, if we identify $\pi_1(S^1)$ with $\mathbb{Z}$, then we obtain a natural identification of $\pi_1(\tilde{S}^1)$ with $n\mathbb{Z}$ (keep this example in mind for the next section where we discuss the subgroup correspondence of coverings). That is, we can think of the fundamental group of the covering space as being obtained by throwing away elements of fundamental group of the base space which do not lift to loops in the covering space (in particular, the fundamental group of the covering space can be identified with a subgroup of the fundamental group of the base space, as we shall see later). By throwing away all elements possible, i.e. everything except the identity, we will obtain a space which is as unraveled as possible. Formally, this means we want the fundamental group of the covering space to be trivial. This leads us to the following definition

Definition 3.3.1 (Universal Cover). *If $p : \tilde{B} \to B$ is a covering such that $\pi_1(\tilde{B})$ is trivial, then $\tilde{B}$ is the universal cover of B.*

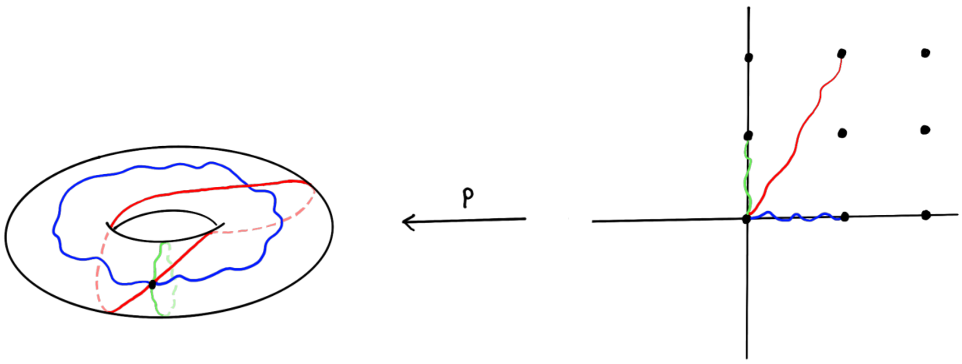
We have already seen how the universal cover of S^1 is $\mathbb{R}$. This extends to the example of the torus, which has as its universal cover the plane $\mathbb{R}^2$ with associated covering map given by $(t, s) \mapsto (e^{it}, e^{is}) \in S^1 \times S^1$. Since the 2-sphere is simply connected, we can also see that the universal cover of the projective plane is S^2. We call this space the universal cover because as we shall see later, any other connected covering space of B is itself covered by $\tilde{B}$.

This universal cover exists only for sufficiently nice spaces. In particular, we have the following theorem:

Proposition 3.3.1 (Existence of Universal Covers). *If B is a path connected, locally path connected, and semi-locally simply connected space, then B has a universal cover $\tilde{B}$.*

We will discuss the definition and relevance of the terms locally path connected and semi-locally simply connected once they become relevant for the construction. Before we begin the construction of the universal cover though, we can at least discuss the path connected condition. Recall that the space B is path connected if between any two points $b_1, b_2 \in B$ there is a path connecting them, i.e. a continuous function $\gamma : [0, 1] \to B$ such that $\gamma(0) = b_1$ and $\gamma(1) = b_2$. The notion of a fundamental group is not well-defined on spaces which are not path connected, so since we are assuming $\tilde{B}$ is simply connected, in particular we are assuming it has a fundamental group at all and so it must be path connected. Any covering map $p : \tilde{B} \to B$ would be a surjective continuous map, and these can only map path connected spaces to path connected spaces, meaning B needs to be path connected as well.

Proof. Let us now begin constructing the space $\tilde{B}$. Choose a basepoint $b \in B$. If we consider a (not nullhomotopic) loop in B based at b, we would like for this loop to lift to a path in $\tilde{B}$ which is not a loop, since otherwise $\tilde{B}$ would not be simply connected. A consequence of this is that every element $[\gamma] \in \pi_1(B)$ must have a different endpoint when lifted to the universal cover $\tilde{B}$:



In fact, this is enough to guarantee that every point in $\tilde{B}$ must correspond to a homotopy class of paths (not necessarily loops though) $[\gamma]$ based at b. To see this, let $\tilde{b}$ be a chosen lift of the base point b, and $\tilde{x}$ be any other point in $\tilde{B}$. Then there is a path $\tilde{\gamma}$ going from $\tilde{b}$ to $\tilde{x}$. Since $\tilde{B}$ is simply connected, there is only one choice of this path $\tilde{\gamma}$ up to homotopy. This path must then project by the covering map to a path γ in B going from b to $x = p(\tilde{x})$. Suppose λ is another path in B joining b to x, and the lift $\tilde{\lambda}$ of this path also joins $\tilde{b}$ to $\tilde{x}$. Then by traversing $ga\tilde{m}ma$ and following this by traversing $\tilde{\lambda}$ backwards, we obtain a loop in $\tilde{B}$ based at $\tilde{b}$ which must be nullhomotopic since $\tilde{B}$ is simply connected. Thus $\tilde{\lambda}$ and $\tilde{\gamma}$ are homotopic to one another in $\tilde{B}$, and this homotopy must project down via the covering map to a homotopy between λ and γ in B. Thus $\tilde{x}$ corresponds uniquely to the homotopy class of paths $[\gamma]$.

This discussion is meant to motivate the following construction of the space $\tilde{B}$:

$$\tilde{B} = \{[\gamma] \mid \gamma : [0,1] \to B, \gamma(0) = n\}.$$

That is, we can identify $\tilde{B}$ with the set of all homotopy classes of paths in B. To obtain a topology on $\tilde{B}$, it suffices to define a basis for it, i.e. a collection of open sets $\mathcal{U}$ such that any other open set is a union of sets in $\mathcal{U}$. For instance, a basis for the topology on $\mathbb{R}^2$ is the collection of all open balls. If a collection of open sets forms an open cover of $\tilde{B}$, and has the property that $V \subset U \in \mathcal{U}$ implies $V \in \mathcal{U}$ for any open set V, then this collection automatically forms a basis. This is because any open set of $\tilde{B}$ will possess an open cover by elements of $\mathcal{U}$ and we can express it as the union of its intersections with these open sets in $\mathcal{U}$, each of which must also be in $\mathcal{U}$ be the closure property.

In our case, let $\mathcal{U}$ be the collection of open sets in B for which any $U \in \mathcal{U}$ is path connected and any loop in U is nullhomotopic in B (not necessarily in U, though). The fact that this collection $\mathcal{U}$ forms an open cover of B amounts to the statement that for any point $x \in B$, there exists an open neighborhood U around it which is path connected and such that any loop in U is null-homotopic in B. These properties are respectively referred to as "locally path connected" and "semi-locally simply connected." Since we assumed these as part of the proposition, we may safely conclude that $\mathcal{U}$ is an open cover of B. Moreover, if $V \subseteq U \in \mathcal{U}$ is path connected and open, then any loop in it will also be a loop in U and so will be nullhomotopic in B, meaning $V \in \mathcal{U}$. So $\mathcal{U}$ forms a basis for B (to be sure, our reasoning for why such an open cover is a basis can be modified by representating an arbitrary open set as the union of the *path-connected components* of the intersections of it with the elements of $\mathcal{U}$).

We can extend this basis to a basis $\tilde{\mathcal{U}}$ for the topology of $\tilde{B}$. Let $\tilde{U}_{[\gamma]} \in \tilde{\mathcal{U}}$ if there exists a $U \in \mathcal{U}$ and a path γ joining b to some point $x \in U$ such that [6]

$$\tilde{U}_{[\gamma]} = \{[\gamma \cdot \eta] \mid \eta : [0,1] \to U,\ \eta(0) = \gamma(1)\}.$$

Note that the set $\tilde{U}_{[\gamma]}$ is actually independent of the point x, and once this point is fixed only depends on the homotopy class of γ. Since there is at least one such open set containing each point $[\gamma] \in \tilde{B}$ this collection is an open cover, and by construction it will also share the closure property discussed above (restricted to path-connected open subsets of elements of $\tilde{\mathcal{U}}$). The topology on $\tilde{B}$ is then generated by taking unions of elements of $\tilde{\mathcal{U}}$.

Now that we have the covering space, we next need to define the covering map p. When motivating the construction of this space, we identified a point in the universal cover with the path joining it to $\tilde{b}$ and projected down to B. Therefore, it follows that we should define p by $[\gamma] \mapsto \gamma(1)$, i.e. we map each homotopy class of paths based at b to their common endpoint. We first observe that each point in B is evenly covered by this map, since two paths will share it as an endpoint if and only if they differ by an element of $\pi_1(B)$. Thus the preimage of any point has the same size, namely $|\pi_1(B)|$.

Next, we need to confirm that any point $x \in B$ has an open neighborhood U whose preimage consists of disjoint open sets, each of which map homeomorphically via p to U. In particular, simply take this neighborhood to be in $\mathcal{U}$. Then for any path γ joining b to x we can define the open set $\tilde{U}_{[\gamma]}$ as above. By construction, $\tilde{U}_{[\gamma]}$ maps surjectively to U by p. To see that this map is injective on $\tilde{U}_{[\gamma]}$, suppose that $p([\gamma \cdot \eta_1]) = p([\gamma \cdot \eta_2])$, i.e. $\gamma \cdot \eta_1(1) = \gamma \cdot \gamma \cdot \eta_2(1)$. Then it follows that $\eta_1(1) = \eta_2(1)$, and moreover we know that $\eta_1(0) = \eta_2(0) = \gamma(1)$. Thus $\overline{\eta}_2 \cdot \eta_1$ is a loop based at $x = \gamma(1)$ which is contained in U, and thus is nullhomotopic. It follows that $[\eta_1] = [\eta_2]$ and so $[\gamma \cdot \eta_1] = [\gamma \cdot \eta_2]$.

To see that p is a homeomorphism on $\tilde{U}_{[\gamma]}$ we can finally note that p gives a bijection between a basis for the topology on $\tilde{U}_{[\gamma]}$ and the basis for the topology on U. That is, each $\tilde{V}_{[\gamma']} \in \tilde{\mathcal{U}}$ contained in $\tilde{U}_{[\gamma]}$ corresponds to a unique $V \in \mathcal{U}$ contained in U by setting $V = p(V_{[\gamma']})$, and conversely each $V \in \mathcal{U}$ contained in U corresponds to a unique $\tilde{V}_{[\gamma']} \in \tilde{\mathcal{U}}$ contained in $\tilde{U}_{[\gamma]}$ given by $\tilde{V}_{[\gamma']} = \tilde{U}_{[\gamma']} \cap p^{-1}(V) = \tilde{U}_{[\gamma]} \cap p^{-1}(V)$. This bijective correspondence between the bases then extends to a bijective correspondence between the whole topologies, which with the fact that p is a bijection on $\tilde{U}_{[\gamma]}$ tells us that it is in fact a homeomorphism on this open set.

The fact that the open sets $\tilde{U}_{[\gamma]}$ for the various homotopy classes of paths joining b to x are pairwise disjoint follows from the observation that if $[\gamma_1 \cdot \eta_1] = [\gamma_2 \cdot \eta_2]$ with $[\gamma_1] \neq [\gamma_2]$ then in $\tilde{B}$ the path $\overline{\eta}_1 \cdot \eta_2$ must connect two different pre-images of x (see figure below). It therefore must project by p to a loop based at x which is not null-homotopic. Simultaneously though, this path is contained in U, and every path contained in U must be nullhomotopic, so we reach a contradiction. Thus $\tilde{U}_{[\gamma_1]}$ must be disjoint from $\tilde{U}_{[\gamma_2]}$ for any $[\gamma_1] \neq [\gamma_2]$.

Finally, the continuity of p follows from the fact that any open set $W \subseteq B$ can be expressed as a union of open sets in $\mathcal{U}$, say $W = \cup_{i \in I} U_i$. The preimage of W is then the union of the preimages $p^{-1}(U_i)$, which we have already seen are all open in $\tilde{B}$. That is, for any open set W in B, the preimage $p^{-1}(W)$ is open in $\tilde{B}$. This allows us to conclude that p is indeed a covering map.

At this point we have our covering space and the covering map. To finish this construction all that is left is to show that $\tilde{B}$ is indeed simply connected. Firstly, it is path-connected since the point $[b] \in \tilde{B}$ corresponding to the constant path at b can be connected to any other path $[\gamma]$ via the map $t \to [\gamma_t]$, where γ_t is the path defined by $\gamma_t(s) = \gamma(s)$ for $s < t$ and $\gamma_t(s) = \gamma(t)$ for $s \geq t$. Moreover, we know that any loop in $\tilde{B}$ is the lift of some loop in B. If this loop in B is γ, then the lift of this loop is $t \mapsto [\gamma_t]$ as described above. The fact that this lifted path is a loop in $\tilde{B}$ means that $[\gamma_1] = [b]$, i.e. $[\gamma] = [b]$. In other words, there is a homotopy between γ and the trivial loop at b. This then lifts to a homotopy between the lift of γ and the constant loop at $[b] \in B$. In other words, every loop in $\tilde{B}$ is null-homotopic and so $\tilde{B}$ is simply connected. This completes the construction, and proves that every space B which satisfies the given conditions possesses a universal cover.

□

The condition that a space be path connected, locally path connected, and semi-locally simply connected are rather technical, and for our purposes in this document are a little overkill. This is because we will pretty much only ever be working with manifolds as our topological spaces, and these satisfy the requirements automatically on account of the fact that they are everywhere locally homeomorphic to $\mathbb{R}^n$.

To end this section, we will briefly discuss an additional connection between the fundamental group $\pi_1(B, b)$ and the universal cover $(\tilde{B}, p)$. The space $\tilde{B}$ actually possesses symmetries which are encoded in $\pi_1(B, b)$. These are the symmetries which respect the covering map p, that is they are homeomorphisms $f : \tilde{B} \to \tilde{B}$ such that $p \circ f = p$. We call such a homemorphism f a **deck transformation** of the space $\tilde{B}$. For example, we know that $\mathbb{R}^2$ covers the torus T, and specifically this covering is achieved by quotienting by the translational action of $\mathbb{Z}^2$ on the plane. Therefore, if we pre-compose our quotienting map (which is also our covering map) by one of these translations, it will not affect where any point is mapped to, so the group of deck transformations in this case is precisely $\mathbb{Z}^2$. It is also no coincidence that $\pi_1(T) \cong \mathbb{Z}^2$. This is true in general, in fact. The group of deck transformations of a universal cover $\tilde{B}$ is isomorphic to the fundamental group $\pi_1(B)$ (see [6], prop. 1.39). This characterization of the fundamental group will come in handy later when we discuss orbifolds and orbifold fundamental groups.

3.4 The Subgroup Correspondence

For coverings of a given space B, it turns out that there is a purely algebraic way to classify and characterize them. This is known as the "Galois correspondence" or "subgroup correspondence" of coverings. It relates the coverings of a space B with the conjugacy classes of subgroups of the fundamental group $\pi_1(B, b)$, for some chosen base point $b \in B$. In particular, given a covering $p : E \to B$ let $\tilde{b}$ be the base point of E with $p(\tilde{b}) = b$. We can then construct the homomorphism $p_* : \pi_1(E, \tilde{b}) \to \pi_1(B, b)$ defined by $p_*([\tilde{\gamma}]) = [p \circ \tilde{\gamma}]$. We will take it as given that this is indeed a homomorphism. What we want to show is that the image of this homomorphism determines the covering (E, p), up to an appropriate notion of equivalence of coverings.

We begin with the following proposition [6]:

Proposition 3.4.1. *Let B be a path connected, locally path connected, and semi-locally simply connected space. Then for any subgroup $H \leq \pi_1(B, b)$, there exists a covering $p : E \to B$ and a* *basepoint* $\tilde{b} \in E$ *such that* $p_*(\pi_1(E, \tilde{b})) = H$.

Note that we are saying that these two groups are *equal*, not simply isomorphic. In other words we are thinking of them as sets of particular homotopy classes as well as algebraic structures. In fact, the isomorphism between their algebraic structures will follow from their equivalence as sets, so this is what we aim to prove. What follows is adapted from [6].

Proof. Since B has the requisite properties for possessing a universal cover, we will let $\tilde{B}$ be that universal cover. Define an equivalence relation on the points of $\tilde{B}$ by $[\gamma] \sim [\gamma']$ if and only if $\gamma(1) = \gamma'(1)$ and $[\gamma' \cdot \overline{\gamma}] \in H$. Note that $\gamma' \cdot \overline{\gamma}$ is a loop based at b. This condition is an equivalence relation since G is a subgroup of $\pi_1(B, b)$. That is, it is reflexive since for any $[\gamma]$ we have $\gamma(1) = \gamma(1)$ and $[\gamma \cdot \overline{\gamma}]$ is homotopic to the constant path $[b]$ which is in H since it is the identity. It is symmetric since $[\gamma' \cdot \overline{\gamma}] \in H$ implies that its inverse $[\gamma \cdot \overline{\gamma'}] \in H$ as well. And finally it is transitive since for any $[\gamma' \cdot \overline{\gamma}] \in H$ and $[\gamma'' \cdot \overline{\gamma'}] \in H$ we have $[\gamma'' \cdot \overline{\gamma'}] \cdot [\gamma' \cdot \overline{\gamma}] = [\gamma'' \cdot \overline{\gamma'} \cdot \gamma' \cdot \overline{\gamma}]$ homotopic to $[\gamma'' \cdot \overline{\gamma}]$ must be in H as well since H is closed under concatenation of paths.

Let E be the space obtained from quotienting $\tilde{B}$ by this equivalence relation, and $\tilde{b} \in E$ be the equivalence class containing the constant path $[b]$. We can define $p : E \to B$ by $p([\gamma]) = \gamma(1)$ just as we did for the universal cover. This function is well-defined since all representatives of any equivalence class are required to share the same endpoint. Moreover, if $\gamma(1) = \gamma'(1)$ then for any path η we also have $\gamma \cdot \eta(1) = \gamma' \cdot \eta(1)$, and also $[(\gamma \cdot \eta) \cdot (\overline{\gamma' \cdot \eta})] = [\gamma \cdot \eta \cdot \overline{\eta} \cdot \overline{\gamma'}]$ is homotopic to $[\gamma \cdot \overline{\gamma'}]$. Thus if $[\gamma] \sim [\gamma']$, then for any path η we also have $[\gamma \cdot \eta] \sim [\gamma' \cdot \eta]$. In particular, if one point of an open neighborhood $\tilde{U}_{[\gamma]}$ is identified with a point of $\tilde{U}_{[\gamma']}$ by this equivalence relation, then the whole neighborhoods are identified.

It follows then that the topology on E is generated by the basis formed by quotienting the basis $\tilde{\mathcal{U}}$ for $\tilde{B}$ we saw in the previous section. So just as before, p will be a continuous map from E to B such that each point in B has an open neighborhood whose preimage consists of disjoint open sets that map homeomorphically by p onto it.

To confirm that p evenly covers every point of B, pick a point $x \in B$ and recall that $[\gamma] \in p^{-1}(x)$ if and only if $\gamma(1) = x$. Fix $[\gamma] \in p^{-1}(x)$, and consider another equivalence class represented by $[\gamma']p^{-1}(x)$. We know that $[\gamma' \cdot \overline{\gamma}]$ is in $\pi_1(B, b)$ since these paths share an endpoint, and in particular we will say it lies in the coset Hg for some $g \in \pi_1(B, b)$. We can then observe that $[\gamma'' \cdot \overline{g}] \in Hg$ as well if and only if $[\gamma''] \sim [\gamma']$. The 'if' direction is easy since $[\gamma''] \sim [\gamma']$ means that $[\gamma'' \cdot \overline{\gamma'}] \in H$, so $[\gamma' \cdot \overline{\gamma}] \in Hg$ implies $[\gamma'' \cdot \overline{\gamma'}]\ cdot[\gamma' \cdot \overline{\gamma}] = [\gamma'' \cdot \overline{\gamma}] \in Hg$ as well. For the 'only if' direction note that $[\gamma'' \cdot \overline{\gamma}], [\gamma' \cdot \overline{\gamma}] \in Hg$ implies that

$$[\gamma'' \cdot \overline{\gamma}] \cdot [\overline{\gamma' \cdot \overline{\gamma}}] = [\gamma'' \cdot \overline{\gamma} \cdot \gamma \cdot \overline{\gamma'}] = [\gamma'' \cdot \overline{\gamma'}] \in H,$$

i.e. $[\gamma''] \sim [\gamma']$. We can thus conclude that the elements of $p^{-1}(x)$ are in bijection with the cosets of h in $\pi_1(B, b)$. In other words, $|p^{-1}(x)| = [\pi_1(B, b) : H]$, and as this is true of every point $x \in B$ we can conclude that these points are evenly covered. the map p is therefore a covering map.

All that is left to do is to confirm that $p_*(\pi_1(E, \tilde{b})) = H$. For this, note that $[\gamma] \in p_*(\pi_1(E, \tilde{b}))$ if and only if its lift $\tilde{\gamma}$ is in $\pi_1(E, \tilde{b})$. The endpoint of this lift will be the equivalence class represented by the point $[\gamma]$. However, for this lift to be a loop in E we also require that its endpoint be $\tilde{b}$, which recall was chosen to be represented by $[b]$, the constant path at b. We therefore can conclude that $[\gamma] \in p_*(\pi_1(E, \tilde{b}))$ if and only if $[\gamma] \sim [b]$, which can only occur if $\gamma \in H$. The result follows.

□

This result tells us that the correspondence which associates a covering $p : E \to B$ with the subgroup $p_*(\pi_1(E, \tilde{b})) \leq \pi_1(B, b)$ is a surjective correspondence. Every such subgroup has some covering which is associated to it. To confirm that this is a bijective correspondence, we just need injectivity. Things will not work out so well such that two coverings correspond to the same subgroup only if they are equal 'on the nose', but up to a reasonable notion of equivalence this is indeed true.

The notion of equivalence we are looking for is provided by **isomorhism of coverings**. An isomorphism between two coverings $p_1 : E_1 \to B$ and $p_2 : E_2 \to B$ is simply a homeomorphism $f : E_1 \to E_2$ such that $p_1 = p_2 \circ f$. Thus the two coverings have the same structure to them, even if they don't necessarily map the exact same points to the same outputs. With this in hand, we introduce the next proposition which will complete our classification [6]:

Proposition 3.4.2. *If B is path-connected and locally path-connected, then two path-connected coverings (E_1, p_1) and (E_2, p_2) are isomorphic by a homeomorphism $f : E_1, \to E_2$ that maps the base point $\tilde{b}_1 \in p_1^{-1}(b)$ to the base point $\tilde{b}_2 \in p_2^{-1}(b)$ if and only if $p_{1*}(\pi_1(E_1, \tilde{b}_1)) = p_{2*}(\pi_1(E_2, \tilde{b}_2))$.*

Proof. First assume such an isomorphism exists. Then the fact that $p_1 = p_2 \circ f$ implies that $p_{1*}(\pi_1(E_1, \tilde{b}_1)) = (p_2 \circ f)_*(\pi_1(E_1, \tilde{b}_1)) = p_{2*} \circ f_*(\pi_1(E_1, \tilde{b}_1))$. Since f is a homeomorphism from E_1 to E_2 mapping $\tilde{b}_1$ to $\tilde{b}_2$, it gives a bijection between the homotopy classes of loops based at $\tilde{b}_1$ and those based at $\tilde{b}_2$, and moreover induces an isomorphism between their fundamental groups. Thus we have $p_{2*} \circ f_*(\pi_1(E_1, \tilde{b}_1)) = p_{2*}(\pi_1(E_2, \tilde{b}_2))$, which tells us that $p_{1*}(\pi_1(E_1, \tilde{b}_1)) = p_{2*}(\pi_1(E_2, \tilde{b}_2))$.

