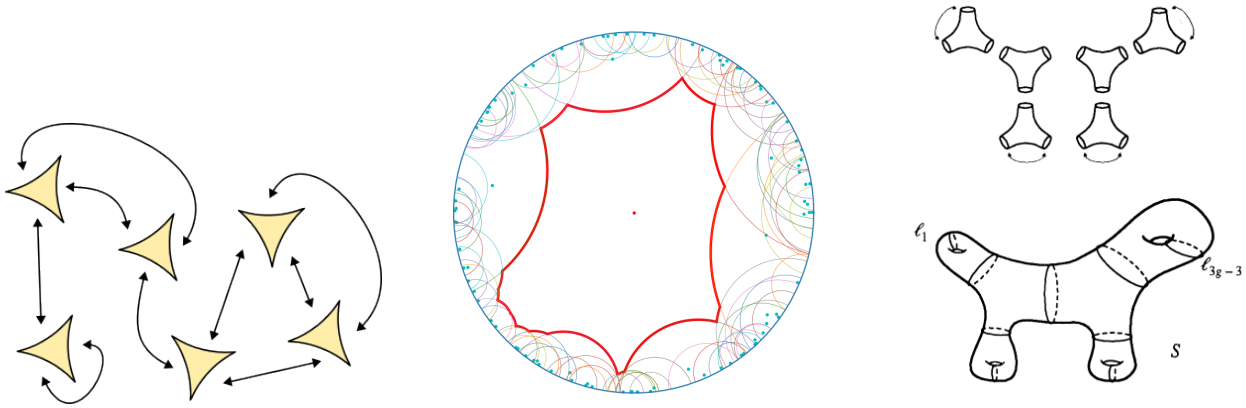
Random surfaces and probability in hyperbolic geometry

Renan Gross

Lecture notes for Cambridge graduate course, Lent 2026



[1]

About this course

Hyperbolic surfaces are a beautiful topic in modern mathematics. Analyzing them requires a mixture of geometry, algebra, combinatorics, probability, and (complex) analysis. There are even connections to number theory(!), and so their study effectively covers large portions the mathematical landscape.

Many problems can be approached from different angles; in this course we will focus mostly on the probabilistic and combinatorial aspects. The course is aimed at students of probability, but does not really require deep knowledge in either probability or geometry (provided that you are willing to accept as fact some of the basic theorems). I hope that students from all fields can enjoy it.

The course is aimed to be easy, and will cover mostly the basics of probability in hyperbolic surfaces, omitting some of the newer and harder (and exciting!) developments (several fundamental problems were solved just in 2025). We will leave some of the more technical aspects as rough sketches, and focus on the more concrete examples. Various exercises, interwoven through these lecture notes, supplement the material and provide some additional hands-on practice.

✧——————————☞

[1] Image taken from *Geometry and Spectra of Compact Riemann Surfaces*, by Buser

Contents

1 Introduction (Lecture 1)

This course will be about randomness in and of hyperbolic geometry. Our main object of study will be the **compact hyperbolic surface**. To begin with, let's talk a bit about what those are, why it is nice to talk about them, and why we would want randomness to be involved at all.

1.1 Hyperbolic surfaces

A hyperbolic surface is an orientable Riemann surface of constant curvature -1. Do not worry if you do not remember exactly the definitions and all theorems in Riemannian geometry; while we will use some of them in this course, our approach will be mostly hands-on, and we will introduce the theorems where we need them. In any case, here is a very quick, very informal recap of what curvature means in differential geometry: suppose we have an oriented surface in $\mathbb{R}^3$, like this torus:

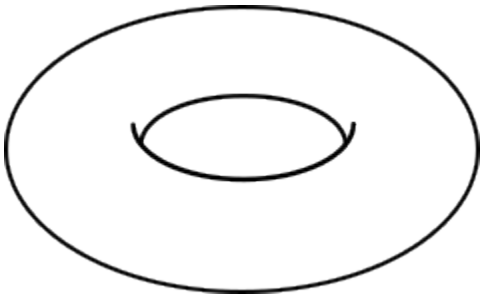


Figure 1.1: The natural "bagel" embedding of a torus in $\mathbb{R}^3$.

When drawn like this, the torus has all three types of curvature: positive, negative and 0. At a point x on the outside rim of the torus, if we look at geodesics emanating from x, both curve in the same direction - there, the curvature is positive. On the top on the torus, we have one geodesic curving down, while the other is "flat" - here there is 0 curvature. And on the inside, the curvature is negative - the geodesics are "curved" in different directions. (This is made precise in differential geometry by looking at the curvature at x of the curve obtained by intersecting planes which pass through a point and contain a tangent vector, calculating the maximum κ_1 and minimum κ_2 among all such planes, and writing $K = \kappa_1\kappa_2$). It turns out that this intuitive geometric interpretation is a fundamental property of the Riemannian surface - curvature is something which can be defined just from the intrinsic distances and angles of the surface, even if we do not embed it in $\mathbb{R}^3$.

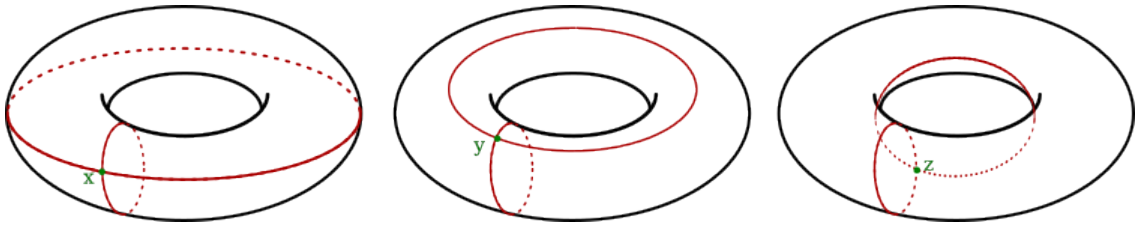


Figure 1.2: In the "bagel" embedding, a torus has points of positive, zero, and negative curvature.

So the torus we have drawn here is definitely not the object of study of this course - it has varying

curvature. What are the constant curvature surfaces?

- For positive curvature, there is only the sphere.
- For curvature 0, there is $\mathbb{R}^2$, which is simply connected, as well as the (flat) torus (which we think of as $\mathbb{R}^2 \mod \mathbb{Z}^2$) and the cylinder (which we think of as $\mathbb{R}^2 \mod \mathbb{Z}$).
- As for curvature -1... that is still a mystery.

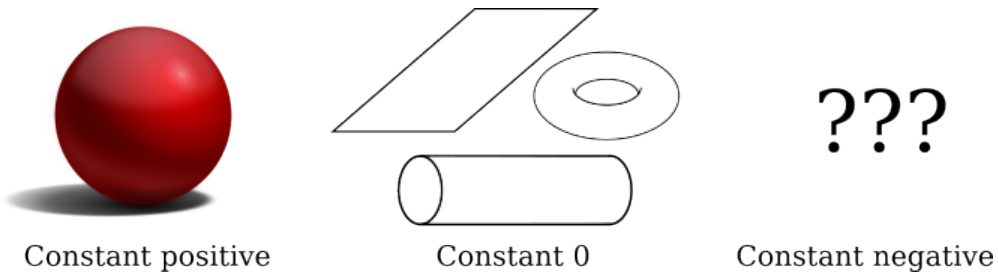


Figure 1.3: At this point in the course, the constant negative curvature surfaces are still a mystery.

There is indeed a infinite, simply connected surface without boundary, called the hyperbolic plane, $\mathbb{H}$. There are also infinitely many compact hyperbolic surfaces, each with finite area, and infinitely many non-compact hyperbolic surfaces, some with finite area and some without. But I cannot draw them here; a theorem by Hilbert says that it is impossible to smoothly embed a complete surface of negative constant curvature in $\mathbb{R}^3$.

Nonetheless, we will draw compact hyperbolic surfaces many many times, like this:

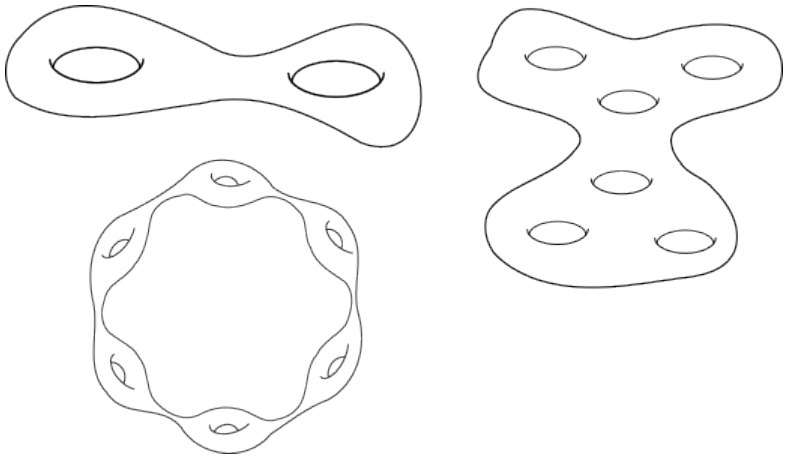


Figure 1.4: Cartoons of compact hyperbolic surfaces.

with the understanding that this is just a cartoon, and that in fact everything has curvature -1 everywhere. All our compact hyperbolic surfaces will look roughly like the above: blobs connected together by small "bridges". The classification of compact surfaces states that this is all there is. However, we are dealing with geometry, not topology. It will turn out that there is a very strict relation between the length of these bridges and their width.

1.1.1 Why we care

There are several reasons we should care about hyperbolic surfaces.

1. Hyperbolic surfaces are the graphs of the continuous world. We will see that surfaces like those drawn above can be decomposed into elementary building blocks, and each elementary building block has very rigid geometry. In a nutshell: if two "blobs" are connected together by a very long "collar", then that collar is:

   (a) very thin in the middle;

   (b) expands exponentially until it has about constant width.

   This makes the blobs act like vertices, and the collars connecting them like edges. But this analogy is very rough and is only valid to a certain extent; essentially, compact hyperbolic surfaces may have the deep structure found in graph theory, and then some. Similarly to how we study the class of "graphs on n vertices", we often clump together and study the class of hyperbolic surfaces with the same number of "blobs", or, more precisely, surfaces of a given genus g (see below for more on this).

2. **The Gauss-Bonnet** theorem states that for a compact Riemannian surface S without boundary,

   $$\int_S K(x)\, dA = 2\pi\chi(S),$$

   where $K(x)$ is the curvature function and χ is the Euler characteristic. The Euler characteristic is a topological property: if the surface S is topologically equivalent to a sphere plus g handles (i.e. a g-holed torus), then

   $$\chi(S) = 2 - 2g.$$

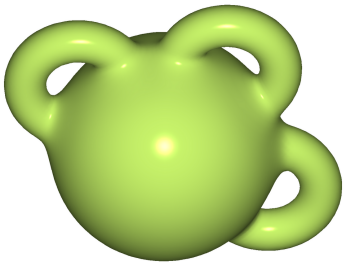


Figure 1.5: A sphere with three handles. Image from Wikipedia by Oleg Alexandrov.

Thus, $\chi(S) = 2$ for the sphere, $\chi(S) = 0$ for a torus, and $\chi(S)$ is negative for any other compact surface. So the moment we are interested in any surface apart from the sphere or torus, that surface must have negative curvature somewhere. Hyperbolic surfaces are then the "simplest" such surfaces, since their curvature is constant -1. The choice of constant -1 doesn't really matter here, and we could have given our course for a general "negative curvature $-K$", at the cost of adding an encumbering and bothersome "K" in half the places. The restriction $K = -1$ is of course severe, and imposes very strong structure on the surfaces, allowing some precise results. However, there is still a lot of depth and richness in these surfaces, as I hope we will see through this course. If this is too restrictive for you, the "next level" of generality, which you can find in many places in the literature, is to talk about surfaces with bounded negative curvature, $-c_1 < K < -c_2$ for some constants c_1, c_2.

This then branches to either $K < 0$ or $|K| < C$ (allowing positive curvature adds a whole new can of worms to the ordeal; the behaviour of a surface at negatively curved and positively curved points is very different).

3. Let S be a compact Riemannian surface of genus greater than 2 with metric g (not necessarily hyperbolic). Then there exists a function $f : S \to (0, \infty)$ such that the metric $f(x)\, g$ is hyperbolic. In other words: every compact surface is conformally equivalent to a hyperbolic surface. Hyperbolic surfaces can be seen as the "simplest" representatives of the equivalence class of conformally equivalent metrics. We will not prove this fact here, but here is a sketch: the surface S has simply connected universal cover $\tilde{S}$. We all know and love the Riemann mapping theorem, which says that a simply connected $U \subseteq \mathbb{C}$ is conformally equivalent to the unit disk. A powerful generalization is the **uniformization theorem**, which states that every simply connected Riemann surface is conformally equivalent to either the unit disk, the complex plane, or the Riemann sphere. When the genus of S is ≥ 2, it can be shown that the cover must be conformally equivalent to the unit disk, which, as we will see, is a model for hyperbolic geometry. This means that under a conformal change of metric, the cover of S can be made to have hyperbolic geometry, and S inherits this metric from the cover.

1.1.2 Interesting quantities

What are the things that people like to look at when studying surface?

1. **The Cheeger constant**. How hard is it to divide the surface into two parts, so that the boundary between the two parts is small? This is a fundamental question related to mixing time, isoperimetry, and various functional inequalities. The Cheeger constant is defined as

$$h(S) = \inf_{\mathrm{Vol}(A) \leq \frac{1}{2}\mathrm{Vol}(S)} \frac{|\partial A|}{\mathrm{Vol}(A)}.$$

Many questions can be asked: given a surface S, what geometric properties give bounds on $h(S)$? What is the largest / smallest $h(S)$ possible as S varies over all surfaces?

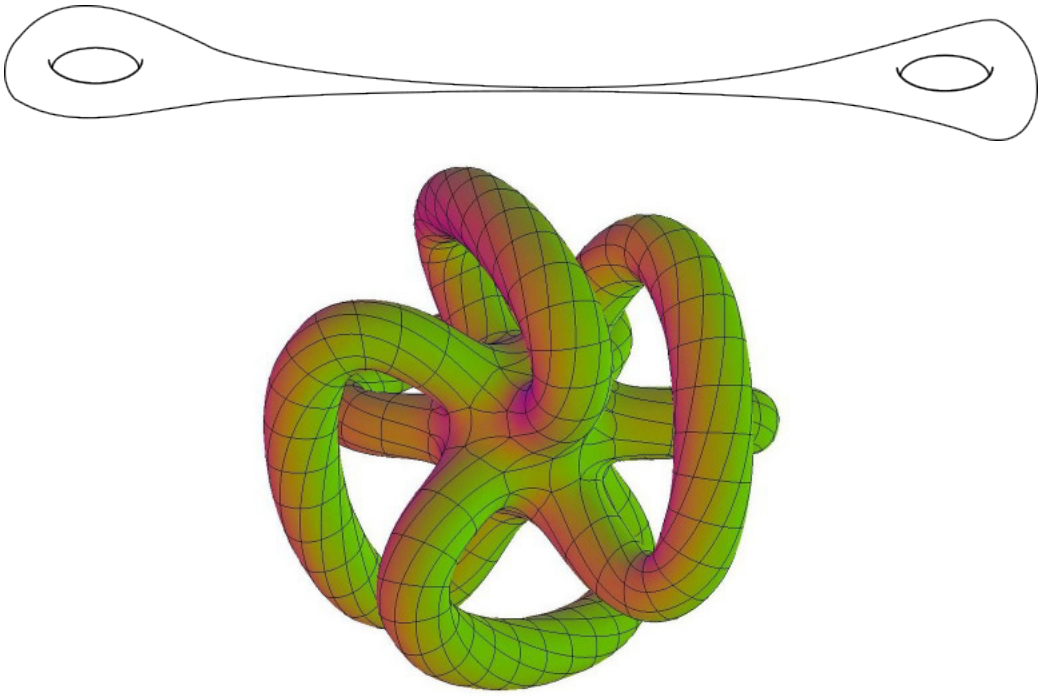


Figure 1.6: An easy-to-cut surface (with small Cheeger constant), and a perhaps-not-so-easy-to-cut surface (with perhaps a larger Cheeger constant).

2. **The spectral gap**. it is possible to define a Laplacian operator on functions on S. For compact surfaces, this operator has eigenfunctions $f_0, f_1, \ldots$ with eigenvalues $\lambda_0 \leq \lambda_1 \leq \ldots$. Just like in graphs, the smallest eigenvalue is $\lambda_0 = 0$, and corresponds on the trivial eigenfunction $f_0 = \text{const}$. The second eigenvalue is therefore known as the *spectral gap*. The Buser-Cheeger inequality states that

$$\frac{h(S)^2}{4} \leq \lambda_1(S) \leq 2h(S) + 10h^2(S).$$

   But in fact the spectrum of hyperbolic surfaces can be studied in greater detail. For example, as the genus $g \to \infty$, $\lambda_1 \leq \frac{1}{4}$ asymptotically. This is reminiscent of a d-regular graph, where the spectral gap is bounded from above by $2\sqrt{d-1}$. Do "most" hyperbolic surfaces have gap close to the maximum $1/4$, or is it obtained only by a lucky few?

3. **Systole**. Compact hyperbolic surfaces are not simply connected. The systole of a surface S is defined as the length of its shortest closed geodesic. If you believed the description of surfaces as "blobs connected by collars, with the collars getting narrower as they get longer", then you should also believe that we can find surfaces of arbitrarily small systole. What about large ones? How large can the smallest closed geodesic be?

4. **Injectivity radius**. This is related to the systole. For a point $x \in S$, the injectivity is

$$\text{Inj}(x) = \sup\{r \geq 0 \mid B(x,r) \text{ is simply connected}\}.$$

   This can be different for different points: points on collars have small injectivity radius, while other points might have very large injectivity radius. For points x on the shortest closed geodesic itself, $2\text{Inj}(x) = \text{Sys}(S)$.

5. **Diameter**. The diameter of a compact hyperbolic surface of a given genus g can be made arbitrarily large. Can it also be made arbitrarily small?

1.2 Randomness

In this course we will use randomness both in and of hyperbolic surfaces. There are two reasons to do this.

First, probability for probability's sake. The hyperbolic plane and other hyperbolic surfaces are just as legitimate a breeding ground for probabilistic questions as more classical, discrete / Euclidean models. How do Brownian motion, point processes, percolation behave in the hyperbolic plane? If hyperbolic surfaces are like graphs, are there models of random surfaces like $G(n,p)$? What can we say about them?

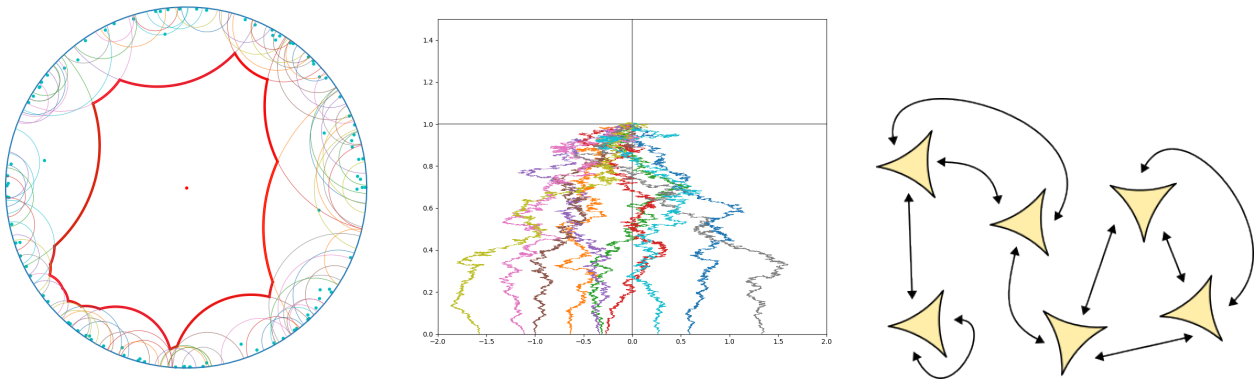


Figure 1.7: Randomness in hyperbolic surfaces: a Poisson point process in the Poincaré disk, Brownian motions in the half-plane, and gluing together hyperbolic triangles.

Second, we can use probabilistic tools to say something meaningful about the objects themselves. This is in the same spirit as the probabilistic method and the various results obtained by it for graphs. For example:

1. Random objects gives us constructions which are hard to find deterministically. For example, by generating a random hyperbolic surface with suitable parameters, it is possible to give a tight lower bound on the smallest diameter among all surfaces; no explicit construction is known in this case.
2. Random processes can be used to prove functional inequalities (this is true in general, not just in hyperbolic geometry).
3. In the case of generating objects at random, if we find good models which we can analyze, we can in some sense get an idea of what a "typical" object looks like.

1.2.1 Popular models of random surfaces

There are several prominent models for random surfaces that people like to study. Here is a very brief overview of the most famous 3:

1. **The Brooks-Makover model** (named after Robert Brooks and Eran Makover): Start with n hyperbolic triangles and connect them together randomly. This results in a hyperbolic surface with some singularities. These can be removed (with some effort). This model is similar to the configuration model in the theory of random graphs, and indeed we will use results from the latter in its analysis. This model will receive the most focus in this course.

2. **The Weil-Petersson model** (named after André Weil and Hans Petersson): We will see that for a fixed $g \geq 2$, surfaces of genus g can be described by a family of $6g-6$ parameters. The Weil-Petersson model gives a "natural" distribution over these parameters, which gives a sort of "uniform distribution" over the space of surfaces of genus g. This model is similar to choosing a uniformly random graph, though perhaps it should be thought of more as a "smart distribution over weighted graphs" due to the fact that the $6g-6$ parameters are continuous, as well as the behavior of the surface as a function of these parameters.

3. **The random cover model** (named after Sir Random Cover): For a fixed integer $k > 0$ and surface S, we may consider the set of all surfaces $\tilde{S}$ which are a k-cover of S (that is, there is a k-to-1 map $f : \tilde{S} \to S$ so that for every $x \in S$ there is a small neighborhood U of x such that $f^{-1}(U)$ is a disjoint union of k neighborhoods). This is a finite set, and we can choose uniformly from all of them. This model is inspired by the "random cover" model for graphs.

These are not the only models: for example, you can find in the literature various random "pairs of pants decompositions", or random polygon gluings. These are often ad-hoc for a specific task, or less well studied.

1.3 Notes about visualization

1. Hilbert's theorem indeed prohibits us from smoothly embedding a hyperbolic surface into $\mathbb{R}^3$. However, Nash's embedding theorem says that any surface of dimension d can be isometrically embedded in dimension $d+1$ (differentiably, but not smoothly). If you are fractal-keen, take a look at the images and papers "visualizing" such an embedding (and others) at the Hevea project.

2. One conceptual reason as to why the hyperbolic plane cannot be embedded in $\mathbb{R}^3$ is that there is not enough space for it. The hyperbolic plane is much larger than $\mathbb{R}^2$, and its geometry can be unintuitive at times. If you want to get a feel for it yourself, take a look at the computer game HyperRogue, in which you obtain treasure and evade monsters in the hyperbolic plane.

2 Hyperbolic geometry (Lecture 2)

In order to say anything about compact hyperbolic surfaces, we must understand the basics of the hyperbolic plane $\mathbb{H}$. We will use a hands-on approach that lets us calculate lengths and areas.

2.1 Half-plane basics

We start with the half-plane model for the hyperbolic plane. In this model, $\mathbb{H}$ is a Riemannian surface, with the half-plane $H = \{x + iy \mid y > 0\}$ as a base set equipped with the metric

$$ds^2 = \frac{dx^2 + dy^2}{y^2},$$

i.e. the metric tensor is given by

$$g(x, y) = \begin{pmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{pmatrix}.$$

Since this is not a course in Riemannian geometry, here is what it means for us:

1. If γ is a curve in H, then its hyperbolic length given by

$$\ell(\gamma) = \int \frac{1}{y} |d\gamma| \quad \left(= \int_\gamma \frac{dz}{\operatorname{Im}(z)} \right)$$

   (or, more explicitly, if $\gamma = x(t) + iy(t)$ is parameterized by t, then

$$\ell(\gamma) = \int_{t_0}^{t_1} \sqrt{\frac{1}{y(t)^2} x'(t)^2 + \frac{1}{y(t)^2} y'(t)^2} dt = \int_{t_0}^{t_1} \frac{1}{y} |\dot{\gamma}(t)| \, dt.$$

   Sometimes we might write $|\gamma|$ for $\ell(\gamma)$.

2. If A is a set in H, then its hyperbolic area is given by

$$\operatorname{Vol}(A) = \int_A \frac{1}{y^2} dx dy \quad \left(= \int_A \frac{dz}{\operatorname{Im}(z)^2} \right).$$

   Sometimes we might write $|A|$ for $\operatorname{Vol}(A)$.

The distance between two points in the plane is given by the shortest way to get between them:

$$\operatorname{dist}(z_1, z_2) = \inf_\gamma \ell(\gamma),$$

where the infimum goes over all curves γ starting at z_1 and ending at z_2. A curve in H which minimizes the distances between any two of its points is called a *geodesic*.

It is not always obvious what the distance should be between two points. Consider even the simple case of $z_1 = i$ and $z_2 = 1 + i$.

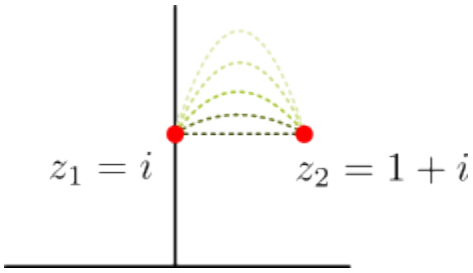


Figure 2.1: Which of the curves (if any) minimizes the length?

Drawing the horizontal line between them, the distance between them is at most 1. But look at the metric: we multiply distances by $\frac{1}{y}$, so a curve might benefit from going upwards a little, moving THERE to the right, and then coming back down. It is possible to find the optimal curve between z_1 and z_2 using some calculus of variations. For general z_1 and z_2, this might be a bit involved, and we will not do it here; we'll have to wait with calculating the distance from i to $1 + i$. Rather, let's do something easier: let $z_1 = i$ and $z_2 = 2i$.

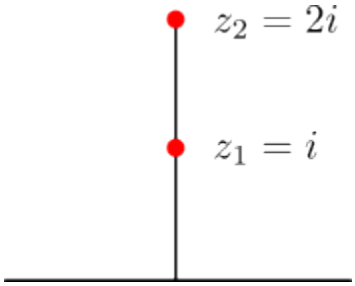


Figure 2.2: This case is much easier than the horizontal one.

The straight line from z_1 to z_2 gives

$$\operatorname{dist}(z_1, z_2) \leq \int_1^2 \frac{1}{y} dy = \log(2) - \log(1) = \log 2.$$

This is not only an upper bound, but also the actual distance: any curve γ connecting z_1 to z_2 must have the imaginary coordinate eventually cross from 1 to 2, and horizontal deviations are just wasting time. Thus, in general,

$$\operatorname{dist}(ai, bi) = \left| \log \frac{b}{a} \right|.$$

In fact, this is all we need in order to find out the distances between any two points. To see this, we need to talk about isometries of the hyperbolic plane.

Let $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ be the Riemann sphere. A Möbius transformation is a function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ of the form

$$f(z) = \frac{az + b}{cz + d},$$

where $a, b, c, d \in \mathbb{C}$. Möbius transformations are analytic functions, and map circles to circles (in the general, Riemann sphere sense - so circles can be mapped to lines).

We will be interested in the following important set of Möbius transformations:

$$\mathrm{SL}(2,\mathbb{R}) = \left\{ z \mapsto \frac{az+b}{cz+d} \mid a,b,c,d \in \mathbb{R}, ad-bc=1 \right\}.$$

Theorem 2.1. *Every $f \in \mathrm{SL}(2,\mathbb{R})$ is a hyperbolic isometry.*

Proof. Plug in any three real numbers to see that $f(\mathbb{R}) = \mathbb{R}$, and so f maps $\mathbb{H}$ to itself. To see that it is an isometry, let γ be any curve in $\mathbb{H}$. By definition,

$$\ell(f(\gamma)) = \int_\gamma \frac{|df(z)|}{\mathrm{Im}(f(z))}.$$

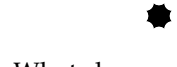
We just have to calculate the numerator and the denominator:

$$\frac{df}{dz} = \frac{a(cz+d) - c(az+b)}{(cz+d)^2} = \frac{ad-cb}{(cz+d)^2} = \frac{1}{(cz+d)^2}.$$

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

Thus

$$\ell(f(\gamma)) = \int_\gamma \frac{|df(z)|}{\mathrm{Im}(f(z))} = \int_\gamma \frac{dz}{\mathrm{Im}(z)} = \ell(\gamma).$$



Exercise 2.2. Show that if f is an orientation preserving isometry of $\mathbb{H}$, then $f \in \mathrm{SL}(2,\mathbb{R})$. What do you think are the non-orientation preserving isometries of $\mathbb{H}$?

Contrast this with the Euclidean situation in $\mathbb{R}^2$, where the orientation preserving isometries are rotations and translations (and their compositions). Playing around with and understanding the symmetries of a space can get you quite a great deal in life.

Proposition 2.3. *Let $z_1 = x_1 + iy_1, z_2 = x_2 + iy_2 \in \mathbb{H}$. There exists $f \in \mathrm{SL}(2,\mathbb{R})$ such that both $f(z_1)$ and $f(z_2)$ are on the imaginary axis.*

Proof. The set $\mathrm{SL}(2,\mathbb{R})$ acts transitively on $\mathbb{H}$: for any z and w, there exists $f_1 \in \mathrm{SL}(2,\mathbb{R})$ such that $f(z) = w$, and in particular, we can send any $z = x + iy$ to i, by considering $(a,b,c,d) = \left(\frac{1}{\sqrt{y}}, -\frac{x}{\sqrt{y}}, 0, \sqrt{y}\right)$ to get

$$f_1(z) = \frac{\frac{1}{\sqrt{y}}(x+iy) - \frac{x}{\sqrt{y}}}{\sqrt{y}} = i$$

(we immediately get transitivity from this, since Möbius transformations are invertible). Second, for any $z = x + iy$, there exists an$f_2 \in \mathrm{SL}(2,\mathbb{R})$ that sends z to the imaginary axis and which fixes i. This f_2 is of

the form $f_2(z) = \frac{(\cos\theta)z+\sin\theta}{-(\sin\theta)z+\cos\theta}$ for some θ. The real part of $f_2(z)$ is equal to

$$\begin{aligned}
\operatorname{Re}(f_2(z)) &= \operatorname{Re}\left(\frac{((\cos\theta)(x+iy)+\sin\theta)(-(\sin\theta)(x-iy)+\cos(\theta))}{|-(\sin\theta)z+\cos\theta|^2}\right) \\
&= \operatorname{Re}\left(\frac{((x\cos\theta+\sin(\theta)+iy\cos\theta))(-x\sin\theta+\cos(\theta)+iy\sin\theta)}{|-|^2}\right) \\
&= \frac{(x\cos\theta+\sin(\theta))(-x(\sin\theta)+\cos(\theta))-y^2\cos\theta\sin\theta}{|-|} \\
&= \frac{\left(1-|z|^2\right)\sin\theta\cos\theta+x\left(\cos^2\theta-\sin^2\theta\right)}{|-|}.
\end{aligned}$$

When $\theta = 0$, the numerator is equal to x, and when $\theta = \pi/2$, the numerator is equal to $-x$, so for some θ we can get 0.

Applying f_1 followed by f_2 (with θ corresponding to $f_1(z_2)$) then gives us the desired result. ✹

This already tells us all we need to know about geodesics!

Theorem 2.4. *A geodesic between two points in the hyperbolic plane is either a vertical line or a segment of a circle which is perpendicular to the real axis.*

Proof. Any two arbitrary points z_1 and z_2 can be sent to the imaginary axis, and in the imaginary axis, the straight line connecting them is a geodesic. Thus, the geodesic between any two points is the image of a straight line under a Möbius transformation: it is either a straight line itself, or an arc segment of a circle! In fact, for points which do not have the same x coordinate, it cannot be a straight line - such a line would meet the real axis at some angle, while imaginary axis meets the real line orthogonally. Möbius transformations are conformal, so angles should be preserved. This forces the geodesic to be a circle which is orthogonal to the real axis - its centre is on the real axis and is the midpoint of the x coordinates of the two intersection points. ✹

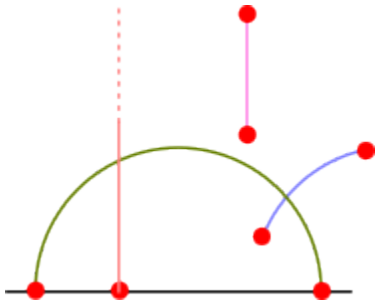


Figure 2.3: Examples of geodesics. Note that we can have geodesics between "points at infinity" as well.

