4. GEODESICS

We gave in the previous section the definition of a geodesic curve. Namely a geodesic is a curve with zero geodesic curvature or equivalently:

Definition 4.1 *A curve $\gamma\colon I \to X$, parameterized by arc length on a surface X, is a **geodesic** if for all $s \in I$ the vector $\ddot{\gamma}(s)$ is normal to the surface at the point $\gamma(s)$.*

Geodesics are also the curves of shortest length on a surface – at least 'locally'. This means that given a geodesic between two points on a surface, varying the geodesic slightly will produce curves of greater length. For example, given two points on a sphere the great circle containing these two points is a geodesic. If the points are not antipodal, then there will be a shorter and longer arc connecting them. However both arcs are geodesics and locally are the shortest paths between the points.

We will see that geodesics are determined by the first fundamental form. Consequently an isometry between two surfaces will map geodesics in the first surface to geodesics in the second.

Theorem 4.2 *Let X be a smooth parameterized surface and γ be a smooth curve on X parameterized by arc length s. Then γ is a geodesic if and only the parameters $(u(s), v(s))$ of $\gamma(s)$ satisfy*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s}(E\dot{u} + F\dot{v}) &= \frac{1}{2}(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2) \\ \frac{\mathrm{d}}{\mathrm{d}s}(F\dot{u} + G\dot{v}) &= \frac{1}{2}(E_v\dot{u}^2 + 2F_v\dot{u}\dot{v} + G_v\dot{v}^2) \end{aligned} \tag{4.18}$$

for all s, where $E\mathrm{d}u^2 + 2F\mathrm{d}u\mathrm{d}v + G\mathrm{d}v^2$ is the first fundamental form of X.

Proof. As $\mathbf{r}_u$ and $\mathbf{r}_v$ are independent tangent vectors then $\ddot{\gamma}(s)$ is normal to the surface if and only if $\ddot{\gamma}(s) \cdot \mathbf{r}_u = 0$ and $\ddot{\gamma}(s) \cdot \mathbf{r}_v = 0$. Now

$$\dot{\gamma}(s) = \dot{u}\mathbf{r}_u + \dot{v}\mathbf{r}_v.$$

Thus

$$\begin{aligned} 0 &= \ddot{\gamma} \cdot \mathbf{r}_u = \frac{\mathrm{d}}{\mathrm{d}s}(\dot{\gamma} \cdot \mathbf{r}_u) - \dot{\gamma} \cdot \dot{\mathbf{r}}_u \\ &= \frac{\mathrm{d}}{\mathrm{d}s}(E\dot{u} + F\dot{v}) - (\dot{u}\mathbf{r}_u + \dot{v}\mathbf{r}_v) \cdot (\mathbf{r}_{uu}\dot{u} + \mathbf{r}_{uv}\dot{v}) \\ &= \frac{\mathrm{d}}{\mathrm{d}s}(E\dot{u} + F\dot{v}) - ((\mathbf{r}_{uu} \cdot \mathbf{r}_u)\dot{u}^2 + (\mathbf{r}_{uu} \cdot \mathbf{r}_v + \mathbf{r}_{uv} \cdot \mathbf{r}_u)\dot{u}\dot{v} + (\mathbf{r}_{uv} \cdot \mathbf{r}_v)\dot{v}^2) \\ &= \frac{\mathrm{d}}{\mathrm{d}s}(E\dot{u} + F\dot{v}) - \frac{1}{2}(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2), \end{aligned}$$

as required. The second geodesic equation follows similarly. ■

Given two points on a surface there need not be a geodesic connecting the two points. For example in $\mathbb{R}^2$ the geodesics are line segments. So in the punctured plane $\mathbb{R}^2\backslash\{\mathbf{0}\}$ there is no geodesic connecting $(1, 0)$ and $(-1, 0)$. Also if a geodesic exists between two points it need not be unique (see the examples of the sphere and cylinder below.) However geodesics always exist locally (Do Carmo p.255):

Theorem 4.3 *Given a point $p \in X$ and a non-zero vector $\mathbf{v} \in T_pX$ then there exists $\varepsilon > 0$ and a unique geodesic $\gamma\colon (-\varepsilon, \varepsilon) \to X$ parameterized by arc-length such that $\gamma(0) = p$ and $\gamma'(0) = \mathbf{v}$.*

The proof of this theorem is beyond this syllabus and in any case largely relates to the analysis of differential equations. The existence and uniqueness of geodesics, at least locally, leads to the notion of *geodesic polar co-ordinates* (mentioned in Remark 3.16). When polar co-ordinates are used to parameterize the plane, from the origin, we obtain a first fundamental form with $E = 1$ and $F = 0$. More generally, we can locally parameterize a surface around a point p, by assigning co-ordinates r and θ to the point of the surface that is distance r from p when measured along the geodesic making an angle θ at p with some fixed tangential direction. When we do this we find that $E = 1$ and $F = 0$ (Do Carmo p.287).