For the converse, we will need a lemma known as the "lifting criterion" which we will not prove here [6]:

Lemma 3.4.1 (Lifting Criterion)**.** *Let $p : E \to B$ be a covering and $f : Y \to B$ a continuous map* *with Y path-connected and locally-path connected. Then there is a lift $\tilde{f} : Y \to E$ (i.e. such that $f = p \circ \tilde{f}$) if and only if $f_*(\pi_1(Y, y_0)) \subseteq p_*(\pi_1(E, \tilde{b}))$.*

In our case, we can take p to be p_2 and f to be p_1. Since we assumed E_1 was path-connected in the theorem and as a covering it inherits local path connectedness from B, this space satisfies the assumptions of the theorem. We also are assuming that $p_{1*}(\pi_1(E_1, \tilde{b}_1)) = p_{2*}(\pi_1(E_2, \tilde{b}_2))$, which in particular tells us that these two groups are subsets of each other. Our lemma thus gives us a lift of p_1 to $\tilde{p}_1 : E_1 \to E_2$, and we similarly have a lift $\tilde{p}_2 : E_2 \to E_1$. Now consider the compositions $\tilde{p}_2 \circ \tilde{p}_1$ and $\tilde{p}_1 \circ \tilde{p}_2$. For the first we have that $p_1 \circ \tilde{p}_2 \circ \tilde{p}_1 = p_2 \circ \tilde{p}_1 = p_1$, meaning that $\tilde{p}_2 \circ \tilde{p}_1$ is a lift of p_1 via p_1 itself. Moreover, we can note that $\tilde{p}_2 \circ \tilde{p}_1(\tilde{b}_1) = \tilde{p}_2(\tilde{b}_2) = \tilde{b}_1$. By the uniqueness of lifts, the fact that the base point is fixed allows us to conclude that in fact $\tilde{p}_2 \circ \tilde{p}_1$ is the identity on E_1. By similar reasoning we also have $\tilde{p}_1 \circ \tilde{p}_2$ is the identity on E_2, so $\tilde{p}_1$ and $\tilde{p}_2$ are inverses of one another. The map $\tilde{p}_1$ is then the desired isomorphism. □

In summary, with this result we have shown that for any path-connected, locally path-connected, semi-locally simply connected space B (in particular, say, the manifolds we will be studying in this work), the collection of isomorphism classes of coverings $p : E \to B$ which preserve the base points are in bijection with the subgroups of $\pi_1(B, b)$.

The connection between coverings and subgroups of the base space's fundamental group does not stop merely with this classification result. For example, recall in the proof of proposition 3.4.1 we found that for the particular covering we defined, its degree was equal to the index of the subgroup $H = p_*(\pi_1(E, \tilde{b}))$ in $\pi_1(B, b)$. However, because we now know the subgroup H actually determines the covering up to isomorphism, we actually have the following corollary:

Corollary 3.4.1. *If $p : E \to B$ is a covering with B path connected, locally path connected, and semi-locally simply connected, then the degree of p is equal to $[\pi_1(B, b) : p_*(\pi_1(E, \tilde{b}))]$.*

To finish this discussion, recall that for all of what we have done so far in this section, we have been restricting ourselves to coverings such that a particular $\tilde{b} \in p^{-1}(b)$ is mapped to b (specifically, $\tilde{b}$ corresponding to the constant path at b). However, if we relax this restriction and ignore which base point we use in E, then the coverings of B are instead classified by conjugacy classes of subgroups of $\pi_1(B, b)$.

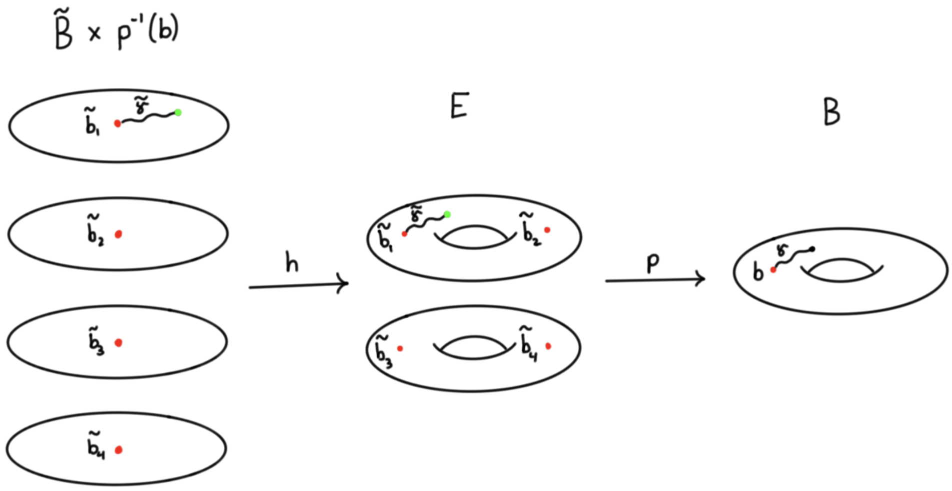
To see this, let $p : E \to B$ be a covering and $\tilde{b}_1$ and $\tilde{b}_2$ be two different possible base points in $p^{-1}(b)$. We wish to show that $p_*(\pi_1(E, \tilde{b}_1))$ is conjugate to $p_*(\pi_1(E, \tilde{b}_2))$ in $\pi_1(B, b)$. Since E is path connected, we can choose a path $\tilde{\gamma}$ going from $\tilde{b}_1$ to $\tilde{b}_2$. Since both endpoints are in $p^{-1}(b)$, we know that $\gamma = p(\tilde{\gamma})$ will be a loop in B based at b, so $[\gamma] \in \pi_1(B, b)$. Let $H_1 = p_*(\pi_1(E, \tilde{b}_1))$ and $H_2 = p_*(\pi_1(E, \tilde{b}_2))$. Then any element of $[\gamma]H_2[\overline{\gamma}]$ will lift to a loop based at $\tilde{b}_1$ in E, since we follow $ga\tilde{m}ma$ from $\tilde{b}_1$ to $\tilde{b}_2$, then some element of $\pi_1(E, \tilde{b}_2$ from this point back to itself, then finally we follow $\bar{\tilde{\gamma}}$ from $\tilde{b}_2$ to $\tilde{b}_1$. When we project this loop down to B via p, it will thus land inside of H_1. We therefore conclude that $[\gamma]H_2[\overline{\gamma}] \subseteq H_1$. By symmetric reasoning, we can conclude that $[\overline{\gamma}]H_1[\gamma] \subseteq H_2$. This two inclusions allow us to conclude that in fact H_1 and H_2 are conjugate to each other, by some element $[\gamma] \in \pi_1(B, b)$.

3.5 Permutation Representation

At this point, we have seen how the (conjugacy classes of) subgroups $\pi_1(B)$, for a manifold B, correspond exactly with the coverings of B. However, each such covering was necessarily connected. One may wonder then if there exists a similar correspondence for disconnected coverings, such as the one for S^1 we saw in Fig. 3.1.1. Though it may seem that there should be too many of these to ever hope for such a classification, there is somewhat miraculously just such a way to do so!

To see this, we will adapt the exposition from Hatcher [6]. A first step is that we want a disconnected version of the universal cover. For now, we will restrict our attention to the task of classifying the n-sheeted covers of a topological space B, and assume that B has a connected universal cover $\tilde{B}$. Note that if $p : E \to B$ is an n-sheeted cover, then E can have at most n connected components, since restricting p to any connected component should still give a covering map and so each connected component contributes at least 1 sheet to the cover of each point of B. We will thus consider the most general disconnected cover of B with at most n components, which will simply be $\tilde{B} \times \{1, \ldots, n\}$. If we have a covering map $p : E \to B$ and choose a base point $b \in B$, we can in fact identify this discrete set $\{1, \ldots, n\}$ with $p^{-1}(b)$.

It should seem quite plausible that any such n-sheeted covering E of B can itself be covered by $\tilde{B} \times p^{-1}(b)$. Indeed this is the case, and to see this we will define a natural covering map $h : \tilde{B} \times p^{-1}(b) \to E$. For this, recall that $\tilde{B}$ can be constructed as the space of all homotopy classes of paths which start at the chosen base point b. Thus each point in $\tilde{B} \times p^{-1}(b)$ looks like $([\gamma], \tilde{b}_i)$ where γ is a path based at b and $\tilde{b}_i$ is a lift of b to E. We then define $h([\gamma], \tilde{b}_i) = \tilde{\gamma}(1)$, where $\tilde{\gamma}$ is the unique lift of γ to E starting at $\tilde{b}_i$:



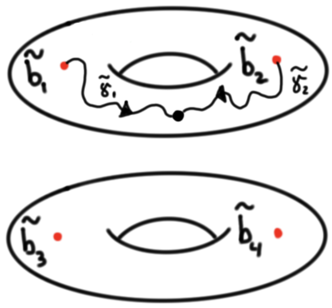
To confirm that h is indeed a covering map, we need to show it is a surjective local homeomorphism. It is surjective since every point $e \in E$ must be in the same connected component as some $\tilde{b}_i$, and so there will be a path α connecting $\tilde{b}_i$ to e. This path can be projected down to B as $p \circ \alpha$, and so we will have $h([p \circ \alpha], \tilde{b}_i) = e$.

To show that h is a local homeomorphism, let $([\gamma], \tilde{b}_i)$ be a point in $\tilde{B} \times p^{-1}(b)$ and let $q_i : \tilde{B} \times \{\tilde{b}_i\} \to B$ be the usual universal covering map which takes this point to $\gamma(1)$. We have already seen that there exists a neighborhood $U^i_\gamma = \{([\eta \cdot \gamma], \tilde{b}_i)\}$ of $([\gamma], \tilde{B}_i) \in \tilde{B}$ which maps homeomorphically to its image $q_i(U^i_\gamma)$ in B. Now take a neighborhood V around $\gamma(1)$ which is homeomorphic to each of its sheets in E, and define $V' = q_i(U^i_\gamma) \cap V$. Then for any choice of sheet $\tilde{V'}$ in $p^{-1}(V')$, we obtain a well-defined map $p^{-1} \circ q$ from $q^{-1}(V')$ to $\tilde{V'}$, and moreover this map is a homeomorphism. If we choose our sheet in E to be the one containing $h([\gamma], \tilde{b}_1)$, then we can observe that

$$p^{-1} \circ q([\eta \cdot \gamma], \tilde{b}_i) = p^{-1}(\eta \cdot \gamma(1)) = \tilde{\eta} \cdot \tilde{\gamma}(1),$$

where $\tilde{\eta} \cdot \tilde{\gamma}$ is the lift of $\eta \cdot \gamma$ to E starting at $\tilde{b}_i$. This is precisely the definition of the map h. Thus we have shown that any point in $\tilde{B} \times p^{-1}(b)$ has a neighborhood which maps homeomorphically via h onto its image in E, i.e. h is a local homeomorphism.

However, the map h is even better than this, as it is in fact a quotient map of $\tilde{B} \times p^{-1}(b)$ by a certain group action. To see this, suppose that we have $h([\gamma_1], \tilde{b}_1) = h([\gamma_2], \tilde{b}_2)$. Then we would have $\tilde{\gamma}_1(1) = \tilde{\gamma}_2(1)$, with these paths lifted to $\tilde{b}_1$ and $\tilde{b}_2$ respectively:



It follows that $\overline{\tilde{\gamma}_2}\cdot\tilde{\gamma}_1$ projects to a loop in B based at b, which will be in some homotopy class in $\pi_1(B)$, say $[\lambda]$. We then have $[\gamma_1] = [\gamma_2]\cdot[\lambda]$, and so these homotopy classes of paths differ by an action of $\pi_1(B)$. Likewise, we can define a right-action of $\pi_1(B)$ on $p^{-1}(b)$ by defining $\tilde{b}_i\cdot[\lambda] = \tilde{\lambda}(1)$, where $\tilde{\lambda}$ is the lift of λ to E starting at $\tilde{b}_i$. In particular for our case, we can observe that $\tilde{b}_1 = \tilde{b}_2\cdot[\lambda]$. We can then conclude that $h([\gamma_1],\tilde{b}_1) = h([\gamma_2],\tilde{b}_2)$ if and only if $([\gamma_1],\tilde{b}_1) = ([\gamma_2],\tilde{b}_2)\cdot[\lambda]$ for some $[\lambda]\in\pi_1(B)$.

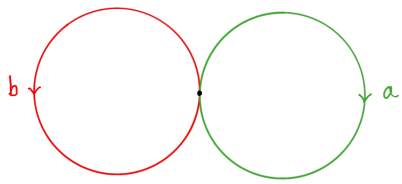
Note that the only action of $[\lambda]$ which depended on the covering map p was the action on $p^{-1}(b)$, since this required lifting λ to E via p. We call this action the *permutartion representation of* $\pi_1(B)$, since it permutes the elements of $p^{-1}(b)$. Let us refer to this action by ψ, so that $\psi:\pi_1(B)\to\mathrm{perm}(p^{-1}(b))$ is a homomorphism. Then we will call the space we obtain from quotienting $\tilde{B}\times p^{-1}(b)$ by this action E_ψ. By construction, we also have h quotienting to a well-defined map $h':E_\psi\to E$, and since this is a bijective local homeomorphism, this is simply a homeomorphism of spaces. Thus something miraculous has happened: we have reconstructed the space E purely from the action of $\pi_1(B)$ on the fiber of b, which if we'd like can be thought of as simply an abstract n-element set as opposed to specifically $p^{-1}(b)$ as a subset of E.

Even more than this, we also have a natural covering map $p':E_\psi\to B$ given by $p'\left([\gamma],\tilde{b}_i\right) = \gamma(1)$, and this covering map is isomorphic to p. That is, we have not only recovered the covering space E, but also the covering map p as well! We now have two claims to prove: that p' is a covering, and that it is isomorphic to p. For the first of these, note that surjectivity of p' follows from the connectedness of B, just as it did when we defined a similar map on the universal cover. To show that p' is a local homeomorphism, take an arbitrary point $([\gamma],\tilde{b}_i)\in E_\psi$, and pick a lift of it to $\tilde{B}\times p^{-1}(b)$ (which we will also write as $([\gamma],\tilde{b}_i)$). Let $h:\tilde{B}\times p^{-1}(b)\to E_\psi$ be the quotienting map, and $q:\tilde{B}\times p^{-1}(b)\to B$ be the usual covering map. Then there is a neighborhood U of $([\gamma],\tilde{b}_i)\in\tilde{B}\times p^{-1}(b)$ which maps homeomorphically via q to its image in B, and a neighborhood V which maps homeomorphically via h to its image in E_ψ. Then define $W = U\cap V$, so that we have restricted homeomorphisms $h:W\to h(W)$ and $q:W\to q(W)$. In particular, with these restrictions in place we can see that $q\circ h^{-1}:h(W)\to q(W)$ is a homeomorphism. It is at this point that we note that $p' \circ h = q$, so that after appropriate restrictions we can also say that $p' = q \circ h^{-1}$ on $h(W)$. Thus we have found a neighborhood of $([\gamma], \tilde{b}_i) \in E_\psi$ which maps homeomorphically via p' to its image in B, menaing p' is a local homeomorphism.

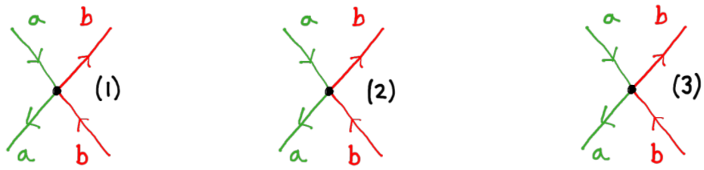
Having now confirmed that p' is indeed a covering, we will show that it is isomorphic to the covering map $p : E \to B$. This follows by the observation that $h : E_\psi \to E$ maps the equivalence class of $([\gamma], \tilde{b}_i)$ to $\tilde{\gamma}(1)$, which p in turn maps to $\gamma(1) = p'([\gamma], \tilde{b}_i)$. Thus $p \circ h = p'$, and since h is a homeomorphism this confirms that p and p' are isomorphic coverings.

In summary then, we have shown that any covering $p : E \to B$ induces a representation of $\pi_1(B)$ in the permutation group of $p^{-1}(b)$, called the permutation representation of p. We can then construct a covering space E_ψ purely from this representation ψ, and the covering we obtain will be isomorphic to p. Thus, the (isomorphism classes of) representations of $\pi_1(B)$ into permutation groups classify all coverings of B (again, up to isomorphism), both connected and disconnected. This is the power of the permutation representation.

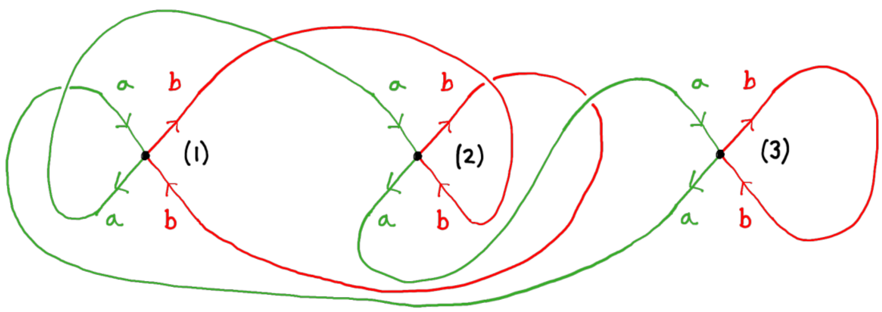
To finish off, let us look at an illustrative example of how the permutation representation can determine a covering, whether connected or not. Consider the space below which is the wedge product of two circles:



Note that above we have labeled the two circles a and b, and also gien them orientations. Thus, if we likewise label edges of a covering space with either a or b and provide orientations, then these will determine a covering map to the base space, since we simply map the a edges to a in the base space, b edges to b in the base space, and do so in such a way as to respect orientations. With this in mind, let's say we wish to construct a 3-sheeted cover of this space. We can first take an open neighborhood of the point of intersection and make 3 copies of it:

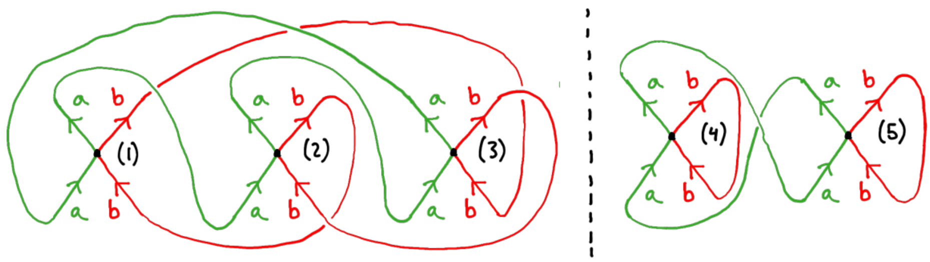


Consider the a segment of the point (1) which is oriented pointing out from the intersection (its 'out' edge). We have to extend this segment to an edge by joining it to another a segment in this figure, and moreover we have to do this in a way which respects orientations, so we must connect it to a segment which is pointing in toward the intersection (its 'in' edge). Suppose we join it to the intersection (2). We now must make the same choice for the 'out' a edge of (2). We cannot join it to itself, since its 'in' edge has been used already. We must therefore join it to either (1) or (3). Let us say we join it to (3). Then the 'out' a edge of (3) is forced to join the 'in' a edge of (1). Because we could only join the 'out' a edge of any point to the 'in' a edge of exactly one other, and vice versa, the end result is essentially gives a permutation of the points (1), (2), and (3) with (1) moving along a to (2), then (2) moving along a to (3), then finally (3) moving along a to (1). Thus we map the edge a to the permutation (123). If we map b to the permutation (12), then the covering space we obtain looks like:



Note that the fundamental group of the base space is the free group F_2, with generators given by the loop going around a and the loop going around b. Thus if we select where the generators go, this extends to a homomorphism from F_2 to S_3, and so this homomorphism does indeed determine the covering.

With this example its also not hard to see how we can determine disconnected coverings as well. For an n-sheeted cover simply partition $\{1, 2, \ldots, n\}$ into some number of subsets $A_1, \ldots, A_k$ and map a and b each to elements in the subgroup $S_{A_1} S_{A_2} \cdots S_{A_k}$. This will correspond to a cover with k connected components. For example, partitioning $\{1, 2, 3, 4, 5\}$ into $\{1, 2, 3\}$ and $\{4, 5\}$ and then mapping a to $(123)(45)$ and b to (132) yields the covering space:



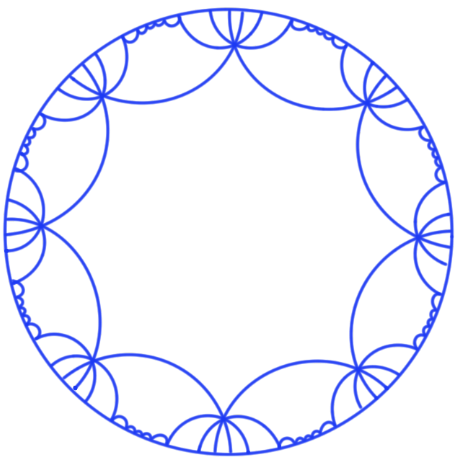
Chapter 4

Hyperbolic Surfaces

4.1 Definition and Constructions

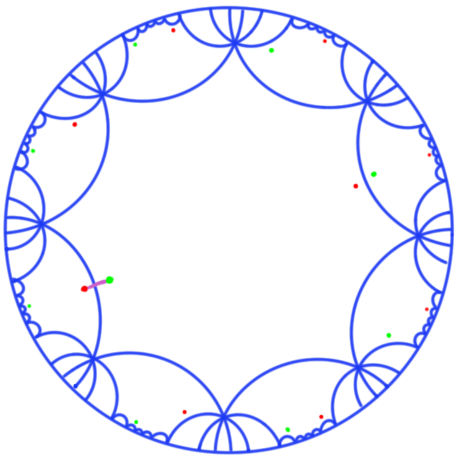
It is not only the disk that possesses a notion of hyperbolic metric. Other topological surfaces can be endowed with constant negative curvature metrics as well. A first example of this is the double torus. A well-known construction from topology is that the torus can be realized as the quotient of $\mathbb{R}^2$ by a translational action of $\mathbb{Z}^2$. Equivalently one can take a square fundamental domain of this action and identify its opposite sides, respecting orientations. Similarly, the double torus can be thought of topologically as an octagon with opposite pairs of sides identified. Unlike the torus however, this representation of the double torus is not a fundamental domain for some group of translations in the euclidean plane.

The hyperbolic plane, however, does possess groups of translations which have octagonal fundamental domains:



The central octagon above is in fact isometric to each of the eight octagons adjacent to it, and can reach any of them via an appropriate translation. Thus, if we call this group of translations which preserve the octagonal tiling Γ, then we can realize our genus 2 surface as the quotient space $\mathbb{H}^2/\Gamma$. A similar construction works for any genus g surface as well, the relevant property of $\mathbb{H}^2$ that allows this being the fact that vertex angles are smaller on average in the hyperbolic plane. Thus we can have convex polygons with arbitrarily many sides but with vertex angles that sum to 2π, thus allowing them to tile the plane together.