Remark 2.5. There is a theorem in differential geometry, stating that for a metric of the form $ds^2 = Edx^2 + Gdy^2$, the Gaussian curvature at a point is equal to

$$K = \frac{-1}{2\sqrt{EG}}\left(\frac{\partial}{\partial x}\left(\frac{G_x}{\sqrt{EG}}\right)+\frac{\partial}{\partial y}\left(\frac{E_y}{\sqrt{EG}}\right)\right).$$

The curvature of the Riemannian surface described above is then equal to

$$K = \frac{-1}{2\sqrt{\frac{1}{y^2}\frac{1}{y^2}}}\left(\frac{\partial}{\partial y}\left(-2\frac{1}{y^3}\frac{1}{\sqrt{\frac{1}{y^2}\frac{1}{y^2}}}\right)\right)$$
$$= \frac{1}{\sqrt{\frac{1}{y^2}\frac{1}{y^2}}}\left(\frac{\partial}{\partial y}\frac{1}{y}\right) = -\frac{1}{\frac{1}{y^2}}\frac{1}{y^2} = -1.$$

So the surface we have described is indeed "THE hyperbolic plane" - a simply connected, infinite surface of constant curvature -1.

2.2 Triangles

Now that we know how geodesics look like, we can start drawing triangles. A triangle has three points, connected by geodesics. Unlike in the Euclidean case, we allow triangles to have points at infinity: a triangle can have, zero, one, two, or all three of its points at infinity. Note that "infinity" in the half plane model can be either at the real axis, or at $\mathrm{Im}\,(z) = \infty$ (the latter is actually just a single point (!), which we sometimes call ∞; after all, the distance between two points on the same horizontal line $\mathrm{Im}\,(z) = y$ is no larger than $(x_2 - x_1)/y$, so they do converge to a single point). Here are some triangles in the upper half-plane:

Figure 2.4: Triangles with 0, 1, 2 and 3 ideal points. There right-most triangle has 3 ideal points, with one of them at $\mathrm{Im}\,(z) = \infty$.

Triangles which have all three points at infinity are called *ideal*. Since the circle segments meet the real axis orthogonally, the interior angle of points at infinity is in fact 0. An ideal triangle has $\sum \alpha_i = 0$! The situation is very different than in the Euclidean plane, and indeed you can convince yourselves that $\sum \alpha_i \leq \pi$ for all hyperbolic triangles.

Ideal triangles are infinite in diameter, but nonetheless they have finite area. In fact, all triangles do.

Theorem 2.6. *Let T be a triangle with interior angles α, β, γ. Then*

$$\mathrm{Vol}\,(T) = \pi - \alpha - \beta - \gamma.$$

Proof. Let's start with a triangle with one point at infinity and two general points

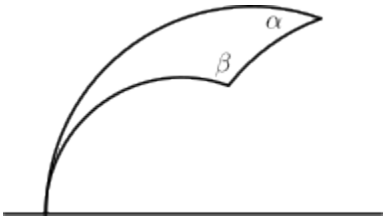


Figure 2.5: A triangle with one ideal point on the real axis.

Actually, by applying a Möbius transformation, we can assume that the triangle is of this form:

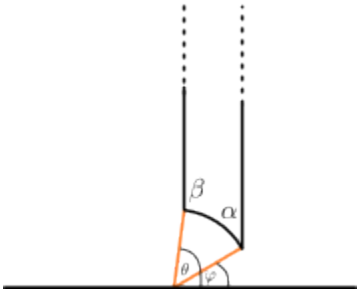


Figure 2.6: A triangle with one ideal point, now at infinity.

We already know that the geodesics are circle segments, and the surface integral is not too bad in this case:

$$\begin{aligned}\mathrm{Vol}\,(T) &= \int_{\cos(\theta)}^{\cos(\varphi)} \int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2}\, dy\, dx \\ &= \int_{\cos(\theta)}^{\cos(\varphi)} \frac{1}{\sqrt{1-x^2}}\, dx.\end{aligned}$$

Denoting $x = \cos(u)$, we have $\frac{dx}{du} = -\sin(u)$ and so

$$\mathrm{Vol}\,(T) = \int_{\theta}^{\varphi} \frac{1}{\sqrt{1-\cos(u)^2}} \left(-\sin(u)\right) du = \theta - \varphi.$$

The angle φ is equal to the angle α of one of the vertices of the triangle:

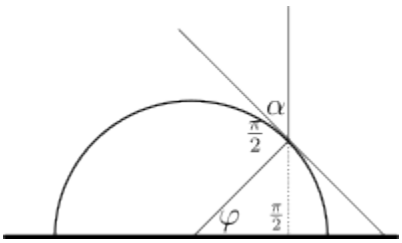


Figure 2.7: The angles φ and α are equal by elementary plane geometry.

while the angle θ is equal to $\pi - \beta$. The total volume of T is therefore

$$\mathrm{Vol}\,(T) = \pi - \alpha - \beta,$$

which is the correct amount for a triangle with a vertex at infinity. To complete the proof for general triangles, consider the following image:

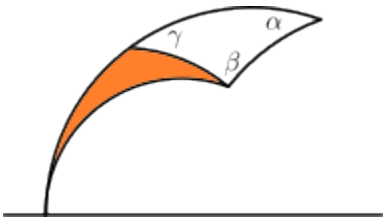


Figure 2.8: A general triangle can be extended to one with an ideal point, for which we already know the area-angle theorem. The result follows by calculating the angles.

Remark 2.7. The largest area that a triangle can have is π; this is in stark contrast to $\mathbb{R}^2$. Heuristically, geodesics which are fired off from the same point but at different angles separate exponentially quickly from each other. When the geodesics must travel a long time to get to their destination (e.g. the two sides of a triangle emanating from the same point), then they must be very close to each other for a very long time. This creates very thin structures.

Remark 2.8. If a triangle is very small (that is, its volume is small), then the sum of its angles must be close to π. In other words - small triangles behave almost Euclidically. This is not surprising - the whole point of a Riemann surface is to be locally Euclidean. Still, it is a useful fact to remember, and is true for basically all local geometric properties, not just the area of triangles.

2.3 The Poincaré disk model

The half plane model is not the only way to think about the hyperbolic plane. There are many others, including the Klein model and the hyperboloid model. We will sometimes use the Poincaré disk model. Here, the space is the open unit disk $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$, with the metric

$$ds^2 = \frac{4\left(dx^2 + dy^2\right)}{\left(1 - \left(x^2 + y^2\right)\right)^2}.$$

You can calculate the curvature to see that it is indeed the same simply connected infinite $\mathbb{H}$ we know and love from the half plane model.

One advantage of this model is that it is radially symmetric, so polar coordinates can be more easily used and are easier to visualize. Further, all the points at infinity are treated the same, unlike the half-plane where ∞ and the real axis are treated differently.

The Poincaré half-plane model and the Poincaré disk model are related by more than just the name Poincaré. The Möbius function

$$f\,(z) = \frac{z - i}{z + i}$$

maps the half-plane to the unit disk, while transferring the metric from $\mathbb{H}$ to that of $\mathbb{D}$, so this is an isometry. As a Möbius transformation, it maps circles to circles, so geodesics in the disk model are again segments of circles which are orthogonal to the boundary.

It will sometimes be more convenient to do calculations in the half-plane, and sometimes more convenient to use the disk model. Here is a calculation made easy in this model.

Theorem 2.9. *The hyperbolic circle of radius R has area*

$$\mathrm{Vol}\left(B\left(R\right)\right) = 4\pi \sinh\left(R/2\right)^2.$$

Proof. A hyperbolic circle of radius $R > 0$ corresponds to a circle of some radius $r < 1$ in the disk model. Finding the area of a circle of Euclidean radius $r < 1$ is just a matter of integration:

$$\begin{aligned}\mathrm{Vol}\left(\{z \in \mathbb{D} \mid |z| \leq r\}\right) &= \int_0^{2\pi}\int_0^r \frac{4}{\left(1-s^2\right)^2} s\, ds d\theta \\ &= \int_0^r \frac{8\pi}{\left(1-s^2\right)^2} s\, ds d\theta \\ &= \frac{4\pi}{1-s^2}\,|_0^r = 4\pi\left(\frac{1}{1-r^2}-1\right) = \frac{4\pi r^2}{1-r^2}.\end{aligned}$$

Finding the $r < 1$ which corresponds to $R > 0$ is also a matter of integration: a point at Euclidean distance r from the origin is at distance

$$R = d\left(0, r\right) = \int_0^r \frac{2}{1-s^2} ds = 2\tanh^{-1}\left(r\right),$$

so

$$r = \tanh\left(\frac{R}{2}\right).$$

Thus

$$\mathrm{Vol}\left(B\left(R\right)\right) = \frac{4\pi \tanh\left(\frac{R}{2}\right)^2}{1-\tanh\left(\frac{R}{2}\right)^2} = 4\pi \sinh^2\left(\frac{R}{2}\right).$$

Remark 2.10. For small values of R, we have $\sinh\left(R/2\right) \approx R/2$, so the area of a circle is approximately the Euclidean πR^2. However, for large values of R, the volume stops growing quadratically, but rather grows like πe^R - it is exponential in R!

Remark 2.11. Using basic hypertrigonometric identities, a useful reformulation to the above is

$$\mathrm{Vol}\left(B\left(R\right)\right) = 2\pi\left(\cosh\left(R\right)-1\right).$$

Exercise 2.12. What is the circumference of a hyperbolic circle of radius R?

Exercise 2.13. Find the area of a circle without using integration.

3 Brownian motion (Lecture 3)

We all love Brownian motion, and are used to playing with it in one, two, and general d dimensions. There are several ways to construct it.

1. The stochastic process definition. Brownian motion B_t is the unique stochastic process with:
   (a) $B_0 = 0$
   (b) B_t has independent Gaussian increments, i.e. $B_{t_1} - B_{t_2}$ is independent from $B_{t_2} - B_{t_3}$ for $t_1 \geq t_2 \geq t_3$, and $B_t - B_s \sim N(0, (t-s) \cdot \mathrm{Id})$.
   (c) B_t has almost surely continuous paths.

   This is a very abstract approach; just from the definition, you have to work a bit to show that it exists, but of course the properties are very, very useful. Still, it is always good to keep in mind an explicit construction.

2. Small incremental steps: Donsker's theorem. This is basically the strong law of large numbers, but for stochastic processes rather than just random variables. For $i = 1, 2, \ldots,$let $X_i \sim N(0, \mathrm{Id})$ be iid standard Gaussians, and let $S_n = \sum_{i=1}^{n} X_i$ be the partial sums. By the law of large numbers,

$$\frac{S_n}{\sqrt{n}} \to N(0, \mathrm{Id})$$

   in distribution. This is very nice, since $B_1 \sim N(0, \mathrm{Id})$, so we actually have

$$\frac{S_n}{\sqrt{n}} \to B_1.$$

   Donsker's theorem states that this convergence applies to an interpolated stochastic process: the process

$$\left(\frac{S_{\lfloor nt \rfloor}}{\sqrt{n}}\right)_{t \in [0,1]} \to (B_t)_{t \in [0,1]}$$

   in distribution.

3. This is not the only "explicit" construction. A crowd favourite is Lévy's construction: start with $B_0 = 0$ and $B_1 = N(0, 1)$, and then recursively interpolate the the values of the process on dyadic points, e.g $B_{\frac{1}{2}} = \frac{B_0 + B_1}{2} + N\left(0, \frac{1}{2}\right)$. Another crowd favourite: the Paley-Wiener Fourier decomposition:

$$B_t = \frac{\sqrt{2}}{\pi} \sum_{n=1}^{\infty} X_n \frac{\sin\left((n - 1/2)\pi t\right)}{n - 1/2}.$$

4. If we have a stochastic process X_t which behaves nicely, we can define an operator from it, called the infinitesimal generator:

$$(A_X f)(x) = \lim_{t \to 0} \frac{\mathbb{E}\left[f(X_t) \mid X_0 = x\right] - f(x)}{t}.$$

   In one dimension, if f is smooth we can take its Taylor expansion:

$$f(X_t) = f(x) + f'(x)(X_t - x) + \frac{1}{2} f''(x)(X_t - x)^2 + O(X_t - x)^3.$$

When $X_t = B_t$, the expectations make our life easier:

$$\mathbb{E}\left[B_t - x \mid B_0 = x\right] = 0$$

$$\mathbb{E}\left[\left(B_t - x\right)^2 \mid B_0 = x\right] = t$$

$$\mathbb{E}\left[\left(B_t - x\right)^3 \mid B_0 = x\right] = 0, \mathbb{E}\left[\left(B_t - x\right)^4 \mid B_0 = x\right] = 3t^2.$$

So

$$\left(A_B f\right)(x) = \lim_{t \to 0} \frac{\frac{1}{2} t f''(x) + O\left(t^2\right)}{t} = \frac{1}{2} f''(x).$$

We can repeat this (with some more clumsy notation) for higher dimensions, and get

$$\left(A_B f\right)(x) = \frac{1}{2} \sum_{i=1}^{d} \frac{\partial f^2}{\partial x_i^2} = \frac{1}{2} \Delta f(x).$$

The Laplacian is the infinitesimal generator of Brownian motion (up to a scaling factor of $1/2$, which won't interest us much).
In many cases, the opposite case also works. Suppose that A is a positive, self-adjoint operator. Here is a sketch, ignoring many details, of a way to obtain a process from it:

(a) Write the spectral decomposition of A: there exists a unitary operator U and a function λ such that

$$\left(U^{-1} A U\right) g(\omega) = \lambda(\omega) g(\omega).$$

This is just a general, fancy way to say "Fourier transform": up to the transformation U (which takes us to and from phase space), all A does is multiply every point in the spectrum by some value.

(b) Having written this decomposition, we define the semigroup of the infinitesimal generator, given by

$$T_t := e^{tA}.$$

This is a new operator, which changes the spectral decomposition of a function in a different way than A does, but of course it relies heavily on it. When $t = 0$, this is just the identity operator. If A just multiplied the spectrum of a function by $\lambda(\omega)$, then T_t just multiplies by $e^{t\lambda(\omega)}$.

(c) Sometimes, the semigroup can be calculated by integrating over a *kernel function*:

$$\left(T_t f\right)(x) = \int_{\mathbb{R}} p_t(x, y) f(y) \, dy.$$

The function $p_t(x, y)$ is called the heat kernel of the semigroup. It is a sort of weighted average over the entire space, since it satisfies (for us) $\int_{\mathbb{R}} p_t(x, y) \, dy = 1$ for all $x \in \mathbb{R}$.

(d) We define a stochastic process by interpreting the heat kernel as the transition density of the process:

$$\mathbb{P}\left[X_t \in A \mid X_0 = x\right] = \int_A p_t(x, y) \, dy,$$

and in general

$$\mathbb{P}\left[X_{t_1} \in A_1, X_{t_2} \in A_2, \ldots, X_{t_n} \in A_n \mid X_0 = x\right] =$$

$$= \int_{A_1, \ldots, A_n} p_{t_1}(x, y_1) \, p_{t_2 - t_1}(y_1, y_2) \ldots p_{t_{n-1} - t_n}(y_{n-1}, y_n) \, dy_1 \cdots dy_n.$$

To see this all in action, let's again consider the simple case of $A = \frac{1}{2}\Delta$. Admittedly, given an arbitrary operator on an arbitrary space, it's a bit hard a-priori to guess what the spectral decomposition will look like. But the Laplacian is a nice operator on $\mathbb{R}$, and $\mathbb{R}$ is a nice space with a nice Fourier transform:

$$\hat{f}(\omega) = \int_{\mathbb{R}} e^{-ix\omega} f(x)\, dx.$$

By integrating by parts, we then have

$$\widehat{\frac{1}{2}\Delta f}(\omega) = \int_{\mathbb{R}} e^{-ix\omega} \frac{1}{2}\Delta f(x)\, dx = -\frac{1}{2}|\omega|^2 \hat{f}(\omega).$$

So the Laplacian squares the spectrum of the function, and $\lambda(\omega) = -\frac{1}{2}|\omega|^2$. In Fourier space, the semigroup then multiplies by $e^{-\frac{1}{2}|\omega|^2}$,i.e

$$\left(\widehat{T_t f}\right)(\omega) = e^{-\frac{1}{2}t|\omega|^2} \hat{f}(\omega).$$

The Fourier transform of a Gaussian is a Gaussian:

$$\widehat{e^{-ax^2}} = C\frac{1}{\sqrt{a}} e^{-\omega^2/a},$$

so the right-hand side is actually the product of two Fourier transforms. By the convolution theorem ($\widehat{f \star g} = \hat{f} \cdot \hat{g}$),

$$T_t f = f \star e^{-\frac{1}{2}t|\cdot|^2} = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{|x-y|^2}{t}} f(y)\, dy.$$

Great! We magically got ourselves into the form that we wanted, i.e.

$$T_f f(x) = \int_{\mathbb{R}} p_t(x,y) f(y)\, dy.$$

Here the heat-kernel is

$$p_t(x,y) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{1}{2}\frac{|x-y|^2}{t}},$$

which is the density of a Gaussian with variance t. In other words, when defining the process, we have independent Gaussian increments! A bit more analysis can give us that the paths are continuous (this is basically the same analysis as when constructing Brownian motion from its basic definition). After a bit of abstract and practical work, we arrive at a new definition of Brownian motion: Brownian motion is the stochastic process obtained via the above method from the operator $\frac{1}{2}\Delta$. (Is it a coincidence that this is also its infinitesimal generator?)
We should definitely remark that the heat-kernel is the fundamental solution to the heat equation,

$$\partial_t f = \frac{1}{2}\Delta f,$$

with $\lim_{t\to 0} \int p_t(x,y)\, u(y)\, dy = u(x)$. Solving a partial differential equation is a much more straightforward approach to the work above.

5. This may be considered cheating, but if you somehow forgot what Brownian motion is, but still miraculously remembered how to perform stochastic calculus, we can define Brownian motion as the solution to the stochastic differential equation $dX_t = dB_t$, or, in integral form,

$$X_t = \int_0^t 1 \cdot dB_s.$$

So far, all this was about Brownian motion in $\mathbb{R}^d$. Now let's do hyperbolic geometry.

1. On the surface of it (ha!), the first construction - that of willing Brownian motion into existence by definition - fails. We cannot add together points in the hyperbolic plane, it is not a vector space. However, we can shoot out geodesics. Define a random walk of step size ε as follows: Let $X_0 = 0$ be some point in the hyperbolic plane. Given X_t, let θ_{t+1} be a uniformly random angle, and X_{i+1} be the point obtained by walking a distance ε in the direction θ_{i+1} from X_i. By taking $\varepsilon = \frac{1}{\sqrt{n}}$ and looking at $X_{\lfloor nt \rfloor}$ (interpolating for fractional times), we can get Brownian motion in the hyperbolic plane (or any Riemannian manifold, actually; see [1]).

2. Perhaps more abstractly, we can in general consider Brownian motion on the tangent plane of a point in a Riemannian manifold, and project it down to the manifold itself (see [2])

3. All of our functional analysis tools apply in the case of Riemannian manifolds as well. In Riemannian geometry, the Laplacian is defined as
$$\Delta f = \operatorname{div}(\nabla f).$$
This is very abstract; if we actually want to calculate the Laplacian in a real life case, we usually have to work with local coordinates and a metric g defined with those coordinates. It can then be shown that, in general,
$$\Delta f = \frac{1}{\sqrt{|g|}} \sum_{i,j} \partial_i \left(\sqrt{|g|} g^{ij} \partial_j f \right).$$
Let's see how that applies to our case, where the hyperbolic plane is identified with the half-plane, and
$$ds^2 = \frac{dx^2 + dy^2}{y^2},$$
i.e.
$$g = \begin{pmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{pmatrix}.$$
Then $|g| = \left| \det \begin{pmatrix} \frac{1}{y} & 0 \\ 0 & \frac{1}{y} \end{pmatrix} \right| = y^{-4}$
$$\begin{aligned} \Delta f &= \frac{1}{\sqrt{|g|}} \sum_{i,j} \partial_i \left(\sqrt{|g|} g^{ij} \partial_j f \right) \\ &= y^2 \sum_i \partial_i^2 f. \end{aligned}$$

Exercise 3.1. If you feel comfortable with differential geometry, prove the equality $\operatorname{div}(\nabla f) = \frac{1}{\sqrt{|g|}} \partial_i \left(\sqrt{|g|} g^{ij} \partial_j f \right)$.

Armed with this Laplacian, we can repeat the whole heat kernel procedure and eventually get a heat kernel $p_t(x, y)$, allowing us to define a diffusion process. Alternatively, we could have tried to solve the heat equation on $\mathbb{H}$ directly, using the definition of the Laplacian. It turns out (and this is no small matter), that
$$p_t^{\mathbb{H}}(x, y) = \frac{\sqrt{2}}{(4\pi t)^{3/2}} e^{-\frac{t}{4}} \int_{d(x,y)}^{\infty} \frac{s e^{-\frac{s^2}{4t}}}{\sqrt{\cosh s - \cosh(d(x, y))}} ds.$$

Compare this with the Euclidean heat kernel in two dimensions:

$$p_t^{\mathbb{R}^2}(x,y) = \frac{1}{4\pi t} e^{-\frac{1}{4}\frac{|x-y|^2}{t}}.$$

The hyperbolic plane is a Riemannian surface, so it locally looks Euclidean. So we expect that for very small distances and very small times, the heat kernels are comparable.

Exercise 3.2. Show that for fixed x, y and small times t, $p_t^{\mathbb{H}}(x,y) \approx p_t^{\mathbb{R}^2}(x,y)$. Figure out how small the "small times" should be in comparison to $d(x,y)$.

4. Since we have the half-plane model for the hyperbolic plane, we can use the metric to see how a diffusion is affected directly. Intuitively, if we think of a particle (X_t, Y_t) moving about in the upper half-plane, its movements are "impeded" as it gets closer to the real line, by a factor of $\frac{1}{Y_t}$. We then arrive at the stochastic differential equation

$$dX_t = Y_t dB_t^x, \quad dY_t = Y_t dB_t^y.$$

We can solve the equation just for Y_t, since it does not depend on X_t. For this we use Ito's formula for functions of diffusion processes. Let's remind ourselves of the details. If Z_t is a diffusion process given by $dZ_t = \mu_t dt + \sigma_t dB_t$, then for nice enough functions $f(x)$, the process $f(Z_t)$ is also a diffusion process, with differential given by

$$df(t, Z_t) = \frac{\partial f}{\partial x}(Z_t)\, dZ_t + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(Z_t)\,(dZ_t)^2.$$

This is the general formula. We now take the simple function $f(x) = \log(x)$, and apply it to Y_t:

$$\begin{aligned} df(Y_t) &= \frac{1}{Y_t} dY_t + \frac{1}{2}\left(-\frac{1}{Y^2}\right)\left(Y_t^2 dt\right) \\ &= dB_t - \frac{1}{2} dt. \end{aligned}$$

So $\log(Y_t)$ is simple Brownian motion with drift $-\frac{1}{2}$, and $Y_t = e^{B_t - \frac{1}{2}t}$. Every Brownian motion in the half-plane model eventually drifts down towards the real line (this is infinity).

5. Alternatively, there is the Ito isometry: if Z_t is a stochastic process, then

$$\mathbb{E}\left[\left(\int_a^b Z_t dB_t\right)^2\right] = \mathbb{E}\left[\int_a^b Z_t^2 dt\right].$$

Using $Z_t = Y_t$, we have $dY_t = Y_t dB_t$, and so we get

$$\begin{aligned} \mathbb{E}\left[(Y_b - Y_a)^2\right] &= \mathbb{E}\left[\left(\int_a^b dY_t\right)^2\right] = \mathbb{E}\left[\int_a^b Y_t^2 dt\right] = \mathbb{E}\left[\int_0^b Y_t^2 dt - \int_0^a Y_t^2 dt\right] \\ &= \mathbb{E}\left[\left(B_{\int_0^b Y_t^2 dt} - B_{\int_0^a Y_t^2 dt}\right)^2\right]. \end{aligned}$$

Now, the process Y_t is a Gaussian process with independent increments. We see above that its second moment is equal to that of time changed Brownian motion! We actually have

$$Y_t = B_{\int_0^t Y_t^2 ds}.$$

Thus Y_t is a time change of Brownian motion (a random time change, mind you).

4 Constructing surfaces (Lecture 4)

The hyperbolic plane is a closed, unbounded, simply connected surface. From it, we can obtain other hyperbolic surfaces, both compact and noncompact.

4.1 The pseudosphere

In the half-plane model, look at the set

$$A = \{\mathrm{Im}\,(z) \geq 1\} \mod z \mapsto z+1,$$

i.e. we identify points whose x coordinates differ by an integer. A set of representatives for this set is

$$C = \left\{\mathrm{Im}\,(z) \geq 1, -\frac{1}{2} < \mathrm{Re}\,(z) \leq \frac{1}{2}\right\}.$$

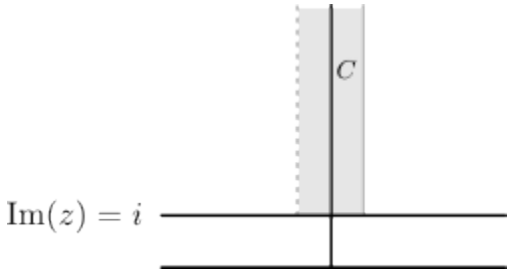


Figure 4.1: A fundamental domain for $z \mapsto z+1$.

This is a fundamental domain for the action $z \mapsto z+1$: it contains one point from every orbit of the action of $z \mapsto z+1$ on $\mathbb{C}$ (we can also consider the set $\left\{\mathrm{Im}\,(z) \geq 1, -\frac{1}{2} < \mathrm{Re}\,(z) < \frac{1}{2}\right\}$; it has the advantage of being open, and it *almost* contains a point from every orbit: its closure does, and contains at most 2 points from every orbit. It often doesn't really matter whether you include the boundary or not).

We can equip this set with the exact same underlying hyperbolic metric:

$$ds^2 = \frac{dx^2 + dy^2}{y^2},$$

and the distance between two points is now

$$d\,(x,y) = \inf_{\gamma} \int \frac{1}{\gamma\,(t)} \gamma'\,(t)\,dt,$$

where the infimum is over all curves γ that connect any representative of x to any representative of y. This is called a *pseudosphere*.

If we were dealing with the Euclidean plane $\mathbb{R}^2$, we could easily describe what C is: it is a cylinder of constant width. Well, a half cylinder, to be precise, with one end cut off, and the other going to infinity:

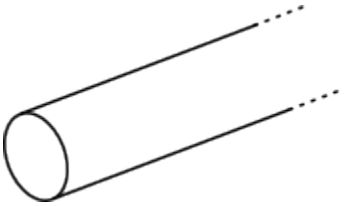


Figure 4.2: A cylinder in $\mathbb{R}^3$.

However, in the half-plane model, the higher the imaginary component, the shorter the distance becomes. Two points at height y are at distance no more than $\frac{1}{y}$ from each other (and in fact only slightly less). Thus, the end that goes off to infinity eventually shrinks off to a point. This is called a cusp.

Figure 4.3: A cartoon of a cusp in $\mathbb{R}^3$. Note that the boundary is NOT a geodesic.

Two cool facts about the cusp:

1. The cusp has no closed geodesics: In a Euclidean cylinder, we can draw a straight line that cuts the cylinder into two parts and reconnects to itself:

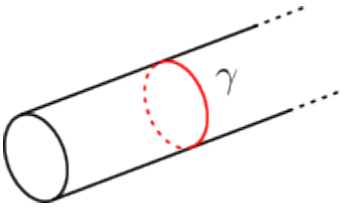


Figure 4.4: The geodesic γ is closed - it ends up in the same place and the same direction where it started.

If we think about tiling $\mathbb{R}^2$ with the fundamental domain C of a cylinder, such a geodesic looks like this:

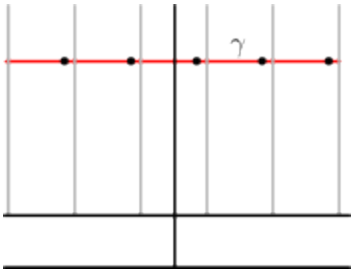


Figure 4.5: The geodesic γ is a straight line in $\mathbb{R}^2$.

However, geodesics in $\mathbb{H}$ are circular arcs, while the points that are equivalent to each other all have the same imaginary component. There is no way for a hyperbolic geodesic to hit the same point more than twice - there are no closed geodesics.

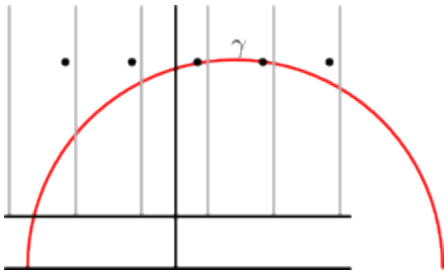


Figure 4.6: A geodesic can't hit more than two points in the orbit of $z \mapsto z+1$.

Exercise 4.1. Show that a bininfinite geodesic may intersect itself. Does every geodesic intersect itself?

2. The pseudosphere is infinite - its tail end is unbounded. However, its area is given by

$$A(S) = \int_1^\infty \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{y^2} dx dy = \int_1^\infty \frac{1}{y^2} dy = 1.$$

Note that instead of blocking of the imaginary component at 1, we could taken $A = \mathbb{H} \mod (z+1)$. In this case the cusp would extend to infinity in both directions, and "spread out" towards infinity at the bottom. The volume would be infinite.

4.2 The Bolza surface

In the Poincaré disk model, look at the equilateral octagon centered at the origin. Well, I say "the equilateral" octagon, but there are infinitely many of them: one for every possible angle θ at the vertices. For this particular construction, we choose $\theta = \pi/4$. If we are still not used to hyperbolic geometry, this statement by itself can cause a bit of discomfort. However, we can rest assured that such an octagon exists. For example, we can always take an ideal regular ocatgon, say in the Poincaré disk. This octagon has an angle of 0 at every vertex.

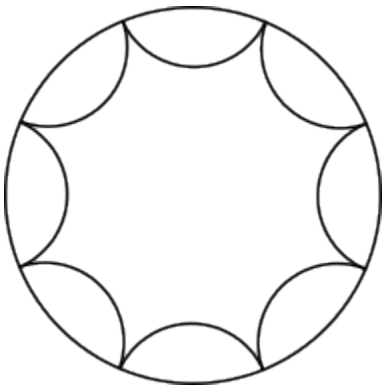


Figure 4.7: An ideal octagon.

When we shrink the side lengths of the octagon, the angles start growing; as the side-length goes to 0, the octagon becomes more similar to a Euclidean hexagon, so each angle goes to $3\pi/4$. By continuity, there must be an intermediate side-length that gives $\pi/4$.

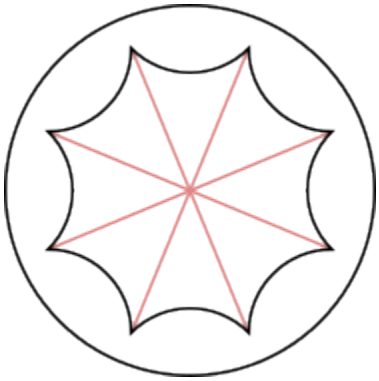


Figure 4.8: An equilateral octagon, here divided into 8 triangles.

Now identify opposite edges of the octagon with each other. This time, we get a compact surface S - no boundaries or infinities involved - called the Bolza surface (after Oskar Bolza). The Bolza surface serves as the QED symbol in the proofs of these lecture notes. Here are an additional two facts.

1. It is made of 8 triangles, each with angles $2\pi/8$, $\pi/8$, and $\pi/8$ (see 4.8). Since the area of a hyperbolic triangle with angles α, β, γ is equal to $\pi - \alpha - \beta - \gamma$, the total area of the Bolza surface is

$$A(S) = 8 \cdot \left(\pi - \frac{2\pi}{8} - \frac{\pi}{8} - \frac{\pi}{8}\right) = 4\pi.$$

2. In topology class, we learn about the classification of closed (topological) surfaces: every orientable surface is either equal to the sphere or to a g-holed torus, where g, the number of holes, is called the genus of the surface. Perhaps we are already convinced that the Bolza surface is not topologically equivalent to the sphere or the torus (since its curvature is -1). What is its genus then? Well, consider the original polygon, with its 8 sides and 8 vertices, and treat is as a planar map: there are 8 edges, 8 vertices, and 2 faces. Once we identify pairs of edges together, we get a new graph, with 4 edges, 1 face, and 1 vertex.