For many surfaces geodesics are not just locally defined but many be extended indefinitely – such surfaces are called *complete surfaces*. Note that this may mean that the geodesic wraps back on to itself as with a great circle on a sphere. This notion of completeness coincides with the notion of complete metric spaces. Given a connected surface X it can be shown that, given two points a, b of the surface, there is a piecewise-smooth curve between the points. The *intrinsic distance* $d(a, b)$ between the points can then be defined as the infimum

$$d(a, b) = \inf_{\gamma} \mathcal{L}(\gamma)$$

where $\mathcal{L}(\gamma)$ is the length of a curve γ and the infimum is taken over all piecewise smooth curves γ in X which connect a and b. The function d is a metric on X and the *Hopf-Rinow theorem* states that:

Theorem 4.4 ***(Hopf-Rinow)*** *(Off-syllabus) For a connected, smooth geometric surface X the following are equivalent:*

(a) (X, d) is a complete metric space.

(b) A geodesic can be indefinitely extended.

The following examples are left to Sheet 3, Part B, Exercise 1.

Example 4.5 *(a) The geodesics on a sphere are arcs of great circles.*

(b) The geodesics on a cylinder are the meridians, the latitudes and helices. So between two points of the cylinder, that do not lie on the same meridian or parallel, there are infinitely many geodesics between the points.

Example 4.6 *(a) Prove that a meridian on a surface of revolution is a geodesic.*

(b) When is a parallel of latitude a geodesic on such a surface?

Solution. Suppose that the surface of revolution is generated by rotating the curve $y = f(x)$ about the x-axis and parameterize it as

$$\mathbf{r}(x, \theta) = (x, f(x)\cos\theta, f(x)\sin\theta), \qquad x \in \mathbb{R},\ \theta \in (-\pi, \pi).$$

By Sheet 2, Part A, Exercise 2, the first fundamental form equals

$$(1 + f'(x)^2)\mathrm{d}x^2 + f(x)^2\mathrm{d}\theta^2$$

and the geodesic equations are

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}((1+f'(x)^2)\dot{x}) &= f'(x)(f''(x)\dot{x}^2+f(x)\dot{\theta}^2),\\ \frac{\mathrm{d}}{\mathrm{d}s}(f(x)^2\dot{\theta}) &= 0.\end{aligned}$$

(a) Along a meridian $\dot{\theta}=0$ and $\dot{x}=(1+f'(x)^2)^{-1/2}$. The second equation is then trivially true and substituting into the first equation we find

$$\frac{\mathrm{d}}{\mathrm{d}s}((1+f'(x)^2)\dot{x}) = \dot{x}\frac{\mathrm{d}}{\mathrm{d}x}(1+f'(x)^2)^{1/2} = \frac{f'(x)f''(x)}{1+f'(x)^2} = f'(x)f''(x)\dot{x}^2$$

as required.

(b) A parallel is given by the equation $\dot{x}=0$. Thus the two geodesic equations now read as $f(x)f'(x)\dot{\theta}^2=0$ and $f(x)^2\ddot{\theta}=0$. As the geodesic is a circle parameterized by arc-length then $\dot{\theta}$ is a non-zero constant and $\ddot{\theta}=0$. As $f(x)>0$, then the equations hold if and only if $f'(x)=0$. ■

We now prove an earlier comment on geodesics namely that they are locally curves of least length. That is, however a geodesic between two points is perturbed, we produce curves of greater length

Theorem 4.7 *Let $\gamma\colon [a,b]\to X$ be a smooth geodesic in X. Let γ_δ, where $\delta\in(-\varepsilon,\varepsilon)$, be a family of smooth curves*

$$\gamma_\delta\colon [a,b]\to X$$

with $\gamma_0=\gamma$ and $\gamma_\delta(a)=\gamma(a),\gamma_\delta(b)=\gamma(b)$ for all $\delta\in(-\varepsilon,\varepsilon)$ and let $\mathcal{L}(\delta)=\mathcal{L}(\gamma_\delta)$. Then $\mathcal{L}'(0)=0$.