Notice that constructing a surface S in this way embellishes it with a hyperbolic metric inherited from $\mathbb{H}^2$. This can either be understood as meaning that every point of our surface has constant curvature -1 at every point, or equivalently as meaning that every point has a neighborhood which is isometric to an open subset of $\mathbb{H}^2$. To be precise, this hyperbolic metric is defined in the following fashion: consider the pre-images of two points $x, y \in S$ in $\mathbb{H}^2$:



Then to obtain the metric on S we define

$$d_S(x, y) = \inf_{\tilde{x}, \tilde{y}} \{d_{\text{hyp}}(\tilde{x}, \tilde{y})\}.$$

In words, it is the minimum distance between any two pre-images of x and y. As this minimum will be realized for some pair $\tilde{x}$, and $\tilde{y}$, then we can join these points in $\mathbb{H}^2$ by a unique geodesic which will then project onto a geodesic in S. Thus, any two points $x, y \in S$ can be joined by a unique geodesic whose length is the distance between x and y, meaning S is a geodesic space.

With this class of examples now in mind, let us endeavor to codify a definition of "hyperbolic surface." Let S be some closed surface, and let's see how we can most generically construct a hyperbolic metric for S. We can start by noticing that if $U \subseteq S$ is an open disk, it is easy to give it a hyperbolic metric. Simply choose some open disk $V \subseteq \mathbb{H}^2$ and a homeomorphism $\phi : U \to V$. Then U can be given a hyperbolic metric d_U by simply pulling back the hyperbolic metric on V via ϕ. That is, by setting $d_U(x, y) := d_{\text{hyp}}(\phi(x), \phi(y))$. If we have an open covering $\{U_i\}_{i \in I}$ of S by open disks, then we can choose homeomorphisms $\phi_i : U_i \to V_i \subseteq \mathrm{H}^2$ as above and give each U_i a hyperbolic metric. Each pair (U_i, ϕ_i) is called a chart and the collection of all these pairs, $\{(U_i, \phi_i)\}_{i\in I}$, is called an atlas of charts for S. This is what will allow us to endow S with a hyperbolic metric.

Additional care must be taken, however, in the cases where U_i and U_j have non-empty intersection. Of course, if we have ϕ_i and ϕ_j being equal on $U_i \cap U_j$ for all such intersections, then the ϕ_i's can be combined into a single homeomorphism from S to a subset of $\mathbb{H}^2$, which is not possible for general surfaces. Thus there must be some intersections $U_i \cap U_j$ on which ϕ_i and ϕ_j act differently. Keep in mind though that ϕ_i and ϕ_j must still pullback to the same metric on $U_i \cap U_j$. This is equivalent to $\phi_i \circ \phi_j^{-1}$ being an isometry between the images $\phi_i(U_i \cap U_j)$ and $\phi_j(U_i \cap U_j)$. To see this, note that for all $x, y \in U_i \cap U_j$ we must have

$$d_{\text{hyp}}(\phi_i(x), \phi_i(y)) = d_{U_i\cap U_j}(x, y) = d_{\text{hyp}}(\phi_j(x), \phi_j(y)),$$

relabeling $\phi_j(x)$ as z and $\phi_j(y)$ as w, this can be rewritten as

$$d_{hyp}(\phi_i \circ \phi_j^{-1}(z), \phi_i \circ \phi_j^{-1}(w)) = d_{\text{hyp}}(z, w),$$

which will be true for all $z, w \in \phi_j(U_i \cap U_j)$. So $\phi_i \circ \phi_j^{-1}$ must be an isometry between the images of the intersection.

The preceding considerations can actually be thought of as the definition of a hyperbolic surface [7]:

Definition 4.1.1 (Hyperbolic Surface)**.** *A hyperbolic surface S is a collection of pairs $\{(U_i, \rho_i)\}_{i\in I}$ such that $\{U_i\}_{i\in I}$ is an open cover of S, $\rho_i : U_i \to \mathbb{H}^2$ is a homeomorphism onto its image, and for any U_i and U_j with non-empty intersection $\rho_i \circ \rho_j^{-1} : \rho_j(U_i \cap U_j) \to \rho_i(U_i \cap U_j)$ is an isometry.*

The metric on such a surface S is obtained by pulling back the metric of $\mathbb{H}^2$ to each open set U_i, and the final condition that the transition maps $\rho_i \circ \rho_j^{-1}$ be isometries is there to ensure that these local pullback metrics form a consistent metric across S.

One might wonder, in light of the generality of this definition, whether there may be other examples of hyperbolic surfaces other than those provided above via quotients of $\mathbb{H}^2$ by Fuchsian groups. An amazing fact is that there are no other hyperbolic surfaces. That is, if S is a closed, orientable surface which is hyperbolic in the sense of the above definition, then there must exist a Fuchsian group $\Gamma \leq \text{Isom}^+(\mathbb{H}^2)$ such that S is isometric to $\mathbb{H}^2/\Gamma$. This a consequence of the "Uniformization Theorem," one statement of which is that [5]

Theorem 4.1.1 (Uniformization Theorem)**.** *If $D \subseteq \mathbb{H}^2$ is an open set, S a hyperbolic surface and $\rho : D \to S$ an isometry onto its image, then ρ can be extended to a surjective local isometry $\hat{\rho} : \mathbb{H}^2 \to S$.*

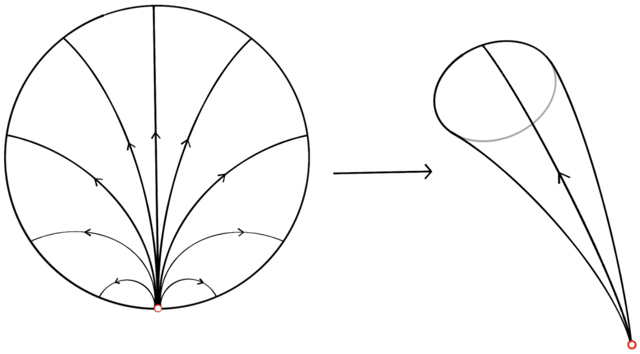
A local isometry between metric spaces X and Y is any map $\phi : X \to Y$ for which every point $x \in X$ has an open neighborhood U such that ϕ restricted to U is an isometry onto its image. In our case, letting x range over the pre-images of a point $x' \in S$, we can obtain a collection of disjoint open sets in $\mathbb{H}^2$ each of which map homeomorphically onto an open set of x'. Thus, $\hat{\rho}$ is in fact a covering map. Notice also that $\mathbb{H}^2$ is simply connected, meaning it is the universal cover of S. Thus the covering map takes the form of a quotient of $\mathbb{H}^2$ by some group of deck transformations. This group of deck transformations must be a group of isometries in order that the covering map is a local isometry, and it must also be isomorphic to the fundamental group $\pi_1(S) \cong \langle a_1, \ldots, a_{2g} \mid a_1 a_2 \cdots a_{2g} = a_{2g} \cdots a_2 a_1 \rangle$.

This group has no elements of finite order, so the isometries in this group must either be translations or horocycles. Horocyclic isometries can be ruled out by the fact that our surface is compact and hence has a complete metric, so we cannot have sequences of points which converge to a point at infinity. Thus our group of isometries consists of $2g$ axes of translations, the fundamental domain for which will be a $4g$-gon with opposite sides identified by one of these $2g$ translations. In other words, our surface emerges as a $4g$-gon in $\mathbb{H}^2$ with opposite sides identified by translation. So all hyperbolic surfaces can be obtained by the construction described at the start of this section.

4.2 Geodesics on Hyperbolic Surfaces

A crucial feature of hyperbolic geometries on surfaces, is that every free homotopy class of closed curves on the surface has a unique geodesic representative. This provides a link between a surface's hyperbolic geometry and its topology, and is a direct consquence of the fact that any hyperbolic surface can be locally isometrically covered by $\mathbb{H}^2$. To see why this is the case, consider an essential closed curve γ on S. Then $[\gamma] \in \pi_1(S)$ acts on $\mathbb{H}^2$ by some isometry, and so generates a cyclic subgroup $\langle [\gamma] \rangle$. Quotienting $\mathbb{H}^2$ by this cyclic action thus forms what is topologically a cylinder.

Let us consider what this cylinder could look like. We first of all know that this cyclic subgroup is not finite, since $\pi_1(S)$ contains no elements of finite order. This means $[\gamma]$ either acts by a parabolic action, or by a proper translation. If $[\gamma]$ corresponds to a parabolic action, then the cylinder will have a cusp at infinity where the fixed point of the action was:



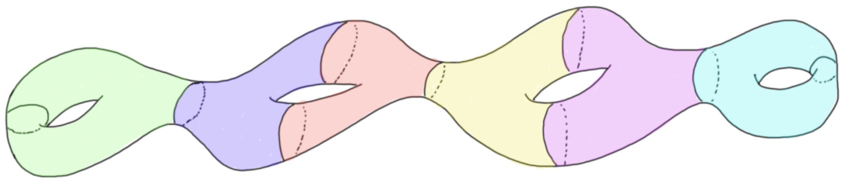
Let us lift γ to the cylinder and freely homotope it to this cusp. We can then observe that the length of these homotoped copies of γ approach 0 as γ approaches the cusp. However, essential closed loops on S cannot have arbitrarily small length. To see this, consider that every point $s \in S$ has what is known as an *injectivity radius*, i.e. the supremum of all possible radii disks containing s such that they are still simply connected. This injectivity radius is a continuous function from S to $\mathbb{R}_{\geq 0}$, and as S is compact some point must achieve the minimum of this function. This minimum must be non-zero, as every point s must have some neighborhood around it homeomorphic to a disk. Finally, observe that the length of any essential closed loop on S must be at least twice the injectivity radius of any point along it. Thus our γ cannot be homotoped to loops of arbitrarily small length.

Thus, the only possibility is that the action we are quotienting by is a hyperbolic translation. This action must have a unique geodesic axis which remains fixed under it, and so this axis will quotient to an essential closed geodesic on our cylinder. This unique geodesic is also in the same free homotopy class as γ. We can complete the picture by quotienting our cylinder by the rest of $\pi_1(S)$, and so conclude that the free homotopy class of γ has a unique geodesic representative on S. The coordinatization of all hyperbolic metrics on a surface S will depend on this fact.

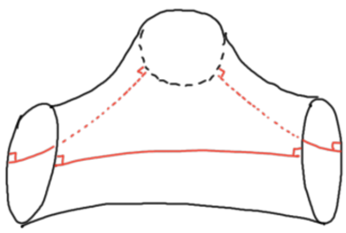
4.3 Teichmüller Space of a Surface

As was pointed out above, all hyperbolic metrics of a surface S come from a quotient of $\mathbb{H}^2$ by a Fuchsian group $\Gamma \cong \pi_1(S)$. Since for a genus g surface the fundamental domain for this translational action will be a $4g$-gon, this suggests it should be possible to find a finite number of parameters that uniquely determine a hyperbolic metric on a surface S, since a hyperbolic polygon is determined up to isometry by a finite number of parameters. Namely the side lengths and vertex angles, for example.

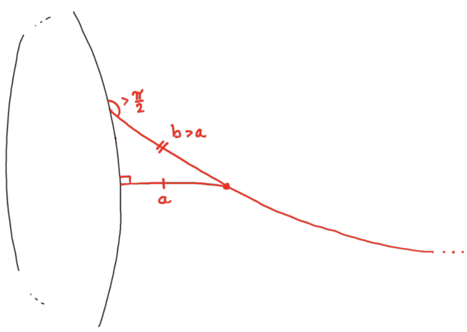
It will not be convenient to work with the $4g$-gon representation to find these parameters, however. Instead, we will decompose our surface S into a collection of *pants*, i.e. spheres with three open disks removed. This is known as the *pants decomposition* of S. Any such surface will have a collection of closed curves that, when we cut along them, divide the surface into sub-manifolds of this type, as illustrated below:



In general, for a surface of genus $g \geq 2$ a total of $3(g-1)$ such curves will be necessary to divide the surface into pants sub-manifolds. As shown in the previous section, each of these closed curves has a unique geodesic representative, so we may choose to cut along these geodesics in particular. Each such pants surface is then a hyperbolic surface with geodesic boundary. The three boundary circles can then in turn be connected by unique shortest geodesic paths:

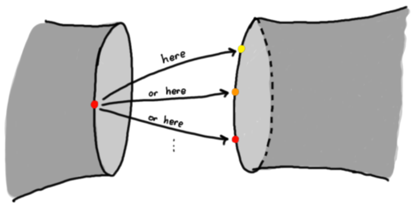


Note that this shortest geodesic will have to meet the geodesic boundaries at right angles. Otherwise near the boundary we could locally re-route the path away from the geodesic and form a shorter path, which is a contradiction:



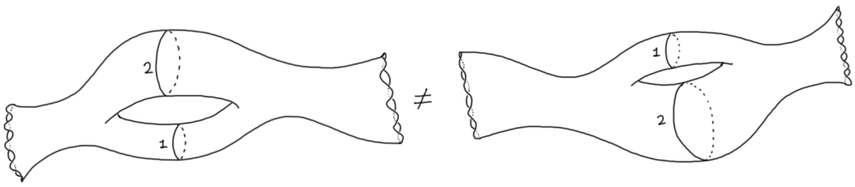
We have thus divided our surface into a collection of pants, each of which is itself isometric to two hyperbolic hexagons identified together along three non-adjacent edges. In particular, every vertex angle of these hexagons as a right angle. It is a theorem from hyperbolic geometry (see appendix, section 2) that such a hexagon is determined up to isometry by the lengths of three non-adjacent sides. We could take these three sides to be the sides that the top and bottom hexagons are identified along. Then since these identifications are done isometrically, we can conclude that the top and bottom hexagons are isometric. If instead we now take the three non-adjacent sides to lie along the boundary of the pants surface, then the top and bottom hexagon contribute equal length to each boundary component because they are isometric. This means that the length of the boundary components uniquely determines the hyperbolic metric of the pants surface. So the collective $3(g-1)$ geodesics we cut along each provide a parameter for determining the metric on S.

But there is another consideration. When we identify the boundaries of these pairs of pants to form the surface S, there is some freedom in how this is done. In particular, given a point p on the boundary of one hexagon, we can choose which point on the corresponding identified point to join it with:

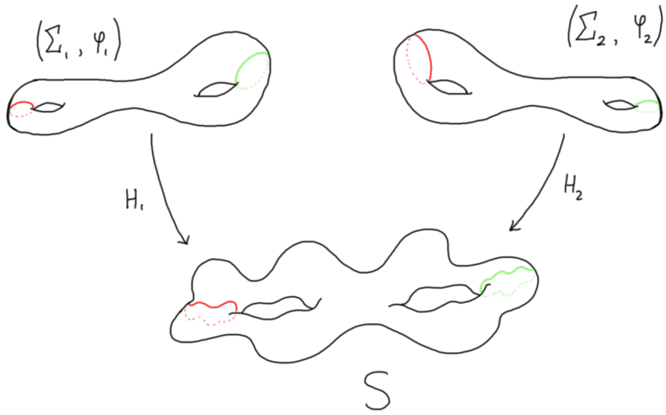


After this choice is made, the other points are determined by the fact that this mapping must be an isometry and that our surface must be orientable. Thus each curve we cut along gives an additional "twisting" parameter, which corresponds to a point in the space S^1. Once all length and twisting parameters of our pants decomposition are specified, the metric on S will be completely determined. Thus, the space of all hyperbolic metrics on S is parameterized by $\mathbb{R}_{>0}^{3(g-1)} \times (S^1)^{3(g-1)}$, which is a $6g-6$ dimensional space.

In general, the space of all constant curvature metrics on a surface S (up to isotopy, as will be elaborated on shortly) is referred to as the "Teichmüller space" of S, or $\mathcal{T}(S)$. However, there is a subtlety in this construction of the Tiechmüller space. Note that we have implicitly given the closed geodesics we cut along an order by taking the cartesian product of the spaces $\mathbb{R}_{>0}$. A surface S where the length of the first geodesic was 1 while the length of the second was 2, would be considered distinct from a hyperbolic surface where the lengths were instead 2 and 1, respectively:



in order to distinguish between these two metrics then, we will need a standardized way of labeling our free homotopy classes. To do this, we will fix some purely topological representative of our surface S. Then for each hyperbolic metrization of S, say (Σ, φ), we supply a homeomorphism $H : \Sigma \to S$. Then we can identify the homotopy classes of Σ with those of S, and do so in a consistent way across all hyperbolic metrics on S:



However, there are many choices for what this homeomorphism H could be, and we would not wish for two triples $(\Sigma_1, \varphi_1, H_1)$ and $(\Sigma_2, \varphi_2, H_2)$ to be different merely because they possess different homeomorphisms to S. So we must develop a notion of equivalence between these triples.

For this, note that if we have $H_1 : \Sigma_1 \to S$ and $H_2 : \Sigma_2 \to S$, then we can also create the function $H_2^{-1} \circ H_1 : \Sigma_1 \to \Sigma_2$. This is a homeomorphism between our hyperbolic surfaces, and importantly we know that it identifies their free homotopy classes in a consistent way. That is, if we label one homotopy class on S as a and another class as b, then H_1 allows us to say which homotopy class of Σ_1 we should label as a and which we should label as b, and likewise for Σ_2 via H_2. Then what we label as a on Σ_1 will be sent by $H_2^{-1} \circ H_1$ to what we labeled as a on Σ_2, and similarly for b. We can now check if the length of a on Σ_1 is the same as the length of a on Σ_2, for instance. If this is the case for all free homotopy classes, then we can say that Σ_1 and Σ_2 are the same hyperbolic surface, in a precise sense.

Thus, even though $H_2^{-1} \circ H_1$ is almost certainly not itself an isometry, in order for Σ_1 and Σ_2 to be equivalent we need it to map a closed geodesic on Σ_1 to a closed loop on Σ_2 which is homotopic to a closed geodesic of the same length. That is, we require that $H_2^{-1} \circ H_1$ be *isotopic* to an isometry. With this in mind, we can state the formal definition of the Teichmüller space of a surface S:

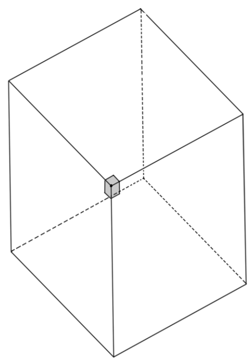
Definition 4.3.1 (Teichmüller Space)**.** *For a closed, orientable surface S, its* ***Teichmüller space*** *$\mathcal{T}(S)$ is the set of all equivalence classes of triples (Σ, φ, H), where Σ is a surface, φ is a hyperbolic metric on Σ, $H : \Sigma \to S$ is a homeomorphism, and two triples $(\Sigma_1, \varphi_1, H_1)$ and $(\Sigma_2, \varphi_2, H_2)$ are considered equivalent if $H_2^{-1} \circ H_1$ is isotopic to an isometry.*

This definition seems more restrictive than is necessary, since it differentiates between surfaces that one would normally think of as being isometric and thus the same. Such as two surfaces that differ only by a rotation or a reflection. But in fact the Teichmüller space turns out to be much more convenient to parameterize, as we saw above with the pants decomposition method [8]. And for that reason we will use Teichmüller space to talk about the "space of hyperbolic metrics on S."

4.4 Cone Surfaces

Cone surfaces are slight generalizations of the metrized surfaces we have seen thus far, in that they allow a finite collection of their points to have local angles which are not 2π. These are called the "**cone points**" of the surface. Intuitively, one could imagine a space where you stand in one place and start to turn around, but before you have done a full 2π rotation your line of sight is already back to where it began. Or it may take turning more than 2π radians for your line of sight to return to the start.

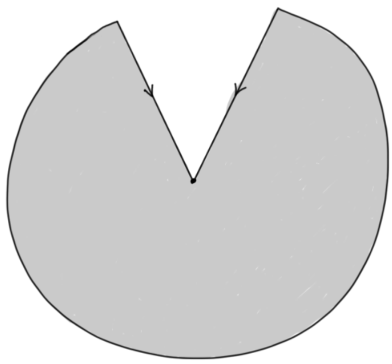
A good example of this phenomenon to start with is provided by polyhedra. If one considers a vertex of a cube, say, then the three right angles adjacent to it can be seen to add up to $\frac{3\pi}{2}$ radians, which is less than 2π:



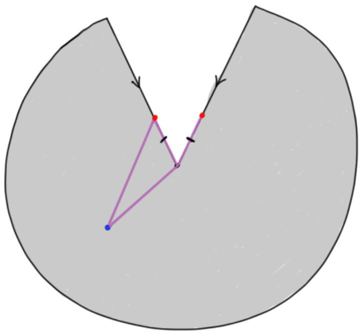
The presence of these cone points is, in fact, what allows for the existence of these polyhedra. As we will see in the next section on the Gauss-Bonnet theorem, the average curvature over a surface is a topological invariant, namely it is proportional to the Euler characteristic. This result seems to be in contradiction with the existence of such polyhedra though, since the sphere has an average curvature of 1, but a polyhedron like the cube seems to have an average curvature of 0 since each of its faces are flat. And if each point of our polyhedron did have a full angle of 2π around it, this would indeed be impossible. So it is helpful to think of cone points as possessing a sort of "combinatorial curvature" all their own, so that all of the positive curvature one would expect a topological sphere to have is contained at these points. We define this combinatorial curvature in terms of the angle deficit at the cone point. That is, if the cone angle at a cone point p is θ, then we define its combinatorial curvature as

$$\kappa(p) = 2\pi - \theta.$$

The presence of cone points is also important for how they affect the geodesics of our surface. In particular, for cone points with cone angles which are less than 2π there is no way to continue a geodesic which passes through them. To see this, consider the local region around such a cone point, which can be envisioned as an open disk with a wedge removed and with its bounding edges identified:

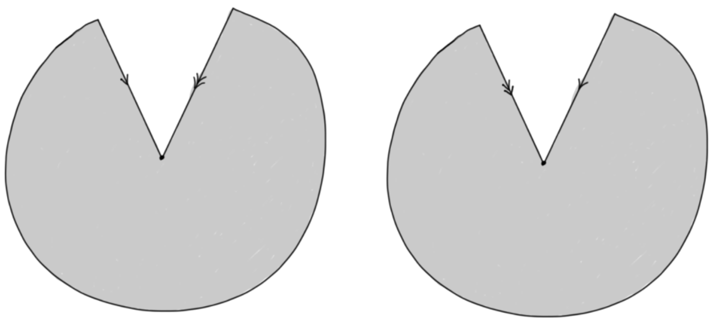


Where we cut out this wedge and make our identifications is arbitrary though. Thus for any two points near our cone point we may always choose our cut so that one of the points lies on it. There are then two apparent straight line paths from the first point to the two representatives of the second point, but the triangle inequality tells us that the path which does not pass through the cone point will always be shorter:

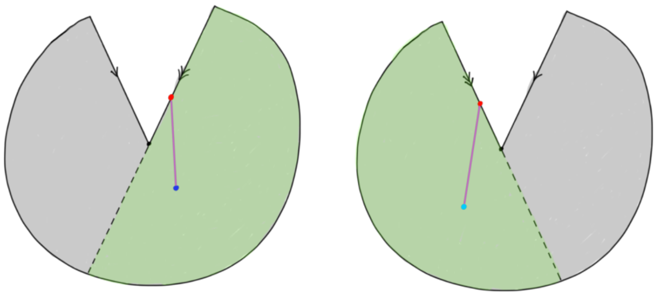


Thus the shortest path between any two points near the cone point never passes through the cone point itself. If we had a geodesic which passed through the cone point, there would have to be a neighborhood around the cone point for which this geodesic path was the shortest path connecting any two points on it in that neighborhood. As we have just shown this is impossible, we conclude that no geodesic path passes through this cone point, i.e. contains it in its interior.