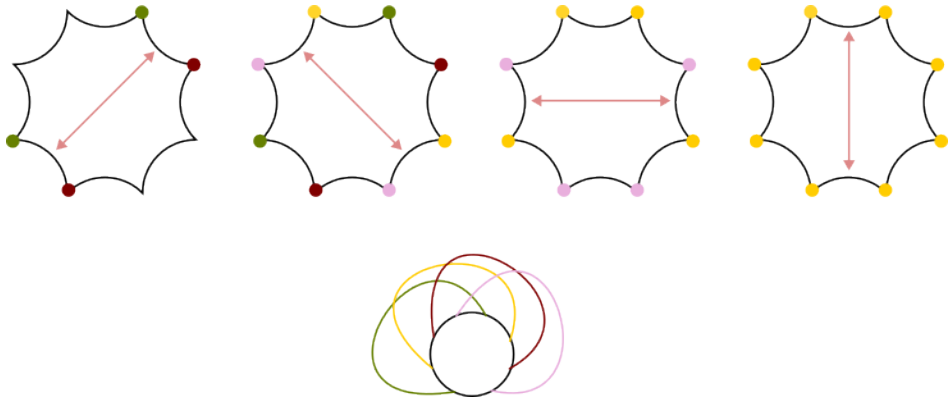


Figure 4.9: Identifying the sides of the polygon results in a map with 4 edges, 1 face, and one vertex.

By the Euler characteristic theorem,

$$\chi = V - E + F = 1 - 4 + 1 = -2,$$

where χ is the Euler characteristic of the surface, which is equal to $2 - 2g$. Thus

$$-2 = 2 - 2g \implies g = 2.$$

3. Knowing the genus, here is an alternative way to calculate the area of the Bolza surface.

   Theorem 4.2 (The Gauss-Bonet theorem)**.** *Let S be a compact Riemannian surface, perhaps with boundary. Let K be the curvature function in S and k be the geodesic curvature of ∂S. Then*

$$\int_S K dA + \int_{\partial S} k ds = 2\pi\chi(S).$$

   The Bolza surface has no boundary, and since it is hyperbolic, it has constant curvature -1 everywhere. Its genus is $g = 2$, so its Euler characteristic is $\chi = 2 - 2g = -2$. Thus

$$-1 \cdot A(S) = 2\pi \cdot (-2) \implies A(S) = 4\pi.$$

   In fact, if a hyperbolic surface M has genus g, then its area will always be

$$A(M) = 2\pi(2 - 2g).$$

4. The Bolza surface definitely has closed geodesics. For example, look at the horizontal radius line perpendicular to the polygon's sides. But other, more complicated closed geodesics exist. For example, this one:

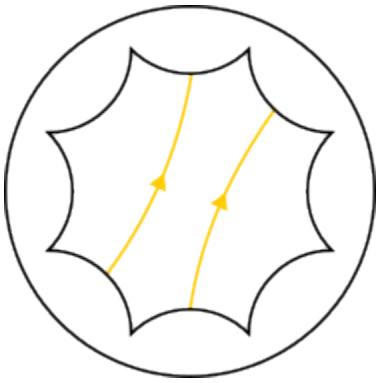


Figure 4.10: Finding closed geodesics in the Bolza surface is not as easy as it is on the torus, where all you need is a rational slope.

Definition 4.3. The *systole* of a compact hyperbolic surface S is the length of its shortest closed geodesic.

Exercise 4.4. Find the systole of the Bolza surface.

Theorem 4.5 ([3])**.** *Among all compact hyperbolic surfaces of genus* 2*, the Bolza surface has the largest systole.*

Exercise 4.6. The Bolza surface was the result of identifying opposing sides of a regular hyperbolic octagon. We can obtain a different surface by identifying the sides of the octagon in different ways. For example, consider the following polygon, based on the work of Zieschang-Vogt-Coldewey. What is the systole of this surface?

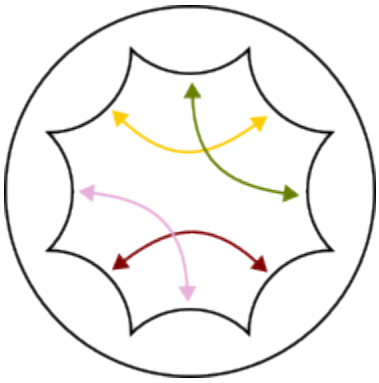


Figure 4.11: This surface is different from the Bolza surface.

4.3 The pair of pants

So far, we have seen two compact hyperbolic surfaces, both of which had genus 2 and were obtained by gluing sides of a regular hyperbolic octagon. It will turn out that actually, all genus 2 surfaces are the result of side gluing of some hyperbolic octagon, but not necessarily regular. Unfortunately, it is not a very simple matter to enumerate all the possible octagons and side matchings which give all compact genus 2 surfaces. The issue becomes even more severe when moving to higher genus, where instead of octagons we deal with $2g$-gons. Luckily, there is a more fashionable way to address the problem.

Let H be a right angled hyperbolic hexagon. Even though it may sound odd to our Euclidean ears, this creature does indeed exist, like the octagon before it. In fact, the set of all right-angled hexagons is a three-parameter family: if we label the opposite sides of the hexagon by the pairs $a-\alpha$, $b-\beta$, $c-\gamma$, then there exists a hexagon for every choice of non-negative a, b, c (including 0). In this case, the parameters α, β, γ are given by

$$\cosh(\alpha) = \frac{\cosh(b)\cosh(c) + \cosh(a)}{\sinh(b)\sinh(c)}$$

(and cyclically for the rest).

Definition 4.7. Let H and H' be two copies of a right angled hexagon with sides $a, b, c, \alpha, \beta, \gamma$ (resp. $a', b', c', \alpha', \beta', \gamma'$). The hyperbolic surface S obtained by gluing together α and α', β and β', and γ and γ' is called a *pair of pants*. Sometimes, it is also called Y piece. The lengths $2a, 2b, 2c$ are called the *cuff lengths*, or *waist lengths*. The surface S has three geodesic boundary components, with respective lengths $2a$, $2b$ and $2c$.

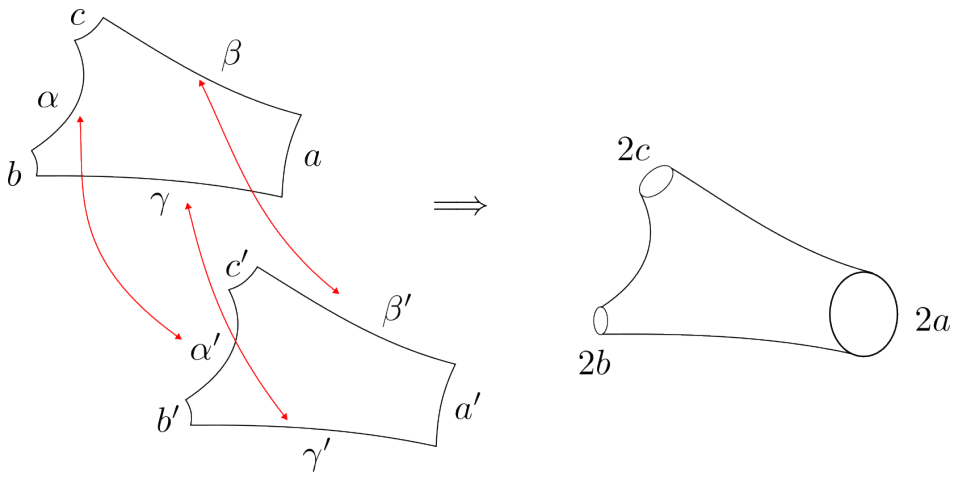


Figure 4.12: Gluing together the edges of two right-analged hexagons gives surface with three boundary components, called a *pair of pants*.

The name "pair of pants" stems from the shape of the surface S that we get if we sort of ignore the hyperbolic geometry. If a is large and $b = c$ are small, we obtain a shape resembling pants (the British equivalent "pair of trousers" never caught on, I'm afraid). However, we stress that a, b and c are interchangeable, and a pair of pants doesn't necessarily have to look like recognizable pants. For example, when $a = b = c = 0$, we get an ideal pair of pants:

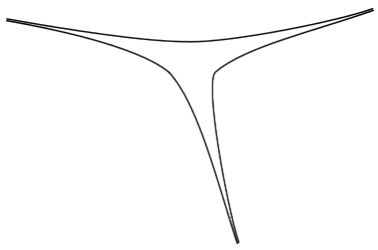


Figure 4.13: A pair of pants with 0 cuff lengths.

When $a = b = c$ are very large, the analogy to clothing is not convincing anymore.

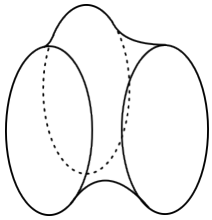


Figure 4.14: A cartoon of a pair of pants with large cuff lengths. The resemblance to actual pants is no longer there.

Topologically, the pair of pants has no holes - it is a genus 0 surface with three boundary components. Its Euler characteristic is therefore $\chi = 2 - 0 - 3 = -1$. Since the boundary curves are geodesics, their geodesic curvature is 0, and so by the Gauss-Bonnet theorem, the area of a pair of pants is equal to

$$\int_S -1 dA + \int_{\partial S} 0 ds = 2\pi (-1) \implies A(S) = 2\pi.$$

This is surprising, in the sense that it is irrespective of the side lengths a, b, c! (but it is not *really* surprising, since a pair of pants is just two right-angled hexagons glued together, and the area of a hyperbolic polygon depends only on its interior angles).

Exercise 4.8. Find the area of a hyperbolic n-gon with interior angles $\alpha_1, \ldots, \alpha_n$.

4.4 Gluing pairs of pants

With pairs of pants at our disposal, we can build up different surfaces. Let $G = (V, E)$ be a connected three-regular multigraph on n vertices, where we also allow self loops. For each edge e, let $\ell_e > 0$ be and $\tau_e \in [0, 2\pi]$ be two numbers; we call ℓ the *length* and τ the *twist*. Since G has n vertices, there are $3n/2$ edges, and so $3n$ parameters in total (note that n must be even, since there is an integer number of edges). For every vertex v with edges e_1, e_2, e_3, let S_v be a pair of pants with cuffs of length ℓ_{e_1}, ℓ_{e_2} and ℓ_{e_3}, respectively. The surface S generated by $\left(G, L = \{\ell_e\}_{e \in E}, T = \{\tau_e\}_{e \in E}\right)$ is given by gluing together the pairs of pants S_v: if u is connected to v by an edge e, then the cuffs corresponding to e in both S_v and S_u are glued to each other. There is a matter of orientation to consider here, though. Both cuffs have length ℓ_e, but the boundaries are isometric to circles, and we can rotate them by any angle in $[0, 2\pi]$. This is of course where τ_e comes into play. Each pair of pants is composed of two hexagons. We arbitrarily choose one, and mark the midpoints of the edges a, b, c. This is the origin of the cuffs. Now, when gluing two cuffs together, we glue so that the origin of one is glued to the point at an angle τ_e away clockwise from the origin of the other (convince yourself that in this construction, it does not matter which one is "one" and which one is "the other").

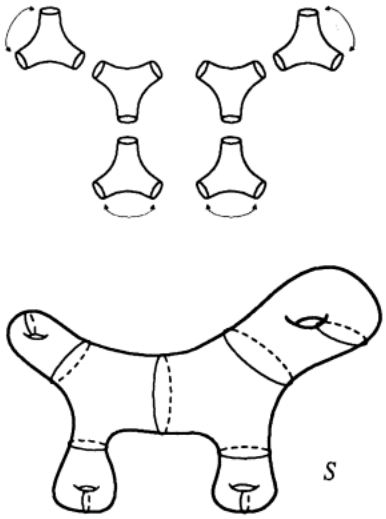


Figure 4.15: Gluing together hyperbolic pair of pants.

What is the genus of the resulting surface? We think of S as the result of gluing together the cuffs of the various pairs of pants, one by one, in some given order. Let H be a spanning tree of G, and start by first gluing the cuffs according to the edges H. At the end, we have a hyperbolic surface which uses all n pairs of pants, but only used up $n-1$ of edges of the graphs. There are $3n/2-(n-1) = n/2+1$ edges remaining, and for each such edge, gluing the cuffs of the corresponding pairs of pants creates a hole in the surface, i.e. increases the genus by 1. So after we have finished gluing them all together, we are left with a compact hyperbolic surface of genus $g = n/2 + 1$.

In other words, in order to create a surface of genus g, all we need is a graph with $n = 2g - 2$ vertices, as well as $6g - 6$ parameters ($3g - 3$ of length parameters, and $3g - 3$ twist parameters). This also works the other way around:

Theorem 4.9. *Let S be a compact hyperbolic surface of genus $g \geq 2$. Then we can find $3g-3$ disjoint closed geodesics $\gamma_1, \ldots, \gamma_{3g-3}$ on it, so that cutting the surface along the γ_i decomposes the surface into pairs of pants.*

In fact, this follows from an even stronger statement:

Theorem 4.10. *Let $g \geq 2$, and let G be a fixed connected cubic graph on $2g - 2$ vertices. Every compact hyperbolic surface S of genus g is generated by (G, L, T) for some L, T.*

The parameters L, T are called the Fenchel-Nielsen coordinates of S (with respect to the fixed graph G), named after Werner Fenchel and Jakob Nielsen.

Remark 4.11. The pair of pants construction gives us yet another way to calculate the area of a compact hyperbolic surface. Each pair of pants has area 2π; since a surface of genus g is made from $2g - 2$ pairs of pants, its area must be $A(S) = 2\pi(2g - 2)$.

4.5 A word on groups

We have already seen that the cusp is the quotient of $\mathbb{H}$ by the map $z \mapsto z + 1$ in the upper half-plane. We have also seen that the Bolza surface is an octagon whose sides are glued to each other - but this too is

actually $\mathbb{H}\backslash\Gamma$ for a group Γ, whose generators send opposite edges of the octagon to each other. In fact, this is true for every compact hyperbolic surface: every such S is equal to $\mathbb{H}\backslash\Gamma$ for some torsion free group of isometries Γ (torsion free = no group element has finite order apart from the identity). This fact lets us perform calculations, or at least approximations, for various quantities on compact surfaces.

5 Diameter via randomness (Lectures 5-6)

The diameter of a metric space is the supremum distance between two of its points. In the space of graphs, different graphs can have vastly different diameters. For example, some 3-regular graphs have large diameters, like the 2-thick cycle, while others, like the 3-regular tree, have a diameter proportional to $\log n$.

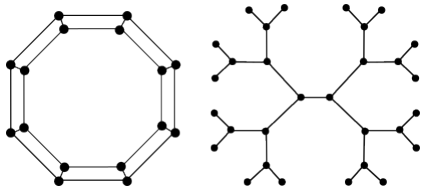


Figure 5.1: The thick cycle with n vertices has diameter $n/4$, while the tree has diameter proportional to $\log n$ (a finite tree is not 3-regular, but the leaves can be matched with each other while keeping the diameter the same)

What about hyperbolic surfaces?

First, from the pairs of pants construction, for every cubic graph G, we can get a compact hyperbolic surface with comparable diameter by considering the surface S generated by G, and, say, lengths 1 and no twists.

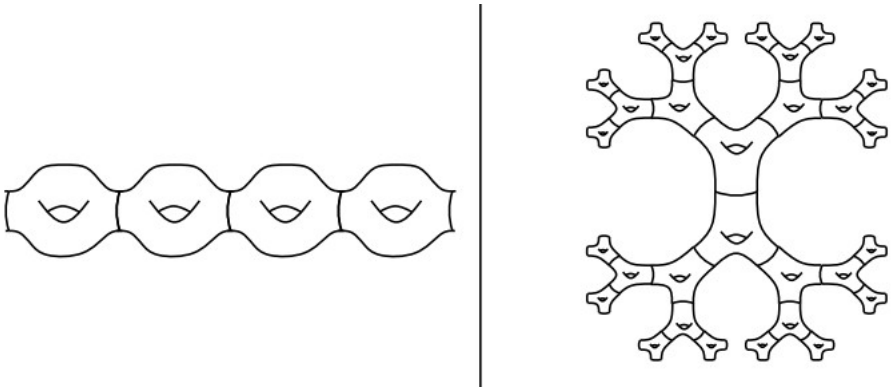


Figure 5.2: Cycles and trees can appear in hyperbolic surfaces as well. Image taken from [4].

And indeed, when all the cuff lengths are constant in the hyperbolic pant decomposition, perhaps there is not much difference between surfaces and graphs.

But this does not have to be the case. For example, as we have seen, the pant leg goes to infinity when the cuff length goes to 0 (this is just a consequence of how a hyperbolic right angled hexagon looks like). So even a simple genus 2 surface can have arbitrarily large diameter:

Figure 5.3: Both easy to cut AND having a large diameter.

Can we have an arbitrarily small diameter? In light of the pair of pants decomposition, perhaps the first instinct would be to take the size of the cuff to infinity, and not to 0, thus avoiding the very long pant legs. Unfortunately, this is wrong: the hexagon is symmetric: if we take $a, b, c \to \infty$, the result will be that $\alpha, \beta, \gamma \to 0$, and our surface will just be squeezed the other way.

This is wrong for another reason, though: if we do this, then sure, the distance along "graph connectivity" will be small, but then there is a big cost to travel to opposite points on the boundary within the same pair of pants. But we may ask - perhaps if we add some twists we can get around this fact?

Well, no, and the attempt is wrong for yet another reason: there is a lower bound on the diameter. As we have seen, a genus g surface S has area $A(S) = 2\pi(g-2)$. If we lift the surface to the hyperbolic plane $\mathbb{H}$, it must be contained in the hyperbolic ball of radius $\operatorname{diam}(S)$ (around any point in the lift). The hyperbolic ball of radius r has area $2\pi(\cosh(r) - 1)$, so we must have

$$2\pi(\cosh(\operatorname{diam}(S)) - 1) \geq 2\pi(2g-2),$$

yielding

$$\begin{aligned}\operatorname{diam}(S) &\geq \cosh^{-1}(2g-1) \\ &\geq \log(2g-1) \\ &= \log(g) + O_g(1).\end{aligned}$$

So similarly to cubic graphs, as $g \to \infty$, the diameter must be at least $\log(g)$. This raises the question:

Question 5.1. *How small can the smallest diameter of a genus g surface be?*

Here is an answer by Thomas Budzinski, Nicolas Curien and Bram Petri.

Theorem 5.2 ([5])**.** *Let D_g be the smallest diameter of a genus g surface. Then*

$$\lim_{g\to\infty} \frac{D_g}{\log g} = 1.$$

In other words, the simple volume bound is asymptotically correct.

Since we already have the lower bound, in order to prove the theorem, all we have to do is find a sequence of surfaces S_g of genus $g \to \infty$ such that $\operatorname{diam}(S_g) = (1 + o(1))\log(g)$. Easier said than done, of course. Any ideas?

I myself do not have any particular surface in mind that satisfies this requirement. But neither did Budzinski, Curien and Petri when they proved the theorem. Instead, they used the good old probabilistic method - generating a hyperbolic surface at random, and showing that with positive probability, the generated surface satisfies this diameter requirement (in fact, as is common, this positive probability is actually asymptotically 1).

The random hyperbolic surface comes from a very simple model. As we saw, given a cubic graph G, we can generate a hyperbolic surface by gluing together the cuffs of pairs of pants according to the edges of the graph, with parameters L and T. A random version of this would involve taking a random triplet (G, L, T), and considering the surface generated from that triplet. All three parameters can be random, and perhaps the easiest way to do this is keep the length and twist parameters constant, while taking G to be random. And this is indeed what we'll do for the proof of this theorem: the pairs of pants will all have cuff length equal to a constant a (so $L = (a, \ldots, a)$), and will be connected to each other with 0 twist (so $T = (0, \ldots, 0)$) according to a uniformly random cubic graph on n vertices (with n even).

The set of all cubic graphs on n vertices is finite, so it is very easy to say "pick a graph uniformly at random". However, unless we use some sort of algorithmic / explicit way of constructing a graph from this distribution, it will be very hard to say something meaningful about its properties.

A standard way of generating a cubic graph is using the half-edge model, also called the "configuration model". We start with n isolated vertices, where each vertex has 3 (ordered) "half-edges":

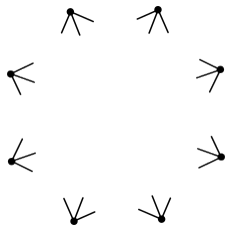


Figure 5.4: In the configuration model, n vertices (8 in this case) each have d half-edges (3 in this case), which are then randomly connected to each other.

We then iteratively pick two unused half-edges and fuse them together. By the end, we have a three-regular graph!

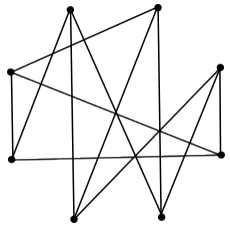


Figure 5.5: A sample graph from the configuration model.

Well, a multigraph, actually, and one that can have self-loops as well. If we were actually interested in 3-regular simple graphs, we would have had to condition on this event not happening. It is possible to prove that the probability of getting a non-simple graph stays bounded away from 0 as $n \to \infty$. For us, this doesn't really matter, since two cuffs of a single pair of pants can be glued together without any problem. Perhaps a more serious problem for our model is that the resulting graph can be disconnected.

Exercise 5.3. Let G_n be a random cubic graph drawn according to the half-edge model. Show that as $n \to \infty$, $\mathbb{P}[G_n \text{ is connected}] \to 1$.

5.1 Warm-up: the diameter of 3 regular graphs

In the configuration model, a randomly generated graph locally looks like a tree. That is, if you pick an arbitrary vertex and look at the subgraph induced by all vertices at a fixed distance d from it, you will get a tree with high probability - without any internal connections whatsoever. Here is one way to see this. Rather than generating all the half-edge pairings in the configuration model all at once, we can generate them one by one, and in any order that we want, as long as we do not "peek into the future" and let our choice be affected by the future, unknown pairings. So starting with the vertex that is being considered, we can first pair its three half-edges, and then pair the half edges of its neighbors, and so on, in a breadth-first

search fashion. For the first move, there are 3 half-edges, and the probability that any of them connect to the same vertex (or the original root) is proportional to $1/n$, so with probability $1 - c/n$, after one step in this "edge reveal" process, we have a tree (albeit a very small one). Next, conditioned on the revealed graph being a tree so far, there are now 6 half-edges to pair up, and again the probability to stay a tree is $1 - c/n$ (although, with a larger c this time).

Doing this k times and taking the product of the probabilities, the probability to be a tree at the end is of the form

$$\left(1 - \frac{c_k}{n}\right)^k.$$

For a constant k, this goes to 1 as n goes to infinity (and in fact we can union bound over all vertices, so this will be true for every vertex). Now, this is a very crude analysis, and a lot of the literature dealing with random graphs involves doing it with much higher precision. For example, one can ask what is the largest k where this still happens with high probability. Obviously, having $k = cn$ is not an option, for even if we replace the large c_k by 1, we would get $\left(1 - \frac{1}{n}\right)^{cn} \approx e^{-c}$, only a constant probability of remaining treelike. But what about $k = \sqrt{n}$? Or $k = \log n$? Or anything in between? More delicately, one can also ask about the number of "defects", i.e. how many times a half-edge is connected to one of the vertices already explored, or about the number of cycles of large length, or other non-tree-like properties of the explored graph (the property depends on that particular application you are investigating).

We say (slightly) more about this, in due time. But for now, to understand where the analysis is going, let us assume that every k-neighbourhood of every vertex is a tree, for the particular choice of $k = \frac{1}{2}\log_2 n + \log_2 \log n$. **This is a wrong assumption!** in real life, a k-neighbourhood of this size *would* have some cycles. However, it is true enough in spirit so as to show the gist of the argument.

Given this assumption, the number of leaves at the end of the tree is equal to

$$\#\text{Leaves} = 3 \cdot 2^k = 3\left(\sqrt{n}\log n\right).$$

This actually implies that the diameter of the graph is, with high probability, bounded by $2k + 1 \approx \log_2 n$. Indeed, if we take any two arbitrary vertices, we can run the vertex exploration for one, and then the vertex exploration for the other. The number of leaves in the tree surrounding the first vertex is of order $\sqrt{n}\log n$, and so is the number of leaves in the tree surrounding the second vertex. If we run the edge reveal process for just one more step, there is a vert high probability of connecting these two leaf sets: the probability to not connect is bounded above by

$$\left(1 - \frac{\sqrt{n}\log n}{n}\right)^{\sqrt{n}\log n} = \left(1 - \frac{\log n}{\sqrt{n}}\right)^{\sqrt{n}\log n} \leq e^{-(\log n)^2} \leq \frac{1}{n^3},$$

say. Union bounding over all n^2 pairs of vertices, we can get from any vertex to another with just $2k + 1 = (1 + o(1))\log_2 n$ steps.

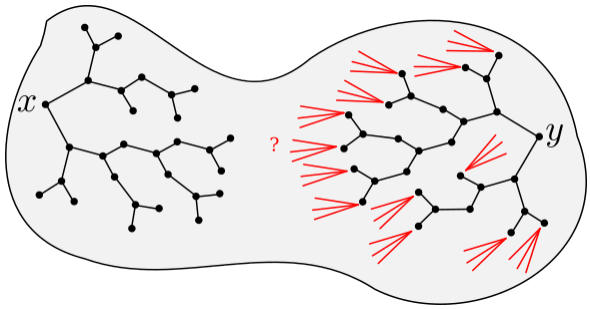


Figure 5.6: The subtree rooted at y has about $\sqrt{n}\log n$ vertices. With high probability, one of its leaves will be connected in the next step to the subtree rooted at x.

5.2 Trees and hyperbolic surfaces

In our case, we are not dealing with trees and abstract vertices, but rather with hyperbolic pairs of pants. But we can do a similar thing: lets reveal the pairs of pants one by one, until we have revealed $\sqrt{n}\log n$ of them. Assume (and again, this is **NOT satisfied** in real life) that the revealed pairs of pants form a tree, without any cycles. A pair of pants with all cuffs equal to a has some fixed diameter, d_a. Since the graph diameter is $\approx 2k = \log_2 n$, the diameter of the resulting surface is bounded above by $d_a \cdot 2k = d_a \log_2 n$. We can then minimize over all a, yielding a hyperbolic surface whose diameter is bounded by

$$D \leq \left(\min_a d_a\right) \log_2 n.$$

Exercise 5.4. Calculate $\min_a d_a$, or at least give lower and upper bounds.

This gives you something which is up to a constant factor optimal. And that's not bad! (though it will turn out that $\min_a d_a$ is not small enough to give the optimal bound of $\log n$ (note also the difference between $\log$ and $\log_2$)). But we can do better than that.

Instead of going from the discrete to the continuous (i.e. using the fact that balls of radius R in random graphs have approximately 2^R vertices), we will go from the continuous to the discrete, and count how many pairs of pants there are at distance R from the already-glued-up surface. This is idealized in the following lemma about distances in hyperbolic trees.

Let T be the infinite hyperbolic surface, obtained by gluing together pairs of pants with cuff lengths a according to the infinite 3-regular tree, without twists. For every pair of pants P, let x_p be the centre of the top part of the pair of pants (i.e. the centre of one of the hexagons; by centre, we mean the point which is equidistant from the boundaries of the pair of pants).

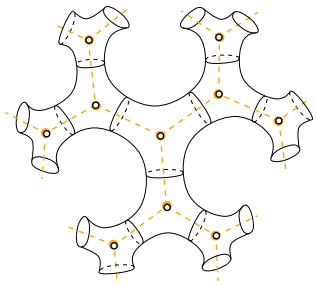


Figure 5.7: The hyperbolic tree with cuff length a. Image taken from [5].

Let $N_a(R)$ be the number of midpoints contained in a ball of radius R around the midpoint of a fixed "root" pair of pants.

Theorem 5.5. *For any $a > 0$, there are constants c_a and $\delta_a \in (0,1)$ such that*

$$N_a(R) \sim c_a e^{\delta_a R} \quad \text{as } R \to \infty,$$

with $\delta_a \to 1$ as $a \to \infty$.

This theorem lets us bound the diameter of the hyperbolic surface directly, without going through the diameter of the cubic graph at its base. Indeed, suppose we have revealed $\sqrt{n}\log n$ pairs of pants in our exploration process. Since the neighourhood of every pair of pants is a (hyperbolic) tree, this means that all of these points are contained in the ball of radius $R + d_a$ around the root. So

$$\sqrt{n}\log n \sim c_a e^{\delta_a (R+d_a)},$$

i.e.

$$\frac{1}{2}\log n + \log\log n = \log c_a + \delta_a R + d_a,$$

implying that

$$R = \frac{1}{\delta_a}\left(\frac{1}{2}\log n + \log\log n - c_a'\right).$$

The diameter of the surface is then bounded by $2R + 2d_a$, which is asymptotic to

$$\frac{1}{\delta_a}\log n.$$

Since $\delta_a \to 1$ as $a \to \infty$, we get a sequence of hyperbolic surfaces which approach the lower bound! So all that remains is to prove the theorem.

Proof. Remember that a pair of pants is the result of gluing together pairs of sides of two right-angled hexagons. For us, since the tree is glued without any twists between the cuffs, we can actually ignore half of the pairs of pants involved. That is, let's consider just the "top half" of the tree: these are the hexagons which are glued to each other according to the tree structure.

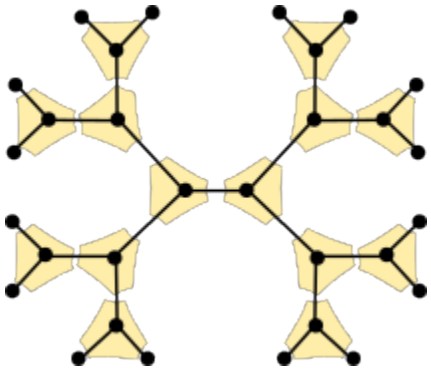


Figure 5.8: The "top" hexagons are glued according to the tree structure.

In the hyperbolic plane, these hexagons are the result of starting with an initial hexagon at (say) the centre of the Poincaré disk, and reflecting about its edges iteratively.

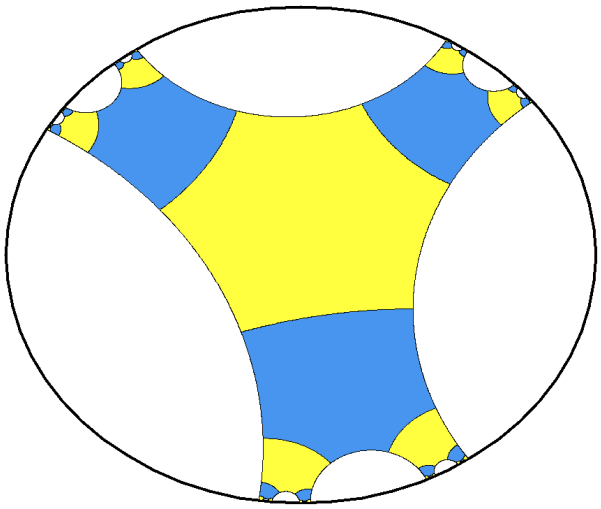


Figure 5.9: The "upper half" of the hyperbolic tree in the Poincaré disk, for $a = \cosh^{-1}(2)$. What is special about this value?

The reason we can restrict to this case is that the shortest path between any two middle points of pairs always stays in this top half: if not, then we could reflect the parts of the path in the bottom part relative to the cut separating the top and bottom, and obtain a path which is of the same length. So when counting how many points are accessible at distance R, we can ignore paths to points that go through the bottom.

We are left with, essentially, counting how many centres of the hexagons there are in the tiling T of the tiling image above.

Here is a sketch on how to do this. Consider the circle of hyperbolic radius R in the Poincaré disk. What is the intersection of this circle with the tree? The number of pairs of pants at the boundary are also roughly proportional to the number of pairs of pants inside the disk of radius R, since every pair of pants at the boundary has two "outgoing" pairs of pants emanating from it. Roughly speaking, each pair of pants at the boundary contributes a constant amount of width, which depends on the parameter a of course. Thus, we can cover the set $\partial B(R) \cap T$ with about $N(R)$ intervals of size c_a. For R large, in the Poincaré disk, the Euclidean disk $\partial B(R)$ is just a circle of size almost 1. Since it has hyperbolic circumference $\ell(\partial B(R)) = 2\pi \sinh(R) \approx \pi e^R$, this means that we can cover the intersection of the boundary of the Euclidean disk with $N(R)$ intervals of size proportional to e^{-R}.

What happens when we take R to infinity? In the Poincaré disk, the tree T reaches the boundary $\partial\mathbb{H}$. In fact, it is a fractal, similar to the Cantor set, and as such has a fractal dimension δ (for our case, we'll take the box counting dimension): if $M(\varepsilon)$ is the number of intervals of size ε required to cover the boundary set, then

$$\delta := \lim_{\varepsilon \to 0} \frac{\log M(\varepsilon)}{\log(1/\varepsilon)}.$$

But we have just seen that we can cover the Euclidean disk (of radius almost 1) with $N(R)$ intervals of size about e^{-R}. This in fact serves as both a lower and upper bound. Thus, up to many "roughly"s and "constants", we get

$$\delta = \lim_{R \to \infty} \frac{\log N(R)}{\log(e^R)},$$

or, in other words,

$$N(R) \asymp e^{R\delta}.$$

This δ is the fractal dimension of the boundary of the tree T. Following our intuition on Cantor sets, you can see that as $a \to \infty$, the boundary of takes a larger and larger "chunk" out of $\partial\mathbb{H}$, and the fractal dimension $\delta \to 1$ as $a \to \infty$.