Proof. (Proof non-examinable) Let $R(\delta,t)=E\dot{u}^2+2F\dot{u}\dot{v}+G\dot{v}^2$ where $\gamma_\delta(t)=\mathbf{r}(u(\delta,t),v(\delta,t))$ and the dot denotes differentiation with respect to t. Then

$$\mathcal{L}(\delta)=\int_a^b \sqrt{R}\ \mathrm{d}t$$

giving

$$\mathcal{L}'(0)=\frac{\mathrm{d}}{\mathrm{d}\delta}\bigg|_{\delta=0}\int_a^b\sqrt{R}\,\mathrm{d}t=\int_a^b\frac{\partial\sqrt{R}}{\partial\delta}\bigg|_{\delta=0}\mathrm{d}t=\frac{1}{2}\int_a^b\frac{1}{\sqrt{R}}\frac{\partial R}{\partial\delta}\bigg|_{\delta=0}\mathrm{d}t, \tag{4.19}$$

by differentiation under the integral sign. Now

$$\begin{aligned}\frac{\partial R}{\partial\delta} &= \{E_u\dot{u}^2+2F_u\dot{u}\dot{v}+G_u\dot{v}^2\}\frac{\partial u}{\partial\delta}\\ &+ \{E_v\dot{u}^2+2F_v\dot{u}\dot{v}+G_v\dot{v}^2\}\frac{\partial v}{\partial\delta}\\ &+ 2(E\dot{u}+F\dot{v})\frac{\partial\dot{u}}{\partial\delta}+2(F\dot{u}+G\dot{v})\frac{\partial\dot{v}}{\partial\delta}.\end{aligned}$$

As $\gamma = \gamma_0$ is a geodesic then substituting in the geodesic equations (4.18)

$$\begin{aligned}\left.\frac{\partial R}{\partial \delta}\right|_{\delta=0} &= 2\left[\frac{\mathrm{d}}{\mathrm{d}t}(E\dot{u}+F\dot{v})\frac{\partial u}{\partial \delta}+(E\dot{u}+F\dot{v})\frac{\partial \dot{u}}{\partial \delta}\right.\\ &\quad\left.+\ \frac{\mathrm{d}}{\mathrm{d}t}(F\dot{u}+G\dot{v})\frac{\partial v}{\partial \delta}+(F\dot{u}+G\dot{v})\frac{\partial \dot{v}}{\partial \delta}\right]_{\delta=0}\\ &= 2\frac{\mathrm{d}}{\mathrm{d}t}\left\{(E\dot{u}+F\dot{v})\left.\frac{\partial u}{\partial \delta}\right|_{\delta=0}+(F\dot{u}+g\dot{v})\left.\frac{\partial v}{\partial \delta}\right|_{\delta=0}\right\}. \qquad (4.20)\end{aligned}$$

We may assume without loss of generality that $\gamma = \gamma_0$ is parameterized by arc length so that $R(0,t) = 1$. Hence, substituting (4.20) into (4.19),

$$\begin{aligned}\mathcal{L}'(0) &= \int_a^b \frac{\mathrm{d}}{\mathrm{d}t}\left\{(E\dot{u}+F\dot{v})\left.\frac{\partial u}{\partial \delta}\right|_{\delta=0}+(F\dot{u}+g\dot{v})\left.\frac{\partial v}{\partial \delta}\right|_{\delta=0}\right\}\mathrm{d}t\\ &= \left[(E\dot{u}+F\dot{v})\left.\frac{\partial u}{\partial \delta}\right|_{\delta=0}+(F\dot{u}+g\dot{v})\left.\frac{\partial v}{\partial \delta}\right|_{\delta=0}\right]_{t=a}^{t=b}.\end{aligned}$$

However $u(\delta,a), u(\delta,b), v(\delta,a)$ and $v(\delta,b)$ are all constant giving $\partial u/\partial\delta = \partial v/\partial\delta = 0$ when $t = a$ and $t = b$ and hence $\mathcal{L}'(0) = 0$. ■

Example 4.8 *What are the geodesics in the hyperbolic plane $\mathbb{H}$? (See Examples 3.28 and 3.54.)*

Solution. Method 1: If we substitute $E = G = y^{-2}$ and $F = 0$ into the geodesic equations (4.18) then we find

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\dot{x}}{y^2}\right) = 0, \qquad \frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\dot{y}}{y^2}\right) = \frac{-(\dot{x}^2+\dot{y}^2)}{y^3}.$$

The first equation yields $\dot{x} = cy^2$ for some constant c. So the half-lines $x =$ constant are then geodesics corresponding to $c = 0$. Assume that $c \neq 0$. The second equation may be rewritten as

$$\frac{\ddot{y}y-\dot{y}^2}{y^2} = \frac{-\dot{x}^2}{y^2},$$

or equivalently

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{\dot{y}}{y}\right) = -c\dot{x}.$$

Integrating we find that $\dot{y} = (b-cx)y$ for some constant b. Now

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\dot{y}}{\dot{x}} = \frac{b-cx}{cy},$$

and solving this differential equation gives

$$\frac{1}{2}c(x^2+y^2) - bx = a,$$

for some constant a, which is the equation of a semicircle in $\mathbb{H}$ which cuts the x-axis orthogonally.