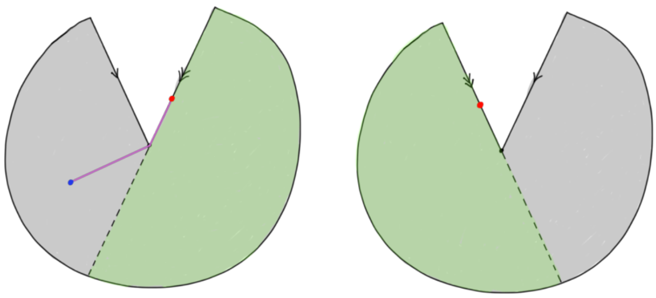
The opposite phenomenon occurs if the cone point has cone angle which is greater than 2π. Namely, any geodesic passing through a cone point can be extended in multiple ways. To see this, we will similarly model a small neighborhood around our cone point via open disks with wedges removed, but now we will use two disks appropriately identified to create an angle larger than 2π:



Now if we have 2 points on this surface which we want to find the geodesic between, there are two cases. We will again choose our cuts so that one of them passes through one of our cone points. Then there will be a region of angle 2π, highlighted in green below, where indeed there is a unique geodesic connecting our chosen points and which does not pass through the cone point:



And by the same reasoning as above, any path connecting these points which passes through the cone point will always be longer. However, for any other point outside this region, the shortest path which joins the two points must pass through the central cone point:



So if we connect a point on this surface via a geodesic to a cone point with excess angle, there will be an infinitude of geodesic paths which extend this initial geodesic. This observation will be important later when we discuss geodesic currents and the Liouville current in particular, and why the support of this current should contain any information about the metric on our surface.

4.5 The Gauss-Bonnet Theorem

A nice property of the hyperbolic plane is that the area of any triangle is easy to calculate, and thus so too is the area of any polygon as these can be decomposed into triangles. This extends to surfaces as well, where the Gauss-Bonnet theorem provides an easy way to compute their area. In the simplest case, let S be a closed Riemannian surface without boundary and with cone points $V_{\text{cone}} = \{p_1, \dots, p_n\}$. Then the Gauss-Bonnet theorem states

$$\sum_{j=1}^{n}(2\pi - \theta_j) + \int_S K dA = 2\pi\chi(S),$$

where θ_j is the cone angle at p_j and K is the Gaussian curvature at each regular point of S (the "Riemannian" adjective above is there to ensure this curvature is defined). Notice that if K is constant and -1, then the integral above will simply evaluate to $-\text{Area}(\varphi)$, where φ is the hyperbolic metric on S. Though the general proof of the Gauss-Bonnet theorem will not be presented here, in the special case where our surface has constant negative curvature it is not too hard to prove.

All that needs to be done is to choose a triangulation T of our surface which has each cone point of S as a vertex. Let T have V vertices, E edges, and F faces. The total area of S is the sum of the areas of its triangular faces, each of which takes the form $\pi - \alpha - \beta - \gamma$, where α, β, and γ are the vertex angles of the triangular face. If we denote the k-th vertex angle of the i-th vertex in T by α^i_k, then it follows that

$$\text{Area}(\varphi) = \pi F - \sum_{i=1}^{V} \sum_{k=1}^{n_i} \alpha_i^k,$$

where n_i is the number of vertex angles which meet the i-th vertex. The sum of these vertex angles must be 2π for every non-cone point vertex of our triangulation, and must sum to θ_j at the corresponding cone point vertex. The summation thus simplifies to

$$\begin{aligned} \pi F - \sum_{i=1}^{V} \sum_{k=1}^{n_i} \alpha_i^k &= \pi F - \sum_{v_i \notin V_{\text{cone}}} 2\pi - \sum_{v_i = p_j \in V_{\text{cone}}} \theta_j \\ &= \pi F - \sum_{v \in V} 2\pi + \sum_{p_j \in V_{\text{cone}}} (2\pi - \theta_j) \\ &= \pi F - 2\pi V + \sum_{p_j \in V_{\text{cone}}} (2\pi - \theta_j). \end{aligned}$$

At this point, note that since every face in this triangulation is adjacent to precisely 3 edges, it follows that $2E = 3F$, or equivalently $2E - 2F = F$. We may substitute this into our formula for the area to find that

$$\begin{aligned} \text{Area}(\varphi) &= 2\pi E - 2\pi F - 2\pi V + \sum_{p_j \in V_{\text{cone}}} (2\pi - \theta_j) \\ &= -2\pi\chi(S) + \sum_{j=1}^{n} (2\pi - \theta_j) \end{aligned}$$

which is equivalent to the statement of the Gauss-Bonnet theorem for cone surfaces of constant negative curvature. This fact will be essential for proving the main result of this thesis in chapter 8.

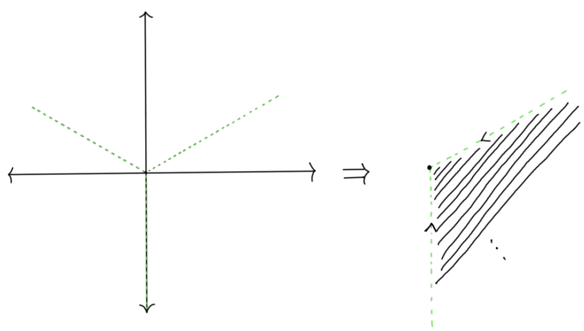
Chapter 5

Orbifolds

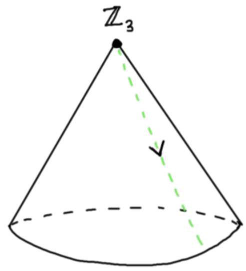
5.1 Preliminary Examples

As we saw in the chapters on covering spaces and on hyperbolic surfaces, a general way to construct a manifold is to take a "simpler" space like $\mathbb{R}^2$ or $\mathbb{H}^2$ and quotient it by some group of homeomorphisms. These group actions needed to be chosen carefully, though. In particular they needed to act freely and properly discontinuously on our space. However, a space like $\mathbb{R}^2$ can have many sorts of symmetric tilings, and the groups which act on the space to preserve these tilings do not always act freely. It will therefore be advantageous to develop a more general theory of quotient spaces which come from groups that do not act freely on a space. This is the theory of orbifolds, which we develop here.

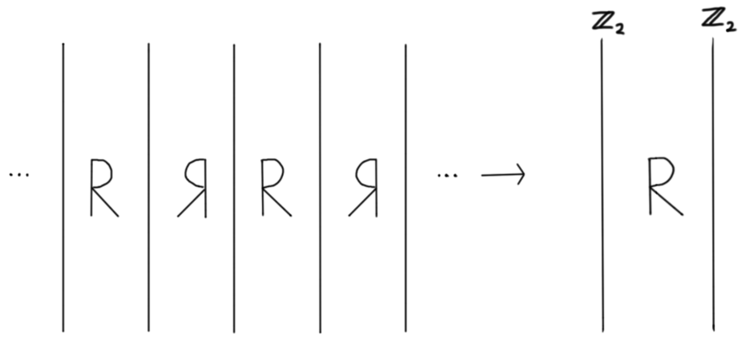
Let us begin with some examples. A very basic example is to take $\mathbb{R}^2$ and quotient it by $\mathbb{Z}_3$, acting via rotation about the origin:



We see that the "wedge" of the plane that results has its two bounding rays identified with one another. One can thus imagine stretching this wedge back around the origin until the two rays meet up again. We therefore have a homeomorphism from our wedge to the whole plane. However, it seems rash to write these two surfaces off as the same. They are indeed homeomorphic, but we would like a way to record the fact that the above "wedge" space was obtained from a certain group quotient, whereas the original plane $\mathbb{R}^2$ was not. We can accomplish this by marking the fixed point of our group action with its stabilizer (which in this case is the whole group):



We will develop a more precise notion of this later in the next section. Another way we can act on the plane is by reflections, for example:



This picture shows the plane being acted on by the infinite dihedral group D_∞, and the resulting quotient strip with marked edges that can be though of as reflectors, or mirrors. Note that these edges are only marked with the group $\mathbb{Z}_2$, since this is their stabilizer subgroup. This structure we have created, consisting of the underlying space of the strip as well as the markings on the fixed edges which carry information about the group action, is called an **Orbifold**. Orbifolds can thus be thought of as manifolds with an added structure, or if one prefers, as a generalization of manifolds. That is, if the motto of a manifold is that it is locally modeled on $\mathbb{R}^n$, then the motto of an orbifold is that it is locally modeled on a *group quotient* of $\mathbb{R}^n$.

Let us look at another, more interesting example before looking at the formal definition of an orbifold. We can tile the plane by isosceles squares as shown in the image below (ignore the diagonal lines for now):

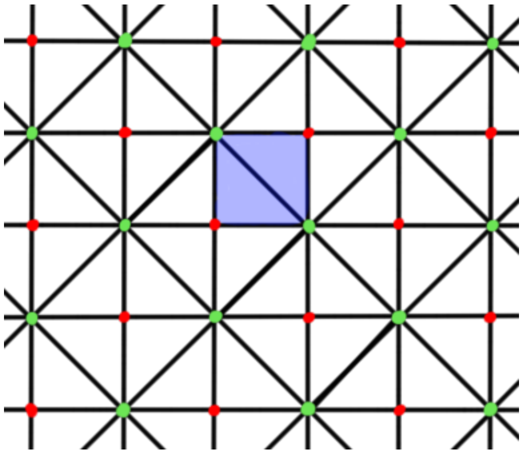
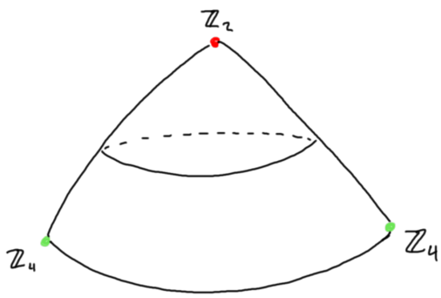


Fig. 5.1.1: Fundamental domains for a rotational action on $\mathbb{R}^2$

One group of symmetries for this pattern is, of course, the translational action by $\mathbb{Z}^2$. However, we could also act on this tiling by the group which is generated by degree 4 clockwise rotations at every other vertex of the grid (the green points in the picture above). Call the group generated by these rotations R. Note that the green vertices are not the only fixed points of the action of R on the plane. If one rotates around a green vertex and then rotates around another green vertex diagonally opposite it, then this action will be equivalent to rotating around a red point by π radians.

If we focus on a particular fundamental domain of this action, say the blue highlighted square above, then we may observe that by rotating around one green point we map the top edge to the left edge, or the right edge to the bottom edge, depending on which vertex we rotate around. Therefore when we form the quotient space $\mathbb{R}^2/R$, we must identify these edges together, and in so doing form what is topologically a sphere:



Thus we mark the two green points on our quotient with their stabilizer subgroup, which will be the group $\mathbb{Z}_4$ acting rotationally, and the red point by its stabilizer which is likewise $\mathbb{Z}_2$. At each of these marked orbifold points, we are locally quotienting by a rotation about that point, just like we did in the first example we saw of an orbifold. And just as in that example, we can observe that this local quotient by a rotation will create a cone point at the center. Thus the green vertices each have an angle of only $\frac{\pi}{2}$ around them, and the red vertex has an angle of π

around it. And in general, the cone angle at each orbifold point α is $\dfrac{2\pi}{|\Gamma(\alpha)|}$, where $\Gamma(\alpha)$ is the stabilizer of α. This occurs whenever our group acts via isometries on our metric space.

5.2 Definition

Here we present the formal definition of an orbifold, following that presented by Thurston [9]:

Definition 5.2.1 (Orbifolds)**.** *An orbifold $\mathcal{O}$ is a Hausdorff, second-countable topological space $X_{\mathcal{O}}$ together with a collection of 4-tuples $\{U_i, \tilde{U}_i, \Gamma_i, \phi_i\}_{i\in I}$. The U_i's form an open covering of $X_{\mathcal{O}}$ which is closed under finite intersections, each $\tilde{U}_i$ is a connected open subset of $\mathbb{R}^n$, Γ_i is a finite group acting faithfully on $\tilde{U}_i$, and ϕ_i is a homeomorphism $\phi_i : U_i \to \tilde{U}_i/\Gamma_i$.*

Additionally, whenever $U_i \subset U_j$ there should exist an injective homomorphism $f_{ij} : \Gamma_i \to \Gamma_j$ and an inclusion map $e_{ij} : \tilde{U}_i \to \tilde{U}_j$ such that for all $\gamma \in \Gamma_i$ we have $e_{ij}(\gamma \cdot x) = f_{ij}(\gamma) \cdot e_{ij}(x)$ (i.e. e_{ij} is equivariant with respect to f_{ij}), and for which the following diagram commutes:

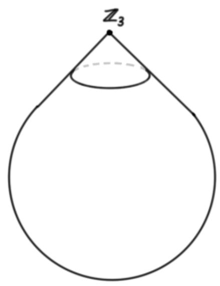
$$
\begin{array}{ccc}
\tilde{U}_i & \xrightarrow{\;e_{ij}\;} & \tilde{U}_j \\
\downarrow{\scriptstyle q_i} & & \downarrow{\scriptstyle q'_j} \\
\tilde{U}_i/\Gamma_i & \xrightarrow{\;e_{ij}/\Gamma_i\;} & \tilde{U}_j/f_{ij}(\Gamma_i) \\
\uparrow & & \downarrow{\scriptstyle q''_j} \\
{\scriptstyle \phi_i} & & \tilde{U}_j/\Gamma_j \\
\big| & & {\scriptstyle \phi_j}\uparrow \\
U_i & \hookrightarrow & U_j
\end{array}
$$

Where q_i, q'_j, and q''_j are quotienting maps.

Upon careful reading of this definition, one first observes that it is very arduous. So let us unpack it and see what justifies each of its components. To begin, the space $X_{\mathcal{O}}$ is referred to as the "underlying space" of $\mathcal{O}$. In the last example of the previous section, for instance, this space would be the sphere. We require $X_{\mathcal{O}}$ to be Hausdorff and second-countable purely to rule out pathological examples of spaces which we are uninterested in such as the "long line," much in the same way that we require these properties of manifolds to ensure they are not pathological either. The remainder of the conditions in the first paragraph simply spell out in detail what we mean when we say an orbifold is "locally modelled on a group quotient of $\mathbb{R}^n$." The atlas of charts $\{U_i\}_{i\in I}$ provides a sense of "locality" on $X_{\mathcal{O}}$, while $\tilde{U}_i$, Γ_i, and ϕ_i are used to endow U_i with the structure of a quotient of $\mathbb{R}^n$.

The primary difficulty of the definition lies in the commutative diagram. The top rectangle of the diagram is simply capturing the equivariance of e_{ij}. The bottom rectangle meanwhile is ensuring that the action of Γ_j on $\tilde{U}_j$ is simply a refinement of the action of Γ_i on $\tilde{U}_i$. Meaning that quotienting $\tilde{U}_j$ by Γ_j can be broken down into first a quotient of $\tilde{U}_j$ by $\Gamma_i \cong f_{ij}(\Gamma_i)$, and then a quotient by the equivalence relation that identifies two orbits in $\tilde{U}_j/f_{ij}(\Gamma_i)$ if they are orbits of the same point.

Note that this definition is actually slightly more general than what our examples from before would suggest. We only need that our orbifold is locally equivalent to a group quotient of $\mathbb{R}^n$, but this does not mean that the whole orbifold must come from a single group quotient of some manifold. For instance, consider an orbifold which consists of a sphere with one point marked by a cyclic group acting rotationally:



It turns out that this 'teardrop' orbifold does not come from a group quotient of any manifold, as we shall see later after developing a better understanding of orbifold coverings and orbifold fundamental groups. Nonetheless, one can observe that this orbifold does satisfy the formal definition given above.

5.3 Equivalence of Orbifolds

Now that we have introduced a new kind of mathematical object, the orbifold, it is first necessary to discuss precisely when two orbifolds are equivalent. As orbifolds contain both topological information as well as information about group actions, one would want such a notion of equivalence to take the form of a map which preserves both. That is, it should be a homeomorphism of the underlying spaces which maps each orbifold point to an orbifold point which is a quotient by an isomorphic group action. Formalizing this into a definition, we have

Definition 5.3.1 (Orbifold Equivalence)**.** *Two orbifolds $\mathcal{O}$ and $\mathcal{O}'$ are equivalent if there exists a homeomorphism $f : X_{\mathcal{O}} \to X_{\mathcal{O}'}$ such that every point $x \in X_{\mathcal{O}}$, x and $f(x)$ have neighborhoods $U \cong \tilde{U}/\Gamma$ and $V \cong \tilde{V}/\Gamma$ respectively such that the following diagram commutes for all $\gamma \in \Gamma$:*

$$\begin{array}{ccc} \tilde{U} & \xrightarrow{\tilde{f}} & \tilde{V} \\ \downarrow{\scriptstyle \gamma} & & \downarrow{\scriptstyle \phi(\gamma)} \\ \tilde{U} & \xrightarrow{\tilde{f}} & \tilde{V} \end{array}$$

where $\tilde{f}$ is a lift of $f : \tilde{U}/\Gamma \to \tilde{V}/\Gamma$ and $\phi : \Gamma \to \Gamma$ is an isomorphism.

This simply sets in stone what we already intuitively understood from the examples and definition of the previous section: that an orbifold is completely determined by its underlying topology and the group information it is endowed with. Because the underlying space of most orbifolds, or at least those which we will study here, is simply a closed orientable surface, and because we are typically quotienting by groups of isometries which only come in so many "flavors" (reflections, rotations, translations, etc.), it becomes possible to collate all of the topological and group information of such an orbifold into a single convenient notation: **the orbifold signature**. Following Thurston's conventions [9], the orbifold signature will, in general, look like $(g; r_1, r_2, \ldots, r_n; d_1, d_2, \ldots, d_m)$, where g is the genus of our underlying space, the r_i's are the orders of the cyclic groups acting on orbifold points by rotation, and the d_j's are the orders of the dihedral groups acting on orbifold points by rotations and reflections.

The above definition is sufficiently general to work for all orbifolds, but for our purposes in this work it will be convenient to use a slightly more restrained definition. In particular, we will assume our orbifolds emerge as a properly discontinuous group quotient of some simply-connected manifold. This manifold will thus be the universal cover of our orbifold $\mathcal{O}$. If we refer to this universal cover as $\tilde{\mathcal{O}}$, then we obtain the following alternative definition of equivalence

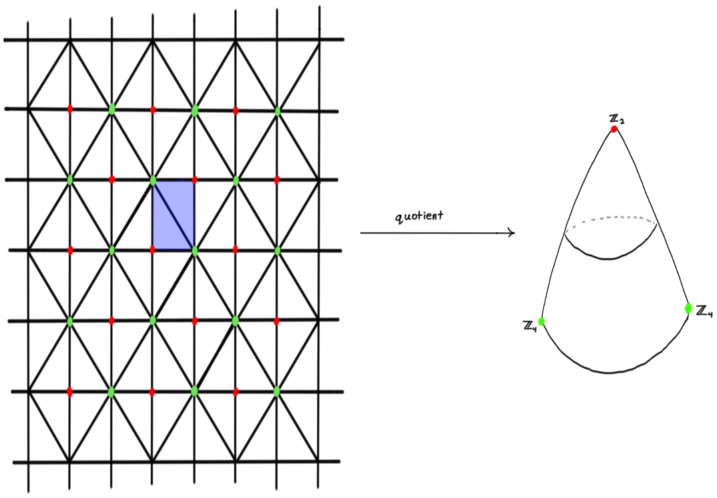
Definition 5.3.2. *Two orbifolds $\mathcal{O}_1 \cong \tilde{\mathcal{O}}_1/\Gamma_1$ and $\mathcal{O}_2 \cong \tilde{\mathcal{O}}_2/\Gamma_2$ are equivalent if there exists a map $f : \mathcal{O}_1 \to \mathcal{O}_2$, a homeomorphism $\tilde{f} : \tilde{\mathcal{O}}_1 \to \tilde{\mathcal{O}}_2$ between there universal covers, and an isomorphism $\phi : \Gamma_1 \to \Gamma_2$ such that the following diagram commutes:*

$$\begin{array}{ccc} \tilde{\mathcal{O}}_1 & \xrightarrow{\tilde{f}} & \tilde{\mathcal{O}}_2 \\ \downarrow{\scriptstyle \gamma} & & \downarrow{\scriptstyle \phi(\gamma)} \\ \tilde{\mathcal{O}}_1 & \xrightarrow{\tilde{f}} & \tilde{\mathcal{O}}_2 \\ \downarrow{\scriptstyle q_1} & & \downarrow{\scriptstyle q_2} \\ \mathcal{O}_1 & \xrightarrow{f} & \mathcal{O}_2 \end{array}$$

where q_1 and q_2 are the quotienting maps.

The top rectangle of this definition simply ensures that $\tilde{f}$ is equivariant with respect to the group action we are quotienting by, and the bottom rectangle ensures that $\tilde{f}$ is the lift of some map between the base orbifolds. So in fact, this alternative definition is simply a convenient way to form an equivalence map as described in the first definition, via a homeomorphism of the universal covers.

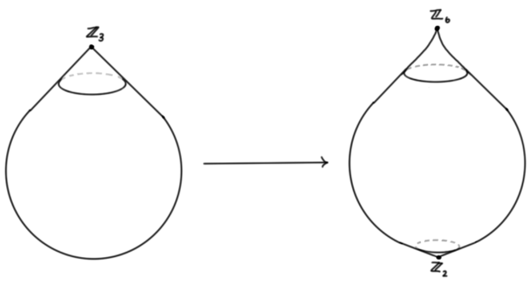
To see an example of this, consider the square grid we saw in 5.1. We can apply a homeomorphism to this grid which stretches out the rectangular regions in one direction. This doesn't change the nature of our group action however, and so when we quotient by this "elongated" action we should obtain an equivalent orbifold:



5.4 Orbifold Coverings

Orbifolds are not merely generalizations of manifolds in terms of their local structure. There also exist generalizations to orbifolds of certain concepts meant to study manifolds.

A first example of this is the notion of a covering. The examples shown in the first section were directly inspired from ordinary coverings of manifolds. So it is very natural to extend the notion of a "covering" to orbifolds, such that the examples listed above are instances of such orbifold coverings. Another example is provided by the "teardrop" orbifold discussed in the last section:



Here we have quotiented by the action on the covering orbifold which rotates it by π radians around the axis through the single orbifold point. We already mentioned that this teardrop orbifold does not come from a group quotient of any manifold. In fact, it does not come from a quotient of any other orbifold either. In this sense it is akin to the universal covers we saw in the chapter on covering spaces.

In normal coverings we want each point in the pre-image of a chosen point of the base space to have a neighborhood which maps homeomorphically via the covering map to a neighborhood of our chosen point. Likewise for orbifold coverings, we want each point in the pre-image of a chosen point of the base orbifold to have a neighborhood which maps by some local group quotient to a neighborhood of our chosen point. Thus, following Thurston [9], we can define an orbifold covering in the following way:

Definition 5.4.1 (Orbifold Covering)**.** *An orbifold covering of an orbifold $\mathcal{O}$ by another orbifold $\tilde{\mathcal{O}}$ is a map $p : X_{\tilde{\mathcal{O}}} \to X_{\mathcal{O}}$ such that*

(i) For all $\tilde{x} \in X_{\tilde{\mathcal{O}}}$, there exists a neighborhood $\tilde{U}/\Gamma$ containing $\tilde{x}$ for which p restricted to $\tilde{U}/\Gamma$ is isomorphic to a quotienting map $p : \tilde{U}/\Gamma \to \tilde{U}/\Gamma'$, with $\Gamma \leq \Gamma'$.

(ii) For every $x \in X_{\mathcal{O}}$ there exists an open neighborhood $\tilde{V}/\Gamma$ containing x such that each component U_i of its preimage $p^{-1}(\tilde{V}/\Gamma)$ is isomorphic to $\tilde{V}/\Gamma_i$ for some $\Gamma_i \leq \Gamma$.