All of this can be made precise, and δ also has an analytic interpretation: if we look at the sum

$$P(s) = \sum_{\text{centres of pairs of pants } x} e^{-s \cdot d(0,x)},$$

we can define $\delta = \inf\{s : P(s) < \infty\}$. This sort of assumes exponential growth on behalf of $N(R)$, but captures it well: by grouping the centres of pairs of pants x into intervals $[R, R+1]$, the sum is equal, up to constants, to

$$P(s) \approx \sum_{R=0}^{\infty} N(R)\, e^{-sR}.$$

If $N(R)$ has growth faster than e^{sR}, then the sum will be infinite, while if not, the sum will be finite. For $s > \delta$, this then lets us define a measure on the space of points in the Poincaré disk,

$$\frac{1}{P(s)} \sum_{\text{centres of pairs of pants } x} e^{-s \cdot d(0,x)} \delta_x,$$

where δ_x is the Dirac delta distribution. As $s \to \delta$ from above, most of the mass is concentrated on further and further points, which in the limits gives a measure on the boundary. ✹

5.3 The real world

The above was the simple, clean case, which does not correspond to the real world. In truth, if we start exploring the cubic graph until we have revealed $\sqrt{n}\log n$ vertices, we will certainly have "defects" in our perfect tree model: some of the half-edges revealed will connect to pairs of pants which have already been visited. This means that volume of the revealed pairs of pants will have a smaller volume than that of the infinite tree of pairs of pants. Still, it is possible to carry out a more careful analysis, considering how many defects there are and at what times they happened. The technical lemma in the original paper is this.

Lemma 5.6. *For any ε, suppose that there are less than $3/\varepsilon$ defects at time smaller than $n^{1/2-\varepsilon}$, and less than $\log^3 n$ defects until we have discovered $\sqrt{n}\log n$ vertices. Then*

$$R \leq \frac{1}{2}\left(\frac{1}{\delta_a} + \varepsilon\right)\log n.$$

Compare this with the naive bound we had for the trees

$$R = \frac{1}{\delta_a}\left(\frac{1}{2}\log n + \log\log n - c_a'\right).$$

Of course, this lemma relies on the fact that the exploration indeed does not contain a lot of defects, a combinatorial result which involves a slightly more careful analysis of the probability to make an error at every step.

6 The Brooks Makover model (Lectures 7-10)

6.1 The model

The decomposition of a compact hyperbolic surface into pairs of pants suggests that 3-regular graphs may play an important part in the understanding their behaviour. As we have seen, a natural model for generating a random compact hyperbolic surface is to pick a random cubic graph, and use this graph as a base for connecting the pairs of pants. Then, randomly pick (according to some distribution) the $3g-3$ circumference parameters and $3g-3$ twist parameters. Voila! You have a random compact hyperbolic surface. Our previous diameter calculations have done this, with the length and twist parameters being constant (so, non-random).

Before we even begin any analysis, there a couple of complications to this approach.

1. How do we choose the distribution according to which we pick the length and twist parameters? Well, maybe for the twists, it is natural to pick a uniformly random rotation. But is there some canonical distribution for the lengths? If not, then any distribution we choose is arbitrary, and while nice, perhaps won't tell us what a "typical" compact surface looks like. This slightly lowers the motivation for studying such a model, especially if the choice of lengths makes a large difference (the world could have been different; for example, in first passage percolation, there is evidence / hope that the various models are largely indifferent to the particular weight distribution (under some assumptions, of course)).
2. As we mentioned before, if we fix the underlying graph, then when we vary the $6g-6$ twist and length parameters over their entire possible range $\mathbb{R}_{>0}^{3g-3} \times \mathbb{R}^{3g-3}$, we just get all the compact hyperbolic surfaces. In this sense, the randomness of the underlying 3-regular graph is lost. This is too bad, since the structure of random 3-regular graphs is quite well understood - there must be something in the choice of parameters which "invalidates" the structure.

Here is one way to overcome these difficulties: what if, instead of connecting pairs of pants to each other, we connect triangles? This is the Brooks-Makover model, first introduced in [6].

Let $G=(V,E)$ be a 3-regular graph on $2n$ vertices, together with an orientation $\mathcal{O}$: this is a function $\mathcal{O}: V \to E^3$ which assigns for each vertex a cyclic ordering of its edges. We then take $2n$ isometric equilateral hyperbolic triangles - one for each vertex - label their sides by $1,2,3$, and assign side i to the edge $\mathcal{O}(v)_i$. If two vertices are connected by an edge $e=(v_1,v_2)$, then we glue together the sides of the triangles corresponding to that edge. The end result is some surface that we denote by $S^O(G,\mathcal{O})$.

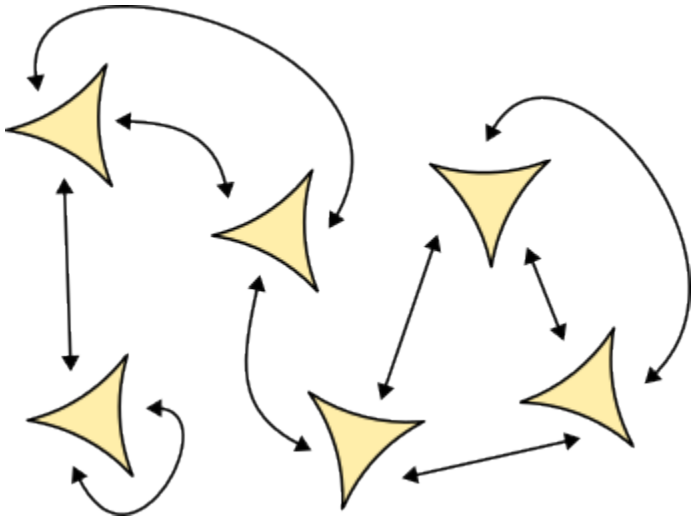


Figure 6.1: Gluing triangles together gives you a surface.

Question: Is the resulting surface even orientable?

Answer: Yes (exercise. Think about how triangle sides must meet each other with respect to the orientation $\mathcal{O}$).

Question: Does this lead to a compact hyperbolic surface?

Answer: Not usually. Suppose that the triangles have angle α, and that the graph G contains a cycle of length m. Then with the right orientation $\mathcal{O}$, the triangles could be connected like so:

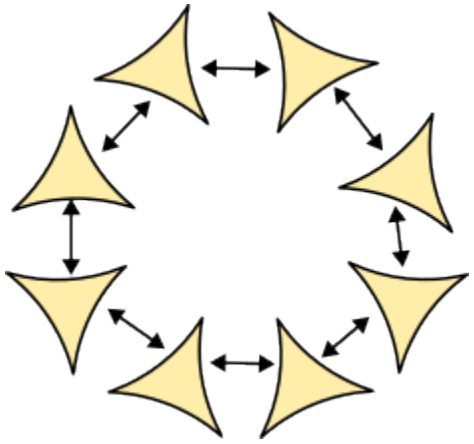


Figure 6.2: A cycle in G corresponds to several triangles meeting at the same point.

The triangles all share a common vertex. The angle around this vertex is αm. Depending on α and m, this angle is generally not going to be 2π. Thus, the resulting surface is hyperbolic almost everywhere, but has some singular points (at most $3n$ of them). Note that the singular points could have angles either greater or lesser than 2π, so we may have different types of singularities at different places.

Question: Are there actually any graphs which can lead to a compact hyperbolic surface in this way? In other words, are there any compact hyperbolic surfaces which can be triangulated by equilateral triangles?

Answer: Yes, but not many. For every $k \geq 7$, it is possible to tile the hyperbolic plane by equilateral triangles, with k triangles meeting at every vertex. This gives rise to a group of symmetries Γ_k. Any freely acting subgroup Γ of Γ_k gives rise to some surface which can be triangulated. But this is certainly not the vast majority of surfaces.

In any case, this is already problematic, and it actually sheds light on a different problem in the model: how do we choose the size of the equilateral triangles? In the Euclidean plane, all equilateral triangles are congruent - the have the same angles, and choosing different ones just means scaling up the entire surface by some factor. But here we have an infinite amount of triangles to choose one. Which one is canonical?

Here, the answer is actually simple: the ideal triangle is canonical. This is the triangle whose edges have infinite length, and whose angles are all 0. Still, it has finite area, $A(T) = \pi$. When we connect the triangles together, we make sure that they are aligned at the "midpoint". What is the midpoint of an infinite line? In the Poincaré disk model, it is the midpoint of the Euclidean circular arc. In the half-plane model, the points correspond to (say) i, $i+1$, and $\frac{i+1}{2}$ (this can also be obtained by using the Möbius isometry).

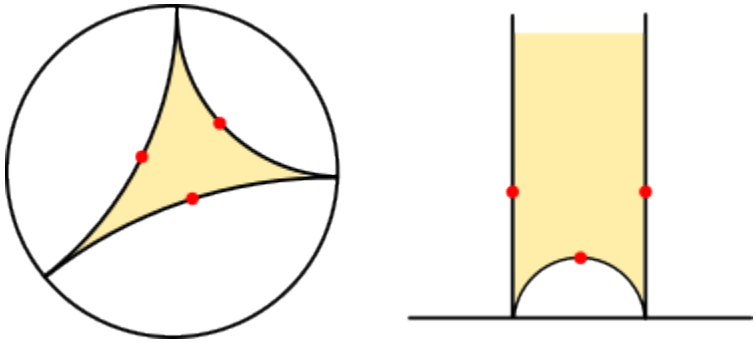


Figure 6.3: The "middle" points of the edges of an ideal triangle, in the disk and half-plane models.

In addition being a natural candidate to the problem of "which triangle to choose", the ideal triangle also simplifies the singularity problems: now, all the singularities are the same, and there is no chance for a vertex to accidentally have the correct angle of 2π: whenever two triangles meet, their vertices will have a cusp singularity. We thus obtain a non-compact hyperbolic surface, which contains some number of cusp singularities. Very roughly, we can think of it like this:

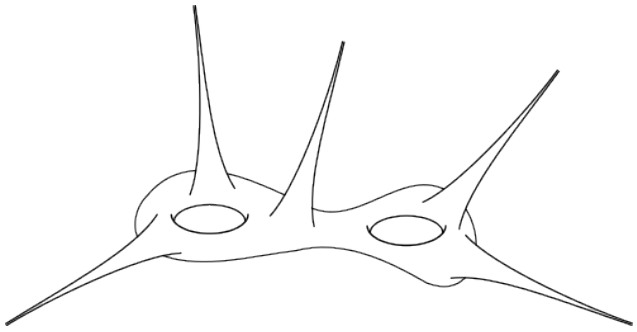


Figure 6.4: A surface with singularities.

If you are anything like me, when you look at this picture, you really want to take a pair of scissors and just snip off the singularities. This will give you a compact surface. And indeed, this is sort of what Brooks and Makover do.

Remark 6.1. The surface S^O is made out of $2n$ glued-together ideal triangles, and therefore has surface area $2\pi n$. If S^O were compact, this would immediately tell us its genus - it would also be the gluing-up of n pairs of pants. However, it has cusps, which correspond to a 0-length boundary in a pair of pants. The genus is therefore given by

$$g = 1 + \frac{n - \#\text{cusps}}{2}.$$

Remark 6.2. The surfaces that you get when gluing together triangles are called Belyi surfaces, named after the Soviet mathematician Gennady Belyi, who studied algebraic curves over the complex numbers (we have previously grazed the connection between the topics: in short, a curve is a one-dimensional complex manifold. This can be treated as a two-dimensional surface with conformal structure, and by the uniformization theorem, there is a unique hyperbolic surface for every conformal equivalence class). Correspondingly, a Brooks-Makover random surface is sometimes called a "random Belyi surface". If we want a "society-neutral" name, perhaps "random triangular gluing model" will suffice, especially since it suggests the obvious generalization of gluing together other polygons.

6.2 Compactification

In our surface, a cusp is formed by concatenating in a cycle m ideal triangle. In the half-plane model, this looks like this:

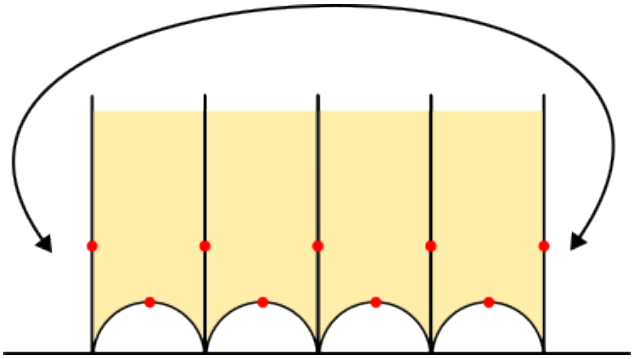


Figure 6.5: Four ideal triangles connected to each other in the half-plane. The leftmost and rightmost sides are also glues together, creating a cusp.

Recall that all ideal triangles are isometric to each other.

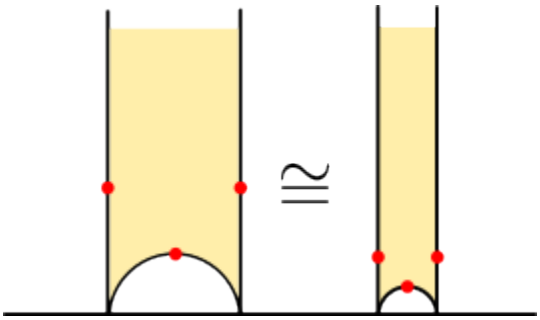


Figure 6.6: There is only one ideal triangle, so the two triangles are isometric to each other.

So for convenience, we can choose the m triangles to span a Euclidean distance of 1 in the x axis. This will determine the height y to which the semicircles go - the larger the k, the more we have to squeeze the triangles in, and so the smaller the y. If there are k triangles, the diameter of the k semicircles is $1/k$, and so $y = 1/2k$. Thus, around each cusp there is a neighborhood which is isometric to the region

$$C^y = \left\{ z \in \mathbb{C} \mid \mathrm{Im}\,(z) \geq \frac{1}{y} \right\} / \left(z \sim z + 1 \right)$$

in the half plane model. We have seen this object before - it is the pseudosphere.

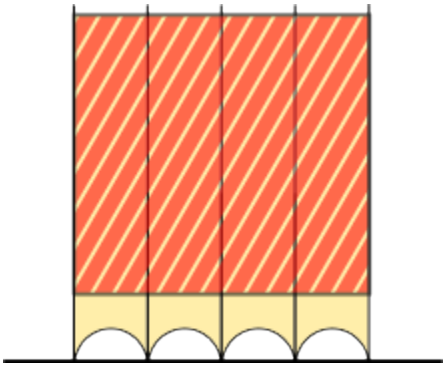


Figure 6.7: The shaded region is the pseudosphere, and can be thought of as a neighborhood of ∞. Each cusp in the surface contains such a sphere, with size depending on the number of triangles glued in the cycle.

Each C^y is conformally equivalent to a punctured disk, i.e. there exists a conformal map from C^y to the punctured disk. In fact, this conformal map extends to the entire disk by a one-point compactification: we superficially add a point p at infinity and and say that p maps to 0, by definition.

Exercise 6.3. Find an explicit conformal mapping from C^y to the punctured disk. Use this to find an explicit hyperbolic metric on D with a cusp at 0.

This lets us define a compact surface from S^O, as follows. Suppose that there are k cusps, with neighborhoods $C_1, \ldots, C_k$. For each cusp i, we replace the conformal map from C_i to the punctured disk by that of $\tilde{C}_i$ (the compactified $C_i \cup \{p\}$) to the non-punctured unit disk. In effect, we cut out C_i and replace it

by a non-punctured disk. This will give us a **compact** complex surface S^C. On this S^C, there are points $p_1, \ldots, p_k$ so that $S^C - \{p_1, \ldots, p_k\}$ is conformally equivalent to S^O (of course, S^C itself is not conformally equivalent to S^O).

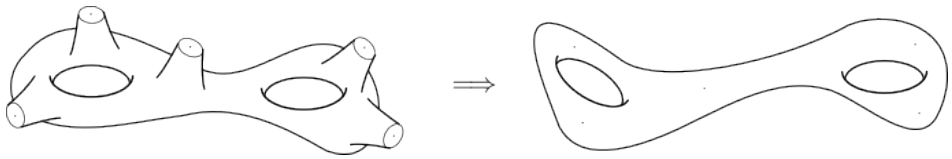


Figure 6.8: A compact surface is obtained by snipping out the cusps and replacing the holes with disks (left). This surface is conformally equivalent to the original S^O, and is not (yet) hyperbolic. By the uniformization theorem, there exists a hyperbolic surface conformal to it (right); this is S^C.

By the uniformization theorem, if the genus is greater than 1, there exists a unique hyperbolic metric on S^C. Voila! We have compactified S^O, and obtained a (canonical) compact surface from the graph G. The main question now is how to analyze the geometric properties of the compact surface, and specifically to relate them to the properties of G. This will be done in two steps:

1. The geometric properties of S^C can be controlled by those of S^O.
2. The geometric properties of S^O can be controlled by the geometric properties of G.

6.3 Relating S^C to S^O

Let's look at the cusp neighborhood C^{y_i}. It has a single boundary curve, $y = \frac{1}{y_i}$. This is a called a "horocycle" (strictly speaking, a horocycle is a curvature 1 curve whose orthogonal geodesics all converge to the same ideal point. But we don't need that here - we'll just use lines parallel to the real axis). The length of this curve is

$$\int_{x=0}^{x=1} \frac{1}{1/y_i} = y_i.$$

Thus, the more triangles there are which meet the cusp (i.e. the larger m), the larger y_i is, and the larger the length of the horocycle. The cusp is "larger" in some sense (well, in a very easy sense - it has larger area). The main requirement which lets us relate S^C and S^O is that S^O has large cusps.

Definition 6.4. We say that S^O has cusps of length $\geq L$ if we can choose the cusp neighborhoods C_i to be disjoint such that $y_i \geq L$ for each i.

Theorem 6.5. *For every ε, there exists numbers L, r and y such that if the cusps of S^O have length $\geq L$, then outside the union of cusp neighborhoods $\cup C^y \subseteq S^O$ and outside of the union of balls $B_r(p_i) \subseteq S^C$, the metrics ds^2_C and ds^2_O satisfy*

$$\frac{1}{1+\varepsilon} ds^2_O \leq ds^2_C \leq (1+\varepsilon)\, ds^2_O.$$

Further, the image of the neighborhood C^y is contained in $B_r(p_i)$. All numbers L, r, y increase to ∞ as $\varepsilon \to 0$.

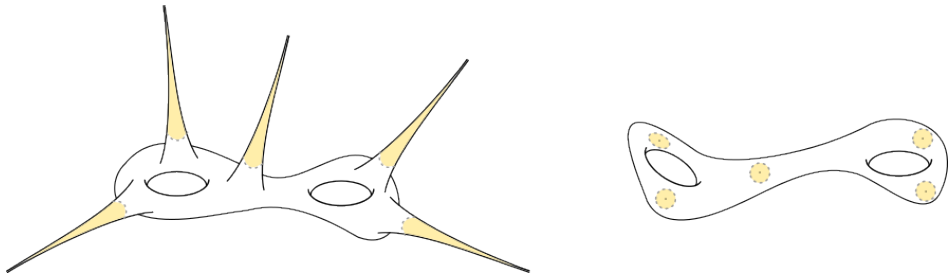


Figure 6.9: Outside the cusp neighbourhoods and circles, the hyperbolic metric on the two surfaces is almost the same.

This is the main the geometric content comparing the two surfaces. We will not prove it here, but rather take it as a given, since the proof is a bit too technical, a bit too low-level, and in any case relies on a result from differential geometry relating the metric and curvature which we will not prove. The basic intuition is this: of course, the hyperbolic metric on the cusp is going to be very different from the hyperbolic metric on the ball. But if the cusp is large, then there is a lot of time until the effect is actually noticeable, it is really only noticeable near the centre. We can imagine starting with the ball metric on the boundary, then slowly changing it into the cusp metric as we progress towards the centre. This metric will not be hyperbolic, of course. Since the curvature depends on the derivatives of the change of metric, if there is enough space we can do this while changing the curvature by only a small amount; a generalization of the Ahlfors-Schwarz-Pick lemma (a holomorphic function from $\mathbb{H}$ to a hyperbolic surface is contracting) then states that the original metrics are not much different from this intermediate metric.

The theorem allows us to compare between lengths of curves between the two surfaces.

Theorem 6.6. *For every $\varepsilon > 0$, there exists an L such that if S^O has cusps of length $\geq L$, then the shortest closed geodesics of S^O and S^C satisfy*

$$\operatorname{syst}\left(S^C\right) \geq \frac{1}{1+\varepsilon}\operatorname{syst}\left(S^O\right).$$

Proof. In fact, we can show something stronger. For every simple closed geodesic γ in S^C, there is a simple closed geodesic γ' in S^O whose image in S^C is homotopic to γ, and such that

$$\ell(\gamma) \geq \frac{1}{1+\varepsilon}\ell(\gamma').$$

The theorem follows from this, since

$$\operatorname{syst}\left(S^C\right) = \inf_{\gamma \subseteq S^C} \ell(\gamma) \geq \inf_{\gamma \subseteq S^C} \frac{1}{1+\varepsilon}\ell(\gamma') \geq \inf_{\eta \subseteq S^O} \frac{1}{1+\varepsilon}\ell(\eta) = \frac{1}{1+\varepsilon}\operatorname{syst}\left(S^O\right).$$

Given our ε, we get an L, an r and a y as in Theorem 6.5. First, let's think about the simplest case, where $\gamma \subseteq S^C$ completely avoids the balls $B_r(p_i)$.

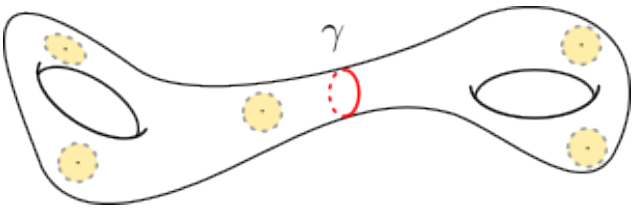


Figure 6.10: The geodesic γ only stays in places where the metric is comparable.

Then γ only stays in places where the metric of S^C is comparable to that of S^O: indeed, when measured under ds_O, the length can decrease a factor at most $\sqrt{1+\varepsilon}$ (the preimage couldn't have come from the cusps, since their image is contained in the ball). Of course, γ might not be a geodesic in S^O, but the geodesic γ' homotopic to it can only be shorter. So

$$\ell(\gamma') \leq \sqrt{1+\varepsilon}\ell(\gamma)$$

in this case.

The actual interesting case is when γ ventures into $B_r(p_i)$, so that the metrics become incomparable. In this case, we have to nudge γ a bit. The idea is to surround the balls $B_r(p_i)$ by much larger balls $B_{r_2}(p_i)$. If γ crosses through $B_{r_2}(p_i)$ and has to go through $B_r(p_i)$, we can push it out of the way by following the boundary of $B_r(p_i)$:

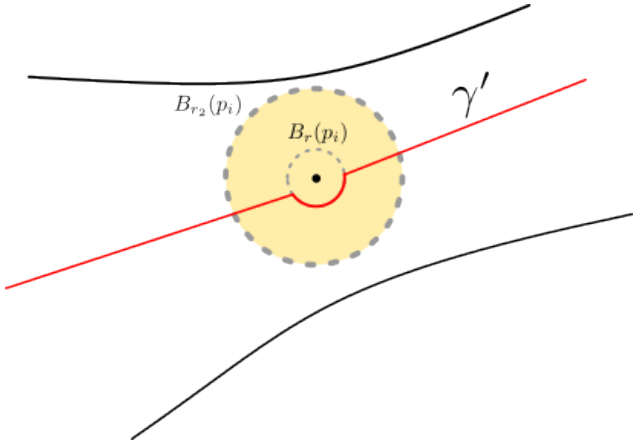


Figure 6.11: The geodesic γ has to go through the ball $B_r(p_i)$. Since we have no idea what goes on in there, we nudge it a bit so that it stays out. The new curve is no longer a geodesic, of course.

When $r_2 >> r_1$, this nudge is negligible compared to the length of the rest of the geodesic, and only contributes at most a multiplicative $(1+\delta)$ increase (where $\delta = \ell(\partial B_r(p_i))/(2r_2 - r)$). Since, when we take $\varepsilon \to 0$ we have that $r \to \infty$, we can therefore choose a much smaller ε than the one given to us, giving us r_2 arbitrarily large, so δ can be arbitrarily small. If γ is a simple closed geodesic and $B_{r_2}(p_i)$ is homeomorphic to a disk, it cannot be contained strictly inside B_{r_2}. So it must go in and out of $B_{r_2}(p_i)$, and after shifting it we obtain a curve γ' which does not go into any $B_{r_1}(p_i)$. Now we can apply the reasoning as above and get another $\sqrt{1+\varepsilon}$ factor, finishing the proof.

(What if $B_{r_2}(p_i)$ is not homeomorphic to a disk? Well, then it is very close to being one, in that $B_{r_2/(1+\varepsilon)^{3/2}}(p_i)$ is a disk. See [7] for details). ✹

Exercise 6.7. What can be said about the other direction? Is it true that $\text{syst}(S^O) \geq \frac{1}{1+\varepsilon}\text{syst}(S^C)$?

Theorem 6.8. *For every $\varepsilon > 0$, there exists an L such that if S^O has cusps of length $\geq L$, then the Cheeger constants of S^O and S^C satisfy*

$$h(S^C) \geq \frac{1}{1+\varepsilon}h(S^O).$$

Proof. Let γ be a curve in S^C which divides it into two parts, $A \cup B = S^C$, and which realizes the Cheeger constant:

$$h\left(S^C\right) = \frac{\ell\left(\gamma\right)}{\min\left(\mathrm{Vol}\left(A\right), \mathrm{Vol}\left(B\right)\right)}.$$

We assume for now that γ is not null-homotopic.

Similarly to the systole theorem, the happy case is when γ doesn't pass through any $B_r\left(p_i\right)$. In this case, the length of the curve can increase by a factor at most $\sqrt{1+\varepsilon}$ from that on S^O, while the volume of $S^C \setminus \cup B_r\left(p_i\right)$ can decrease by a factor of no more than $1+\varepsilon$. What about the volume of the balls themselves? It can be shown that the balls have a larger volume under dS_O than in dS_C (this is done via the Schwarz-Pick lemma, which says that a holomorphic map of the unit disk into itself always decreases the distances in the hyperbolic metric, i.e. $\left|\frac{f(z_1)-f(z_2)}{1-\overline{f(z_1)}f(z_2)}\right| \leq \left|\frac{z_1-z_2}{1-\overline{z_1}z_2}\right|$; the hyperbolic distance in the disk model is $2\tanh^{-1}\left(\left|\frac{z_1-z_2}{1-\overline{z_1}z_2}\right|\right)$). So in this happy case, we would accrue an error of $\left(1+\varepsilon\right)^{3/2}$, so that

$$h\left(S^C\right) \geq \frac{1}{\left(1+\varepsilon\right)^{3/2}} \frac{\ell\left(\gamma\right)}{\min\left(\mathrm{Vol}\left(A\right), \mathrm{Vol}\left(B\right)\right)} \left(\text{in } S^O\right) \geq \frac{1}{\left(1+\varepsilon\right)^{3/2}} h\left(S^O\right).$$

When γ goes through $B_r\left(p_i\right)$, we would like to use the same trick as before: we look at a larger ball of radius r_2, and nudge γ around $B_r\left(p_i\right)$. As we saw before, this can increase the length by at most a factor of $1+\delta$, which we can choose to be arbitrarily small for large L. But now we have a problem: when we nudge γ to $\tilde{\gamma}$, it changes the volumes as well. We might find ourselves in a situation where the volume of one set is decreased considerably:

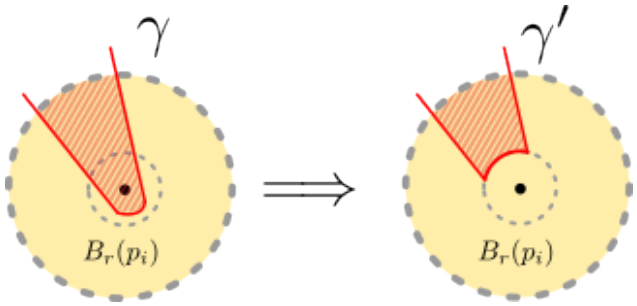


Figure 6.12: The volume of the sets separated by γ can change considerably when we move the curve to wrap around $B_r\left(p_i\right)$. In fact, the intersection of the new set with $B_{r_2}\left(p_i\right)$ may be an arbitrarily small fraction of the intersection of the old set with $B_{r_2}\left(p_i\right)$.

This can, on the face of it, change the Cheeger constant. However, we are actually given a choice of how to nudge γ around $B_r\left(p_i\right)$. If we go around the other direction, we end up adding the volume of $B_r\left(p_i\right)$ instead of subtracting it:

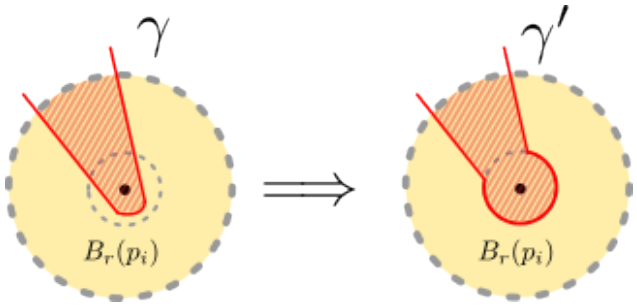


Figure 6.13: We can always choose to nudge γ so that the volume is added to the smaller set.

If we always nudge γ so that the volume of $B_r(p_i)$ is added to the smaller of the two sets $A_i = A \cap B_{r_2}(p_i)$ and $B_i = B \cap B_{r_2}(p_i)$, then the volume of the larger set cannot be changed by much: if (say) A_i was the larger of the two, then $\mathrm{Vol}(A_i) \geq \frac{1}{2}\mathrm{Vol}(B_{r_2}(p_i))$, and the new set $\tilde{A}_i$ separated by $\tilde{\gamma}$ has its volume bounded by

$$\mathrm{Vol}\left(\tilde{A}_i\right) \geq \left(1 - 2\frac{\mathrm{Vol}(B_{r_1}(p_i))}{\mathrm{Vol}(B_{r_2}(p_i))}\right)\mathrm{Vol}(A_i).$$

(the same holds trivially for $\tilde{B}_i$). Adding up the volumes for all i, we obtain that the volumes change a multiplicative factor of no more than $(1-\tilde{\varepsilon})$, which can be made as small as we wish by controlling the ratio of radii r_1/r_2. Thus,

$$h\left(S^C\right) = \frac{\ell(\gamma)}{\min(\mathrm{Vol}(A), \mathrm{Vol}(B))} \geq \frac{1}{1+\varepsilon}\frac{\ell(\gamma')}{\min(\mathrm{Vol}(A), \mathrm{Vol}(B))} \text{ (in } S_O) \geq \frac{1}{1+\varepsilon}h\left(S^O\right).$$

Exercise 6.9. Prove the case where γ is null-homotopic.

Exercise 6.10. What can be said about the other direction? Is it true that $h\left(S^O\right) \geq \frac{1}{1+\varepsilon}h\left(S^C\right)$?

6.4 Relating S^O to G.

Theorem 6.11. *There exists a constant c such that*

$$h\left(S^O\right) \geq c \cdot h(G)$$

for all 3-regular G.

Proof. Let η be a set of curves in S^O which achieves the Cheeger constants, i.e it separates S^O into two parts S_1 and S_2

$$h\left(S^O\right) = \frac{\ell(\eta)}{\min(\mathrm{Vol}(S_1), \mathrm{Vol}(S_2))}.$$

We can always assume that it is S_1 that minimizes the denominator. It can be shown (see [8]) that for a hyperbolic surface with cusps, a collection of curves which enclose a given area and minimizes the perimeter is made of either

1. Horocycles around cusps.
2. Geodesics, or neighboring curves to geodesics (curves of constant distance to geodesics). These have curvature ≤ 1.

The first case is easy to deal with: suppose a horocycle sits at height α in the half-plane model. Then the area of the cusp above it is

$$\text{Area} = \int_0^1 \int_\alpha^\infty \frac{1}{y^2} dx dy = \int_\alpha^\infty \frac{1}{y^2} dy = \frac{1}{\alpha},$$

while the its length is

$$\ell = \int_0^1 \frac{1}{\alpha} dx = \frac{1}{\alpha}.$$

If the minimum volume is the cusps, then the Cheeger constant is given by $\frac{\sum \frac{1}{\alpha_i}}{\sum \frac{1}{\alpha_i}} = 1$. But the minimum volume cannot be the interior, for we can always push up the boundaries along the cusps, to both increase the minimum area and decrease the perimeter, until the volumes are equal. Since a 3-regular graph has Cheeger constant no more than 3, the result follows.

So we only have to deal with geodesics and neighbors of geodesics. The idea here is actually rather simple: the curve η cannot go through just a tiny bit of triangle - if it starts going in a triangle, it must go through a lot of it. Thus, the length $\ell(\eta)$ is at least proportional to the number of triangles through which it passes. This lets us consider the Cheeger inequality in the graph.

Here is a sketch of this argument; in this sketch we'll only deal with such curves that stay inside the "compact" part of the manifold, so for simplicity, we'll imagine that S^O is made by gluing together not n ideal triangles, but rather n truncated triangles:

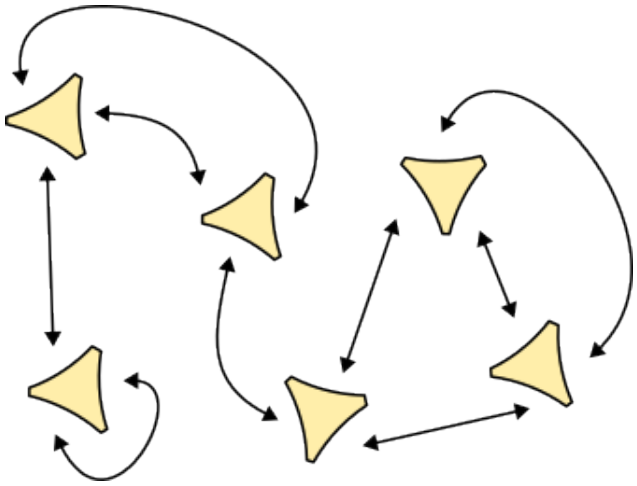


Figure 6.14: Imagine connecting truncated triangles (i.e. hexagons) instead of ideal triangles. Formally, the surface is now a compact hyperbolic surface with perhaps several boundary components. We assume that all curves stay away from the boundary.

We can pick a fundamental domain for S^O in $\mathbb{H}$ that is also comprised of these truncated triangles, and the graph G is just the connectivity graph between them. Denote the set of triangles that meet η by C, and by A and B the triangles which are contained in S_1 and S_2, respectively.

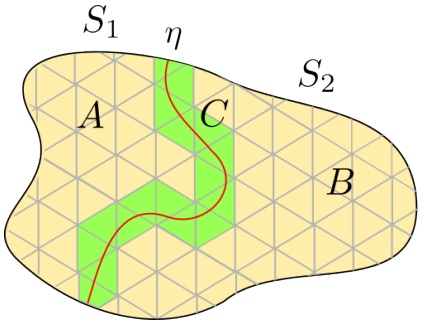


Figure 6.15: The set of triangles is partitioned into three parts, depending on the intersection with S_1, S_2, and η. Triangles intersection only S_1 are in A, triangles intersection only S_2 are in B, and triangles intersecting both are in C (these must meet η).

Then the length $\ell(\eta)$ is proportional to $\#C$. To see this, cover a truncated triangle by a finite number M of small balls $\{B_\delta(x_k)\}_{k=1}^M$, and repeat this covering for all triangles. If η intersects a ball $B_\delta(x)$, then the length of $\eta \cap B_{2\delta}(x)$ is at least some constant K (think about the geodesic case, then extend to bounded curvature. The intuition of the plane is ok here).

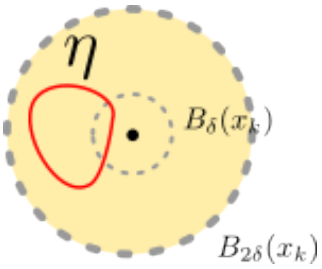


Figure 6.16: If the curve η intersects $B_\delta(x_k)$ at all, then it must have a large intersection with $B_{2\delta}(x_k)$. This is because its curvature is bounded from above - it cannot "close in" on itself too quickly. In the best case, when its curvature is 0 (i.e. it is a geodesic) then it must leave $B_{2\delta}(x_k)$, and so its intersection with it has length at least δ.

From compactness, there is a number L_1 so that every point in S^O is covered by at most L_1 balls of radius δ (and so also balls of radius 2δ); also, a ball of radius 2δ can only be in at most L_2 different triangles (irrespective of G). So

$$\ell(\eta) \geq \frac{|\{B_\delta(x_k) \cap \eta \neq \emptyset\}| \cdot K}{L_1} \geq \frac{K}{L_1 L_2} |C|.$$

So there is some constant c such that

$$\ell(\eta) \geq c \cdot |C|.$$

Assume first that $|C| \geq |V|/10 (= 2n/10)$. Then

$$\frac{\ell(\eta)}{\mathrm{Vol}(S_1)} \geq \frac{c \cdot |C|}{\mathrm{Vol}(S^O)} = \frac{c \cdot |C|}{2\pi n} \geq \tilde{c},$$

i.e. the Cheeger constant is greater than a constant. As before, we are done in this case.

On the other hand, if $|C| \leq |V|/10$, we look at the following partition of G: we take as V_1 the set of all vertices corresponding to $A \cup C$, V_2 the set corresponding to B.

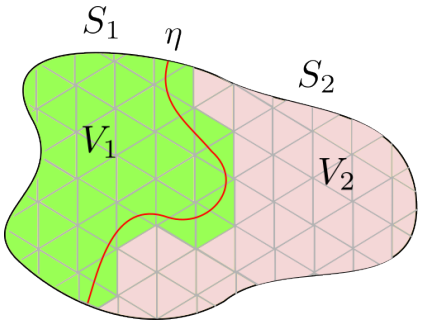


Figure 6.17: A natural way to partition G from the partition of S^O.

Since the Cheeger constant of G is the infimum over all partitions, we have

$$h(G) \leq \frac{\partial(V_1, V_2)}{\min(|V_1|, |V_2|)}.$$

The number of edges $\partial(V_1, V_2)$ is at most $3|C|$, while the size $|S_1|$ is at most $\pi|A \cup C|$, so

$$\frac{\ell(\eta)}{\mathrm{Vol}(S_1)} \geq c\frac{|C|}{\pi|A \cup C|} \geq c\frac{\partial(V_1, V_2)}{|V_1|}.$$

What to do about the minimum? This only takes care of one part of the denominator in the definition of the Cheeger constant for G. Could it be that $|V_2| \leq |V_1|$? Yes, but by assumption,

$$|A| \leq |B| + |C|$$

(since $\mathrm{Vol}(S_1) \leq \mathrm{Vol}(S_2)$), and

$$|C| \leq (|A| + |B| + |C|)/10,$$

so

$$4|C| \leq |B|,$$

which leads to

$$|A| \leq |B| + \frac{1}{4}|B| = \frac{5}{4}|B| \implies |B| \geq \frac{4}{5}|A|.$$

So we can only be off by a constant factor, and we get

$$\frac{\ell(\eta)}{\mathrm{Vol}(S_1)} \geq \tilde{c}h(G)$$

in this case. ✹

Exercise 6.12. Show that there exists a constant C such that

$$h(S^O) \leq Ch(G).$$

Exercise 6.13. Take care of the case where geodesics might enter a cusp area.

6.5 The properties of a random 3-regular graph

Up until now, we have related the properties of the hyperbolic surface to that of its underlying graph. It is now time to analyse how a random cubic graph looks like. We have already talked about the diameter of random cubic graphs when discussing the smallest diameter of a hyperbolic surface. We saw (in an exercise, so perhaps not all of us saw...) that a random graph on $2n$ vertices in the configuration model will be connected with probability approaching 1 as $n \to \infty$. So the two remaining important questions are now:

1. What is the Cheeger constant of G?
2. For a given L, what is the probability that S^O has cusps $\geq L$? As we saw above, this will allow us to control the Cheeger constant of S^C using the Cheeger constant of G.

We start with the Cheeger constant. A constant bound from below is possible by a direct combinatorial proof.

Theorem 6.14. *There exists a constant C so that if G_{2n} is drawn according to the half-edge model,*

$$\lim_{n\to\infty} \mathbb{P}\left[h\left(G_{2n}\right) > C\right] = 1.$$

Proof. Showing that the Cheeger constant is large can often be rather difficult. Often, it involves bounding the spectral gap (using whatever complicated technique required). In this case, there is a straightforward combinatorial proof. Simply put, if the Cheeger constant is small, then there must be a set which only has a small number of edges going outside of it. But there is only a limited number of ways to connect the set using a small number of edges (when compared with the number of ways to do so using a large number of edges).

More formally, for an integer $u \in [0, n]$ and $s \in [0, Cu]$, let $P\left(u, s\right)$ be the probability that a configuration contains a set of vertices $U \subseteq V$ with $|U| = u$ and with s edges between U and $V \backslash U$. Under such an event, we have $h\left(G\right) \leq \frac{s}{u} \leq C$, and indeed, the event that $h\left(G\right) \leq C$ is the union of all such events. The theorem will then follow if we can show that

$$\sum_{u=1}^{n} \sum_{s=0}^{Cu} P\left(u, s\right) = o\left(1\right).$$

For a given choice of u and s, we have

$$P\left(u, s\right) = \binom{2n}{u} \binom{3u}{s} \binom{3\left(2n-u\right)}{s} s!\left(3u-s-1\right)!!\left(3\left(2n-u\right)-s-1\right)!! \frac{1}{\left(6n-1\right)!!}.$$

(here, for k odd, $k!! = k \cdot (k-2) \cdot (k-4) \cdots 3 \cdot 1$). Let's break this up:

1. The term $\binom{n}{u}$ is the number of ways to choose a subset of size u from the $2n$ vertices.
2. The term $\binom{3u}{s}$ is the number of ways to choose half-edges in the set that we have chosen.
3. The term $\binom{3(2n-u)}{s}$ is the number of ways to choose the remaining half-edges, which will connect to our s half-edges.
4. The term $s!$ is the number of ways to match the s half-edges in u to the s half-edges not in u.
5. The term $(3u-s-1)!!$ is the number of ways to match the edges in u which do not leave u amongst themselves.

6. The term $(3(2n-u)-s-1)!!$ is the number of ways to match the edges not in u amongst themselves.
7. The term $\frac{1}{(6n-1)!!}$ is the total number of ways to match the half-edges.

This expression is monotone increasing in s:

$$\begin{aligned} P(u,s) &\propto \frac{1}{s!\,(3u-s)!}\frac{s!}{s!\,(3(2n-u)-s)!}(3u-s-1)!!\,(3(2n-u)-s-1)!! \\ &= \frac{1}{s!\,(3u-s)!!\,(3(2n-u)-s)!!} \end{aligned}$$

So

$$\frac{P(u,s)}{P(u,s-1)} = \frac{(s-1)!\,(3u-s+1)!!\,(3(2n-u)-s+1)!!}{s!\,(3u-s)!!\,(3(2n-u)-s)!!}.$$

The ratio of two double factorials can be calculated to be $\frac{k!!}{(k-1)!!} = \Theta\left(\sqrt{k}\right)$. So this is about equal to

$$\frac{P(u,s)}{P(u,s-1)} \approx \frac{\sqrt{3u-s+1}\sqrt{3(2n-u)-s+1}}{s}.$$

Since $s \leq Cu$, this is greater than 1 for C small enough. This means that it's enough to show that

$$\sum_{u=1}^{n} u \cdot P(u,Cu) = o(1).$$

For any u smaller than some fixed constant (say, $u \leq 100$), we have $P(u,Cu) = o(1)$ (as $n \to \infty$). Indeed,

$$P(u,Cu) \approx C_u n^u (6n-3u)^s (6n-3u-s-1)!!\frac{1}{(6n-1)!!}.$$

The ratio $\frac{(6n-3u-s-1)!!}{(6n-1)!!}$ is of order at most $\frac{1}{n^{3u}}$, which is enough to defeat the n^{u+s} at the beginning of the right-hand side. For other u, we can't ignore the constants any longer; it can be shown that $P(u,Cu) = o\left(n^{-2}\right)$ when C is taken to be small enough. Thus

$$\sum_{u=1}^{n} u \cdot P(u,Cu) \leq n\sum_{u=1}^{n} o\left(n^{-2}\right) = o(1)$$

as needed. ●

Exercise 6.15. Show that $\frac{(2n)!!}{(2n-1)!!} \approx \sqrt{\pi n}$ asymptotically. What is $\frac{(2n+1)!!}{(2n)!!}$?

Exercise 6.16. Show that $P(u,Cu) = o\left(n^{-2}\right)$.

6.6 The large cusps condition

Right now we are in very good shape. We have shown that with high probability, a random 3-regular graph has Cheeger constant bounded from below by a constant. In turn, the Cheeger constant of the open surface S^O is also bounded from below by a constant. And we have shown that if S^O has large cusps, then the Cheeger constant of the compact surface S^C is comparable to that of S^O. Diagrammatically, under the large cusp condition,

$$h(G) > c \implies h\left(S^O\right) > c \implies h\left(S^C\right) > c.$$

All that is left to do is show that S^O satisfies the large cusp condition for a prescribed L (i.e. we choose some small $\varepsilon > 0$, say $\varepsilon = 0.1$, at hope that S^C has cusps $\geq L$, where L is the number needed for the comparison theorems to work). Luckily for us (and we wouldn't be talking about this model otherwise), this is true.

Theorem 6.17. *Let $L > 0$. Then*

$$\lim_{n\to\infty} \mathbb{P}\left[S^O \text{ has cusps} \geq L\right] = 1.$$

How do we get a cusp, anyway? A cusp is a collection of triangles which all meet at a point, and which are all connected together in a cycle. So every cusp originates from some cycle in the graph G_{2n}. (The opposite is not true - some cycles in G_{2n} do not give rise to cusps:

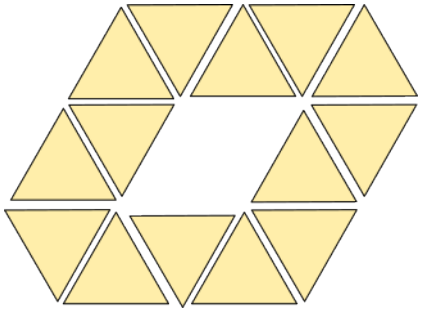


Figure 6.18: This cycle of triangles does not correspond to a cusp.

In fact, if we consider the orientation $\mathcal{O}$ of the graph, then a cusp originates from cycles in which either all the choice of "which edge to go to" are left turns, or all the choices are right turns (and since we can reverse the direction in which we traverse the cycle, we can assume that all the turns are left turns)). For each such cusp, we can choose the "canonical horocycle" as the horocycle with which we measure the lengths of the cusps. This is simply the horocycle at height 1. So to show that S^O has cusps of size $\geq L$, it would suffice to show that with high probability, G_{2n} does not have short left turn cycles. Unfortunately, this is not true.

Theorem 6.18. *For every L, there is a constant c_L such that*

$$\lim_{n\to\infty} \mathbb{P}\left[G_{2n} \text{ has LHT cycles of length} = L\right] \geq c_L.$$

Proof sketch. In fact, it is possible to show something stronger: as $n \to \infty$, the number of cycles X_L of length L (not necessarily LHT cycles) is asymptotically distributed like a Poisson distribution with mean

$$\lambda_L = \frac{2^L}{2L}.$$

Here is a very light sketch. We start with computing the expectation of the number of cycles. Given L vertices in some order, what is the probability that they form a cycle? If there is already a half-edge incoming into a vertex, there are 3 choices for half-edge to which it connects, and then another 2 choices for the outgoing half-edge from that vertex. So there are 6^L choices overall. The probability that these particular ones are actually matched is

$$\frac{1}{(6n-1)}\frac{1}{(6n-3)}\cdots\frac{1}{(6n-(2L-1))},$$

so the probability that these L in order vertices form a cycle is $\frac{6^L}{(6n-1)(6n-3)\cdots(6n-(2L-1))}$. Now we choose all such sets ($2n\,(2n-1)\,(2n-2)\cdots(2n-L+1)$ options), divide by rotational symmetry and direction choice ($2L$ choices), and get that

$$\mathbb{E}\left[X_L\right] = \frac{1}{2L} 2n\,(2n-1)\cdots(2n-L+1)\,\frac{6^L}{(6n-1)\,(6n-3)\cdots(6n-(2L-1))}.$$

Since L is constant, this is asymptotic to

$$\frac{2^L}{2L}.$$

So at least we got the expectation right. The higher moments are left as an exercise.

Exercise 6.19. Show that the higher moments of X_L converge to those of the Poisson distribution, i.e.

$$\mathbb{E}\left[X_L\,(X_L-1)\cdots(X_L-k-1)\right] = \lambda_L^k.$$

Using the same technique for oriented cycles, we add another factor of 2^{-L+1} for each cycle (either all left hand turns, or all right-hand turns), and the number of LHT cycles is asymptotic to a Poisson random variable with parameter $\frac{1}{L}$. For a fixed L, there is a constant probability that this variable is non-zero, and the theorem is proved. ✹

The simple approach does not work. Still, not all is lost. True, there may be some (maybe even many) short cycles in G_{2n}, so that taking the canonical horocycle wouldn't work. But we don't have to take the canonical horocycles when figuring out whether or not the surface has cusps of length $\geq L$. Consider even just a single ideal triangle: if it is connected in such a way that the bottom is "filled in" with other triangles with no cusps, the horocycle can be chosen much lower:

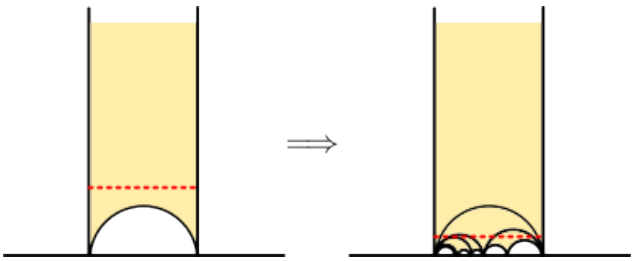


Figure 6.19: If the bottom of an ideal triangle is filled in with more ideal triangles which aren't glued to each other in odd ways, then the horocycle can be chosen to be lower.

In other words, if the cusps are all isolated and far away from each other, even if they themselves originated from a small number of triangles, we might still have cusps of length $\geq L$.

Proof of Theorem 6.17. Let L_1, L_2, d be positive integers, and let $Q_n\left(L_1, L_2, d\right)$ be the event that G_{2n} has two closed paths γ_1 and γ_2 of lengths L_1 and L_2 respectively, that are at distance d apart. Then

$$\lim_{n\to\infty} \mathbb{P}\left[Q_n\left(L_1, L_2, d\right)\right] = 0.$$

To see this, note that the number of cycles of length L_1 is asymptotically Poisson, so for every ε there is an $N\left(\varepsilon\right)$ such that with probability $1-\varepsilon$, the number of paths of length L_1 is less than $N\left(\varepsilon\right)$. Then

$$\mathbb{P}\left[Q_n\left(L_1, L_2, d\right)\right] \leq \mathbb{P}\left[Q_n\left(L_1, L_2, d\right) \cap \left\{X_{L_1} \leq N\left(\varepsilon\right)\right\}\right] + \varepsilon.$$

Now,

$$\mathbb{P}\left[Q_n\left(L_1,L_2,d\right)\cap\left\{X_{L_1}\leq N\left(\varepsilon\right)\right\}\right]=\mathbb{E}\left[\sum_{S\subseteq V,|S|=L_1}\mathbf{1}_{S\text{ is a cycle}}\mathbf{1}_{\exists\text{ a cycle of length }L_2\text{at distance }d\text{ from }S}\mathbf{1}_{X_{L_1}\leq N(\varepsilon)}\right]$$
$$\leq N\left(\varepsilon\right)\max_{S}\mathbb{P}\left[\mathbf{1}_{\exists\text{ a cycle of length }L_2\text{at distance }d\text{ from a given cycle }S}\right].$$

Given the vertices S, we can reveal the half-edges one by one and check if there is another cycle of length L_2 at distance at most d from S. Since there are only a constant number of vertices to reveal (say, 2^{L_2+d}), the probability that any path will close in on itself goes to 0 (it is of order $1-\left(1-\frac{1}{\left(N-2^{L_2+d}\right)}\right)^{L_2}$. So the probability goes to 0 as $N\to\infty$, and we are left with $\lim_{n\to\infty}\mathbb{P}\left[Q_n\left(L_1,L_2,d\right)\right]\leq\varepsilon$. This is true for every ε, so $\lim_{n\to\infty}\mathbb{P}\left[Q_n\left(L_1,L_2,d\right)\right]=0$.

With this in hand, we fix an L, and go over the canonical horocycles one by one. Whenever there is one that has length $\leq L$, we increase its size (i.e. lower the axis-parallel line) until its length becomes L.

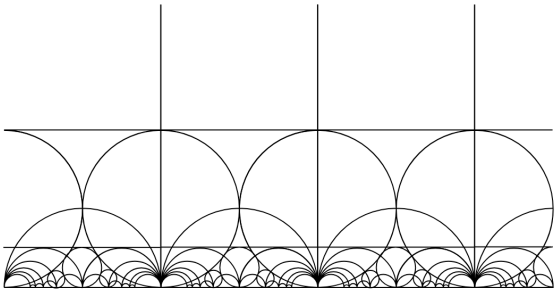


Figure 6.20: Increasing the size of the horocycle. Image taken from [6].

1. When we do this, the horocycle can intersect at most a finite number M of other triangles. If we ensure that the distance between cycles of length $\leq L$ in G is at most $2M$, then doing so will not cause any intersection between other horocycles of length $\leq L$.
2. It is nonetheless possible for the new horocycle to intersect a long canonical horocycle (of length $>L$). We would have to shrink this larger horocycle in order to avoid collision. In order to make sure that we do not shrink it to size $<L$, we take actually take L_2 to be larger than L_1, so that any long canonical horocycles we meet are long enough to be shrunk.

Exercise 6.20. Find an explicit relation between L and L_2 which ensures that the above two points can hold.

6.7 Expected genus

As we mentioned before, the surface S^O is made out of $2n$ glued-together ideal triangles. The cusps correspond to 0-length boundary in a pair of pants decomposition, and the genus of the surface is therefore given by

$$g=1+\frac{n-\#\text{cusps}}{2}.$$

A-priori, the number of cusps could be anything from 2 to n, and so the genus could be anything from 0 to $n/2$. This is one of the problems with the Brooks-Makover model: you don't get to control the genus of the resulting surface. Still, it's possible to say something intelligent about it.

Theorem 6.21. *The expected number of cusps satisfies*

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\#\text{cusps}\right]}{\log 2n} = 1.$$

Proof. We already saw above that that X_L, the number of cycles of length L in $(G, \mathcal{O})$, has expectation

$$\mathbb{E}\left[X_L\right] = \frac{2^L}{2L},$$

and so the number of LHT cycles has expectation

$$\mathbb{E}\left[LHT_L\right] = \frac{1}{L}.$$

The expected number of LHT cycles is

$$\mathbb{E}\left[LHT\right] = \sum_{L=1}^{2n} \mathbb{E}\left[LHT_L\right] = \sum_{L=1}^{2n} \frac{1}{L} = \log 2n + \gamma + O\left(\frac{1}{n}\right).$$

It should be of no surprise to you that $\text{Var}\left[\#\text{cusps}\right] = O\left(\log\left(n\right)\right)$ as well, and so the genus is strongly concentrated around its expected value of $n/2 - \log\left(n\right)$. In fact, using the Poisson distribution of cycles, it can be shown that the genus is asymptotically normal:

Theorem 6.22 ([9])**.** *Asymptotically as $n \to \infty$, we have*

$$g \to 1 + \frac{n}{2} - N\left(\log 2n, \log 2n\right)$$

in distribution.

6.8 Conclusion

The Brooks-Makover model gives us a model for random hyperbolic surfaces with what we might call "good" geometric properties (good in the "mixing, well-connected" sense):

1. The Cheeger constant satisfies
$$\mathbb{P}\left[h\left(S^C\right) > c\right] \to 1$$
2. Consequently, the first eigenvalue satisfies
$$\mathbb{P}\left[\lambda_1\left(S^C\right) > c\right] \to 1$$
3. The shortest geodesic can be shown to satisfy
$$\mathbb{P}\left[\text{sys}\left(S^C\right) > c\right] \to 1$$

4. The diameter is small (we haven't shown this).

$$\mathbb{P}\left[\operatorname{diam}\left(S^c\right) \leq c \log\left(g\right)\right] \to 1.$$

The results shown in these notes are all "preliminary", based on Brooks' and Makover's original paper, and the constants they give are far from optimal. The model was devised in the late 90s and early 2000s by building on Brooks' earlier work, and is still being studied today. More advanced techniques can be used to yield better constants. In particular:

1. It has been shown that the spectral gap of a Brooks-Makover surface is optimal - it approaches $\frac{1}{4}$ as $g \to \infty$ [10].
2. It has been shown that the diameter of a Brooks-Makover surface is asymptotic to $2\log\left(g\right)$ - this is twice as large as the optimal diameter for a hyperbolic surface [11].

Remark 6.23. The set of all surface S^C is dense in the space of all compact hyperbolic surfaces in the following sense. If S is a surface of genus g, then there is a graph G and an orientation $\mathcal{O}$ such that S and $S^C\left(G,\mathcal{O}\right)$ have a pair of pants decomposition with the same underlying graph (which might be different from G!), and the $6g-6$ parameters get arbitrarily close to each other. This denseness is not a trivial statement, and relies on a result by Riemann which states that every compact surface is (when seen as a complex surface) the zero set of an irreducible polynomial $\Phi\left(x,y\right)=0$, together with a theorem of Belyi about algebraic curves (which essentially states that the irreducible polynomials which generate the surfaces $S^C\left(G,\mathcal{O}\right)$ are dense in the space of all polynomials). Note that for a fixed n there is only a finite number of graphs and orientations on n vertices; to better and better approximate a genus g surface, the number of vertices must increase to infinity.

7 The Weil-Petersson model (Lectures 11-13)

We have spent quite a lot of time on the Brooks-Makover model, motivated by the fact that it lets us avoid questions like "what distribution to take on the $6g-6$ paramters", and "what happens to the graph structure?". However, it turns out that these issues can be tackled head on, leading to a different model of random surfaces.

For a fixed underlying graph, every choice of $6g-6$ length and twist parameters specifies a surface. However, some choices of will yield the same surface, with the question of which ones depending on what you mean by "same" (i.e. what symmetries you are allowed to apply to the surface in order to call it "the same"). Two natural choices lead to two different spaces, called the Teichmüller space (named after Oswald Teichmüller) and the moduli space (named after Lord Module).

The Weil-Petersson model can be thought of as picking a "uniformly random" surface from the space of all surfaces of a particular genus. This is done by imbuing the space with a metric (the Weil-Petersson metric), which in turns give the space a volume measure. This measure is (surprisingly!) finite, and so can be renormalized to give a probability measure.

Understanding both the moduli space and Teichmüller space is complicated, and we will not go into too much detail about it. Suffice to say, successful analysis of the Teichmüller and moduli spaces was a large part of Maryam Mirzakhani's Fields medal.

Remark 7.1. Teichmüller was a devout Nazi. He protested against Jewish teachers in his youth, joined the SA, volunteered to fight the Soviets, and died in the Eastern front. Such is life. His mathematics stand regardless.

7.1 Warm-up: the torus

We'll start with a simple question: what are all the possible flat tori? How can we parameterize the space of of all tori in a meaningful way?

First of all, a caveat. A torus is the quotient of $\mathbb{R}^2$ by a lattice Γ, and unlike $\mathbb{H}$, the Euclidean plane $\mathbb{R}^2$ can be scaled by a constant λ. So if we have a flat torus T, we can always scale all the distances by any $\lambda > 0$ and again obtain a flat torus (of different area). So in what follows, we will modulate out this $\mathbb{R}_{>0}$ factor, and keep it in mind until the end.

We can of course consider $\mathbb{C}$ instead of $\mathbb{R}^2$. When we say that a torus is the quotient of $\mathbb{C}$ by a lattice Γ, we mean that two points in $\mathbb{C}$ are equivalent if they differ by a vector in the lattice. A lattice is defined by two independent vectors which generate it,

$$\Lambda = \{az_1 + bz_2 \mid a, b \in \mathbb{Z}\}.$$

So far we have always drawn the fundamental domain of the torus as an axis-aligned rectangle, but this does not have to be the case: the lattice can be rotated, or made of parallelograms.

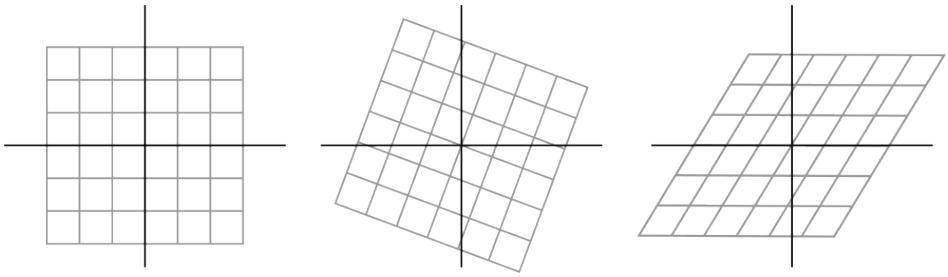


Figure 7.1: Lattices; each one defines a torus.

Exercise 7.2. Two tori are isometric if and only if their two lattices are isometric.

Keeping in mind the result about scaling and rotation, we can always make the following three assumptions about z_1 and z_2:

1. The vector z_1 is shorter than z_2 (if not, just swap them).
2. The vector z_1 is actually just $1 + 0i$ (if not, scale both vectors until z_1 has length 1, then rotate it to meet the real axis).
3. The vector z_2 always points towards the upper half-plane (if not, take $z_2 \mapsto -z_2$; since $b \in \mathbb{Z}$ this does not change the lattice).

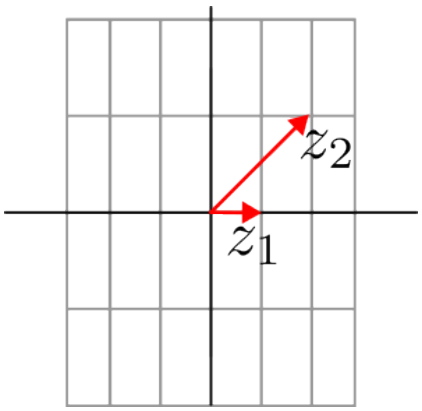


Figure 7.2: We may always choose the lattice vectors so that $z_1 = (1, 0)$, $|z_2| \geq 1$, and $\mathrm{Im}\,(z_2) > 0$.

Specifying Λ is then just a matter of stating what z_2 is. A-priori, z_2 could be any point in the upper half-plane of magnitude at least 1. However, there are still some more symmetries involved. For example, if $\mathrm{Im}\,(z_2) > 1$, then all points of the form $z_2 + a$ for integer a have magnitude at least 1 as well, and so the lattices generated by them are equivalent to the original lattice. We can therefore assume then that for such z_2, the real part is always constrained to be in (say) $\left[-\frac{1}{2}, \frac{1}{2}\right]$.

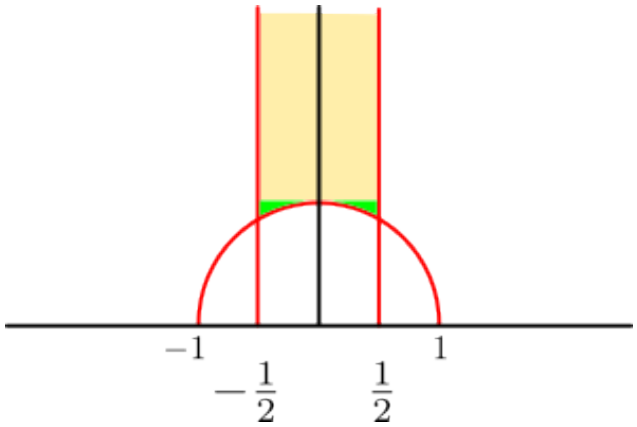


Figure 7.3: When $\mathrm{Im}(z_2) > 1$, we may always translate it by an integer real part, so we can assume that it lies within the slab $|\mathrm{Re}(z)| \leq \frac{1}{2}$. When $\mathrm{Im}(z_2) < 1$, such translations may reduce its magnitude to less than 1, breaking our assumptions on z_1 and z_2, and a bit more work is needed.

What about points with $\mathrm{Im}(z_2) < 1$?