Note that an orbifold covering is not necessarily a covering map for the underlying spaces.

5.5 Orbifold Fundamental Groups

Having a notion of orbifold coverings and of universal covers also allows us to define a natural notion of an orbifold fundamental group. Consider the last example of the first section, which was a sphere with three orbifold points. The fundamental group of this underlying space is trivial, but simply defining the fundamental group of the orbifold to be trivial as a result would needlessly ignore the symmetry information we have so carefully stored in its structure. Thus, to generalize the notion of the fundamental group to orbifolds we will recall what we know of covering spaces, namely that the fundamental group of a manifold is isomorphic to the group of deck transformations of its universal cover.

Now, it is a theorem of Thurston's [9] that every orbifold has a universal cover, i.e. a covering orbifold which also covers every other covering orbifold of the given orbifold. This means every orbifold has a fundamental group, since we can define its group of deck transformations analogously to how we defined it in the manifold case. In particular, a deck transformation of a covering orbifold $\tilde{\mathcal{O}}$ for the orbifold $\mathcal{O}$ is an orbifold equivalence $f : \tilde{\mathcal{O}} \to \tilde{\mathcal{O}}$ for which $p \circ f = p$, with $p : \tilde{\mathcal{O}} \to \mathcal{O}$ the orbifold covering map.

Just as with fundamental groups of manifolds, there exist tools for computing orbifold fundamental groups. Chief among these is a generalization of Seifert-Van Kampen to Orbifold [10]:

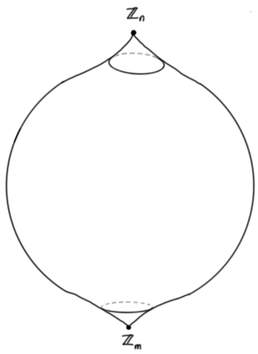
Theorem 5.5.1 (Van Kampen for Orbifolds)**.** *For a connected orbifold $\mathcal{O}$ and open sets $U, V \subseteq X_{\mathcal{O}}$ for which $X_{\mathcal{O}} = U \cup V$ and $W = U \cap V$ is connected, we have*

$$\pi_1(\mathcal{O}) \cong \pi_1(U) *_{\pi_1(W)} \pi_1(V).$$

The expression in this definition for $\pi_1(\mathcal{O})$ is the "amalgamated free product" of $\pi_1(U)$ and $\pi_1(V)$. That is, it is the quotient of the free product $\pi_1(U) * \pi_1(V)$ by the smallest normal subgroup containing $\{i_U^*(\gamma)i_V^*(\gamma)^{-1} \mid \gamma \in \pi_1(W)\}$, where i_U^* and i_V^* are the induced homomorphisms on the fundamental groups induced by the continuous inclusion maps $i_U : W \to U$ and $i_V : W \to V$. This amalgamation is done to avoid double counting of loops in our fundamental group, since any loop contained in W would of course be equivalent to itself, but would be counted as formally different when considered as an element of $\pi_1(U)$ and $\pi_1(V)$ when we take the free product. We must therefore specify that $i_U^*(\gamma) = i_V^*(\gamma)$ for any such loop γ, and this is accomplished by the specified group quotient.

An example of an orbifold fundamental group is that for the quotient of the plane by a rotational action, as described in the first section of this chapter. In this case, the universal cover for this orbifold is of course the plane, and deck transformations for this covering take the form of the rotational actions. Thus, just by the definition of the fundamental group it is not hard to see that the fundamental group for such an orbifold is simply $\mathbb{Z}_n$, where n is the degree of the rotation.

This observation then allows us to compute the fundamental groups of more complicatd orbifolds thanks to Van Kampen's theorem. Take for instance an orbifold which is a sphere with two orbifold points, each being locally quotients by some rotation group $\mathbb{Z}_n$ and $\mathbb{Z}_m$ respectively:



We can split it into the open subsets U and V as shown in the picture above, in which case $U \cap V$ will be a cylinder with no orbifold points and so will have $\mathbb{Z}$ as its fundamental group. Moreover, each of U and V can be considered as quotients of the plane by $\mathbb{Z}_n$ and $\mathbb{Z}_m$ respectively, so we thus have

$$\pi_1(U) \cong \mathbb{Z}_n \cong \langle a \mid a^n \rangle,$$

$$\pi_1(V) \cong \mathbb{Z}_m \cong \langle b \mid b^m \rangle,$$

$$\pi_1(U \cap V) \cong \mathbb{Z} \cong \langle c \rangle.$$

The induced homomorphisms i_U^* and i_V^* are specified by $i_U^*(c) = a$ and $i_V^*(c) = b$, meaning we will be quotienting the free product of $\pi_1(U) * \pi_1(V)$ by the relation ab^{-1}, i.e. we are specifying that $a = b$. The group presentation for our fundamental group is then

$$\pi_1(\mathcal{O}) \cong \langle a, b \mid a^n, b^m, ab^{-1} \rangle \cong \langle a \mid a^n, a^m \rangle \cong \langle a \mid a^{\gcd(m,n)} \rangle.$$

In particular, this means that if n and m are coprime, then the resulting orbifold will have trivial fundamental group and hence will be its own universal cover. The fundamental group for orbifolds which take the form of spheres possessing orbifold points that are locally quotients by $\mathbb{Z}_n$ can then be calculated inductively by this same process, for any number of orbifold points.

5.6 Orbifold Euler Characteristic

Later on, it will be useful for us to have a notion of an orbifold Euler characteristic at our disposal. This is because, as we shall see later, hyperbolic orbifolds abide by the Gauss-Bonnet formula just as regular hyperbolic surfaces do, for an appropriate notion of Euler characteristic. To define this Euler characteristic, we take inspiration from an important property of Euler characteristics of regular manifolds.

Suppose $p : \tilde{S} \to S$ is a covering map of degree d for a compact surface S. Then it follows that

$$\chi(\tilde{S}) = d \cdot \chi(S).$$

To see why this is true, recall that p being a covering map means that each $x \in S$ has an open neighborhood $U_x \subseteq S$ for which $p^{-1}(U_x)$ consists of d disjoint sheets each being homeomorphic to U_x. The set of all such open neighborhoods $\{U_x\}_{x \in S}$ forms an open covering of S, and as S is compact we can select a finite subcover $\{U_{x_i}\}_{i \leq n}$. Now construct a triangulation T of S which is sufficiently fine so that each triangular face, each edge, and each vertex is contained in one of these open sets U_{x_i}. The pre-image of T is a triangulation of $\tilde{S}$, and in order to calculate $\chi(\tilde{S})$ we will want to know how many faces, edges, and vertices are contained in $p^{-1}(T)$.

Let F be the number of faces of T. Then as each face is contained in some open set U_{x_i}, and $p^{-1}(U_{x_i})$ consists of d homeomorphic disjoint sheets, any given face of T will have d pre-images in $\tilde{S}$. So the number of faces of $p^{-1}(T)$ will be $d \cdot F$. A similar line of reasoning for edges and vertices tells us that there are $d \cdot E$ edges and $d \cdot V$ vertices in $p^{-1}(T)$, where E and V are the number of edges and the number of vertices in T, respectively. We can then compute that

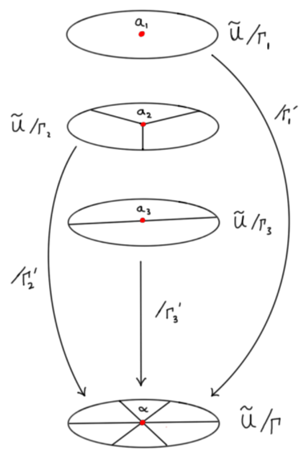
$$\chi(\tilde{S}) = d \cdot F - d \cdot E + d \cdot V = d \cdot (F - E + V) = d \cdot \chi(S).$$

We can now motivate a definition for orbifold Euler characteristic by asserting that if an orbifold covering $q : \tilde{\mathcal{O}} \to \mathcal{O}$ has degree d, then

$$\chi(\tilde{\mathcal{O}}) = d\chi(\mathcal{O})$$

An issue arises though since our reasoning above required that each vertex and edge of our triangulation had precisely d pre-images. But this is not necessarily true of an orbifold covering, which recall is not necessarily a covering map of the underlying spaces. To remedy this, we will first make a small simplifying assumption that our orbifold arises from a group quotient of a manifold by some action on the manifold by a finite group G. Then the degree of this covering map will be $|G|$ at any point of the manifold on which G acts freely. At any fixed points α of the action though, the degree of the map will be $\frac{|G|}{|\Gamma(\alpha)|}$ by the orbit-stabilizer theorem, where $\Gamma(\alpha)$ is the stabilizer subgroup of α. Therefore, in order to ensure that the Euler characteristic of the manifold is $|G|$ times the Euler characteristic of the orbifold it covers, we will need this fixed point α to have a "weight" of $\frac{1}{|\Gamma(\alpha)|}$ in our calculation of the Euler characteristic above.

We will let this observation inspire us now to define the Euler characteristic more generally. That is, in general even if our orbifold covering does not arise from a finite group quotient of some manifold, the definition of orbifold coverings tells us that it should locally look something like:



We will give the vertex/edge α the weight $\frac{1}{|\Gamma(\alpha)|}$ as described above, and similarly for each pre-image of α. We therefore need the sum of the weights of the pre-images of α to be precisely

d times the weight of α itself, where d is the degree of the covering map. In symbols,

$$\sum_{a_i \in p^{-1}(\alpha)} \frac{1}{|\Gamma_i(a_i)|} = \frac{d}{|\Gamma(\alpha)|}$$
$$\Updownarrow$$
$$\sum_{a_i \in p^{-1}(\alpha)} \frac{|\Gamma(\alpha)|}{|\Gamma_i(a_i)|} = d.$$

From here, we would like to say that $\frac{|\Gamma(\alpha)|}{|\Gamma_i(a_i)|} = |\Gamma_i'(a_i)|$. Then, because the total number of sheets of our covering over α, i.e. its degree, is given by the sum of the sizes of these groups $\Gamma_i'(a_i)$, we obtain what we want. To see why this is the case, recall that by quotienting $\tilde{U}$ by $\Gamma_i(a_i)$, we take a point x an equivocate it with every other point in its orbit under $\Gamma_i(a_i)$. If we choose x so that $\Gamma_i(a_i)$ acts freely on it, then the size of this orbit will be precisely $|\Gamma_i(a_i)|$. We then quotient $\tilde{U}/\Gamma_i(a_i)$ by the group $\Gamma_i'(a_i)$. By the same reasoning, if we choose a point of $\tilde{U}/\Gamma_i(a_i)$ on which $\Gamma_i'(a_i)$ acts freely, then we are equivocating it with $|\Gamma_i'(a_i)|$ points. In total then, we have taken the point $x \in \tilde{U}$ and equivocated it with $|\Gamma_i(a_i)||\Gamma_i'(a_i)|$ points. But by the definition of orbifold coverings, the resulting space must be isomorphic to $\tilde{U}/\Gamma(\alpha)$. Mapping to this quotient space equivocates x with $|\Gamma(\alpha)|$ points. Thus we must have $|\Gamma_i(a_i)||\Gamma_i'(a_i)| = |\Gamma(\alpha)|$, from which the result follows.

Now we can use these weightings for our vertices and edges to calculate the Euler characteristic of our orbifold. We will choose our triangulation T so that each edge e and each vertex v of T has its own stabilizer subgroup $\Gamma(e)$ and $\Gamma(v)$, respectively. Moreover, we will let T_2 be the set of 2-cells or faces of our triangulation, T_1 be the set of all 1-cells or edges, and T_0 be the set of all 0-cells or vertices. Then following from the reasoning above, we define the Euler characteristic of our orbifold to be

$$\chi(\mathcal{O}) := \sum_{f \in T_2} 1 - \sum_{e \in T_1} \frac{1}{\Gamma(e)} + \sum_{v \in T_0} \frac{1}{\Gamma(v)}.$$

As one final simplification to this expression, we can add and subtract 1 for every face, edge, and vertex of our triangulation, giving:

$$\begin{aligned} \chi(\mathcal{O}) &= |T_2| - |T_1| + |T_0| + \sum_{e \in T_1} (1 - \frac{1}{\Gamma(e)}) - \sum_{v \in T_0} (1 - \frac{1}{\Gamma(v)}) \\ &= \chi(X_{\mathcal{O}}) + \sum_{e \in T_1} (1 - \frac{1}{\Gamma(e)}) - \sum_{v \in T_0} (1 - \frac{1}{\Gamma(v)}). \end{aligned}$$

So we can view the Euler characteristic of $\mathcal{O}$ as simply a "correction" to the Euler characteristic of its underlying space in order to account for the symmetry structure.

As additional confirmation that this is a good definition for the Euler characteristic of an orbifold, we note that this satisfies the Gauss-Bonnet theorem derived in chapter 4. If we suppose our orbifold has no edges with non-trivial local group, and we endow our orbifold with a constant negative curvature metric such that each local group acts via isometries, then as we have pointed out before each of the orbifold points becomes a cone point of our surface with cone angle $\frac{2\pi}{|\Gamma(\alpha)|}$. The Gauss-Bonnet theorem for cone surfaces then tells us that

$$\begin{aligned}
\text{Area}(\mathcal{O}) &= -2\pi\chi(X_{\mathcal{O}}) + \sum_{v\in T_0}(2\pi - \frac{2\pi}{|\Gamma(v)|}) \\
&= -2\pi\left(\chi(X_{\mathcal{O}}) - \sum_{v\in T_0}(1 - \frac{1}{|\Gamma(v)|})\right) \\
&= -2\pi\chi(\mathcal{O}).
\end{aligned}$$

Chapter 6

Geodesic Currents

6.1 The Space of Geodesics (via Topology)

Consider a topological surface S of genus $g \geq 2$. We know that the universal cover of this surface is homeomorphic to the unit disk $\mathbb{D}^2$, and we can give this disk a hyperbolic metric to form $\mathbb{H}^2$. We saw in section 3 of chapter 2 how this metric leads to a natural definition of a boundary for $\mathbb{H}^2$. That is, each point Ω of the boundary is identified with an equivalence class of geodesic rays in $\mathbb{H}^2$ which remain a bounded distance apart when given a unit-speed parameterization. This could then be used to define immediately a notion of the space of geodesics on S, denoted $\mathcal{G}(S)$. However, it will be valuable to instead develop a purely topological definition for this space.

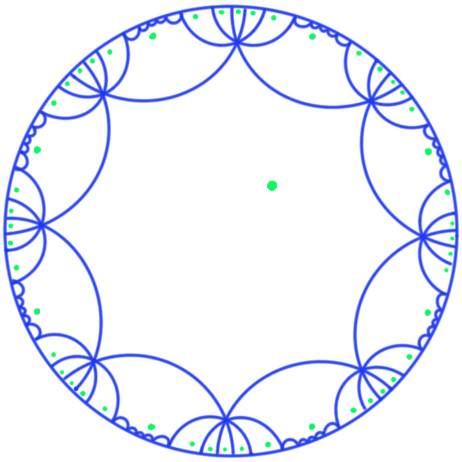
To do this, let us initially be agnostic about the metric on S, and simply refer to it as d. So to be clear, d need not even be a hyperbolic metric. However, regardless of the metric we place on S, its compactness will ensure that any two such metrics d_1 and d_2 will be "quasi-isometric," that is there will exist a map $f : (S, d_1) \to (S, d_2)$ such that, for some constants $A \geq 1$, $B \geq 0$, and $C \geq 0$,

(i) For all $x, y \in S$, we have $\frac{1}{A} d_1(x, y) - B \leq d_2(f(x), f(y)) \leq A d_1(x, y) + B$, and

(ii) For all $z \in S$ there exists $x \in S$ such that $d_2(z, f(x)) \leq C$.

In other words, the metric d_2 must only differ from the metric on d_1 by a constant multiple, and by some constant difference. Note that we are assuming that d_1 and d_2 generate the same topology on S, so d_1 and d_2 are not allowed to be arbitrarily erratic metrics.

It is still the case that $\pi_1(S)$ acts on the universal cover $\tilde{S}$ by the covering space theory developed in chapter 3. And it will act by isometries if we lift a given metric d on S to the universal cover $\tilde{S}$. This will allow us to employ a powerful fact geometric group theory known as the Milnor-Schwarz lemma. Firstly, pick a point $p \in \tilde{S}$ and consider its orbit under $\pi_1(S)$:



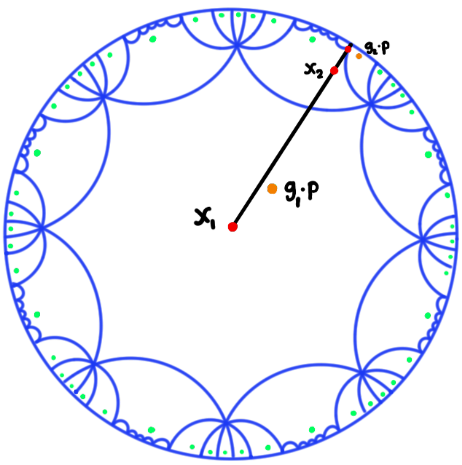
Then the Milnor-Schwarz lemma states the following [11]:

Proposition 6.1.1 (Milnor-Schwarz)**.** *If a group of isometries G acts properly discontinuously on a (proper, geodesic) metric space X such that the quotient space X/G is compact, then G is finitely generated with generating set T and for any point $p \in X$ the map*

$$f : (G, d_T) \to (X, d_X), \quad g \mapsto g \cdot p$$

is a quasi-isometry,

where d_T is the word metric on G. That is, $d_T(g_1, g_2)$ is the minimal number of generators in T needed to get from g_1 to g_2, or equivalently the minimum number of generators needed to form $g_1 g_2^{-1}$. Condition (ii) from the definition of quasi-isometry is what will be most important for us here. Consider a geodesic ray which approaches some boundary point $\Omega \in \tilde{S}$, and choose a sequence of points $\{x_i\}$ on this geodesic which approach this point on the boundary:



For each point x_i, choose an element of the orbit of p, say $g_i \cdot p$, which is within a distance of C of x_i. Then the sequence of points $\{g_i \cdot p\}$ stays within a bounded distance of this geodesic ray, and so approaches the same point on the boundary. This allows us to associate the boundary point Ω with the sequence $\{g_i\}$. And moreover, as any two metrics which generate the topology on S are quasi-isometric with each other, this sequence of group elements is not dependent on the choice of metric on S, only on its topology. We also obtain a natural action of $\pi_1(S)$ on the boundary of $\tilde{S}$ thought of in this topological light, given by left-multiplication by an element $h \in \pi_1(S)$, i.e. $h \cdot \{g_i\} = \{hg_i\}$. We refer to this space of boundary points as S^1_∞.

The importance of this discussion is to emphasize the fact that S^1_∞ is a purely topological object, and independent of the metric we place on S. However, when we do place a hyperbolic metric on S, then S^1_∞ can also be thought of as the space of equivalence classes of geodesic rays that stay bounded distance apart. As before in chapter 2, we can form the space of geodesics $\mathcal{G}(\mathbb{H}^2)$ by constructing the set of unordered pairs of distinct boundary points, $(S^1_\infty \times S^1_\infty - \Delta)/(x, y) \sim (y, x)$ (Note that since S^1_∞ is topologically a circle, this space of geodesics will be topologically a Möbius strip). Now though, we have an action of $\pi_1(S)$ on this space given by simply extending the action on S^1_∞ discussed above. Quotienting by this action yields the space of geodesics on S, which we denote as $\mathcal{G}(S)$.

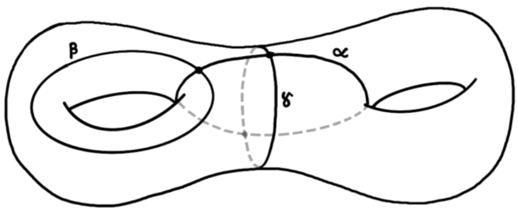
6.2 Geodesic Currents and Intersection Numbers

A **Geodesic Current** on a surface S is any $\pi_1(S)$-invariant Radon measure on $\mathcal{G}(\mathbb{H}^2)$. A Radon measure being any measure for which the measure of any open set U is the supremum of the measures of the compact subsets of U, and for which every point $x \in \mathcal{G}(S)$ is contained in an open set of finite measure. A measure μ being $\pi_1(S)$-invariant means that $\mu(g \cdot A) = \mu(A)$ for any $g \in \pi_1(S)$ and measurable $A \subseteq \mathcal{G}(\mathbb{H}^2)$.

A basic example of such a measure is the dirac measure associated to a simple closed geodesic curve γ on S [12]. This measure, which we will denote as δ_γ, assigns to any subset of $\mathcal{G}(\mathbb{H}^2)$ a measure equal to the number of lifts of γ it contains. Since the lifts of γ are invariant under the action of $\pi_1(S)$, this measure will also be $\pi_1(S)$-invariant. It also happens to be a Radon measure, since any lift of γ will only be contained in an open set U if it is contained in some compact subset of U (remember, $\mathcal{G}(S)$ is topologically just a Möbius strip, so our usual intuitions apply here). Moreover each point of $\mathcal{G}(\mathbb{H}^2)^2$ either has a neighborhood containing no lifts of γ or is itself a lift of γ, so every point has a local neighborhood whose measure is either 0 or 1 and so is finite. We can also define a measure for any non-simple closed loop which is a multiple of γ by simply setting $\delta_{\gamma^n} = n\delta_\gamma$.

More than just being a simple first example, the dirac measures δ_γ turn out to be a useful tool for characterizing all other geodesic currents of S. For one thing, it is a theorem of Bonahon that the collection of real linear combinations of these dirac measures is dense within the space of all geodesic currents of S, typically denoted $\mathcal{C}(S)$ [12]. In other words, for any geodesic current μ, there exists a sequence of multiples of dirac functions $\{c_i\delta_{\gamma_i}\}$ whose partial sums converge to μ, in the sense that for any (integrable) real-valued function f on $\mathbb{G}(\mathbb{H}^2)$ its integral with respect to the partial sums of dirac measures converges to its integral with respect to μ (this is known as the weak^* topology on $\mathcal{C}(S)$).

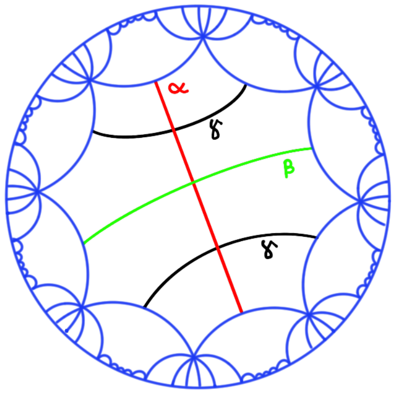
More than this though, there is a more precise, geometric connection between arbitrary geodesic currents and the simple dirac measures. This is provided by the notion of the "intersection number" of two geodesic currents. To explain this we begin with the notion of the geometric intersection number of curves. For any two simple closed curves α and β on a surface S, we define their geometric intersection number $I(\alpha, \beta)$ to be the minimum number of intersection points possible among all representatives of α and β in their respective homotopy classes. For example, in the picture below the curves α and β have a geometric intersection number of 1, while β and γ have a geometric intersection number of 2



We can reformulate this concept into the language of geodesic currents. Consider the space of pairs of geodesics in $\mathbb{H}^2$, $\mathcal{G}(\mathbb{H}^2) \times \mathcal{G}(\mathbb{H}^2) = \mathcal{G}(\mathbb{H}^2)^2$. It turns out that the action of $\pi_1(S)$ on this space is free and properly discontinuous, meaning that the quotient map to $\mathcal{G}(S)^2$ is actually a covering map. In this case, it is possible to directly define a measure on $\mathcal{G}(S)^2$ in terms of a $\pi_1(S)$-invariant measure on $\mathcal{G}(\mathbb{H}^2)^2$. Let μ be a $\pi_1(S)$ invariant measure on $\mathcal{G}(\mathbb{H})^2$. Consider an open set $U \subseteq \mathcal{G}(S)^2$ for which its pre-image in $\mathcal{G}(\mathbb{H}^2)^2$ consists of disjoint sheets and let $\tilde{U}$ be one of those sheets. Then we define the measure μ^* on $\mathcal{G}(S)^2$ by $\mu^*(U) = \mu(\tilde{U})$. Since the open sets U of this form generate the topology on $\mathcal{G}(S)^2$, we can use μ^* to measure any open subset of $\mathcal{G}(S)^2$. This is the "local push-forward" of the measure μ on $\mathcal{G}(\mathbb{H}^2$ to $\mathcal{G}(S)^2$ [12].