Proposition 7.3. *Let $D = \{|z| \geq 1\} \cap \left\{|\mathrm{Re}(z)| \leq \frac{1}{2}\right\}$. Then for every $z_2 \notin D$ there exists a $z \in D$ so that z_2 and z generate the same lattice.*

Proof. If $z_2 \notin D$, look at $u = z_2 - a$, where $a \in \mathbb{Z}$ is such that $|\mathrm{Re}(u)| \leq \frac{1}{2}$. Then z_1 and u generate the same lattice. If $|u| \geq 1$, we are done. Otherwise, u is shorter than z_1, which is against our assumptions above: we can normalize by the length of u (i.e. multiply by $1/|u|$), rotate so that u lies on the axis (i.e. multiply by $u^{-1} = \overline{u}$ in the new, normalized form), and flip z_1 so that it is in the upper half-plane (i.e. multiply by -1). The roles of the two vectors have now switched. This gives us a new, equivalent lattice, but where the vector not on the x axis now has a larger imaginary component than before. In fact, when we perform the entire normalize-then flip operation, we are essentially replacing z_2 by $-1/z_2$: before swapping the indices, the new value of z_1 is

$$z_1 = -\frac{1}{|z_2|} \cdot \frac{\overline{z_2}}{|z_2|} = -\frac{\overline{z_2}}{z_2 \overline{z_2}} = -\frac{1}{z_2}.$$

And when we shift by z_1, we essentially apply the map $z \mapsto \pm 1$. Thus, z_2 is mapped by this procedure to an element of $\mathrm{PSL}(2, \mathbb{Z})$ - the Möbius transformations with determinant 1 and integer coefficients, which is generated by these actions:

$$\mathrm{PSL}(2, \mathbb{Z}) = \left\{ \frac{az + b}{cz + d} \mid a, b, c, d \in \mathbb{Z}, ad - bc = 1 \right\}.$$

We have already seen that for $f \in \mathrm{SL}(2, \mathbb{R})$, we have

$$\mathrm{Im}(f(z)) = \frac{\mathrm{Im}(z)}{|cz + d|^2}.$$

Here, the maps we apply are $z \mapsto z + n$ followed by $z \mapsto -1/z$, so $f = -\frac{1}{z+n}$,

$$\mathrm{Im}(f(z)) = \frac{\mathrm{Im}(z)}{|z + n|^2} = \frac{\mathrm{Im}(z)}{\mathrm{Im}(z + n)^2 + \mathrm{Re}(z + n)^2}.$$

The parameter n was chosen so that $\mathrm{Im}(z + n)^2 + \mathrm{Re}(z + n)^2 < 1$, and so applying $f(z)$ increases the imaginary part as long as $|z| \leq 1$, so in the limit we must reach the unit circle. ✹

Exercise 7.4. Prove that lattices generated by $z_2 \in \mathrm{int}\,(D)$ are distinct.

The domain D is a fundamental domain for the *moduli space* of the two dimensional torus (up to scaling, which we can always reintroduce). The moduli space itself has more structure than just a subset of $\mathbb{C}$: the vertical lines are identified, as well as the circle segments. In fact, we have seen that it is the quotient of the upper half-plane by the Möbius group action of $\mathrm{PSL}\,(2,\mathbb{Z})$. But $\mathrm{PSL}\,(2,\mathbb{Z})$ is a subgroup of $\mathrm{SL}\,(2,\mathbb{R})$, the set of (orientation-preserving) isometries of the hyperbolic plane! In other words, we can equip the moduli space D with a hyperbolic structure, and obtain a hyperbolic surface. This is called the *modular surface*. Some remarks:

1. D has finite area - it is contained in (say) the slab $\left\{\mathrm{Im}\,(z) \geq \frac{1}{2}, |\mathrm{Re}\,(z)| \leq \frac{1}{2}\right\}$, which has area 2. This gives a natural probability measure on the space of all tori, modulo a scalar factor. But you can always restrict yourself to (say) tori of area 1.
2. D is a hyperbolic surface, but it is not smooth or compact. It has a singularity at the point i as well as the points $\frac{1}{2} + i\frac{\sqrt{3}}{2}$. These are points of high symmetries, corresponding to tori which have rotational symmetries (the point i is the square torus; the point $\frac{1}{2} + i\frac{\sqrt{3}}{2}$ is a symmetric rhombus).

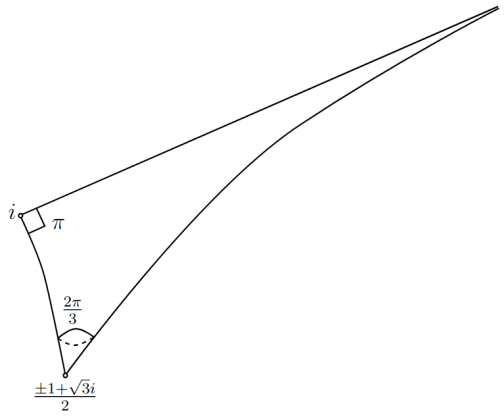


Figure 7.4: The moduli space of the torus. Note the two singular points, where the geometry is that of a cone, as well as the infinitely long (but finite-measured) tail. Note that the line connecting i to infinity is not a "crease"; this part of the space is like a cylinder. The image is better thought of as taking a half-infinite cylinder and pinching to a line the boundary. Image taken from [12].

It is often useful to work with smooth structures, without singularities. Doing this requires removing some of the symmetries of the torus. One way to do this is to "mark" the torus, so that even if two lattices are isometric, they will not be considered unless their markings are also the same under this isometry.

A natural discrete analogue of this is random graphs. Generating a uniformly random graph from the space of all labeled graphs is easy - this is the $G\left(n, \frac{1}{2}\right)$ model. However, generating a uniformly random graph from the space of all equivalence classes is much harder. Almost all graphs sit as singletons in their own equivalence class - they have no automorphisms but the trivial one. But of course, some do, and their existence makes exact computation very difficult. In other words - putting labels on the vertices made the analysis of the space of graphs considerably easier, and no real cost.

The situation is the same for hyperbolic surfaces. In this case, we have no vertices to label; the "marking" of a surface X will be a homeomorphism from a fixed topological surface of the same genus to X. But before we go into the precise definition for hyperbolic surfaces, let's see such a marking for tori and how it affects their moduli space.

Throughout this course, when we were careless about the types of tori we wanted, we just drew a square and identified its sides. We'll take that as our "base surface". The two sides give two loops on the torus; these are the loops which generate the fundamental group.

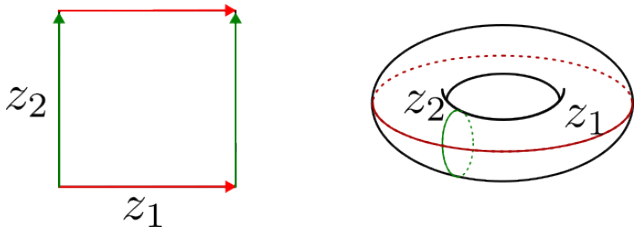


Figure 7.5: The standard way of gluing the sides of a square. The sides correspond to geodesics on the torus (the image on the right is just a cartoon; in reality, the torus is flat and both geodesics have the same length).

We can take another parallelogram with the same area as this standard one, but with z_2 having twice the x component:

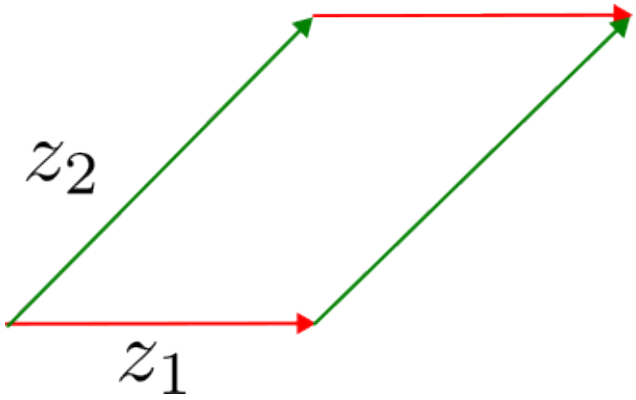


Figure 7.6: A different fundamental domain of the same torus as above.

These are clearly isometric, as tori, since the lattices are isometric. But if we look at the loops defined by the sides z_1 and z_2 of the parallelogram, we see that they are different.

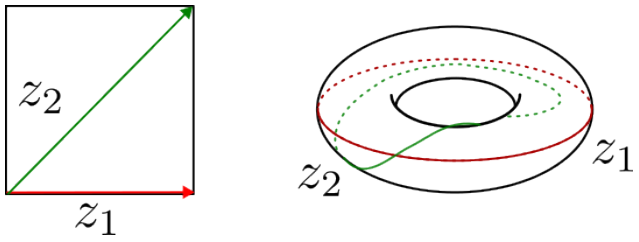


Figure 7.7: The geodesic corresponding to z_2 now wraps around both axes.

The standard curves wrap around only once around the torus, while the new one combines a rotation around the other axis - they are not isotopic!

Thus, every $z \in H$ represents a marked torus, where the homemomorphism from the standard topological torus is one that sends the edge boundary loops of the standard square to those of the parallelogram defined by z (or rather, the equivalence class of such homeomorphisms...). We see that under this representation, the map $z \mapsto z+1$ *no longer* sends a torus to its equivalent - they may be isometric, but the resultant homeomorphisms are not isotopic (in technical terms, they introduce a *Dehn twist*). Having lost this equivalence, all the above moduli space considerations become moot, and we are back to just the basic rotation / scale invariance. In other words: the space of *marked* tori is the upper half-plane $\mathbb{H}$. This is called the *Teichmüller space* of the torus. Some remarks:

1. Unlike the moduli space, the Teichmüller space is a smooth, simply connected surface. Like the moduli space, it has a hyperbolic structure on it (this makes sense - the moduli space is obtained from the Teichmüller by ignoring the markings, i.e. by adding symmetries. The group acting on these markings is called the *mapping class* group).
2. The moduli space of the torus has a finite area, and so we could naturally generate a random torus from it. Under the hyperbolic metric, the Teichmüller space is infinite, so there is no natural probability measure on it.

In a sense, we were very lucky. The Teichmüller space just happened to be, topologically, equivalent to the half plane. We just happened to be fairly well acquainted with a metric on the half plane (the hyperbolic metric), and this metric just happened to be the one for which the mapping class group $\mathrm{PSL}(2,\mathbb{Z})$ acts by isometries, and the fundamental domain tiles $\mathbb{H}$, so we got a hyperbolic orbifold structure on the moduli space. Finally, the quotient just happened to have a finite mass under the projected metric. How nice!

7.2 The Weil-Petersson metric

We already know that a hyperbolic surface of genus g is parameterized by a connectivity graph G and $6g-6$ parameters (and in fact, we can forget about the graph!). The moduli space of genus g surfaces, denoted by $\mathcal{M}_g$, is then just the equivalence classes modulo isometries. Like the torus, this will be a $6g-6$-dimensional space almost everywhere. However, at some symmetric points, the space will fold upon itself in a non-smooth manner, resulting in singularities (which may be of various dimensions). Instead of working with this very difficult beast, we turn to the Teichmüller space.

Definition 7.5. Let S_g be a fixed genus g topological surface.

1. A marked surface (X, φ) is a compact hyperbolic surface X together with a homeomorphism $\varphi : S \to X$.

2. Two marked surfaces (X, φ) ,(Y, ψ) are equivalent if there is an isometry $m : X \to Y$ such that $m \circ \varphi$ is isotopic to ψ.

3. The Teichmüller space of genus g, denoted $\mathcal{T}_g$, is the space of all marked surfaces modulo the equivalence relation.

We can again think of the homeomorphism by the way it acts on closed loops.

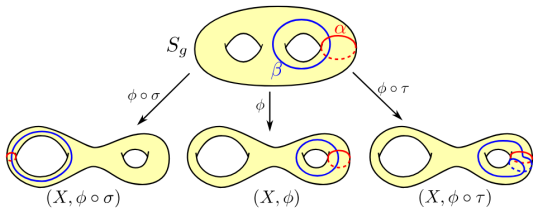


Figure 7.8: The Teichmüller space contains surfaces marked with a homeomorphism. This homeomorphism tells us where curves in the base surface go to. Image taken from [13].

The Teichmüller space is also a $6g-6$ dimensional space. One way of to get these $6g-6$ parameters is to consider an arbitrary pair of pants decomposition for S; this decomposition gives us $3g-3$ closed curves on S which divide it into pairs of pants. The homeomorphism from S to X maps the pairs of pants of S to those of X, and the lengths and twists parameters follow accordingly. It is important to note, however, that we do not have to use a "canonical" pair of pants decomposition: any set of $3g-3$ closed curves which dissect S into disjoint parts will do, even if the curves wind quite a lot around S. For any choice of curves, the lengths and twist parameters are the *Fenchel-Nielsen* coordinates relative to this choice. While we ordinarily think of the twist parameters as being in $[0, 2\pi]$ (or perhaps, the twist τ_i as being in $[0, \ell_i]$), for the sake of this construction we think of them as in $\mathbb{R}$; this is exactly like in the torus case, where parallelograms with a difference of $(1, 0)$ were identified in the moduli space but were distinguished in the Teichmüller space. It can be proved (basically by following the definitions around and being careful with loops) that there is a bijection between $\mathcal{T}_g$ and the Fenchel-Nielsen coordinates, which are simply vectors in $\mathbb{R}^{3g-3} \times \mathbb{R}_{>0}^{3g-3}$. This is not that different from the torus, where the Teichmüller space was equal to the half-plane $H = \mathbb{R} \times \mathbb{R}_{>0}$.

Being homeomorphic to $\mathbb{R}^{3g-3} \times \mathbb{R}_{>0}^{3g-3}$, there are many metrics we can put on the Teichmüller space. Perhaps the greatest success in the study of Teichmüller space of hyperbolic surfaces is the following non-trivial theorem.

Theorem 7.6. *There exists a metric on $\mathcal{T}_g$, called the Weil-Petersson metric, such that the mapping class group are isometries. This means that the moduli space inherits the metric in a natural way.*

The standard way to describe this metric is a bit much to chew on, and relies on facts from differential geometry and complex analysis which we have not and will not cover. Still, in many papers on hyperbolic spaces, all that is said concerning the Weil-Petersson metric is that "it exists", and I think it is worthwhile to see a bit of the work required into making it exist.

The construction goes roughly like this. A point $X \in \mathcal{T}_g$ is (basically) a compact hyperbolic surface of genus g. This hyperbolic surface can also be considered as a one-dimensional Riemann surface (i.e. by multiplying the metric by a conformal factor, we get a complex structure). A quadratic differential on X is a function φ acting the tangent space TX, and written by $\varphi = f(z)\, dz^2$, where f is some mermomorphic

valued function. The idea is that φ assign some function $f(z)$ to every point of the surface, and if we use some other coordinate system z^* to locally describe the surface, then

$$\varphi^*(z^*) = \varphi(z)\left(\frac{dz}{dz^*}\right)^2.$$

Practically, for $z \in X$, if $v \in T_z X$ is a tangent vector written by $v = a\frac{\partial}{\partial z} + b\frac{\partial}{\partial \overline{z}}$, then

$$\varphi(v) = f(z)\, a^2.$$

If $\varphi = f(z)\, dz^2$ and $\psi = g(z)\, dz^2$ are two quadratic differentials on X, we can define an inner product between them:

$$\langle \varphi, \psi \rangle = \int_X f(z)\, \overline{g}(z)\, dA$$

and this in turns provides a metric g_Q on the space of quadratic forms. It is a fact of Teichmüller theory (which requires more complex analysis than we would like to show here) that the space of quadratic forms on $\mathcal{T}_g$ at a point X is equal to the cotangent space. This is in fact a *cometric* on the Teichmüller space - a metric is supposed to tell us how to measure angles and distances between vectors in the tangent plane; this metric on quadratic forms tells us how to measure angles and distances between functionals on vectors in the tangent plane. It is a fact from differential geometry that the inverse of a metric tensor is also a metric, defined on the dual space. So by inverting g_Q we get an actual metric on the Teichmüller space itself.

In the case of the torus the underlying metric was just the hyperbolic metric, but the Weil-Petersson metric is much more complicated. For example, the curvature of the space is much wilder:

Fact 7.7. *Let $K(X)$ be the sectional curvature of $\mathcal{T}_g$ at point X. Then*

$$\sup_{X \in \mathcal{T}_g} \sup_{\sigma_p} K(\sigma_p, X) = 0$$

and

$$\inf_{X \in \mathcal{T}_g} \inf_{\sigma_p} K(\sigma_p, X) = -\infty.$$

Note that this is no the curvature of the surfaces $X \in \mathcal{T}_g$ themselves; these are always hyperbolic, meaning they have curvature -1. Rather, it is the curvature of the Teichmüller space itself, as a manifold. Note the term "sectional" standing in front of curvature - this is the Gaussian curvature of the surface spanned by a two dimensional tangent plane σ_p. In two dimensions there is only one type of curvature, but $\mathcal{T}_g$ is a $6g-6$ dimensional space, and in higher dimensions there are several different notions of curvature, e.g. scalar, sectional, Ricci (in the end, the metric tensor describes the manifold completely; these are just attempts to capture some meaningful information about the manifold without describing the metric).

It is not clear at this point how to work with this metric at all, and how to say something meaningful about it or about the moduli space. Salvation comes from the fact that the metric has a nice representation in Fenchel-Nielsen coordinates, for ANY choice of separating curves on S.

Theorem 7.8. *The Weil-Petersson metric gives rise to a symplectic 2-form which is also invariant under the mapping class group. This form is given by*

$$\omega_g^{WP} = \sum_{i=1}^{3g-3} d\ell_i \wedge d\tau_i,$$

where ℓ_i and τ_i are the Fenchel-Nielsen coordinates. (This is irrespective of the underlying surface S / graph G used to define those coordinates!)

Do not be alarmed if you do not remember what a symplectic form is, nor if you do not remember what the wedge $\wedge$ means here. In short, it is a skew-symmetric counterpart of the Riemannian metric: if a Riemannian metric is a way to calculate the inner product between two vectors in the tangent space of a point on the manifold, then a symplectic 2-form is a way to calculate an "antisymmetric" inner product between two vectors in the tangent space. Of course, this doesn't correspond to angles and lengths, so we don't have our Euclidean intuition when analyzing such forms. The gist is that we can turn the 2-form ω_g^{WP}, which gets just two vectors, into a volume form, by looking at

$$d\mathrm{Vol}_g^{WP} = \frac{1}{(3g-3)!}\left(\omega_g^{WP}\right)^{\wedge(3g-3)}.$$

The volume gets as an input n vectors, and returns the volume of a parallelepiped spanned by those vectors. It is a way to measure volume on manifolds. When the symplectic form arises from a Riemannian metric (such as in this case), then the volume given by $d\mathrm{Vol}_g^{WP}$ is the same as that given by the Riemannian metric.

Essentially, this theorem (which we will definitely not prove), states that we can integrate over the space using standard integrals, no matter what Fenchel-Nielsen coordinates we choose. This gives us some understanding of the Weil-Petersson metric. But even if we understand the Weil-Petersson metric very well (and at this point, we certainly do not), understanding $\mathcal{M}_g$ still involves understanding how the mapping class group acts on $\mathcal{T}_g$. This is difficult. For example, for the torus, we worked a bit, but eventually arrived at a nice fundamental domain $D = \{|z| \geq 1\} \cap \left\{|\mathrm{Re}\,(z)| \leq \frac{1}{2}\right\}$; this is not the case for hyperbolic surfaces, and no fundamental domain is known for $\mathcal{M}_g$. Still, it possible to say quite a lot about a random Weil-Petersson hyperbolic surface, for example by using work by Mirzakhani. For starters, using this theorem it is possible to show the following very-useful-for-random-surfaces fact.

Theorem 7.9. *The moduli space $\mathcal{M}_g$ of all compact hyperbolic surfaces of genus g has finite Weil-Petersson volume.*

Thus, it is possible to choose a "random" hyperbolic surface from the set of all surfaces. This is the Weil-Petersson model for random surfaces. We have mentioned in the introduction of these lecture notes that these can be seen as a sort of "uniform" random surface. The "uniform" comes from the fact that we do not weigh the surfaces in any way, and the fact that the volume form seems to give equal weight to every Fenchel-Nielsen parameter. However, this is no strict analogy with uniform random graphs (which are modeled by $G\left(n, \frac{1}{2}\right)$). The choice of a random Fenchel-Nielsen parameter is done modulo the mapping class group, and it is not immediately obvious what properties a chosen surface might have. It should also be noted that, like the moduli space of tori, the space $\mathcal{M}_g$ is non-compact. The main reason for this is surfaces with cuff-lengths going to 0 - think of a sequence of surfaces which, in the limit, have cusps. However, for every fixed ε, if we look at $\mathcal{M}_{g,\varepsilon}$, the space of all surfaces whose systole is at leas ε, then that space *is* compact.

7.3 Volumes and recursions

We'll start our moduli space exploration with a toy example, where instead of working with general compact hyperbolic surfaces of genus g, we'll look at the moduli space of a *punctured torus*.

A topological torus, as we already know, has genus $g = 1$, and so has Euler characteristic $\chi(T) = 2 - 2g = 0$. For any metric on the torus, the integral over its curvature is 0, and it therefore does not admit a hyperbolic metric. By removing a single point from it, however, we create a punctured torus, reducing the Euler characteristic to -1 and making negatively curved metrics possible; indeed, there are infinitely many hyperbolic metrics on it. Alternatively, a punctured torus is just a hyperbolic pair of pants P where

two of the cuffs have finite > 0 length and are connected to each other with some twist, while the third cuff has length 0 and is in fact a cusp.

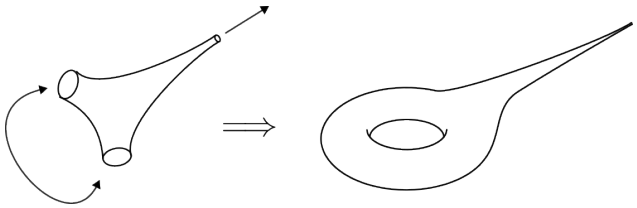


Figure 7.9: A cartoon of a punctured torus. This is a single pair of pants with two cuffs glued together and a cusp.

The punctured torus is not compact, and so marks a departure from what we have dealt with so far in the course (even in the Brooks-Makover model, we were hasty to compactify the cusped surfaces!). However, its treatment is simpler than that of the simplest compact hyperbolic surface, which would be of genus $g = 2$. The reason for this is that a genus 2 surface requires $6g - 6 = 6$ parameters to define, while the punctured torus requires only 2; integration and calculation in the Teichmüller space is therefore easier.

Theorem 7.10. *Let $\mathcal{M}_{1,1}$ be the moduli space of hyperbolic punctured tori. Then the Weil-Petersson volume of $\mathcal{M}_{1,1}$ is*

$$\mathrm{Vol}\,(\mathcal{M}_{1,1}) = \frac{\pi^2}{6}.$$

Proof. The proof is a beautiful application of using different curves to bisect a surface X into two parts. The idea is as follows. The covering Teichmüller space is the space of marked surfaces (X, φ), where $\varphi : S \to X$ is a homeomorphism from a topological punctured torus to the concrete punctured torus X with a hyperbolic metric. For any simple closed (separating) curve $\alpha \subseteq S$ which corresponds to a geodesic on X, the Fenchel-Nielsen coordinates along that curve are just (ℓ, τ) - the length of the curve, and the twist between the two parts that it disconnects:

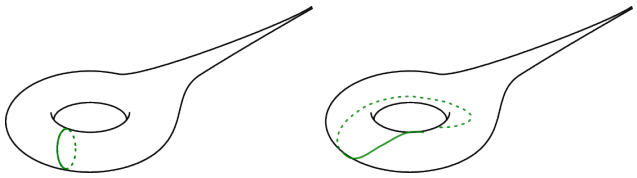


Figure 7.10: Both of these curves are geodesics, and so can be used as a base curve for Fenchel-Nielsen coordinates.

Note that it is entirely possible for two different choices of length and twists to give identical hyperbolic surfaces; that's the whole crux of going down from Teichmüller space to the moduli space. To get around this, we instead look at the collection of punctured surfaces with marked geodesics. Formally, let

$$\mathcal{M}^*_{1,1} = \{(X, \gamma) \mid X \in \mathcal{M}_{1,1}, \gamma \text{ a simple closed geodesic in } X\}$$

If we take the Fenchel-Nielsen coordinates relative to γ, it is impossible to get the same element in $\mathcal{M}^*_{1,1}$ with two different choices of lengths and twists around; any different choice in length will of course give different γ, while it can be shown that if the same γ is used, different twists will give different surfaces.

Now, for a specific curve $\alpha \subseteq S$, we have no problem integrating in Teichmüller space, due to the "symplectic volume" theorem above. That is, integration of a function $f(X) = f(\ell, \tau)$ in Teichmüller space (where ℓ and τ are the Fenchel-Nielsen coordinates relative to α) essentially boils down to

$$\int_0^\infty \int_0^\infty f(x,t)\ dtdx.$$

This can be projected down to $\mathcal{M}_{1,1}^*$ by observing that for a given curve $\alpha \subseteq S$ and Fenchel-Nielsen coordinates (ℓ, τ), it is clear that (ℓ, τ) and $(\ell, \tau + \ell)$ give the same surface. Integrating a function f over $\mathcal{M}_{1,1}^*$ means computing

$$\int_{\mathcal{M}_{1,1}^*} f(X)\, dX = \int_0^\infty \int_0^x f(x,t)\ dtdx.$$

The quantity that we would actually like to find is just

$$\int_{\mathcal{M}_{1,1}} 1\, dX.$$

The two integrals are NOT equal, however:

$$\int_{\mathcal{M}_{1,1}} 1\, dX \neq \int_0^\infty \int_0^x 1\, dtdx,$$

since on the right-hand the surfaces have a marked geodesic, and so it overcounts the left-hand side (indeed, it computes to ∞, while those who are in-the-know expect the left-hand side to be finite). There is a way to convert between the integral $\int_{\mathcal{M}_{1,1}} 1\, dX$ and an integral of the form $\int_0^\infty \int_0^x f(x,t)\ dtdx$, but for a different function than $f = 1$, using the currently-ex-machina McShane identity:

Theorem 7.11. *Let X be a hyperbolic once-punctured torus. Then*

$$\sum_\gamma \frac{1}{1 + e^{\ell(\gamma)}} = \frac{1}{2},$$

where the sum goes over all simple closed geodesics in X.

On the one hand, although perhaps you could not have guessed this exact formula yourself, by itself it should not be too surprising. For any fixed R, there is only a finite number of simple closed geodesics of length $\leq R$. If there are not too many of them, then the sum $\sum_\gamma e^{-\ell(\gamma)}$ should converge (and indeed, remember our fractal dimension calculation when we computed the diameter of random hyperbolic surfaces - the number of centres of hyperbolic pairs of pants in the covering tree at radius R was roughly $e^{\delta R}$ with $\delta < 1$; each such point corresponds to a closed geodesic which passes through the centre (not necessarily simple, of course, and there may be geodesics not passing through the centre - this is just intuition). That the actual expression to be summed over is $\left(1 + e^\ell\right)^{-1}$ rather than e^ℓ is a technicality.

On the other hand, this identity is marvelous: the constant $\frac{1}{2}$ is the same *for all tori*.

Writing $f(r) = \frac{1}{1+e^r}$, consider the integral of the sum

$$\int_{\mathcal{M}_{1,1}} \sum_{\gamma \subseteq X} f(\ell(\gamma))\, dX,$$

where γ is a simple closed geodesic in X. The sum over all γ means that we are in fact integrating over $\mathcal{M}_{1,1}^*$!

$$\int_{\mathcal{M}_{1,1}} \sum_{\gamma \subseteq X} f\left(\ell\left(\gamma\right)\right) dX = \int_{\mathcal{M}_{1,1}^*} f\left(\ell\left(\gamma\right)\right) d\left(X,\gamma\right) = \int_0^\infty \int_0^x f\left(x\right)\, dtdx.$$

Now, on the one hand, by the McShane identity, for a fixed surface X we have $\sum_{\gamma \subseteq X} f\left(\ell\left(\gamma\right)\right) = \frac{1}{2}$, and so the left-hand side is equal to

$$\frac{1}{2}\mathrm{Vol}\left(\mathcal{M}_{1,1}\right).$$

On the other hand, the integral can be computed explicitly (though not expressed as elementary functions)

$$\int_0^\infty \int_0^x \frac{1}{1+e^x}\, dtdx = \int_0^\infty \frac{x}{1+e^x} dx.$$

To calculate the value of the integral, write

$$\begin{aligned}\frac{x}{1+e^x} &= x\frac{e^{-x}}{1+e^{-x}} \\ &= xe^{-x}\cdot\left(1-e^{-x}+e^{-2x}-e^{-3x}+\ldots\right) \\ &= x\sum_{n=1}^{\infty}\left(-1\right)^{n+1}e^{-nx}.\end{aligned}$$

We can interchange sum and integral, so

$$\int_0^\infty \frac{x}{1+e^x} = \sum_{n=1}^{\infty}\left(-1\right)^{n+1}\int_0^\infty xe^{-nx}dx = \sum_{n=1}^{\infty}\frac{\left(-1\right)^{n+1}}{n^2} = \frac{\pi^2}{12}$$

(it is perhaps very tempting to use the known result $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$ and just say that the alternating sum is half of that, but of course this is not the way to prove the last equality). In any case, this gives the desired $\mathrm{Vol}\left(\mathcal{M}_{1,1}\right) = \pi^2/6$. ✹

Calculating the volume of the entire moduli space $\mathcal{M}_{1,1}$ (or rather, $\mathcal{M}_g$ for general compact hyperbolic surfaces) is the first step in computing probabilities under the Weil-Petersson model. After all, probabilities are just ratios of volumes; for example, the probability of a random hyperbolic surface of genus g to have systole smaller than ε is just given by

$$\frac{\mathrm{Vol}\left(\mathcal{M}_{g,\varepsilon}\right)}{\mathrm{Vol}\left(\mathcal{M}_g\right)}.$$

So of course we need to know the denominator to calculate this probability.

This type of result is a simple, "baby" version of the general framework of Maryam Mirzakhani, who invented a method to calculate these volumes. Roughly speaking, she generalized the McShine identity to other hyperbolic surfaces, and constructed an elaborate recursive scheme where one splits up a high-genus surface along some marked geodesics, obtaining an expression of the volume of $\mathcal{M}_g$ as a function of lower-order moduli spaces. One important aspect of the analysis is that it goes through bordered hyperbolic surfaces, i.e. the surfaces are made of pairs of pants where some of the cuffs have not been glued to each other, so the surface is left with geodesic boundary. This is essential for the analysis of unbordered surfaces as well. Mirzakhani then gives recursive equations for $\mathcal{M}_{g,n}\left(L_1,\ldots,L_n\right)$, the moduli space of surfaces with n geodesic boundary components, where component i has length L_i.

Theorem 7.12 (Generalized McShane identity). *Let X be a hyperbolic surface with n boundary geodesics $\beta_1, \ldots, \beta_n$ with lengths $L_1 \ldots, L_n$. Then*

$$\sum_{\alpha_1,\alpha_2} D\left(L_1, \ell\left(\alpha_1\right), \ell\left(\alpha_2\right)\right) + \sum_{i=2}^{n} \sum_{\gamma} R\left(L_1, L_i, \ell\left(\gamma\right)\right) = L_1,$$

where (α_1, α_2) go over all unordered pairs of simple closed geodesics which together with β_1 bound a pair of pants, and γ goes over all simple closed geodesics which together with β_1 and β_i bound a pair of pants. The functions D and R are given by

$$D\left(x, y, z\right) = 2 \log\left(\frac{e^{\frac{x}{2}} + e^{\frac{y+z}{2}}}{e^{\frac{-x}{2}} + e^{\frac{y+z}{2}}}\right)$$

and

$$R\left(x, y, z\right) = x - \log\left(\frac{\cosh\left(\frac{y}{2}\right) + \cosh\left(\frac{x+z}{2}\right)}{\cosh\left(\frac{y}{2}\right) + \cosh\left(x - z\right)}\right).$$

Very simple!! but it basically comes out of the plane hyperbolic geometry (the D and R are going to be lengths of geodesics connecting various points on the boundaries of pairs of pants).

The recursive formula obtained from this is a bit complicated; put very simply, it is of the form

$$\frac{\partial}{\partial L_1} L_1 V_{g,n}\left(L_1, \ldots, L_n\right) = \mathcal{A}_{g,n}^{con}\left(L\right) + \mathcal{A}_{g,n}^{discon}\left(L\right) + \mathcal{B}_{g,n}\left(L\right),$$

where each term on the right hand side is a rather monstrous sum / integral originating from the different effects that cutting out a pair of pants with one or two boundary components has on the surface X:

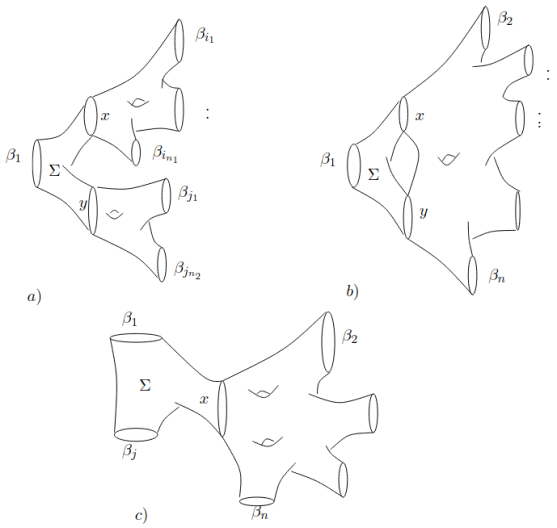


Figure 7.11: Cutting out a pair of pants can give different topologies. Image taken from [14].