Now construct the subset $DG(\mathbb{H}^2) \subseteq \mathcal{G}(\mathbb{H}^2)^2$ which consists of all pairs of geodesics which intersect, and let $DG(S)$ be the image of $DG(\mathbb{H}^2)$ under the quotienting map. For two closed loops on S, α and β, we can construct the product measure $\delta_\alpha \times \delta_\beta$ by defining $\delta_\alpha \times \delta_\beta(A \times B) = \delta_\alpha(A)\delta_\beta(B)$ for any measurable subsets $A, B \subseteq \mathcal{G}(\mathcal{H}^2)^2$. We can now observe that if we take $\delta_\alpha \times \delta_\beta^*$

to be the local-pushforward measure on $\mathcal{G}(S)^2$, then $\delta_\alpha \times \delta^*_\beta(DG(S))$ recovers the geometric intersection number $I(\alpha, \beta)$. For instance, when lifted to a fundamental domain in $\mathbb{H}^2$ the above curves may look like:

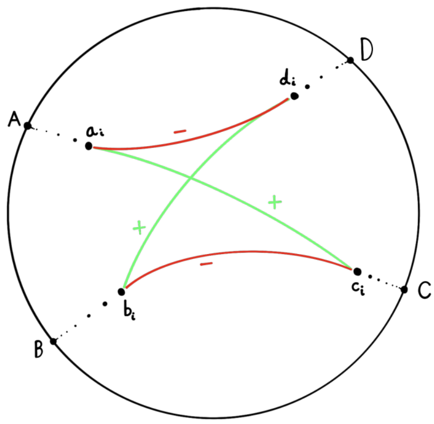


Note that an equivalent way of characterizing the space $DG(S)$ is as the collection of all pairs of geodesics which intersect in some chosen fundamental domain of $\mathbb{H}^2$ representing S.

From this perspective, it is clear how we can generalize the concept of an intersection number to any geodesic currents. Let μ and ν be $\pi_1(S)$-invariant measures on $\mathcal{G}(\mathbb{H}^2)^2$, then define their intersection number $i(\mu, \nu) = \mu \times \nu^*(DG(S))$. It is a theorem of Bonahon's [13] that this function is a continuous, bi-linear form on $\mathcal{C}(S)$. Importantly, this construction provides a way to characterize any geodesic current via a finite set of data. Namely, the values of $i(\delta_\gamma, \mu)$ over all closed curves γ in S completely determines the measure μ [1]. This will be relevant for seeing the naturalness of the Liouville current in the next section.

6.3 Liouville Currents

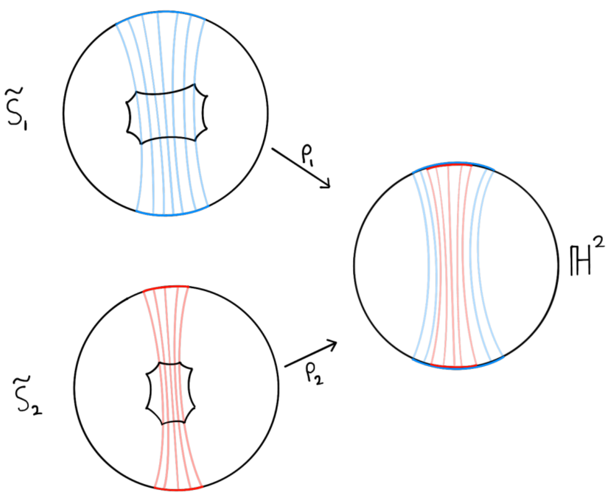
Let us first describe the Liouville current explicitly as a measure on $\mathcal{G}(\mathbb{H}^2)$. The simplest definition for it is likely via its cross-ratio definition [14]. For this, consider two disjoint intervals $[A, B]$ and $[C, D]$ in S^1_∞ and choose sequences of points $\{a_i\}$, $\{b_i\}$, $\{c_i\}$, and $\{d_i\}$ in $\mathbb{H}^2$ which converge to A, B, C, and D respectively. For each term in these sequences we can construct the geodesic segments $\overline{a_i d_i}$, $\overline{a_i c_i}$, $\overline{b_i d_i}$, and $\overline{b_i, c_i}$:



We then define the cross ratio of the boundary points A, B, C, and D as

$$C(A, B, C, D) = \frac{1}{2} \lim_{i \to \infty} \left(|a_i c_i| - |a_i d_i| + |b_i d_i| - |b_i c_i| \right).$$

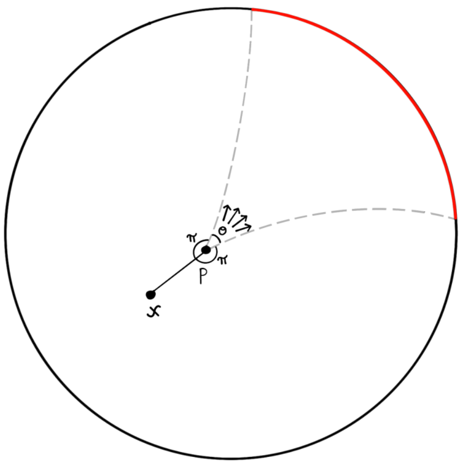
Using this, we define the Liouville current L on $\mathbb{H}^2$ by setting $L([A, B] \times [C, D]) = |C(A, B, C, D)|$. Note that this measure is invariant under any isometry of $\mathbb{H}^2$. As it stands, this particular measure on $\mathcal{G}(\mathbb{H}^2)$ is not at all dependent on any hyperbolic metric φ on our surface S. To get a proper geodesic current from this, consider some open set of geodesics on S, say U. Recall that there is a purely topological notion for these geodesics, as we discussed in the first section of this chapter. So if we have two hyperbolic surfaces (S_1, φ_1) and (S_2, φ_2) which are homeomorphic to S then as a set this collection of geodesics on S_1 is the same as that on S_2. However, because $\pi_1(S_1)$ and $\pi_1(S_2)$ act by different groups of translations on $\mathbb{H}^2$, the lifts of these geodesics to the universal covers $\tilde{S}_1$ and $\tilde{S}_2$ will look different:



As $\tilde{S}_1$ and $\tilde{S}_2$ are isometric to $\mathbb{H}^2$, we can map these collections of geodesics into $\mathbb{H}^2$ and evaluate their measure with respect to the Liouville current here. We will then find that $L_{\varphi_1}(U) \neq L_{\varphi_2}(U)$, where L_{φ_1} and L_{φ_2} represent the value of the Liouville metric for the different lifts of our set of geodesics. Thus, L_φ is what we call the Liouville current for (S, φ) and it is indeed dependent on the metric φ, unlike the Liouville current L on $\mathbb{H}^2$.

The Louivile current actually has a much stronger relationship with the geometry on S than even the above explicit definition would suggest. It is the case that for any closed loop α on S, the intersection number $i(\delta_\alpha, L_\varphi)$ equals the length of α with respect to φ, $\ell_\varphi(\alpha)$ [12]. In other words, the Liouville current determines the marked length spectrum of S, and vice versa. We saw in the introduction that Otal's theorem assures that the marked length spectrum determines the metric φ, and so it follows that in fact L_φ must determine the metric φ as well. This is the fact that the CST seeks to generalize to, and indeed strengthen for hyperbolic cone surfaces.

We can define the Liouville current for hyperbolic cone surfaces using the same definition as above, however a fundamental change occurs to the Liouville current in that it loses some of its support. That is, there will now be open sets of geodesics which have measure 0 according to the Liouville current. To see this, we first recall that geodesic segments cannot be uniquely continued through cone points of excess cone angle. Rather, if one connects a regular point x to a cone point p then there will be a whole range of geodesic continuations, corresponding to a range of boundary points at infinity that these geodesics will tend toward [14]:



So consider a geodesic which passes through a cone point p such that its endpoints lie within two such intervals on the boundary, i.e. it turns by an angle greater than π at p (such a geodesic is referred to as 'singular'). Choose two points x_1 and x_2 on this geodesic which lie on opposite sides of p. Let $[A, B]$ be the interval of endpoints for the possible continuations of the segment $\overline{x_2 p}$ and $[C, D]$ the interval of endpoints for the possible continuations of $\overline{x_1 p}$. Then every geodesic connecting an endpoint in $[A, B]$ to an endpoint in $[C, D]$ has to pass through p. In particular, choose points a_i, b_i, c_i, and d_i such that their distances to p are all ℓ_i and they lie along the geodesic rays connecting p to A, B, C, and D respectively. Once can then compute the cross ratio in this case and see that it evaluates to 0. Thus we conclude that $L([A, B] \times [C, D]) = 0$. In other words, given a singular geodesic we have found a neighborhood of it which has measure 0. Thus, the support of the Liouville current L_φ consists only of the non-singular geodesics of S.

As in the case of surfaces without cone points, the geodesic current L_φ determines the hyperbolic cone metric φ. However, the CST states something even stronger. One doesn't need to know the Liouville current explicitly, in almost all case they only need to know its support, i.e. the space of non-singular geodesics, to determine φ.

Chapter 7

An Application to Hyperbolic Billiards

Having now reviewed the mathematical theory surrounding the CST, we are in a good position to use it in solving other problems. In particular, the original paper by Leininger et al. provides an application of their theorem to symbolic dynamics of hyperbolic polygons. In particular, they prove a result they refer to as the "Hyperbolic Bounce Theorem," which we will introduce below.

7.1 The Problem

Let P be a polygon in the hyperbolic plane, i.e. a simply connected, bounded subset of $\mathbb{H}^2$, whose boundary is piece-wise geodesic. We can label its sides cyclically, by say A, B, C, etc.

Now imagine we have a billiards ball bouncing around inside this polygon, which we may imagine as our “billiards table." That is, it travels along a geodesic segment while it is in the interior of P, but once it contacts an edge of the polygon, its direction of travel is changed such that the angle of incidence with the edge is equal to the angle of reflection. The path that this point traces out we then refer to as a billiards trajectory.

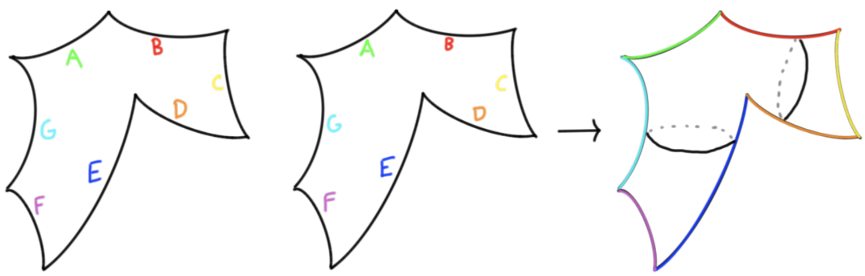
For a billiards trajectory in this polygon, we can record the sequence of sides that it hits (both going forward and backward in time), e.g.:

$$...ABACDBABABCDADBCBDA...$$

This is a “bounce sequence" of P. The collection of all possible bounce sequences within a polygon we will call its **“bounce spectrum"**, denoted $B(P)$. The question of interest to us is then whether the bounce spectrum of a polygon determines its shape up to isometry.

7.2 Unfolding Polygons

To tackle this problem using the CST, we will first need to form a cone surface from our polygon P. One immediate way to do this is to make a copy of our polygon and identify the corresponding edges of the polygons:



We refer to this surface as the *double* of the polygon P, or more briefly, DP. It is topologically a sphere, possessing a constant negative curvature hyperbolic metric everywhere, except at the vertices of the polygon which here have become cone points.

Notice that having our billiards trajectory reflect off an edge of the polygon is equivalent to letting it pass through the edge in a straight line and into the copy of the polygon. Therefore billiards trajectories in our polygon become geodesic paths on its double. Another word for the double of our polygon is then an *unfolding*, since whenever our billards ball hits an edge we may imagine unfolding our polygon along this edge to allow it to pass through. Though, we will need to perform more unfoldings than this to be able to apply the CST.

The only remaining deficiency of this surface is that some of its cone angles will be less than 2π. To remedy this, we will first remove the cone points from our doubled polygon to form DP^o. Now we have a non-simply connected surface, and we can consider coverings of it. In particular, if our polygon P has n vertices, then $\pi_1(DP^o) \cong F_{n-1}$, the free group on $n-1$ generators. For convenience however, we will grant F_{n-1} the n generators $a_1, \ldots, a_n$ and simply add in the relation $a_n = a_1 a_2 \cdots a_{n-1}$. Thus a_i corresponds to a loop going around the i-th puncture of DP^o.

If we wish to "unravel" our surface around the i-th puncture k times, such a covering will turn out to correspond to the subgroup

$$H_i = \langle a_1, a_2, \ldots, \widehat{a_i} \ldots, a_n, a_i a_1 a_i^{-1}, \ldots, a_1 a_n a_1^{-1}, \ldots, a_i^{k-1} a_n a_i^{1-k}, a_i^k \mid a_n = a_1 a_2 \ldots a_{n-1} \rangle,$$

where $\widehat{a_i}$ denotes the absence of a_i as a generator in this subgroup. If we suppose that

$k \cdot \theta_i > 2\pi$, where θ_i is the cone angle of the cone point we removed from DP^o to form the i-th puncture, then when we add our cone point back into the surface it will have a cone angle which is greater than 2π. Moreover, since H_i has index k within F_{n-1}, the covering space which it corresponds to will be a k-sheeted cover.

To perform this unraveling at every puncture of our surface, we simply take the intersection of these subgroups: $H := \bigcap_{i=1}^{n} H_i$. Since each H_i has finite index, their intersection will have finite index as well (see Appendix A, section 3), thus the covering corresponding to H is finite-sheeted. Call this covering space X^o. Of course, X^o still contains many copies of the punctures on DP^o, so to fill these in we will simply take the metric completion of X^o, which we will call X. The spaces we have constructed thus far can then be organized into the following diagram:

$$\begin{array}{ccc} X^o & \xrightarrow{\imath} & X \\ \big\downarrow{\scriptstyle q} & & \big\downarrow{\scriptstyle \hat{q}} \\ DP^o & \xrightarrow{\imath} & DP \\ & & \big\downarrow{\scriptstyle \pi} \\ & & P \end{array}$$

7.3 Partial Proof of the Hyperbolic Bounce Theorem

To write the hyperbolic bounce theorem (HBT) down formally, we require the notion of a *pair homeomorphism*. That is, given spaces X and Y, and subspaces $A \subseteq X$ and $B \subseteq Y$, we say $f : X \to Y$ is a pair homeomorphism if f is a homeomorphism and f restricted to A is a homeomorphism from A to B. We then write $f : (X, A) \to (Y, B)$. We can now formally state the HBT:

Theorem 7.3.1 (Hyperbolic Bounce Theorem). *Let P_1 and P_2 be two polygons in $\mathbb{H}^2$. Then $B(P_1) = B(P_2)$ if and only if P_1 and P_2 are isometric, or there exists a pair homeomorphism $f : (P_1, T_1) \to (P_2, T_2)$ where T_i is the skeleton of a tiling of P_i such that adjacent tiles differ by reflection and each vertex angle of P_i is subdivided into angles of the form $\frac{\pi}{2k}$ by the tiles in T_i.*

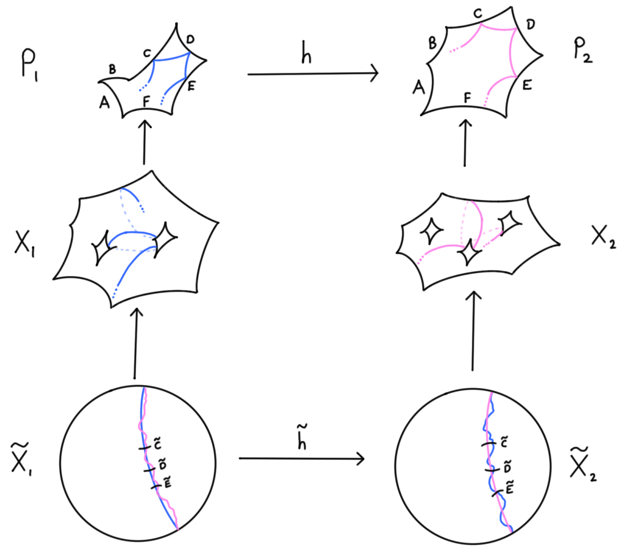
The proof can be broken into two cases. If X_i is rigid, which will occur e.g. if any of the angles of P_i are irrational, then we will only have equivalence of bounce spectra if the two polygons are isoometric. Otherwise if the X_i's are flexible, then the CST may be used to obtain the description of the exceptional polygons in terms of tilings.

Here we focus on the proof in the case that the X_i's are rigid. To begin, suppose we have polygons P_1 and P_2 such that $B(P_1) = B(P_2)$, and construct branched covers X_1 and X_2 as described above. We will first note that the bounce spectrum of a polygon determines its number of sides. Therefore P_1 and P_2 have the same number of sides and vertices, and so the fundamental group for both of their punctured double covers will be F_{n-1}. X_1 and X_2

both correspond to subgroups of F_{n-1}, and so we may take the covering correspondings to the intersection of these subgroups to be our X_i's. In this way, we guarantee that X_1 and X_2 are homeomomorphic.

If we view P_1 and P_2 as being topologically just disks with different metrics attached to them, then there is an identity map h between them which is a homeomorphism and which preserves edge labelings, but which is not an isometry. This homeomorphism then lifts to a homeomorphism $\tilde{h}$ between the universal covers $\tilde{X}_1$ and $\tilde{X}_2$. Because we chose our covering spaces X_1 and X_2 to have negative curvature at their cone points (i.e. all cone angles are greater than 2π), these universal covers will be hyperbolic planes with cone points of their own. Let G_i be the space of non-singular geodesics (i.e. don't pass through cone points) of $\tilde{X}_i$ as described in chapter 6.

Every non-singular geodesic in $\tilde{X}_1$ (colored in blue below) maps by a composition of the covering and folding maps to a billiard's trajectory in the Polygon P_1. If $B(P_1) = B(P_2)$, then for any billiards trajectory in P_1 there is a corresponding billiards trajectory in P_2 (colored in pink below) which hits the identically labeled edges. When we lift this path to the universal cover X_2, we choose the lift so that it passes through at least one edge of the universal cover that the blue geodesic also passes through. It follows that from then on, the blue and pink paths will pass through the same edges in the universal cover of X_1. As the universal cover is tiled by copies of P_2, and since the blue and pink paths pass through the same sequence of edges in both directions, it follows that they will always pass through the same tiles and so remain a bounded distance apart.



Now, this means the blue geodesic in the above picture corresponds to some element of

G_1, say (x, y). $\tilde{h}$ then maps (x, y) to precisely itself. Moreover, since the pink path stays a bounded distance away from the blue path, it must also correspond to the pair of boundary points (x, y). That is, every element of G_1 is an element of G_2, and by reasoning the other way we also have that every element of G_2 is an element of G_1 as well. We then conclude that $G_1 = G_2$. However, we assumed that X_1 and X_2 were rigid cone surfaces, so since the space of non-singular geodesics (i.e. the supports of the liouville currents) are equivalent, we conclude that in fact X_1 and X_2 are isometric. It then follows that P_1 and P_2 must have been isometric as well, completing the proof.

7.4 Exceptions to the HBT

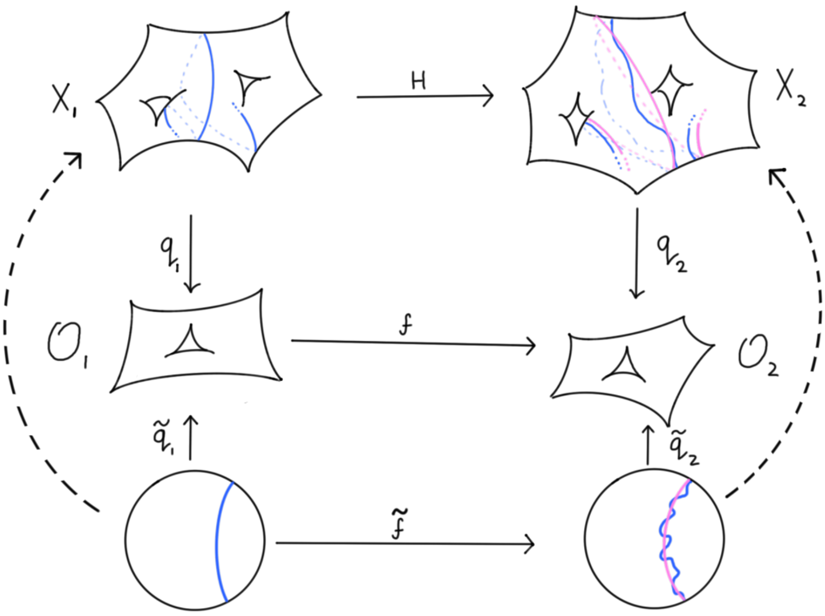
While we will not prove the full characterization of exceptional polygons here, we will demonstrate how the flexibility of X_1 and X_2 allows for such exceptions to occur.

Suppose P_1 and P_2 are not isometric polygons. However their homeomorphic branch covers X_1 and X_2 are flexible, and cover equivalent orbifolds $\mathcal{O}_1 \cong \mathcal{O}_2$. We will further assume that only the cone points are mapped to the orbifold points of these orbifolds, and moreover the stabilizer of each even-ordered orbifold point is $\mathbb{Z}_{2n}$ acting by rotation. Finally, since X_1 and X_2 cover these orbifolds, their universal covers will also cover $\mathcal{O}_1$ and $\mathcal{O}_2$. We collect this information in the following commutative diagram:

$$\begin{array}{ccc} X_1 & \xrightarrow{H} & X_2 \\ \big\downarrow{\scriptstyle q_1} & & \big\downarrow{\scriptstyle q_2} \\ \mathcal{O}_1 & \xrightarrow{f} & \mathcal{O}_2 \\ {\scriptstyle \tilde{q}_1}\big\uparrow & & {\scriptstyle \tilde{q}_2}\big\uparrow \\ \tilde{\mathcal{O}}_1 & \xrightarrow{\tilde{f}} & \tilde{\mathcal{O}}_2 \end{array}$$

where q_i and $\tilde{q}_i$ are the respective covering maps, H is a homeomorphism from X_1 to X_2, f is an orbifold equivalence, and $\tilde{f}$ is a lift of this equivalence to the universal covers. Any geodesic in X_1 which passes through a cone point will map to a geodesic in $\mathcal{O}_1$ since q_1 is local isometry. This in turn will lift to a geodesic in the universal cover of our orbifold, $\tilde{\mathcal{O}}_1$. Because the fundamental group of $\mathcal{O}_1$ acts by even-order rotation on the orbifold point that our cone point maps to, it will also act on the lift of this point due to the equivariant property of $\tilde{f}$. If this subgroup is generated by g of order $2n$, then g^n is a rotation of $\tilde{\mathcal{O}}_1$ by π radians around the pre-image of the orbifold point, and so must map this lifted geodesic to itself. Importantly, the endpoints of this geodesic are preserved, and so they are still contained in $G^c_{\varphi_1}$, as they bound a singular geodesic.

At this point, our situation is summarized in the left-hand side of the following diagram:

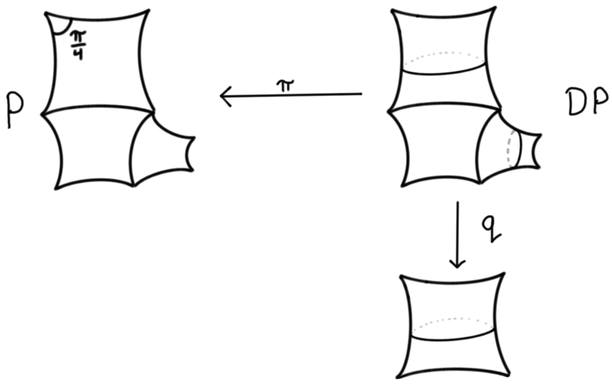


We now consider the image of the blue geodesic above under the topological identity map which changes the metric of our spaces. In the space X_2, the image of the blue geodesic is within some free homotopy class of paths, and as we know from section 4 each homotopy class of paths on a hyperbolic surface contains exactly one geodesic. This is the pink curve sketched above. Note that since X_2 is compact, the blue curve remains a bounded distance away from the pink curve. And this will remain true when we lift to the universal cover (represented by the dashed arrow). Thus they have the same endpoints as each other. In order to conclude $G^c_{\varphi_1} = G^c_{\varphi_1}$, we thus need to show that the pink geodesic is singular as well.

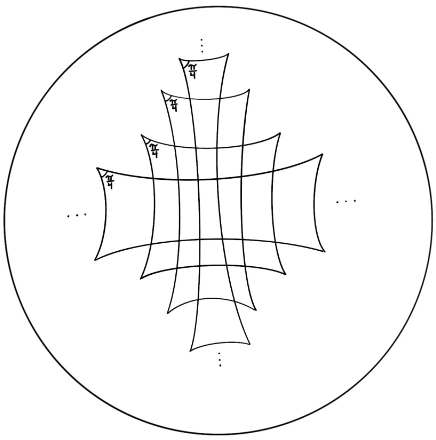
To finish this argument, note that as the group element g^n swaps the endpoints of the pink geodesic, its corresponding group element in $\pi_1(\mathcal{O}_2)$ must do the same to the pink geodesic. However, a continuous map from a line to itself cannot swap the endpoints without having at least one fixed point. This fixed point thus has a group of order 2 acting on it, and thus must cover an even-order orbifold point. However, the pink geodesic in $\mathcal{O}_2$ can only pass through an even-order orbifold point if it passed through a cone point in X_2, by our assumptions. Thus the pink geodesic must pass through a cone point and so is singular, completing the proof.

For an explicit example of the exceptional situation described in the HBT, consider 3 squares glued together in an L-shape, each square having an angle of $\frac{\pi}{2n}$ at each vertex, for $n > 1$ ($n = 2$ illustrated below):

As with any other polygon, we can form its double then unfold it to some hyperbolic surface. But here, we can also go further and form an orbifold covering by first covering the double of a square tile by the double of our polygon P, then quotienting this by reflection to form an orbifold out of our square tile:



However, for any square in the hyperbolic plane, there exists an infinite family of other quadrilaterals which have different side lengths but the same angle at each vertex:



If we use any one of these quadrilaterals in place of the regular quadrilateral we began with, we obtain “stretched" versions of our original orbifold and doubled polygon. They will possess the same angular and combinatorial information. Moreover, only the cone points of the covering surface for this stratched polygon will map to the even-order orbifold points (i.e. any of the orbifold points in this case). Therefore these polygons satisfy the abstract conditions we laid out above, and so even though they are not isometric they will possess the same bounce spectra.

Chapter 8

Surface Genus and Cone Point Count

8.1 A Preliminary Bound

We begin this section by laying out Leininger, et al.'s result that a genus g hyperbolic surface with more than $32(g-1)$ cone points must be rigid. Recall that the Gauss-Bonnet theorem for cone surfaces assures us that for a surface S and a hyperbolic cone metric φ on it,

$$\text{Area}(S) = 2\pi \left(2(g-1) + \sum_i 1 - \frac{\theta_i}{2\pi} \right),$$

where θ_i is the cone angle of the i-th cone point. If we suppose that φ is flexible, then the CST tells us there must exist a locally isometric covering $q : S \to \mathcal{O}$ to an orbifold $\mathcal{O}$ which sends cone points to even-order orbifold points. As this covering is locally isometric, in particular it locally preserves areas and so its degree d must satisfy

$$d = \frac{\text{Area}(S)}{\text{Area}(\mathcal{O})} = \frac{2\pi \left(2(g-1) + \sum_i 1 - \frac{\theta_i}{2\pi} \right)}{\text{Area}(\mathcal{O})}.$$

Since $\theta_i > 2\pi$ for all i, the sum present in the numerator will be strictly negative. We may therefore obtain an upper bound on d by omitting it:

$$d < \frac{2\pi(2(g-1))}{\text{Area}(\mathcal{O})} = \frac{4\pi}{\text{Area}(\mathcal{O})}(g-1).$$

Now let n be the number of cone points for the metric φ, and r be the number of even-order orbifold points on $\mathcal{O}$. We know that each of the n cone points must map to an even-order orbifold point. Moreover if the i-th cone point p_i maps to an orbifold point of order b_j, then its cone angle θ_i must be given by $\frac{2\pi}{b_j}$ (the angle at $q(p_i)$) times the number of sheets that the p_i contributes to the covering of $q(p_i)$. As $\theta_i > 2\pi$, it follows that this number of sheets must be greater than or equal to $b_j + 1$, which in turn must be greater than or equal to 3. We can then reason that the number of cone points which map to a given orbifold point cannot exceed $\frac{d}{3}$, since each cone point contributes at least 3 sheets to the covering. Repeating this reasoning for all r even-order orbifold points, we conclude that

$$n \leq \frac{r}{3}d < \frac{4\pi r}{3\text{Area}(\mathcal{O})}(g-1).$$

In order to get a bound on n which will work for any orbifold $\mathcal{O}$, we will need to optimize the right-hand side over all possible orbifolds $\mathcal{O}$. This involves some light case work, based on the value of r and the genus of the orbifold $\mathcal{O}$. But once this is accomplished, one finds that the value of $\frac{4\pi r}{3\text{Area}(\mathcal{O})}$ is maximized at 32, which gives the desired conclusion.

8.2 Cone points do not Exceed $3(g-1)$

We are now ready to prove our extension of this result, namely

Theorem 8.2.1. *A hyperbolic cone surface S of genus g having more than $3(g-1)$ cone points must be rigid. Moreover, this is the optimal such bound of the form $c(g-1)$.*

We postpone showing optimality until the next section. Here, we will simply prove the bound on the cone points. So, let S be a genus g flexible hyperbolic cone surface with n cone points. By the CST, S must cover a hyperbolic orbifold $\mathcal{O}$ such that each cone point is mapped to an even-ordered orbifold point. For what follows, we will say $\mathcal{O}$ has signature $(g_o; b_1, b_2, \ldots, b_r, c_1, c_2, \ldots, c_s)$, where the b_i's are even and ordered so that $b_1 \leq b_2 \leq \cdots \leq b_r$, and the c_j's are all odd. Note that we are allowing multiplicities among the b_i's, and among the c_j's. Because of this we will slightly abuse notation and refer to the orbifold points and their orders interchangeably. Also, for convenient notation later on, we will define $\hat{p}$ to be the restriction of the covering map to the set of cone points on S.

Since $\mathcal{O}$ is hyperbolic, it must have a negative euler characteristic. Recalling the formula for the euler characteristic of an orbifold, we therefore have

$$|\chi(\mathcal{O})| = 2(g_o - 1) + \sum_{i=1}^{r} \frac{b_i - 1}{b_i} + \sum_{j=1}^{s} \frac{c_j - 1}{c_j}. \tag{8.1}$$

Let us abbreviate the above value as χ. We first observe that if the genus g_o is greater than or equal to 1, then χ is at least $\frac{1}{2}r$, assuming all orbifold points have order 2. Therefore following the argument of the original paper, we can bound the number of cone points in this case by

$$n \leq \frac{r}{3}d \leq \frac{r}{3} \cdot \frac{2(g-1)}{\frac{1}{2}r} = \frac{4}{3}(g-1),$$

where d is the degree of our covering map and n is the number of cone points on S. Since this bound is less than $3(g-1)$, our claim is proven in the case that $g_o \geq 1$. For what follows, we will therefore assume $g_o = 0$.

We continue by choosing an appropriately small value a so that $n \geq a(g-1)$. We will then work toward finding an inequality that a must satisfy, and in so doing we will bound n. To obtain this inequality on a, consider the degree d of our covering map which is given by

$$d = \frac{2(g-1) + \sum_{k=1}^{n} 1 - \frac{\theta_k}{2\pi}}{\chi}, \tag{8.2}$$

where each θ_k is the cone angle of the k-th cone point on S. These cone angles must each be strictly greater than 2π, so the sum in the numerator is strictly negative. We may therefore get an upper bound on d by omitting this sum:

$$d \leq \frac{2}{\chi}(g-1).$$

Now let us generalize and suppose that we have an upper bound on d of the form $d \leq q(g-1)$. The above argument is only useful for establishing that such a q does indeed exist, but the exact form that q takes is irrelevant. If d satisfies such a bound, then the number of cone points that map to an orbifold point b_i, $|\hat{p}^{-1}(b_i)|$, must be bounded above by $\frac{q}{b_i+1}(g-1)$. This is because each such cone point, having cone angle greater than 2π, must contribute at least $b_i + 1$ sheets in order to cover an orbifold point with cone angle $\frac{2\pi}{b_i}$ in a locally isometric way. As the total number of sheets over any point is d, the bound on $|\hat{p}^{-1}(b_i)|$ follows.

We can use this information to update the bound on d. Note that to get an upper bound on d, we would like to minimize the cone angles θ_k as much as possible. If the k-th cone point maps to the orbifold point b_i, then θ_k must be a multiple of $\frac{2\pi}{b_i}$. The smallest such multiple which is strictly greater than 2π is $\frac{b_i+1}{b_i} \cdot 2\pi$. Thus, this cone point adds, at maximum,

$1 - \frac{(b_i+1)\cdot 2\pi}{b_i} \cdot \frac{1}{2\pi} = -\frac{1}{b_i}$ to the numerator of d. Repeating this logic for every cone point on S yields the bound

$$\begin{aligned}
d &\leq \frac{1}{\chi}\left(2(g-1) + \sum_i \sum_{\hat{p}^{-1}(b_i)} -\frac{1}{b_i}\right) \\
&= \frac{2}{\chi}(g-1) + \frac{1}{\chi}\left(\left(-\frac{1}{b_1}\right)\left(n - \sum_{i>1} |\hat{p}^{-1}(b_i)|\right) + \sum_{i>1}\left(-\frac{1}{b_i}\right)|\hat{p}^{-1}(b_i)|\right) \\
&= \frac{2}{\chi}(g-1) + \frac{1}{\chi}\sum_{i>1}\left(\frac{1}{b_1} - \frac{1}{b_i}\right)|\hat{p}^{-1}(b_i)| - \frac{n}{b_1\chi} \\
&\leq \frac{2}{\chi}(g-1) + \frac{1}{\chi}\sum_{i>1}\left(\frac{b_i - b_1}{b_1 b_i}\right)\frac{q}{b_i+1}(g-1) - \frac{a(g-1)}{b_1\chi} \\
&= \left(\frac{2}{\chi} - \frac{a}{b_1\chi} + q\sum_{i>1}\left(\frac{b_i - b_1}{b_1 b_i (b_i+1)\chi}\right)\right)(g-1) \\
&= \left(\frac{2}{\chi} - \frac{a}{b_1\chi} + q\sum_{i}\left(\frac{b_i - b_1}{b_1 b_i (b_i+1)\chi}\right)\right)(g-1).
\end{aligned}$$

In summary then, by assuming a bound for d with a coefficient on $(g-1)$ of q, we obtained a new bound for d with a coefficient which is an affine function of q. We may then iterate this logic as many times as we wish to get a sequence of bounds for d. If the affine function of q has its slope in the range $(-1, 1)$, then by the contraction mapping theorem any sequence of iterations will converge to the unique fixed point of this function. Denoting this fixed point as q^*, we will then have

$$d \leq q^*(g-1). \tag{8.3}$$

Before solving for q^*, we will show that the slope is in the required range. Indeed, each term in the sum will be non-negative since b_1, b_i, and χ are all positive, and $b_i \geq b_1$. Thus this slope is certainly non-negative, and so it will suffice to show that

$$\sum_i \frac{b_i - b_1}{b_1 b_i (b_i+1)\chi} < 1,$$

or equivalently,

$$\sum_i \frac{b_i - b_1}{b_1 b_i (b_i+1)} < \chi.$$

Recall the definition of χ from equation (1), and the fact that $g_o = 0$:

$$\chi = -2 + \sum_i \frac{b_i - 1}{b_i} + \sum_j \frac{c_j - 1}{c_j} = r - 2 - \sum_i \frac{1}{b_i} + \sum_j \frac{c_j - 1}{c_j}$$

The desired inequality therefore becomes

$$\sum_i \frac{b_i - b_1}{b_1 b_i (b_i+1)} < r - 2 - \sum_i \frac{1}{b_i} + \sum_j \frac{c_j - 1}{c_j},$$

Or equivalently,

$$\sum_i \frac{b_i - b_1}{b_1 b_i (b_i+1)} + \sum_i \frac{1}{b_i} < r - 2 + \sum_j \frac{c_j - 1}{c_j}. \tag{8.4}$$

The left-hand side can be simplified as follows:

$$\begin{aligned} \sum_i \frac{b_i - b_1}{b_1 b_i (b_i+1)} + \sum_i \frac{1}{b_i} &= \sum_i \frac{b_i - b_1}{b_1 b_i (b_i+1)} + \sum_i \frac{b_1(b_i+1)}{b_1 b_i (b_i+1)} \\ &= \sum_i \frac{(b_i - b_1) + (b_1 b_i + b_1)}{b_1 b_i (b_i+1)} \\ &= \frac{1}{b_1} \sum_i \frac{b_1 + 1}{b_i + 1} \\ &\le \frac{3}{2} \sum_i \frac{1}{b_i + 1} \end{aligned}$$

A lemma we shall prove later on in this proof is that for hyperbolic, flexible orbifolds like those we are considering here,

$$\frac{5}{3} \sum_i \frac{1}{b_i+1} \le r - 2 + \sum_j \frac{c_j - 1}{c_j}.$$

And as $\frac{5}{3} > \frac{3}{2}$, this gives the desired bound and we can conclude that our affine function of q has slope less than 1. We may thus bound d by its fixed point q^*, which is given by

$$\begin{aligned} q^* &= \frac{2}{\chi} - \frac{a}{b_1 \chi} + q^* \sum_i \frac{b_i - b_1}{b_1 b_i (b_i+1)\chi} \\ \Rightarrow \left(1 - \sum_i \frac{b_i - b_1}{b_1 b_i (b_i+1)\chi}\right) q^* &= \frac{2b_1 - a}{b_1 \chi} \\ \Rightarrow \left(b_1 \chi - \sum_i \frac{b_i - b_1}{b_i (b_i+1)}\right) q^* &= 2b_1 - a \\ \Rightarrow q^* &= \frac{2b_1 - a}{b_1 \chi - \sum_i \frac{b_i - b_1}{b_i(b_i+1)}}. \end{aligned}$$

So we have

$$d \le \frac{2b_1 - a}{b_1 \chi - \sum_i \frac{b_i - b_1}{b_i(b_i+1)}} (g-1).$$

However, this bound on the degree also implies an upper bound on the number of cone points. This is because

$$n = \sum_i |\hat{p}^{-1}(b_i)| \leq \sum_i \frac{d}{b_i+1} = d \sum_i \frac{1}{b_i+1}.$$

Substituting in our bound on d, we get

$$n \leq \frac{(2b_1 - a)\sum\limits_i \frac{1}{b_i+1}}{b_1\chi - \sum\limits_i \frac{b_i - b_1}{b_i(b_i+1)}}(g-1)$$

Recall however that we also chose a so that $n \geq a(g-1)$. It follows that a must satisfy

$$a \leq \frac{(2b_1 - a)\sum\limits_i \frac{1}{b_i+1}}{b_1\chi - \sum\limits_i \frac{b_i - b_1}{b_i(b_i+1)}}.$$

Then by isolating a we get

$$\begin{aligned} a &\leq \frac{2b_1 \sum\limits_i \frac{1}{b_i+1}}{b_1\chi - \sum\limits_i \frac{b_i - b_1}{b_i(b_i+1)} + \sum\limits_i \frac{1}{b_i+1}} \\ &= \frac{2b_1 \sum\limits_i \frac{1}{b_i+1}}{b_1\chi + \sum\limits_i \frac{b_1}{b_i(b_i+1)}} \\ &= \frac{2 \sum\limits_i \frac{1}{b_i+1}}{\chi + \sum\limits_i \frac{1}{b_i(b_i+1)}} \end{aligned}$$

We can also re-write the denominator further by substituting in the formula for χ:

$$\begin{aligned} \chi + \sum_i \frac{1}{b_i(b_i+1)} &= -2 + \sum_i \frac{b_i - 1}{b_i} + \sum_j \frac{c_j - 1}{c_j} + \sum_i \frac{1}{b_i(b_i+1)} \\ &= r - 2 + \sum_j \frac{c_j - 1}{c_j} + \sum_i \left(\frac{1}{b_i(b_i+1)} - \frac{1}{b_i} \right) \\ &= r - 2 + \sum_j \frac{c_j - 1}{c_j} - \sum_i \frac{1}{b_i+1}. \end{aligned}$$

Therefore we have

$$a \leq \frac{2\sum\limits_i \frac{1}{b_i+1}}{r - 2 + \sum\limits_j \frac{c_j - 1}{c_j} - \sum\limits_i \frac{1}{b_i+1}}$$

Our goal now is to find an upper bound on the right-hand side over all possible values of

r, c_j's, and b_i's, and in particular to show that this bound is 3. By re-arranging, we see how the claim that this right-hand expression is bounded by 3 is equivalent to

$$\frac{5}{3}\sum_i \frac{1}{b_i+1} \leq r-2+\sum_j \frac{c_j-1}{c_j}. \tag{8.5}$$

To prove this inequality, we will break into cases over r. First suppose that $r > 4$. Then the left-hand side will be maximized when all b_i's equal 2, and so it will be at most $\frac{5}{9}r$. The right-hand side is minimized by assuming no odd-order orbifold points exist, giving simply $r-2$. As $\frac{5}{9}r < r-2$ for all $r > 4$, the result follows in this case.

Now suppose $r = 4$. We cannot simultaneously have $b_i = 2$ for all i, and also have no odd-order orbifold points, since then χ would be 0 and our orbifold would not be hyperbolic. If there are no odd-order orbifold points, at least one b_i must equal 4, so the left-hand side of (8.5) is maximized at

$$\frac{5}{3}\left(\frac{1}{2+1}+\frac{1}{2+1}+\frac{1}{2+1}+\frac{1}{4+1}\right) = 2.$$

As the right-hand of (8.5) side will evaluate to $r-2 = 4-2 = 2$, the inequality holds in this case. Otherwise, we assume that $b_i = 2$ for all i and there is at least one odd order orbifold point $c_j = 3$. The left hand side is then $\frac{5}{3}\cdot\frac{4}{3} = \frac{20}{9}$, while the right-hand side will be $4-2+\frac{2}{3} = \frac{8}{3}$. As $\frac{20}{9} < \frac{8}{3}$, the inequality holds here as well.

Now assume $r = 3$. Note that our orbifold must itself be flexible in order for it to be locally isometrically covered by a flexible metric of S. An orbifold with 3 orbifold points however is identical to two isometric hyperbolic triangles glued together. As a hyperbolic triangle is determined up to isometry by its vertex angles, there can only be one metric on an orbifold with three given orbifold points (which will determine the vertex angles of the triangles). Thus an orbifold with three orbifold points cannot be flexible. We may thus conclude that our orbifold has at least 4 orbifold points in total. So if $r = 3$, at least 1 orbifold point must be odd.

The left-hand side of (8.5) is then maximized at $\frac{5}{3}\cdot\frac{3}{3} = \frac{5}{3}$. The right-hand side meanwhile will be at minimum $3-2+\frac{2}{3} = \frac{5}{3}$. So the inequality is true for $r = 3$ also (It will be important for the next section to note that we have actually achieved equality in the case of the $(0; 2, 2, 2, 3)$ orbifold).

Now each time we decrease r by 1, the maximum of the left-hand side of (8.5) is decreased by $\frac{5}{3}\cdot\frac{1}{3} = \frac{5}{9}$, while the minimum of the right-hand side of (8.5) will be decreased by $1-\frac{2}{3} = \frac{1}{3}$. As $\frac{1}{3} < \frac{5}{9}$, we see that the inequality is preserved in the $r = 1$ and $r = 2$ cases as well. This then completes the proof of the bound.

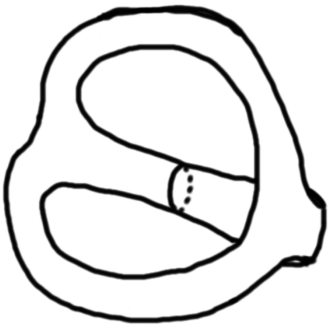
We can thus conclude that for all genus 0 orbifolds, we must have $a \leq 3$. Since this inequality is true for any a which satisfies $n \geq a(g-1)$, we finally obtain

$$n \leq 3(g-1),$$

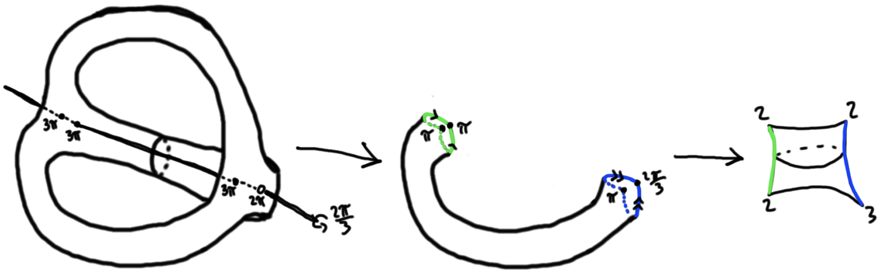
for any flexible hyperbolic cone metric φ on S. Equivalently, if a hyperbolic cone metric has more than $3(g-1)$ cone points, it must be rigid.