This type of reasoning and formulae is sometimes called *topological recursion*. We will not deal with it anymore in this course. Let us just say that despite the size and ferocity of these equations, they can

with persistence be tamed, eventually leading to strong understanding of the geometric properties of the Weil-Petersson model, yielding theorems like this:

Theorem 7.13. *Let $\varepsilon > 0$. Let X_g be a random Weil-Petersson hyperbolic surface. Then as $g \to \infty$,*

$$\mathbb{P}\left[\mathrm{sys}\left(X_g\right) < \varepsilon\right] \asymp \varepsilon^2.$$

Thus, a random hyperbolic surface has a finite probability of having small injectivity radius. This is in contrast to the Brooks-Makover model, where there was a constant c such that $\mathbb{P}\left[\mathrm{sys}\left(S_g\right) > c\right] \to 1$. However, in other aspects, there is similarity:

Theorem 7.14. *Let X_g be a random Weil-Petersson hyperbolic surface. Then as $g \to \infty$,*

$$\mathbb{P}\left[h\left(X\right) \leq \frac{\log 2}{2\pi + \log 2} - \varepsilon\right] \to 0$$

and

$$\mathbb{P}\left[\mathrm{diam}\left(X_g\right) > 40 \log g\right] \to 0.$$

Thus, a random hyperbolic surface has positive Cheeger constant (and therefore also a spectral gap), and up to a constant factor has optimal diameter.

Remark 7.15. There has been much progress since the original calculations of Mirzakhani, both in terms of results and approaches. For example, it has been shown that a random hyperbolic surface has optimal spectral gap - it approaches $\frac{1}{4}$ as $g \to \infty$. This was first proved in [15] using more traditional techniques, then again in [16], using different methods.

8 The random cover model (Lecture 14)

The random cover model is the third and last of the "popular" random hyperbolic surface models. In one sense, it is the most natural of the three - it does not involve an awkward compactification as in the Brooks-Makover model, nor does it require a complicated metric just to define it as in the Weil-Petersson model. However, it requires access to a known surface with good properties, so there is still some small arbitrary choice that we have to make.

The random cover model is inspired by the same-named model for random graphs, which we'll talk about first.

8.1 Random graph covers

Let $G = (V, E)$ be a graph. We say that H is a cover if there exists an onto map $f : H \to G$ which is a local isomorphism - i.e., f is a bijection from v and its neighbors in H to $f(v)$ and its neighbors in G. A cover is said to be of degree n if $\left|f^{-1}(v)\right| = n$ for all $v \in V$.

Given a base graph G, we denote by G_n the random graph, uniformly chosen among all degree n covers of G. This is indeed a valid definition:

1. For every n, there is at least one degree n cover of G: just take n disjoint copies of it (granted, this example is usually not particularly interesting).
2. If G has $|V|$ vertices, every degree n cover has $n|V|$ vertices, so there is only a finite number of possible covers.

Unfortunately, much like saying "choose a random d-regular graph uniformly at random", or "choose a random planar graph", it is often hard to say anything meaningful about this distribution of random graphs if we do not give some algorithmic / systematic way of describing or generating them.

Given $G = (V, E)$, here is one relatively intuitive way to construct a random cover $G_n = (V_n, G_n)$ on $n|V|$ vertices, called the "permutation model":

1. Replace each vertex $v \in V$ by n vertices, $v_1, \ldots, v_n$.
2. If $u \sim v$ in G, then generate a uniformly random perfect matching between $\{v_1, \ldots, v_n\}$ and $\{u_1, \ldots, u_n\}$.

That's it! Very simple.

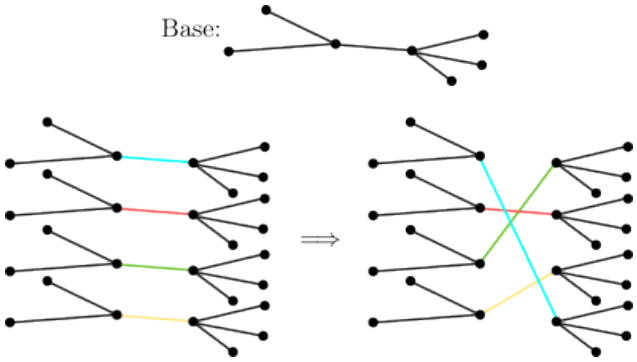


Figure 8.1: In the permutation model for random graph covers, every vertex is replaced by n vertices, and every edge is replaced by a random matching on $K_{n,n}$. In this figure, $n = 4$.

One way to look at this construction is to think of G_n as n copies of G stacked one above another, but with the edges randomly scrambled across the layers. Note that a matching between $\{v_1, \ldots, v_n\}$ and $\{u_1, \ldots, u_n\}$ corresponds to a permutation $\sigma \in S_n$, where v_i is matched with $\sigma(i)$.

In the very worst case, we might accidentally choose all the random matchings to be the identity permutation, i.e. $v_1 \sim u_1, \ldots, v_n \sim u_n$. In this case, we really are just left with n copies of the original graph G. However, perhaps you have the intuition that in most cases, the scrambling of the edges should be enough to connect the copies of G together. You would be right.

Theorem 8.1 (Alon and Linial, 2002)**.** *Let G be a connected graph with minimal degree at least 3. Then G_n is connected with probability $1 - o(n)$.*

Exercise 8.2. Find a family of graphs for which $\mathbb{P}[G_n \text{ is connected}] \neq 1 - o(n)$.

Question: Is this a good way of generating a uniformly random degree n cover?

Answer: No, the probability distribution is not uniformly random! Indeed, what we have is a uniformly random distribution on the set of cover graphs, where the fibre of each vertex $v \in V$ (that is, the set $f^{-1}(v)$) is also labeled. It may very well be that permuting the vertices in the same fibre gives the same graph! Such graphs are over-represented in our model, since there is more than one permutation which yields them. Still, it can be shown (though we won't do this here) that this doesn't really make a difference: only a vanishing proportion of the graphs generated this way have a non-trivial automorphism, and so any statement that holds asymptotically almost surely (that is, with probability tending to 1) in the "uniform degree n cover" model also holds in the "generate permutations at random" model.

What do random covers do to the eigenvalues of a graph (that is, the eigenvalues of its adjacency matrix)? Since G_n has $n|V|$ vertices, it has $n|V|$ eigenvalues - much more than the mere $|V|$ of the original graph. Here are two immediate results:

1. If G was d-regular, then so is G_n. Thus, for d-regular graphs, the highest eigenvalue (d) is preserved.
2. If $f : V \to \mathbb{R}$ is an eigenfunction of G with value λ, then $f_n : V_n \to \mathbb{R}$ defined by $f_n(v_i) = f(v)$ is an eigenfunction of G_n with the same value λ. Thus, the original eigenvalues of G are preserved.

We call the remaining $(n-1)|V|$ eigenvalues the "new" eigenvalues (if an original eigenvalue λ has multiplicity m in G and m' in G_n, then we introduced $m' - m$ new λ eigenvalues). How do these relate to the original eigenvalues? Perhaps the most important theorem known shows that the spectral gap cannot really decrease when taking powers.

Theorem 8.3 (Bordenave and Collins [17]). *Let G be a d-regular graph. Then for every ε, G_n does not have new eigenvalues greater than $2\sqrt{d-1}+\varepsilon$ with probability $1-o_\varepsilon(n)$.*

For a connected d regular graph, the value $2\sqrt{d-1}$ is the largest spectral gap that one can asymptotically get. Thus, in graphs, the random cover method provides a nice way to get larger and larger graphs with near-optimal spectral gap: you just need to start with a seed of one, fixed-size example which has a good gap.

8.2 Random surface covers

What about random surfaces? We say that $\tilde{X}$ is a cover of X if there is a projection $P:\tilde{X}\to X$ such that every $x\in X$ has a neighbourhood U such that $P^{-1}(U)$ is a discrete union of neighbourhoods in $\tilde{X}$, each isometric to U. Can we find a model for a random cover of a surface X?

As a first approach, we can mimic the random cover model by considering a decomposition of a surface S into pairs of pants. This turns it effectively into a 3-regular graph (albeit with possible self-loops - we have to match the half-edges, not the vertices themselves, but this doesn't impact the construction). We replace each pair of pants by n identical copies, and for each glued boundary component, we connect the n copies of the boundary component of one pair of pants to the n copies of the component according to the permutation chosen (and with the same twist parameter).

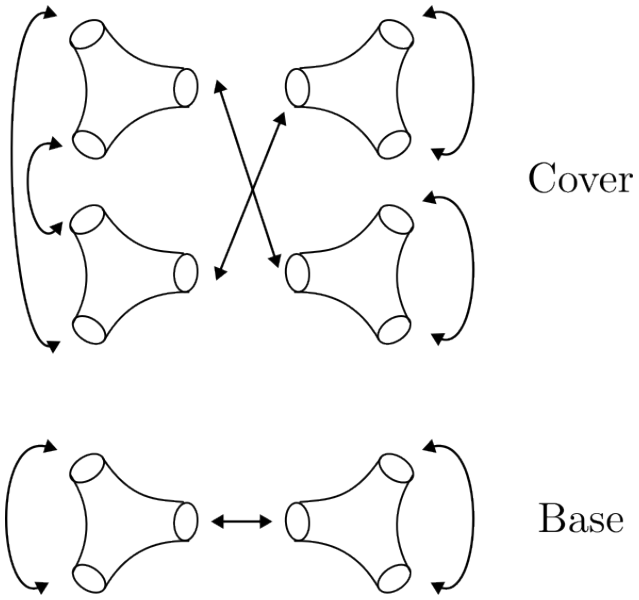


Figure 8.2: We can repeat the permutation cover model for pairs of pants. Here we have a 2-cover of the base surface.

As always, there is the question of which pair of pants decomposition is chosen. But in fact, there is an even more serious (yet related) problem. Consider a closed geodesic in the base surface. As we walk along it, the lift to the cover draws out some curve, which may or may not return to where it started (for the universal cover, it in fact never returns):

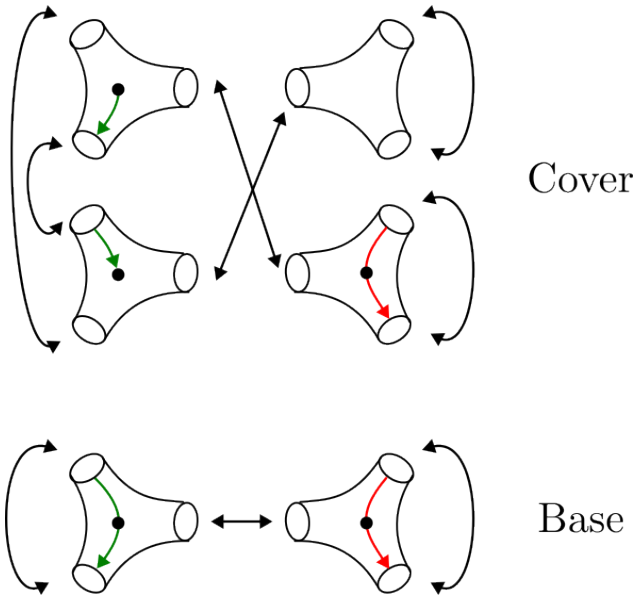


Figure 8.3: Some closed geodesics in the base get lifted to closed geodesics in the cover (red), while others get lifted to geodesic segments (green).

What happens if we go along a geodesic which is exactly the boundary of two pairs of pants? By construction, the lift will always return back to where it started. So this model does not sample from all possible covers - it includes only covers for which the cover of every cuff is n disjoint geodesics. Now, it is true that in every cover of degree n, *some* curves in X must lift to n different geodesics. But by choosing a particular pair of pants decomposition, we force those curves to always contain the pair of pants cuffs. So this model is very far from sampling according to the uniform distribution on random covers.

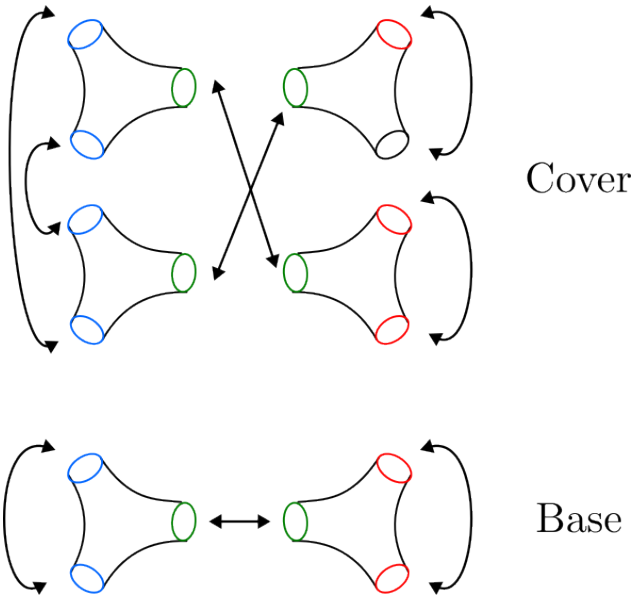


Figure 8.4: In the pair of pants permutation model, cuffs are always lifted to closed geodesics in the cover. There is no intrinsic reason for this to be true for a uniformly random cover.

It might be possible to fix this problem - say, by choosing some distribution on pair of pants decomposition. I would not advise going down this path, especially since there is a pants-free alternative way of thinking about covers and permutations, which does allow us to go over all covers. Here is what is usually done.

Let X be a hyperbolic surface and $\tilde{X}$ a degree n cover. Let $x_0 \in X$ be some point, and let γ be a closed geodesic in X starting at x_0. Label the fibre above x_0 by $\{1, \ldots, n\}$, and let $\tilde{x}_0$ be an (arbitrary) lift of x_0. As we mentioned above, as we follow along γ in X, we also follow a curve in $\tilde{X}$, which may or may not return to the origin $\tilde{x}_0$. In fact, if $\tilde{x}_0$ was in fibre i, then after having traversed along γ, the new point could be in any fibre $j \in \{1, \ldots, n\}$. This defines a permutation $\sigma \in S_n$. If we apply this technique to the (equivalence classes) of closed curves starting at x_0, we get a homomorphism from the fundamental group:

$$\pi_1(X, x_0) \to S_n,$$

called the monodromy homomorphism. It is a (perhaps intuitive) fact, not to be proven here, that the monodromy homomorphism is uniquely determined by $\tilde{X}$, so that two different covers give different monodromy maps.

On the other hand, given such a homomorphism, it is possible to construct a degree n cover as follows. Our surface X is the quotient of $\mathbb{H}$ by some group of isometries Γ_g. Given a monodromy homomorphism $\varphi \in \mathrm{Hom}(\pi_1, S_n)$, we define an action of $\Gamma_g \cong \pi_1(X, x_0)$ on $\mathbb{H} \times [n]$ by

$$T(z, i) = (Tz, \varphi(\gamma)(i)) \quad T \in \Gamma_g.$$

In other words, the point z is still mapped to Tz, but the sheet is chosen according to the permutation φ. It is simple to check that $\mathbb{H} \times [n] / \Gamma_g$ is a compact hyperbolic surface which is in fact a degree n cover of X.

Thus, the degree n covers and the set $\mathrm{Hom}(\pi_1, S_n)$ of homomorphisms are in bijection, so talking about "random covers" can instead be turned to talking about "random monodromy homomorphisms".

Can we even talk about these? The fundamental group $\pi_1(X, x_0)$ is finitely generated. In fact, there is a standard way to write it, as follows. Topologically, a hyperbolic surface of genus g can be obtained by gluing the $4g$ sides of a regular polygon in the following manner:

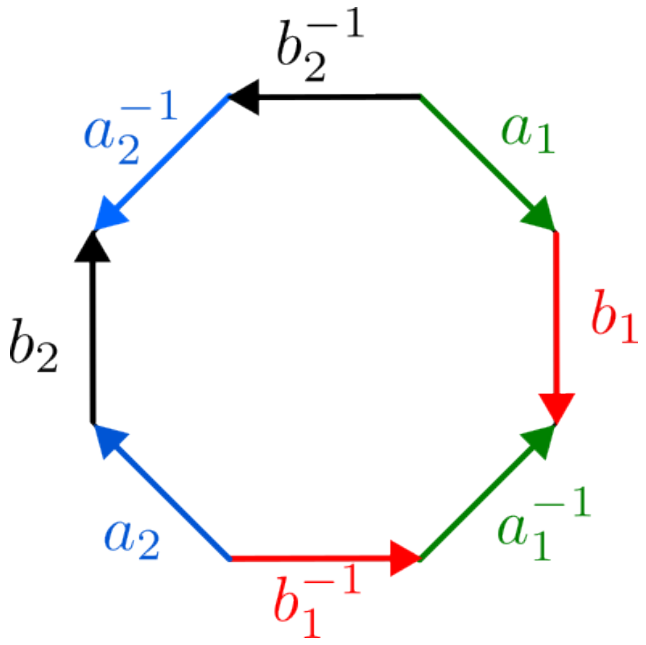


Figure 8.5: This side gluing gives a genus 2 surface.

All the vertices of this polygon are the same vertex, so going along a single polygon side corresponds to some closed loop. This loop is non-contractible, of course - the surface is not simply connected. However, if you go along *all* the sides, you get a loop which can be contracted in the interior of the polygon. It turns out that this is the only relation for the fundamental group, and so π_1 can be written as

$$\pi_1(X, x_0) \cong \Gamma_g := \langle a_1, b_1, \ldots, a_g, b_g \mid [a_1, b_1] \cdots [a_g, b_g] = 1 \rangle .$$

In particular, it is a finitely generated group, and the number of homomorphisms $\mathrm{Hom}(\Gamma_g, S_n)$ is finite. It is therefore possible to choose a uniformly random element from $\mathrm{Hom}(\Gamma_g, S_n)$ (though we won't discuss how you actually do this). This is the *random cover model*.

Exercise 8.4. Let $\gamma_1, \ldots, \gamma_{3g-3}$ be a pair of pants decomposition. Estimate the number of homomorphisms from Γ_g to S_n, as well as the number of homomorphisms from Γ_g to S_n for which $\varphi(\gamma_i) = \mathrm{Id}$ for all pair of pants boundaries γ_i. How many covers would we miss out on if we stuck with the pair of pants cover model?

In some sense, the construction is oblivious to much of the detail concerning hyperbolic surfaces. We do not need to deal with a graph structure , or with a pair of pants decomposition: many properties of the cover will emerge from understanding $\mathrm{Hom}(\Gamma_g, S_n)$, without the need to understand how Γ_g acts on $\mathbb{H}$.

For example, we can again discuss connectivity. It is entirely possible for the random cover to be disconnected; this happens, for example, when $\varphi : \Gamma_g \to S_n$ is the trivial homemomorphism, i.e. $\varphi(T) = \mathrm{Id}$ for all $T \in \Gamma_g$. In this case, $\tilde{X}$ is just n disjoint copies of X. In fact, $\tilde{X}$ is connected if any only if $\varphi(\Gamma) := \{\varphi(T) \mid T \in \Gamma\}$ is a transitive group (i.e. we can go from any sheet to any other sheet).

Fact 8.5 (Liebeck and Shalev [18]). *As $n \to \infty$, $\mathbb{P}[\varphi(\Gamma)$ is transitive$] \to 1$. In fact, $\mathbb{P}[\varphi(\Gamma) \geq A_n] \to 1$.*

How do the closed geodesics of $\tilde{X}$ relate to those of X? In general, they are always longer: a simple closed geodesic in the base X might be just a geodesic segment in $\tilde{X}$. However, there are "more" of them in some sense, since the surface is larger (just think about the trivial case n-copies of X case: every geodesic is simply duplicated n times). We can find out information about the geodesics using the homomorphism φ. Let γ be a geodesic in X generated by a group element $T \in \Gamma$, and let $\sigma = \varphi(T)$. We said that as we follow along γ, if we start at sheet i, the geodesic in $\tilde{X}$ ends at sheet $\sigma(i)$. If i is a fixed point, i.e. $\sigma(i) = i$, then the lift of γ starting at sheet i is also a geodesic, with the same length as γ. In general, the cycle structure of $\varphi(T)$ dictates how many geodesics γ will lift into: the total number will be the number of cycles, and the lengths will be $\ell(\gamma)$ times the lengths of the cycles. Of course, if a γ is lifted to a geodesic of length $2\ell(\gamma)$ by φ, then γ^2, which is the same as γ but going twice around, will be lifted to a curve of length $2\ell(\gamma) = \ell(\gamma^2)$. Thus, counting the lengths of geodesics in $\tilde{X}$ corresponds to counting the number of fixed points of $\varphi(\gamma)$ (for every $\gamma \in \Gamma_g$).

Understanding the lengths of closed geodesics in a surface plays an important part in understanding its spectrum, as we will very lightly see in the next lecture. We therefore have a lot to gain by understanding the set of fixed points of $\varphi(\gamma)$ (for a fixed γ, as $n \to \infty$). A possible type of theorem is the following:

Theorem 8.6 ([19]). *For every $\gamma \in \Gamma_g$, there exists a unique sequence $\{a_i(\gamma)\}_{i=1}^{\infty}$ such that for any $q \in N$,*

$$\mathbb{E}_{g,n}[\#\{\text{fixed points of } \varphi(\gamma)\}] = a_{-1}n + a_0 + \frac{a_1}{n} + \ldots + a_{q-1}n^{-(q-1)} + O_{q,\gamma}\left(n^{-q}\right).$$

It is also possible to say something about the a_i, if more information is known about γ. In fact, for important classes of geodesics, $a_{-1} = 0$ and $a_0 = 1$, so to first order the expected number of fixed points is just 1.

In fact, this understanding leads to bounds on the spectrum of hyperbolic surfaces. A strengthening of the above theorem, combined with the polynomial method for strong convergence, gives the following:

Theorem 8.7 ([20]). *Let X be a hyperbolic surface, and let X_n be a uniformly random cover of degree n. Then as $n \to \infty$, X_n does not introduce any new eigenvalues in the interval $\left[0, \frac{1}{4} - o(1)\right)$.*

In other words, if you start with a fixed base surface with spectral gap $\geq \frac{1}{4}$, taking a random large cover gives you an optimal spectral gap.

We will not say more about this model in this course, even though it is very natural, as the analysis of this model usually requires more group theory than I wanted to introduce in this course. For more information, see the linked series of lectures by Doron Puder. It should be noted that the open surface S^O in the Brooks-Makover model can also be thought of as a cover of some surface (this is not immediate; we would have liked to cover a triangle, but that is a bordered surface. The covered surface is in fact $\mathbb{H}\backslash\mathrm{PSL}(2,\mathbb{Z})$, which we have already seen as the moduli space of tori).

9 Closed geodesics and spectrum (Lecture 15)

When discussing the random cover model, we hinted at how studying the fixed points of a group element can help us understand the numbers and lengths of closed geodesics in the covering surface (this is called the *length spectrum*). When discussing the Brooks-Makover and Weil-Petersson models, we discussed the existence of short closed geodesics. Understanding how the closed geodesics look like is an important part of understanding the spectrum of the Laplacian on a surface.

9.1 Warm-up: graphs

Let G be a d-regular connected graph on n vertices, and let A be its adjacency matrix. As we learn in kindergarten, the trace of A is the sum of eigenvalues:

$$\mathrm{Tr}\left(A\right) = \sum_{i=1}^{n} \lambda_i.$$

If instead we look at A^k, we get

$$\mathrm{Tr}\left(A^k\right) = \sum_{i=1}^{n} \lambda_i^k.$$

It is now common to study the spectrum using these traces. For example, the quantity

$$\frac{1}{n}\mathrm{Tr}\left(A^k\right) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^k$$

is the k-th moment of the spectrum's empirical measure - the measure which puts an atom with mass $1/n$ at every λ_i. By computing the limits of $\frac{1}{n}\mathrm{Tr}\left(A^k\right)$ as $n \to \infty$, this identity can be used to show, for example, that the empirical measure of a random d regular graph on n vertices (in the configuration model, say) converges in distribution to a semicircle law.

The trace has a geometric interpretation. By definition,

$$\left(A^k\right)_{ij} = \sum_{\ell_1,\ldots,\ell_k=1}^{n-1} A_{i,\ell_1} A_{\ell_1,\ell_2} \cdots A_{\ell_{k-1}\ell_k} A_{\ell_k j}.$$

Since $A_{ij} = 1$ only if there is an edge in G between vertex i and j, the entry $\left(A^k\right)_{ij}$ counts how many paths there are that start at i and end at j; the trace $\mathrm{Tr}\left(A^k\right)$ is then equal to the number of rooted cycles of length k in G. This has a probabilistic interpretation as well: if we consider $\left(\frac{A}{d}\right)^k$ instead, we get that $\left(\frac{A^k}{d^k}\right)_{ij}$ is the probability that a simple random walk on the graph starting at i ends up at j after k steps.

Similarly, we can get information about the spectral gap. Since the graph is d regular, the first eigenvalue is always equal to $\lambda_1 = d$. Denoting $\lambda_* = \max\left\{\lambda_2, |\lambda_n|\right\}$, we thus have

$$\mathrm{Tr}\left(A^k\right) = \sum_{i=1}^{n} \lambda_i^k = d^k + \sum_{i=2}^{n} \lambda_i^k \geq d^k + \lambda_*^k,$$

yielding

$$\lambda_* \leq \left(\mathrm{Tr}\left(A^k\right) - d^k\right)^{1/k}.$$

The spectral gap can then be calculated by very carefully calculating the number of closed walks of size k in the graph; if we hope that the graph has a large gap, then hopefully $\mathrm{Tr}\left(A^k\right) - d^k$ is small.

9.2 Trace formulas in hyperbolic surfaces

The heat kernel trace formula is the analogue of the trace method for discrete graphs. It arises as a natural consequence of solving the heat equation on the surface (or, if you wish, of the transition kernel construction for Brownian motion that we spoke about in the beginning of the course). If $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \dots$ are the eigenvalues of the Laplacian on a surface S, then

$$\int_S p_t^S(x,x)\,dx = \sum_{i=0}^{\infty} e^{-t\lambda_i} = 1 + \sum_{i=1}^{\infty} e^{-t\lambda_i},$$

where $p_t^S(x,x)$ is the heat kernel, and is equal to the probability density that Brownian motion starting at x is again found at x after time t. The sum over vertices in the graph case (i.e. the trace) has been replaced by an integral over the surfaces; the Brownian motion transition kernel $p_t^S(x,x)$ replaces A/d, and the eigenvalues are now of the form $e^{-t\lambda_i}$ rather than λ^i.

When t is very very large, the sum on the right-hand side is dominated by $e^{-t\lambda_1}$, which is the largest term. On the other hand, for large times the Brownian motion will mix on the surface, so $p_t^S(x,x) = \frac{1}{\mathrm{Vol}(S)} + E_x$, where E is some error term. If you can control the error term precisely enough, you can then say something meaningful about the eigenvalues, and vice versa.

It should be clear that the heat kernel $p_t^S(x,x)$ has *something* to do with the lengths of closed geodesics, even if the relationship is not immediately clear. Every trajectory of Brownian motion that starts at x and returns to x is either contractible, or is homotopic to some closed geodesic. The probability space can then be partitioned according to which closed geodesic the Brownian motion mimicked, with the intuitive understanding that it cannot deviate too much from it without going into another conjugancy class. For example, when t is very small, Brownian motion will probably not move far away from the origin, and so will have very little opportunities to make loops around anything, and so $p_t^S(x,x)$ in this case will be very similar to $p_t^{\mathbb{H}}(x,x)$. Of course, this statement assumes that the closed geodesics are long enough, so that there is indeed no small waist around which the Brownian motion can go. Here is an example of a more precise statement of this intuition.

Theorem 9.1. *Let S be a compact hyperbolic surface, and let $P_t^S(x,y)$ be the density that Brownian motion, when started at x, is found at y after time t. There exists a constant C such that for every x and all $t \leq 1$,*

$$P_t^S(x,x) \leq C \cdot P_t^{\mathbb{H}}(0,0).$$

At least intuitively, it is no surprise that $P_t^S(x,x) \geq P_t^{\mathbb{H}}(0,0)$. After all, the surface is compact and not simply connected, and the Brownian motion on it has many more opportunities to come back to the origin when compared to motion in the plane. The point of the theorem is that the opposite direction is bounded - the heat kernel of a surface cannot be *too* large compared to the heat kernel on the hyperbolic plane.

Proof. We can write S as the set $\mathbb{H}\backslash\Gamma$, where Γ is a group of isometries. Let $\tilde{x}$ be a lift of x to the hyperbolic plane. Brownian motion B_t on S can be lifted to $\mathbb{H}$ as well, and the motion will return to the origin if and only if $B_t = T\tilde{x}$ for some $T \in \Gamma$. Thus

$$P_t^S(x,x) = \sum_{T\in\Gamma} P_t^{\mathbb{H}}(\tilde{x}, T\tilde{x}).$$

One of the elements in the sum is when $T = \mathrm{Id}$, and we immediately get that $P_t^S(x,x) \geq P_t^{\mathbb{H}}(x,x)$, formalizing our prior intuition. We now define

$$\Gamma(m) = \{T \in \Gamma \mid m < d(\tilde{x}, T\tilde{x}) \leq m+1\},$$

so that

$$\begin{aligned} P_t^S(x,x) &= \sum_{m=0}^{\infty} \sum_{T \in \Gamma(m)} P_t^{\mathbb{H}}(\tilde{x}, T\tilde{x}) \\ &\leq P_t^{\mathbb{H}}(0,0) \#\Gamma(0) + \sum_{m=1}^{\infty} P_t^{\mathbb{H}}(\tilde{x}, y) \#\Gamma(m), \end{aligned}$$

where y is a point at distance m from $\tilde{x}$. We can get an estimate for the number of points in the set $\Gamma(m)$ as follows. Let $r = \mathrm{Inj}(x)/2$. Reminder: the injectivity radius is the largest radius such that a disk in S around x is isometric to a hyperbolic disk. In particular, a disk of radius smaller than $\mathrm{Inj}(x)$ cannot wrap around S, or, more formally, the disk in $\mathbb{H}$ around $\tilde{x}$ cannot contain any point of the form $T\tilde{x}$ for $T \neq \mathrm{Id}$. So if we look at a disk D of radius r around $\tilde{x}$ in $\mathbb{H}$ and translate it by different group elements $T \in \Gamma(m)$, the images must be disjoint. All such disks are contained in a larger disk of radius $m+1+r$, and since they are disjoint, the number of elements must satisfy

$$\#\Gamma(m) \cdot \mathrm{Vol}(B_{\mathbb{H}}(r)) \leq \mathrm{Vol}(B_{\mathbb{H}}(m+1+r)).$$

We already know the area of a hyperbolic circle:

$$\begin{aligned} \#\Gamma(m) &\leq \frac{4\pi \sinh((m+1+r)/2)^2}{4\pi \sinh(r/2)^2} = \frac{\cosh(m+1+r)-1}{\cosh(r/2)-1} \\ &\leq \frac{e^{m+r+1}}{\cosh(r/2)-1} \\ &\leq C\left(1+\frac{1}{r^2}\right) e^m. \end{aligned}$$

The injectivity radius of any point is always bounded below by half the systole, which is a quantity that depends only on S. So there indeed exists a C such that

$$P_t^S(x,x) \leq C\left(1+\frac{1}{r^2}\right)\left(P_t^{\mathbb{H}}(0,0) + \sum_{m=1}^{\infty} P_t^{\mathbb{H}}(\tilde{x}, y)\, e^m\right).$$

The result then follows by bounds on the heat kernel $P_t^{\mathbb{H}}(x,y)$, which decays as the square of the exponential in the distance between the points. ✹

Note that the Cheeger-Buser also gives a relation between the shortest closed geodesic and the eigenvalues: if the systole is larger than ℓ, then any non-contractible cut which separates the surface into 2 parts must have length at least ℓ, and so

$$h(S) \geq \frac{\ell}{\frac{1}{2}\mathrm{Vol}(S)}.$$

This gives

$$\lambda_1 \geq h(S)^2 = \frac{\ell^2}{\mathrm{Vol}(S)^2}.$$

Unfortunately, this bound usually isn't very good: the largest systole possible has size $O(\log g)$, so the largest lower bound you can get from this is $\lambda_1 \geq O\left(\left(\frac{\log g}{g}\right)^2\right)$.

Exercise 9.2. Using the trace formula, give an upper bound on $\int_S \left[p_t^S(x,x) - \frac{1}{\mathrm{Vol}(S)}\right] dx$ for times $t \geq 1$ (this bound should depend on $\mathrm{sys}(S)$ and λ_1).