8.3 Achieving the Bound

In order to demonstrate that the bound above is actually the best we can do, we will construct an infinite family of cone surfaces which achieve the bound. Before this though, we will consider a simple example with a genus 2 surface due to a fellow student, Katherine Chui. Arrange your genus 2 surface so that it has order 3 rotational symmetry:

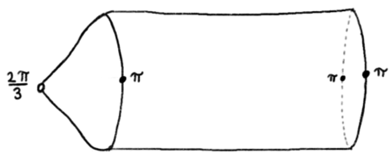


The axis of this order 3 rotation has 4 fixed points, and so when we quotient by its action we will obtain a genus 0 orbifold with 4 orbifold points. We can make 3 of the above fixed points be cone points with cone angle 3π and the other a regular point, so that they cover a $(0; 2, 2, 2, 3)$ orbifold:



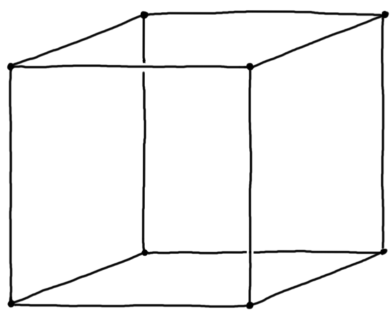
Note that $(0; 2, 2, 2, 3)$ was one of the two orbifold signatures for which it was theoretically possible to achieve the $3(g-1)$ bound. Since the number of cone points is $3 = 3(2-1) = 3(g-1)$, our bound is reached by this example. To see how to obtain more examples like this, imagine taking the surface above and collapsing it to a graph, like so:

Then our order 3 symmetry for the original surface corresponds to an automorphism of this graph which fixes a desired vertex. This suggests that we can devise examples by first considering graphs that have appropriate symmetries, and then inflating them into surfaces with the same symmetries. To understand what kind of graphs we would like, we will first look at the $(0;2,2,2,3)$ orbifold. We can represent it like a cannoli:



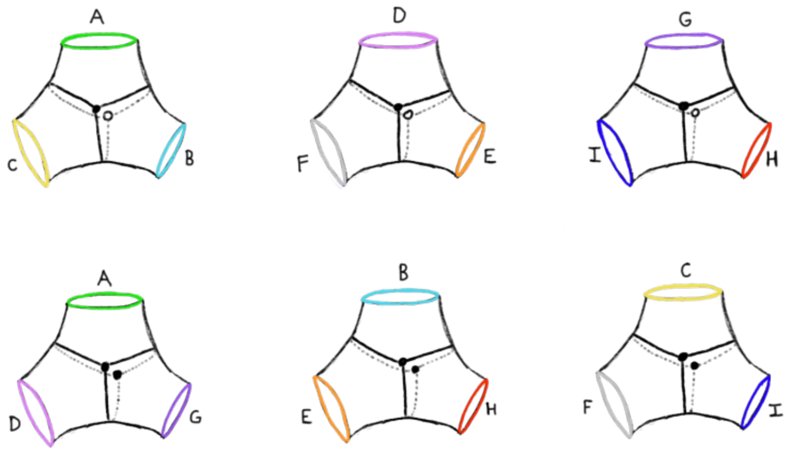
Just as it did in the example above, each copy of this orbifold making up the surface S can be likened to an edge of a graph. Each cone point on S must map to an order 2 orbifold point, and the optimal way to do this is with each cone point mapping to it with a local degree of 3. This means that each end of the orbifold above must be copied 3 times around the orbifold points to form part of our surface S. In terms of our graph though, this means that each vertex should have valence 3, or in other words that our graph is 3-regular. Our graph should also be symmetric with respect to cycling around the three neighbors of any vertex while keeping that vertex fixed, which corresponds to rotating our surface around 2 of these cone points. This action should also be edge transitive so that a single edge forms a fundamental domain for it, which corresponds to our surface covering the given orbifold. Finally, note that the set of vertices can partitioned into two sets: those that will contain 2 cone points when inflated into a surface, and those that will contain 1. And an edge cannot connect two vertices of like-nature, since each copy of our orbifold comprising the surface S has exactly 3 cone points. So to facilitate this we will also require that our graphs be bipartite.

An example that satisfies these properties is the graph of a cube:



Each vertex has valence 3 and degree 3 rotational symmetry. Moreover the action generated by these rotations is edge-transitive. It is also a bipartite graph. One can then check that the surface resulting from inflating this graph covers the orbifold. Moreover, it will have genus 5 and contain $2 \cdot 4 + 1 \cdot 4 = 3 \cdot (5 - 1) = 3(g - 1)$ cone points, as desired.

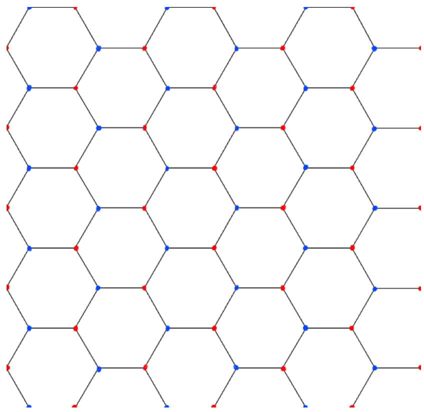
Another example is that of the graph $K_{3,3}$. As a surface, it can be represented as a collection of 3-punctured spheres with boundary components identified according to edge relations of the graph,



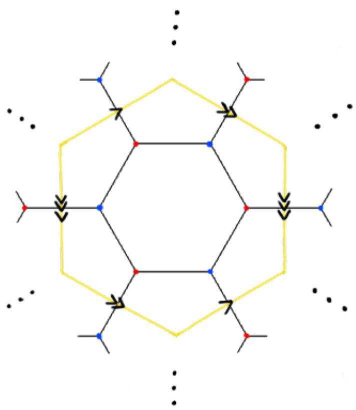
where cone points with cone angle 3π have been marked with filled circles, and regular fixed points are marked with open circles. Note though that the picture above is lying slightly, as each pair of pants containing two cone points actually has area 0 by the Gauss-Bonnet formula. Thus the above picture is purely topological, and the boundary components are not actually geodesics.

Nevertheless, this picture should make clearer how the symmetries of the graph can be realized as hyperbolic isometries of the surface. By representing each vertex as a thrice-punctured sphere and identifying boundary components according to edge relations of the graph, we essentially form a pants decomposition of the surface S. Each pair of pants (of positive area) can then be given a hyperbolic metric, with the symmetries from our graph acting on each pair of pants either by rotation if it is a fixed point, or translation if it is moved to another pair of pants. It follows then that for these examples, vertices in the same orbit must correspond to isometric pairs of pants in the original surface.

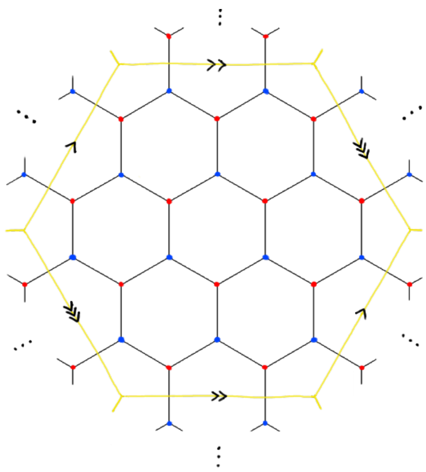
The previous example can also be generalized to an infinite family of surfaces which achieve the bound. To construct these, consider the graph which forms a hexagonal grid in the plane:



Other than the fact that it is infinite, this graph possesses all the properties we desire: degree 3 rotational symmetry about each vertex, edge-transitivity with respect to this action, and bipartiteness. To transform this graph into a finite one, we can isolate a hexagonal domain and perform a quotient. The smallest example of this looks like:



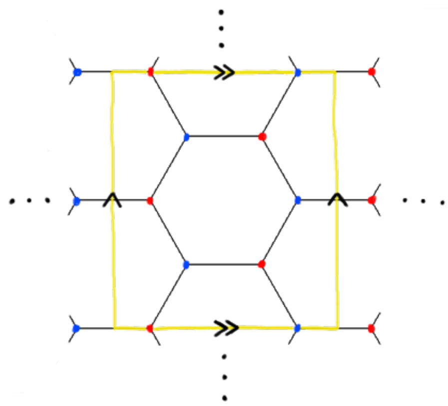
This is in fact just the $K_{3,3}$ example already mentioned. But now we can go further. The next smallest example of this construction is:



which has 24 vertices. In general, the k-th smallest example will have $6k^2$ vertices. When these vertices are inflated to form the surface, each will represent a pair of pants. A genus g

surface can be decomposed into $2(g-1)$ pairs of pants, so $6k^2 = 2(g-1)$ implies that the genus of the surfaces formed by this construction are of the form $3k^2+1$. So this construction yields surfaces of genus $4, 13, 28, 49$, etc.

We should be certain to confirm that these graphs have the desired symmetry, though. As a cautionary example, the cubic graph from before can be realized as a quotient of the hexagonal grid as well, in its case by a rectangular region:



However, one can check that quotients by larger rectangular regions do not yield graphs with the desired automorphisms.

To be sure that the graphs we have constructed actually work, we will need to check that the rotations that generate the group of automorphisms respect the quotient by the hexagonal domain. Let H be the larger hexagonal grid (not the hexagonal graph we are interested in), and let T be the group of translations which preserve H. Then the group T we are quotienting by can be identified with the additive group of eisenstein integers in the complex plane. These are complex numbers of the form $a+b\omega$ with a and b being integers and $\omega = e^{\frac{2\pi}{3}i}$. Two points of the plane, envisioned as complex numbers z and w, are equivalent if and only if $z-w$ is one of these eisenstein integers. Moreover, a degree 3 rotation about a point z_0 can be written algebraically as $r(z) = \omega(z-z_0)+z_0$. Confirming that the rotations of our graph respect the quotient is then simply a matter of checking that $z \sim w$ if and only if $r(z) \sim r(w)$. We first note that

$$r(z) - r(w) = \omega(z-z_0) + z_0 - [\omega(w-z_0)+z_0] = \omega(z-w).$$

Since the eisenstein integers $\mathbb{Z}[\omega]$ are preserved under multiplication by ω, this implies what we want.

So we now have in our hands an infinite family of cone surfaces which achieve our proven bound. This is certainly sufficient to demonstrate tightness, but only in a weak sense. The geni that we have examples for at this point are only quadratically distributed, and it is not clear that in the gaps between the actually bound won't dip even lower.

This is not likely to be the case though, since there is another somewhat well-studied family of examples. These are given by the bipartite symmetric cubic graphs, which are bipartite graphs such that each vertex has three neighbors, and such that the group of automorphisms acts transitively on the *ordered* pairs of adjacent vertices in the graph. We then clearly have bipartiteness, 3-regularity, and edge transitivity satisfied by these examples. To ensure that rotations around vertices are valid automorphisms of such a graph, we appeal to a theorem of Tutte's which asserts that the stabilizer subgroup of any vertex v in a symmetric cubic graph is isomorphic to one of C_3, S_3, $S_3 \times C_2$, S_4, or $S_4 \times C_2$ [15]. Each of these groups contains at least one order 3 element, which must permute the three neighbors of v. The only way this can be done is if these neighbors are permuted cyclically.

So we can see that each such bipartite symmetric cubic graph will provide a valid cone surface which achieves our bound. Moreover, Marston Conder has computed all symmetric cubic (though not necessarily bipartite) graphs with up to 10,000 vertices [16]. Below, we will list the geni that we obtain examples for from this family of graphs, up to genus 100:

$$\begin{aligned}
&4, 5, 8, 9, 10, 11, 13, 14, 16, 17, 20, 21,\\
&22, 25, 26, 28, 29, 32, 33, 37, 38, 40, 41,\\
&44, 46, 49, 50, 53, 56, 57, 58, 61, 62, 64,\\
&65, 68, 73, 74, 76, 77, 80, 82, 85, 92, 94,\\
&97, 98
\end{aligned}$$

In total, we found based on this data that roughly 30% of all geni up to $g = 5,000$ achieve the bound.

Chapter 9

Conclusion

Let us take stock of all that we have seen during the course of this discussion. We have met hyperbolic geometry and its many guises, and analyzed its many non-intuitive properties. Covering spaces were next, revealing deep connections between topological spaces and the algebraic objects we associate to them, namely the fundamental group. Quite beautifully, these two realms of topology and non-euclidean geometry conspired seamlessly to allow us to construct and even parameterize all non-euclidean surfaces. Then, seeking a generalization of the group actions we associated with covering spaces, we introduced ourselves to the notion of orbifolds, which play a prominent role in the CST and in topology/geometry more broadly. We finally learned of geodesic currents, which offer a tidy way of packaging subtle geometric properties of surfaces inside the language of measures and analysis.

At every turn, we have seen deep and unexpected connections between disparate areas of study, from geometry, to topology, to algebra, to analysis. The fact that the CST and other theorems of its kind stand at the nexus of so many insights and nearly two centuries of hard work makes them worthy of study all by itself. Of course, more than just showcasing a wonderful edifice of mathematical theory, theorems like the CST represent significant contributions to it and are testaments to the unexpected rigidity of these hyperbolic objects that geometers have come to study.

The task now is to continue probing this rigidity, shedding light on where it is strong and where it is weak, and in what ways it might present itself. Whether that be in the finite-dimensionality of the Teichmüller space or the support of the Liouville current. For our own part, we have been able to contribute a small piece to this vast puzzle and demonstrate that far fewer cone points than originally shown are sufficient to ensure rigidity of metrics in the sense of the CST.

9.1 Further Directions

Though it took quite a bit of work, our success in reducing the bound on the number of cone points and proving a somewhat weak notion of tightness is still only a small component of the larger project of probing the limits and rigidity of hyperbolic metrics. For those who wish to carry on this task in the future, we highlight here some questions that have been brought up in, but gone unanswered by the present work.

Question 9.1.1. *Are there infinite families of graphs which yield cone surfaces that achieve the bound found herein, such that the geni of these surfaces are linearly distributed?*

Question 9.1.2. *D examples of cone surfaces exist which cover the* $(0; 2, 2, 2, 4)$ *orbifold and achieve the bound on cone point count? And if so, do these examples fill in the gaps in geni left by the examples found here covering the* $(0; 2, 2, 2, 3)$ *orbifold?*

Question 9.1.3. *The CST states that the deformation space of a flexible hyperbolic cone surface* (S, φ) *is parameterized by the Teichmüller space of the orbifold it covers. One can then ask if there are generalizations of the result presented here. Such as whether there exists some constant* $a \in \mathbb{R}$ *for which a surface with more than* $a(g-1)$ *cone points can have at most a 1-dimensional deformation space? Or more generally, given* $n \in \mathbb{N}$*, does there exist a constant* a_n *such that* (S, φ) *possessing more than* $a_n(g-1)$ *cone points implies its deformation space is at most* n*-dimensional? Do these constants approach 0 as* n *grows large?*

Question 9.1.4. *Is there a reason that the* $3(g-1)$ *bound on the number of cone points is equal to the number of closed loops needed to separate the surface into pairs of pants?*

Appendix A

Supplementary Proofs

A.1 Tangents to Hyperboloid Have Positive Quadratic Form

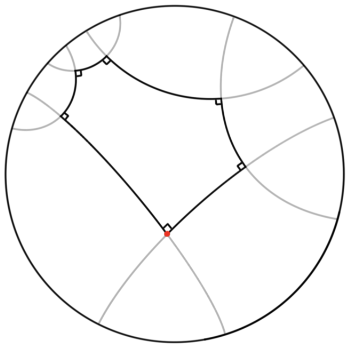
Any tangent to the hyperboloid will in particular be a tangent to a curve that runs along it. We will choose this curve to be a geodesic path whose formula will be $\sinh(t)\vec{w}+\cosh(t)\vec{v}$, where $Q(\vec{v},\vec{v}) = -1$, $Q(\vec{w},\vec{w}) = 1$, and $Q(\vec{v},\vec{w}) = 0$. A tangent vector to this curve will be given by $\cosh(t)\vec{w} + \sinh(t)\vec{v}$, and so we can compute its quadratic form to be.

$$\begin{aligned}
&Q(\cosh(t)\vec{w} + \sinh(t)\vec{v}, \cosh(t)\vec{w} + \sinh(t)\vec{v}) \\
&= \cosh(t)^2 Q(\vec{w},\vec{w}) + 2\cosh(t)\sinh(t)Q(\vec{v},\vec{w}) + \sinh(t)^2 Q(\vec{v},\vec{v}) \\
&= \cosh(t)^2 - \sinh(t)^2 = 1
\end{aligned}$$

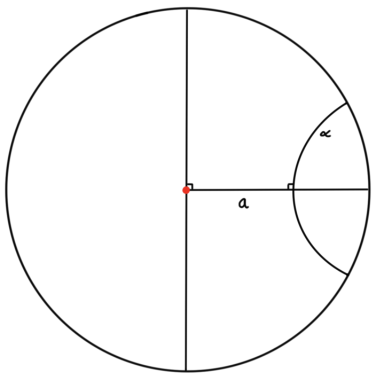
Which is positive. If we scale this tangent vector by λ, then its quadratic form will scale by λ^2 and so will still be positive. This proves the result.

A.2 3 Non-Adjacent Sides Determine a Right-Angled Hyperbolic Hexagon

For this proof, we refer to lecture notes written by Bram Petri [17]. We would like to show that given a triple of positive real values (a, b, c), there exists a unique (up to orientation-preserving isometry) right-angled hyperbolic hexagon such that three non-adjacent sides have length a, b, and c respectively:



Since we are only considering this hexagon up to isometry, we are allowed to choose one of our vertices (e.g. the marked red vertex above) to be the center of the hyperbolic disk. The two sides adjacent to this vertex can then be extended to geodesic rays. Furthermore, we know that a side adjacent to this vertex has length a, so we may also draw the geodesic ray which is at distance a from our chosen vertex and which intersects the corresponding side at right-angles. Call this curve α. Our picture so far thus looks like:



We next consider the side of length c. We know this side will be at right angles with the vertical geodesic we have drawn above, but where it meets this vertical geodesic we do not yet know. So we will simply draw the curve of all points at distance c from the vertical geodesic. Call this curve B:

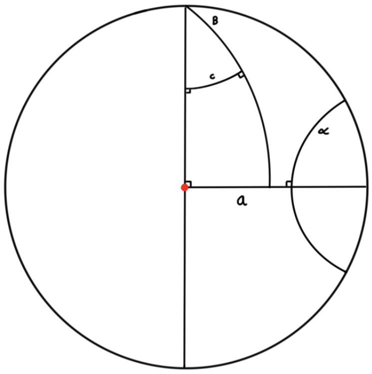
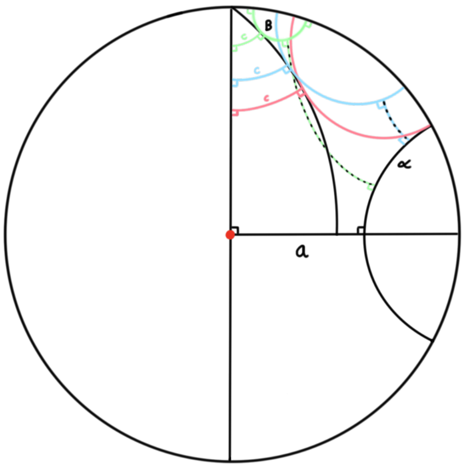
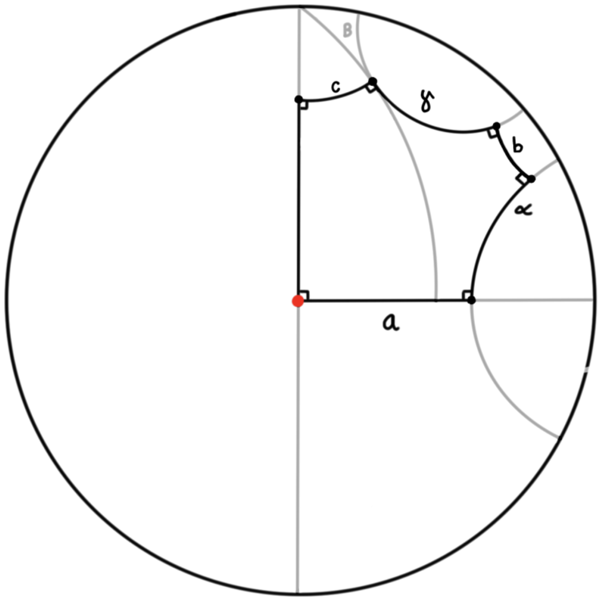


Fig. A.1.1: B is the curve of points a distance c from the vertical geodesic.

Let p be the point where our segment of length c meets B. Note that as we vary the point p on B, the geodesic which lies at right angles to it can be arbitrarily close or far from the geodesic α:



The red geodesic above, for instance, is at distance 0 from α, the blue geodesic is at some greater distance, and the green geodesic is at a greater distance still. Because this distance increases monotonically and continuously as we move along B, there must be some point at which this distance is precisely b. Let γ be the geodesic perpendicular at p which achieves this distance. Then the picture we have produced is:



Thus from our triple of real numbers a, b, and c, we have constructed a right-angled hexagon having these as the lengths of 3 non-adjacent sides. This deals with the existence part of our claim.

For uniqueness, note that up to isometry the picture in fig. A.1.1 above is completely determined by the values of a and c. And since the distance between γ and α increases continuously and strictly monotonically as we move along B, there is exacty one such geodesics γ that can achieve a distance of b from α. Hence, all of the geodesics forming the sides of our hexagon are determined up to isometry by a, b, and c, which gives uniqueness.

A.3 Intersection of Finite Index Subgroups has Finite Index

Suppose that H and K are finite index subgroups of G. Let $I = H \cap K$. We will first consider the index of I in H, $[H : I]$. If this index were infinite, then there would be infinitely many cosets of I in H, say $\{h_1 I, h_2 I, \ldots\}$. If we then consider the (a priori, not distinct) cosets $\{h_1 K, h_2 K \ldots\}$, we can note that $h_1 K \cap h_2 K \neq \emptyset$ implies that there exists $k_1, k_2 \in K$ such that $h_1 k_1 = h_2 k_2$. Or equivalently, $h_2^{-1} h_1 = k_2 k_1^{-1}$ which must then be in $I = H \cap K$. Thus it follows that $h_1 \in h_2 I$, which contradicts the fact that the h_i's were chosen so that they were each in different cosets of I. We are then forced to conclude that each coset $h_i K$ is disjoint from the rest. But this means there are infinitely many cosets of K in G, which contradicts with the fact it has finite index.

Therefore I has finite index in H. It follows that if the cosets of I in H are $\{h_1 I, \ldots, h_n I\}$ and the cosets of H in G are $\{g_1H, \ldots, g_mH\}$, then the cosets of I in G are $\{g_1h_1I, g_1h_2I, \ldots, g_mh_{n-1}I, g_mh_nI\}$. That is, $[G : I] = nm$ which is finite, as desired.

Bibliography

[1] J.-P. Otal, “Le spectre marqué des longueurs des surfaces á courbure négative,” *Ann. Math.*, vol. 131, pp. 151–162, 1990.

[2] V. Erlandsson, C. J. Leininger, and C. Sadanand, “Hyperbolic cone metrics and billiards,” *Advances in Mathematics*, vol. 409, p. 108662, 2022.

[3]

[4]

[5] B. Iversen, *Hyperbolic Geometry*. 1992.

[6] A. Hatcher, *Algebraic Topology*. 2001.

[7] J. Aramayona and C. J. Leininger, “Hyperbolic structures on surfaces and geodesic currents,” 2014.

[8] B. Farb and D. Margalit, *A Primer on Mapping Class Groups*. 2012.

[9] W. Thurston, *Geometry and Topology of 3-Manifolds*. 2002.

[10] F. C. J. Caramello, “Introduction to orbifolds,” 2022.

[11] M. Clay and D. Margalit, *Office Hours with a Geometric Group Theorist*. 2017.

[12] F. Bonahon, “The geometry of teichmüller space via geodesic currents,” *Invent. math.*, vol. 92, pp. 139–162, 1988.

[13] F. Bonahon, “Bouts des variétés hyperbolic de dimension 3,” *Ann. Math.*, vol. 124, pp. 71–158, 1986.

[14] J. Frazier, “Length spectral rigidity of non-positively curved surfaces,” 2012.

[15] M. Conder, A. Hujdurović, and K. Kutnar, “Symmetric cubic graphs via rigid cells,” *Journal of Algebraic Combinatorics*, vol. 53, pp. 881–895, 2020.

[16] M. Conder, “Trivalent (cubic) symmetric graphs on up to 10000 vertices.”

[17] B. Petri, “Random methods in geometry,” 2017.