Another very important formula is the Selberg trace formula, which more implicitly relates the eigenvalues and length spectrum. Write the eigenvalues as $\lambda_i = \frac{1}{4} + r_i^2$; for eigenvalues smaller than $1/4$ (*small eigenvalues*), r_i will be imaginary, while for large eigenvalues, r_i will be real.

Theorem 9.3 (Selberg trace formula)**.** *Let $h : \mathbb{R} \to \mathbb{R}$ be a symmetric smooth real function, and let $\widehat{h}(r) = \int_{\mathbb{R}} h(x) e^{-irx} dx$ be its Fourier transform. Then*

$$\sum_{i=0}^{\infty} \widehat{h}(r_i) = (g-1) \int_{\mathbb{R}} \widehat{h}(r) \tanh(\pi r)\, dr + \sum_{\gamma \text{ primitive}} \sum_{k=1}^{\infty} \frac{h(k\ell(\gamma))}{2\sinh(k\ell(\gamma)/2)}.$$

On the left-hand side, we have an expression that only involves the spectrum; on the right-hand side, we have an expression that only involves the lengths. Thus, having a good understanding of one can lead to a good understanding of the other.

Unlike the trace method for graphs, we cannot just take the k-th square root, and rather must conjure up reasonable h. One way to do this is as follows. Let h be some fixed function supported on $[-1,1]$, and for $L > 0$ large let $h_L(x) = h\left(\frac{x}{L}\right)$. All h_L does is expand the range of h. In terms of Fourier, we have

$$\widehat{h}_L(\omega) = L\widehat{h}(L\omega).$$

For eigenvalues $\lambda_i < \frac{1}{4}$, r_i is imaginary, and we have

$$\widehat{h}_L(r_i) = L\int_{\mathbb{R}} h(x) e^{-irxL} dx = L\int_{\mathbb{R}} h(x) e^{L|r|x} dx.$$

As $L \to \infty$, the contribution of this term gets very large - it has the exponential $e^{L|r|x}$. On the other hand, for eigenvalues $\lambda_i > \frac{1}{4}$, r_i is real, and so

$$\widehat{h}_L(r_i) = L\int_{\mathbb{R}} h(x) e^{-iL|r|x} dx.$$

The term is oscillatory and does not grow.

Thus, for large L, the left-hand side of the Selberg's trace formula sees mostly the contribution from the small eigenvalues. It is dominated by the trivial eigenvalues $\lambda_i = 0$ which corresponds to $r_0 = \frac{i}{2}$, but the next term in its growth is captured by r_i.

Meanwhile, since h_L is supported on $[-L, L]$, the sum in the right-hand only counts the geodesics of length $\leq L$. If we have a very good understanding of the numbers and lengths of geodesics of length $\leq L$, then we can get a good bound on the right hand side. Then, if we also choose the original h correctly, we might be able to show that the contribution of the right-hand side is equal to what we expect from the trivial eigenvalue r_0, plus terms relating to the larger eigenvalues. If we can show that the error terms decay quickly enough as $L \to \infty$, this can be used to show that λ_1 must be large, i.e. there is a spectral gap. Needless to say, this is a complicated technique which requires very exact cancellations, but it was used successfully by Anantharaman and Monk [15] to show that the Weil-Petersson model has optimal spectral gap (it was also used by Magee, Naud and Puder to show that the random cover model has spectral gap $3/16$; more advanced techniques were later used to improve to the optimal $1/4$).

10 Bounding the Cheeger constant via random processes (Lecture 16)

The Cheeger constant

$$h(S) = \inf_{\mathrm{Vol}(A) \leq \frac{1}{2}\mathrm{Vol}(S)} \frac{|\partial A|}{\mathrm{Vol}(A)}$$

is in principle very easy to upper bound: every set A gives you something! But, given a hyperbolic surface, how do we choose the set A? If we know something about the connectivity graph of a pairs of pants decomposition together with bounds on the cuff lengths of the pants, then we can say something: for example, if there happens to be a decomposition where a small number of thin pant legs separate the surface into two roughly equal parts, then of course the Cheeger constant goes to 0.

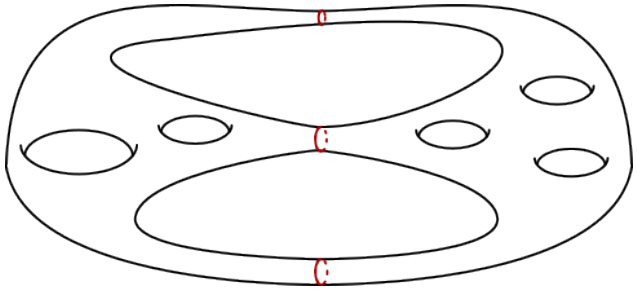


Figure 10.1: A surface with small Cheeger constant, owing to a small number of short geodesics which disconnect the surface.

But what if we are given a surface whose injectivity radius is large everywhere, and which is well-connected in every pairs of pants decomposition?

An obvious choice in this case is to just take a large ball $B(x,r)$ centered at a point x of large injectivity radius, with $r \leq \mathrm{Inj}(x)$. This ball is isometric to a ball in the hyperbolic plane, where the boundary length and volume satisfy

$$|\partial B(x,r)| = 2\pi \sinh(r), \quad \mathrm{Vol}(B(x,r)) = 2\pi(\cosh(r) - 1),$$

so

$$h(S) \leq \frac{2\pi \sinh(r)}{2\pi(\cosh(r) - 1)} \to 1.$$

This shows that $h(S) \leq 1$ asymptotically with $\mathrm{Inj}(S)$. Since having larger injectivity radius should intuitively only make $h(S)$ larger, we expect this result to hold in general; this is indeed true.

Exercise 10.1. Show that as $g \to \infty$, $\sup_{S \in \mathcal{M}_g} h(S) \leq 1$.

Can the bound $h(S) \leq 1$ be improved by choosing a better set A? We have not actually used any real information about the surface with the above choice, except for the injectivity radius at one point.

The answer is YES, and as is often the case, when we don't know how to choose a set, we can just choose one randomly.

As we often do, lets think about the discrete case first. Suppose $G = (V, E)$ is a d-regular graph on n vertices, and that that's all we know about it. A simple bound on the Cheeger constant is obtained by

picking a single vertex. The number of edges emanating from it is d, and so

$$h(G) \leq \frac{d}{1} = d.$$

But we can do better by choosing a larger set. Here is one (rather obvious) choice of choosing a random subset of vertices A: just put half of the vertices of G uniformly at random in A. For every edge, the probability that they are either both in A or both not in A is about $1/2$ ($\frac{\frac{n}{2}-1}{n-1}$, to be precise). The expected number of edges crossing from A to $V \backslash A$ is then roughly equal to $\frac{1}{2}|E| = \frac{1}{2}\frac{dn}{2}$. There thus must exist at least one partition A which has a smaller number of crossing edges than this. Since $|A| = n/2$ exactly, we have

$$h(G) \leq \frac{d}{2}$$

as $n \to \infty$.

Unfortunately, this proof doesn't directly generalize to hyperbolic surfaces. How do you take "half the points" of a continuous surface uniformly at random? And even if you do find a way to make such a construction work, would the set not have a very fractal boundary?

But we don't have to be so direct: we can clump points together, partitioning S into small-but-still-manageable sized chunks, and then try to take those. For example, we can take A to be half of the pairs of pants in a pairs of pants decomposition (this is a 3-regular graph). But which one? How can we control the boundary geodesics?

A clever approach was found by Thomas Budzinski, Nicolas Curien and Bram Petri, using the Voronoi cells of Poisson point processes. They showed the following:

Theorem 10.2 ([21]). *Let M_g be the space of all compact hyperbolic surfaces of genus g. Then*

$$\lim_{g \to \infty} \sup_{S \in M_g} h(S) \leq \frac{2}{\pi}.$$

Since $\pi > 2$ (an easy exercise), this is better than asymptotic the bound of 1 that we got by just taking a large disk.

The construction itself is not too difficult. We start with a Poisson point process of intensity μ in S, i.e. a random set of $x_1, \ldots, x_N$ in S such that:

1. The number of points in a set $A \subseteq S$ is a Poisson random variable with parameter $\mu \mathrm{Vol}(A)$.
2. The number of points in two disjoint sets is independent.

Note that N is itself a random variable (it is the number of points in S).

Given points $x_1, \ldots, x_N$, we can construct their Voronoi cells: the cell C_i corresponding to x_i is the set of all point in S that are closer to x_i than to any other point. The Voronoi cells $C_1, \ldots, C_N$ partition S (with overlaps at the boundary, which we don't care about). Since the x_i are random, this is a random partition - an excellent starting point for choosing our set A. Indeed, given this partition, we let $I \subseteq [N]$ be the random index set obtained by independently including every index with probability $1/2$ (i.e. coloring the cells either BLACK or WHITE), and setting

$$A = \cup_{i \in I} C_i.$$

If N is large enough, then A will contain roughly half the Voronoi cells. If the size of these cells does not fluctuate too much, i.e. there are not many large / small outliers, then with high probability, $\mathrm{Vol}(A) \approx$

$\frac{1}{2}\mathrm{Vol}\,(S)$. As for the boundary, note that for every two adjacent cells, there is a probability of $1/2$ that their shared boundary will be included in the Cheeger constant calculation (one is in A and the other isn't), and probability of half that it will not be. So the expected boundary of A has length $1/2$ of the sums of boundaries of the individual cells. So a bound on the lengths of the boundaries will give a bound on the Cheeger constant!

Remark 10.3. There is no known reason for $\frac{2}{\pi}$ to be the optimal bound. Brooks and Zuk gave some geometric assumptions under which the Cheeger constant is bounded by $1/2$, but they didn't really justify them or explain in what surfaces they think they should hold. Very roughly speaking, this bound corresponds to tiling the surface with disks and coloring them at random (tiling with disks is of course impossible).

10.1 Detour - Poisson processes on the computer

Generating a Poisson process in $\mathbb{R}^2$ is not very difficult. You partition $\mathbb{R}^2$ using (say) a square grid. The number of points in each square is a Poisson random variable, and given that there are n points in the square, they are independent and distributed uniformly in it. Generating random points in the square $[0,1]^2$ is easy - just take two uniformly random points in $[0,1]$.

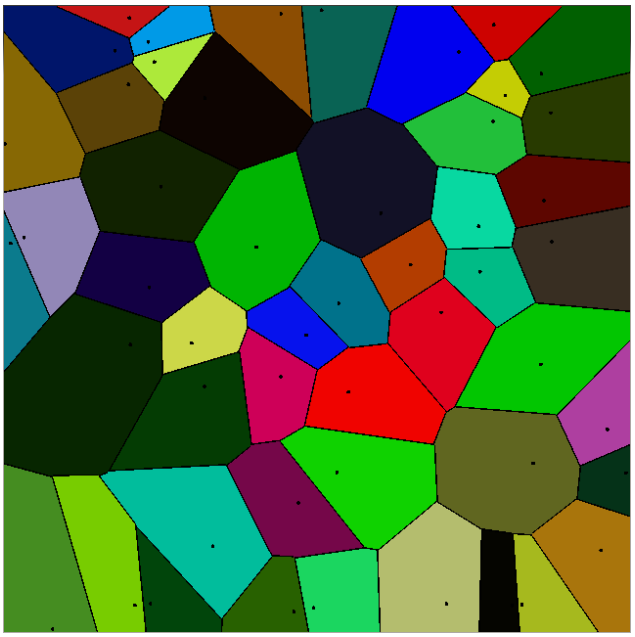


Figure 10.2: A Poisson-Voronoi tessellation of the square.

Generating a Poisson process in $\mathbb{H}$ is still not very difficult, but does require a bit of extra work. For starters, in order to represent the points in memory, you must make a choice for which model of $\mathbb{H}$ you are going to work with. The two models we have used in the course - the half plane and the disk - are both reasonable. We will choose the disk model.

While it is possible to tessellate the disk into polygons (in many ways!), this is not too convenient to work with (and what's more, if you have some tiny polygon at the edge of the disk, how do you distribute points in it?). It will be more convenient to consider concentric circles. A disk of radius $r < 1$ in the Poincaré disk model has finite area, and will contain only finitely many points. Given these n points in the disk, they will

be uniformly distributed in it (relative to the hyperbolic measure, not to the Euclidean one, of course). In particular, their angles $\theta_1, \dots, \theta_n$ relative to the origin are uniform, and all we really have to do is calculate the distance from the origin, $R_1, \dots, R_n$.

Let's order the distances from smallest to largest. What is the distribution of R_1? Well, in order for $R_1 > z$, there must not be any point in the disk of radius z, and since the number of points is a Poisson random variable, this is equal to

$$\mathbb{P}[R_1 > z] = e^{-\mu A(z)} = e^{-\mu 4\pi \sinh\left(\frac{z}{2}\right)^2}.$$

This is perhaps not a familiar probability distribution, but it can definitely become some after we make a change of variables: instead of asking for $\mathbb{P}[R_1 > z]$, let $A(R)$ be the area of a hyperbolic disk of radius R, and ask for $\mathbb{P}[A(R_1) > z]$; then we just have

$$\mathbb{P}[A(R_1) > z] = e^{-\mu z},$$

which is just an exponential random variable with intensity μ. So to generate the first point in our Poisson process, we generate an exponential random variable with rate μ - this will be the area - and calculate the radius of the disk with an area equal to this variable. This is similar to how a Poisson point process on $\mathbb{R}$ has intervals which are exponentially distributed - and in fact, if we suppose that we already generated R_1, then

$$\mathbb{P}[A(R_2) - A(R_1) > z \mid R_1 = a] = e^{-\mu z}$$

as well, i.e. the increments in hyperbolic are exponentially distributed.

We can then easily generate an infinite sequence of points for our Poisson point process, covering the entire hyperbolic plane: let $Z_1, Z_2, \dots$ be iid exponential random variables with rate μ, set

$$R_i = 2\sinh^{-1}\left(\sqrt{\frac{\sum_{j=1}^{i} Z_i}{4\pi}}\right),$$

and let θ_i be uniform in $[0, 2\pi]$. The polar coordinates of the points in $\mathbb{H}$ are (R_i, θ_i), which, in the Poincaré disk model, translates to

$$\left(\tanh\left(\frac{1}{2}R_i\right), \theta_i\right).$$

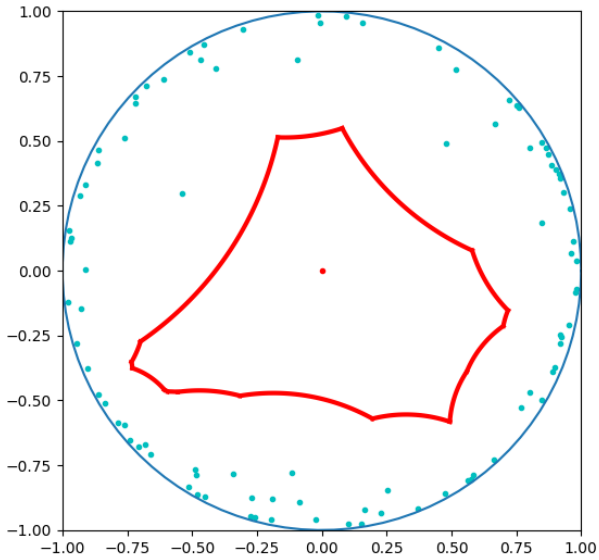


Figure 10.3: A Poisson-Voronoi cell in the hyperbolic plane.

This gives you a way to sample a point process on any compact surface S as well - just draw a fundamental domain anywhere in the plane, and keep the points that fall inside.

10.2 Warm-up: the hyperbolic plane

Before doing the calculations for a compact surface, let's look at Poisson-Voronoi cells in the hyperbolic plane. What is the isoperimetric ratio of a "typical" cell?

Well, what is a typical cell, anyway? We can, for example, just take the cell whose centre is closest to the origin of $\mathbb{H}$, and by homogeneity of the Poisson point process, assume that this is 0 itself. The rest of the point process has the same distribution as the original, in this case (this can be made precise using what is known as "Palm measures" and the Slivnyak-Mecke theorem; we will not talk about it here).

So we assume that the origin of $\mathbb{H}$ is the centre of a Voronoi cell C. What is the volume of this cell?

First, some misguided intuition. Consider a large disk of volume V in $\mathbb{H}$. This disk contains, in expectation, μV points. If the Voronoi cells of these points were entirely contained in the disk (in fact, if they were exactly equal to the disk), then the average volume of each cell would be

$$\frac{V}{\mu V} = \frac{1}{\mu}.$$

Of course, this assumption is not true. If we were dealing with $\mathbb{R}^2$ and not $\mathbb{H}$, it would be easy to appeal to "ignoring boundary effects" - most of the cells would be entirely contained in the disk, with only the cells near the disk's boundary poking out. However, in $\mathbb{H}$ the boundary is proportional to the size of the set, and that's a dangerous game to play.

Luckily, we can carry out an exact calculation. Let's look at some point $x \in \mathbb{H}$ at distance r from 0 . It will be included in C only if there are no points closer to x than 0, i.e. the disk of radius r around x is

empty. We have already calculated the probability of this happening:

$$\mathbb{P}\left[x \in C\right] = e^{-\mu 4\pi \sinh\left(\frac{r}{2}\right)^2} = e^{-\mu 2\pi(\cosh(r)-1)},$$

where the second equality is the hyperbolic identity $2\sinh\left(\frac{r}{2}\right)^2 = \cosh\left(r\right) - 1$; the latter form is more useful for us here. We then have

$$\begin{aligned}\mathbb{E}\left[|C|\right] &= \mathbb{E}\int_{\mathbb{H}} \mathbf{1}_{x\in C_\mu} dx \\ &= \int_{\mathbb{H}} \mathbb{P}\left[x \in C_\mu\right] dx \\ &= \int_{\mathbb{H}} e^{-\mu 2\pi(\cosh(r)-1)} dx.\end{aligned}$$

How to integrate over the hyperbolic plane? The fact that we have used the distance r from the origin to refer to our point strongly suggests that we should use polar coordinates around 0. The Jacobian in this case is $\sinh\left(r\right) dr d\theta$, and so we have

$$\begin{aligned}\mathbb{E}\left[|C|\right] &= \int_0^{2\pi}\int_0^{\infty} e^{-\mu 2\pi(\cosh(r)-1)} \sinh r \, dr d\theta \\ &= 2\pi \int_0^{\infty} e^{-\mu 2\pi(\cosh(r)-1)} \sinh r \, dr \\ &= 2\pi \int_0^{\infty} \frac{d}{dr}\left(-\frac{1}{2\pi\mu} e^{-\mu 2\pi(\cosh(r)-1)}\right) dr \\ &= \frac{1}{\mu}\left[-e^{-\mu 2\pi(\cosh(r)-1)}\right]_0^{\infty} \\ &= \frac{1}{\mu}.\end{aligned}$$

The next thing to calculate is the length of the boundary of the cell. This can be calculated in a similar manner, but the integrals are move involved. The keen student can solve the following.

Exercise 10.4. Show that

$$\mathbb{E}\left[|\partial C|\right] = \frac{8}{\sqrt{\pi\mu}} \int_0^{\infty} e^{-u} \sqrt{u + \frac{u^2}{4\pi\mu}} du.$$

Assuming this result, and ignoring blatant distribution of expectations, the "expected" isoperimetric ratio is

$$\begin{aligned}\frac{\mathbb{E}\left[|\partial C|\right]}{\mathbb{E}\left[|C|\right]} = \frac{\frac{8}{\sqrt{\pi\mu}}\int_0^{\infty} e^{-u}\sqrt{u+\frac{u^2}{4\pi\mu}}du}{1/\mu} &= \frac{8\sqrt{\mu}}{\sqrt{\pi}} \int_0^{\infty} e^{-u}\sqrt{u+\frac{u^2}{4\pi\mu}}du \\ &= \frac{8}{\sqrt{\pi}} \int_0^{\infty} e^{-u}\sqrt{\mu u + \frac{u^2}{4\pi}}du.\end{aligned}$$

That's a fancy integral, which I don't know how to explicitly solve. However, sending $\mu \to 0$, we get

$$\begin{aligned}\frac{\mathbb{E}[|\partial C|]}{\mathbb{E}[|C|]} &\to \frac{8}{\sqrt{\pi}} \int_0^\infty e^{-u} \sqrt{\frac{u^2}{4\pi}} du \\ &= \frac{4}{\pi} \int_0^\infty e^{-u} u \, du \\ &= \frac{4}{\pi}.\end{aligned}$$

This gives us the following "theorem".

Definition 10.5. A sequence S_k of compact hyperbolic surfaces is said to *locally converge* to $\mathbb{H}$ (sometimes also called "Benjamini-Schramm convergence") if for every $r > 0$, the relative measure of points $x \in S_k$ for which $\mathrm{Inj}(x) \geq r$ tends to 1 as $k \to \infty$. In other words, for asymptotically almost every point x in the surface S_k, the disk of radius r around x is isomorphic to a disk in the hyperbolic plane.

Theorem 10.6. *Let S_k be a sequence of compact hyperbolic surface of increasing genus which locally converges to $\mathbb{H}$. Then $\limsup_{k\to\infty} h(S) \leq \frac{2}{\pi}$.*

"Proof". Look at a fundamental domain for S_k in $\mathbb{H}$. As $k \to \infty$, the Poisson Voronoi tessellation of S_k converges to that of $\mathbb{H}$, again in a local sense. In fact, for a fixed rate μ, we can couple the Poisson process on $\mathbb{H}$ and on S_k so that with probability tending to 1, the disks of radius r will look the same around almost all points. The Cheeger constant is then bounded by coloring each cell black or white with probability $1/2$ independently as previously described, giving a Cheeger constant of $h(S) \leq \frac{1}{2} \cdot \frac{\mathbb{E}[|\partial C|]}{\mathbb{E}[|C|]} = \frac{2}{\pi}$. ✹

10.3 Compact surfaces

In real life (insofar as compact hyperbolic surfaces represent real life), one cannot write $\mathbb{E}\frac{|\partial C|}{|C|} = \frac{\mathbb{E}|\partial C|}{\mathbb{E}|C|}$. Further, while the calculations for the length and area in the hyperbolic plane serve as an initial guide, they definitely do not simply carry over to the compact case, where all large enough balls have finite volume, and where the fact that the injectivity radius is finite forces balls in S to behave very differently from those of $\mathbb{H}$ for large radius. Of course, it is still possible to say something meaningful - it's just that the techniques have to take into account the non simple-connectedness and compactness of S. Here is the formal way to do this.

Let $X_1, \ldots, X_N$ be the Poisson points and $C_1, \ldots, C_n$ be the Voronoi cells. Color each cell black or white with probability $1/2$, and let A be the smaller of the two sets.

Lemma 10.7. *For all $\mu > 0$ and $\delta > 0$, if g is large enough and $h(S) \geq \delta$, then*

$$\mathbb{P}\left[\left|\frac{\mathrm{Vol}(A)}{\mathrm{Vol}(S)} - \frac{1}{2}\right| > \varepsilon\right] < \varepsilon.$$

Lemma 10.8. *The boundary satisfies*

$$\limsup_{\mu\to 0} \sup_{g\to\infty} \sup_{S\in\mathcal{M}_g} \frac{1}{\mathrm{Vol}(S)} \mathbb{E}[|\partial \mathrm{Vor}_\mu(S)|] \leq \frac{2}{\pi}.$$

The first lemma states that the Poisson-Voronoi partition cuts the surface S into two roughly equal halves. Is this obvious? It might seem intuitive: for a fixed μ and as $g \to \infty$, we get an increasing number of points, and this is just a simple concentration of measure bound. Still, there is something to prove.

The second lemma states that we can get the required boundary length bound, as long as we take $\mu \to 0$. This is (very slightly) similar to how we had to take $a \to \infty$ when gluing pairs of pants in order to optimize the diameter.

Given these lemmas, the proof of the Cheeger constant is immediate: we let S be a surface of high enough genus, and let μ be small enough so that $\mathbb{E}\left[\left|\partial \mathrm{Vor}_{\mu}\left(S\right)\right|\right] \leq \frac{2}{\pi}\left(1+\delta\right)\left|S\right|$ for some small δ; by Markov's inequality,

$$\mathbb{P}\left[\left|\partial \mathrm{Vor}_{\mu}\left(S\right)\right| > \frac{2}{\pi}\left(1+\delta\right)^{2}\mathrm{Vol}\left(S\right)\right] \leq \frac{1}{1+\delta},$$

i.e there is some small but finite probability for the length of the boundary to be not much larger than $\frac{2}{\pi}\left|S\right|$. On the other hand, there is near certain probability for $\mathrm{Vol}\left(A_{\mu}\right)$ to be arbitrarily close to $\frac{1}{2}\mathrm{Vol}\left(S\right)$; if we take ε such that $\left(1-\varepsilon\right)+\left(\frac{\delta}{1+\delta}\right) > 1$, then there is positive probability for both events to occur. This means that there exists a set of points whose Voronoi tessellation gives the required bound of

$$\frac{\frac{1}{2}\cdot\frac{2}{\pi}\left(1+\delta\right)^{2}\mathrm{Vol}\left(S\right)}{\frac{1}{2}\mathrm{Vol}\left(S\right)\left(1-\delta\right)} = \frac{2}{\pi}+o_{\delta}\left(1\right).$$

The boundary estimate constitutes the main technical difficulty. The issue is that a direct integration, like the one for $\mathbb{H}$, cannot be performed. The calculations are done in a fundamental domain called the Dirichlet domain, which is basically a Voronoi cell for a single point (relative to the images of the point by the group Γ. The proof is too involved to be brought here. We can instead get a glimpse of how it is to work in a compact surface by looking at the area estimate. Roughly, the main outline is:

1. We already know that $\mathbb{E}\left[\mathrm{Vol}\left(A\right)\right] = \frac{\left|S\right|}{2}$, so to show a simple concentration of measure bound, we just need a handle on the variance of $\mathrm{Vol}\left(A\right)$.
2. The variance in the size of A can be written in terms of the probability that two points are in the same Voronoi cell. Instead of looking at the size of A, we can look at the size of A_B, the union of the black cells. Then

$$\begin{aligned}\mathrm{Var}\left(\mathrm{Vol}\left(A_B\right)\right) &= \int_{S^2}\mathbb{P}\left[x,y\in A_B\right]dxdy - \frac{\mathrm{Vol}\left(S\right)^2}{4}\\ &= \int_{S^2}\mathbb{P}\left[C\left(x\right)=C\left(y\right)\text{ and }x\in A_B\,\right]dxdy\\ &\quad+\int_{S^2}\mathbb{P}\left[C\left(x\right)\neq C\left(y\right)\text{ and }x,y\in A_B\,\right]dxdy - \frac{\mathrm{Vol}\left(S\right)^2}{4}.\end{aligned}$$

   Now, whether or not $C\left(x\right)=C\left(y\right)$ is independent of $x\in A_B$, so

$$\mathbb{P}\left[C\left(x\right)=C\left(y\right)\text{ and }x\in A_B\,\right] = \frac{1}{2}\mathbb{P}\left[C\left(x\right)=C\left(y\right)\right];$$

   similarly,

$$\begin{aligned}\mathbb{P}\left[C\left(x\right)\neq C\left(y\right)\text{ and }x,y\in A_B\,\right] &= \mathbb{P}\left[x,y\in A_B\;\middle|\;C\left(x\right)\neq C\left(y\right)\right]\mathbb{P}\left[C\left(x\right)\neq C\left(y\right)\right]\\ &= \frac{1}{4}\mathbb{P}\left[C\left(x\right)\neq C\left(y\right)\right].\end{aligned}$$

   Thus

$$\begin{aligned}\mathrm{Var}\left(\mathrm{Vol}\left(A_B\right)\right) &= \frac{1}{4}\int_{S^2}\mathbb{P}\left[C\left(x\right)=C\left(y\right)\right]+\frac{1}{4}\left(\int_{S^2}\mathbb{P}\left[C\left(x\right)=C\left(y\right)\right]+\int_{S^2}\mathbb{P}\left[C\left(x\right)\neq C\left(y\right)\right]-\mathrm{Vol}\left(S\right)^2\right)\\ &= \frac{1}{4}\int_{S^2}\mathbb{P}\left[C\left(x\right)=C\left(y\right)\right]dxdy.\end{aligned}$$

3. Most pairs of points on the surface are far away from each other (i.e. for any fixed $r > 0$, as $g \to \infty$, most pairs of points will have distance at least r).

4. Most pairs of points which are far away from each other have a significant amount of mass between them, so there are probably Poisson points in the middle, and they are probably not in the same cell. More formally, suppose that $B_r(x)$ and $B_r(y)$ are disjoint. If $C(x) = C(y)$, then it cannot be that both balls contain points from the Poisson process - otherwise each would be in the cell of its (distinct) closest point. So in this case,

$$\mathbb{P}[C(x) = C(y)] \leq e^{-\mu \mathrm{Vol}(B_r(x))} + e^{-\mu \mathrm{Vol}(B_r(y))},$$

and if the volumes of the balls $B_r(x)$ and $B_r(y)$ are large, then this is very small.
Now, it is true that you might be able to find pairs of points in S which will not have a large mass between them. For example, if S has a very short geodesic of length ℓ, it defines a narrow collar of finite volume and length roughly $\log(1/\ell)$. Two points on this collar won't have a lot of mass between them.

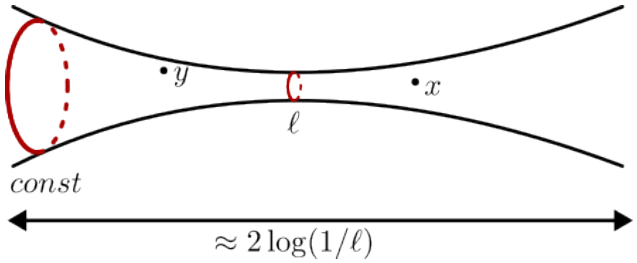


Figure 10.4: A narrow waist contains points which have a large distance but only a small mass between them.

But S cannot be composed of ONLY collars. For every $r_0 > 0$ we choose, the amount of mass of points with injectivity radius $\leq r_0$ is never more than $C \cdot r_0 \cdot \mathrm{Vol}(S)$. And outside of these points, the volume of balls $B_r(x)$ grows at least linearly in r (this is a bit imprecise, but can be made less so. The constant δ from the Cheeger assumption is used to control the size).

5. All of this eventually yields that $\mathrm{Var}(\mathrm{Vol}(A_B)) \leq O\left(\delta^3 |S|^2\right)$, Chebyshev's inequality gives the desired concentration of measure.

References

[1] Erik Jørgensen. *The Central Limit Problem for Geodesic Random Walks*.

[2] Ramesh Gangolli. *On the construction of certain diffusions on a differentiable manifold*.

[3] Paul Schmutz. *Reimann surfaces with shortest geodesic of maximal length*.

[4] Omer Angel, Tom Hutchcroft, Asaf Nachmias and Gourab Ray. *Hyperbolic and parabolic unimodular random maps*.

[5] Thomas Budzinski, Nicolas Curien and Bram Petri. *On the minimal diameter of closed hyperbolic surfaces*.

[6] Robert Brooks and Eran Makover. *Random Construction of Riemann Surfaces*.

[7] Robert Brooks. *Platonic surfaces*.

[8] Colin Adams and Frank Morgan. *Isoperimetric curves on hyperbolic surfaces*.x

[9] Alex Gamburd. *Poisson-Dirichlet distribution for random Belyi surfaces*.

[10] Yang Shen, Yunhui Wu. *Nearly optimal spectral gaps for random Belyi surfaces*.

[11] Thomas Budzinski, Nicolas Curien and Bram Petri. *The diameter of random Belyi surfaces*.

[12] Bram Petri. *Introduction to Teichmüller theory*.

[13] Laura Monk. *Geometry and Spectrum of typical hyperbolic surfaces*.

[14] Maryam Mirzakhani. *Simple geodesics and Weil-Petersson volumes of moduli spaces of bordered Riemann surfaces*.

[15] Nalini Anantharaman and Laura Monk. *Friedman-Ramanujan functions in random hyperbolic geometry and application to spectral gaps II*.

[16] Will Hide, Davide Macera, Joe Thomas. *Spectral gap with polynomial rate for Weil-Petersson random surfaces*.

[17] Charles Bordenave and Benoît Collins. *Eigenvalues of random lifts and polynomials of random permutation matrices*.

[18] Martin Liebeck, Aner Shalev. *Fuchsian groups, coverings of Riemann surfaces, subgroup growth, random quotients and random walks*.

[19] Michael Magee and Doron Puder. *The Asymptotic Statistics of Random Covering Surfaces*.

[20] Michael Magee, Doron Puder, and Ramon van Handel. *Strong convergence of uniformly random permutation representations of surface groups*.

[21] Thomas Budzinski, Nicolas Curien and Bram Petri. *On Cheeger constants of hyperbolic surfaces*.