Method 2: Alternatively we could consider what the isometries of $\mathbb{H}$ might be and use the fact that geodesics are mapped to other geodesics by isometries. For ease of notation we now introduce a complex variable $z = x + iy$ so that the first fundamental form on $\mathbb{H}$ is now given by

$$\frac{-4|\mathrm{d}z|^2}{(z-\bar{z})^2}.$$

Then I claim the map

$$w\colon z \mapsto \frac{az+b}{cz+d},$$

where a, b, c, d are real numbers satisfying $ad - bc = 1$, is an isometry of $\mathbb{H}$. From standard theorems concerning Möbius transformations we can see that w maps the upper half plane onto the upper half plane; as a, b, c, d are real, the real axis is mapped to the real axis and the imaginary part of the image of i equals

$$\mathrm{Im}\left(\frac{ai+b}{ci+d}\right) = \frac{\mathrm{Im}\left((ai+b)\left(d-ci\right)\right)}{c^2+d^2} = \frac{ad-bc}{c^2+d^2} = \frac{1}{c^2+d^2} > 0.$$

To check w is an isometry we need to prove that $\mathbb{H}$ when parameterized by w and z has the same first fundamental form. Firstly note

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

So

$$\frac{-4|\mathrm{d}w|^2}{(w-\bar{w})^2} = \frac{\frac{-4|\mathrm{d}z|^2}{|cz+d|^4}}{\left(\frac{az+b}{cz+d} - \frac{a\bar{z}+b}{c\bar{z}+d}\right)^2} = \frac{-4|\mathrm{d}z|^2}{((az+b)(c\bar{z}+d) - (cz+d)(a\bar{z}+b))^2}.$$

The denominator in the final expression above factorises as $(ad-bc)^2(z-\bar{z})^2$ showing that

$$\frac{-4|\mathrm{d}w|^2}{(w-\bar{w})^2} = \frac{-4|\mathrm{d}z|^2}{(z-\bar{z})^2}$$

and consequently w is an isometry.

Note now that $x = 0, y = e^{-s}$ is a solution to the geodesic equations for $\mathbb{H}$ (Sheet 3, Part A, Exercise 1) and so the positive imaginary axis is a geodesic. As we show below, there is a Möbius map of the same form as w which maps any other half line or semicircle orthogonal to the positive imaginary axis, showing that these too are examples of geodesics. From Theorem 4.3 we know that these are all the geodesics of $\mathbb{H}$.

Given another half-line $\mathrm{Re}\, z = k$ then the Möbius map $z \mapsto z - k$ (where $a = 1, b = -k, c = 0, d = 1$ so that $ad - bc = 1$) takes the half-line to the positive imaginary axis. For the semicircle perpendicular to the real axis, meeting at p and q (where $p < q$), the Möbius map

$$z \mapsto \frac{1}{\sqrt{q-p}}\left(\frac{p-z}{z-q}\right)$$

takes the semicircle to the positive imaginary axis. Again we check

$$ad - bc = \left(\frac{-1}{\sqrt{q-p}}\right)\left(\frac{-q}{\sqrt{q-p}}\right) - \left(\frac{p}{\sqrt{q-p}}\right)\left(\frac{1}{\sqrt{q-p}}\right) = \frac{q-p}{q-p} = 1.$$

■

Remark 4.9 *(**Historical context**) The hyperbolic plane $\mathbb{H}$ is of interest because it is an example of a non-Euclidean geometry. A Euclidean geometry is one that satisfies certain axioms including the **parallel postulate** which states that:*

- *given a line l and a point p not on l then there is a unique line through p (known as a parallel) which does not meet l.*

If we read 'geodesic' for 'line' in the above, then we see that given a line l in $\mathbb{H}$ and a point p not on the line then there are infinitely many lines through p not meeting l. (In Figure 4.1 M_1, M_2, M_3 are parallels of L through P.)

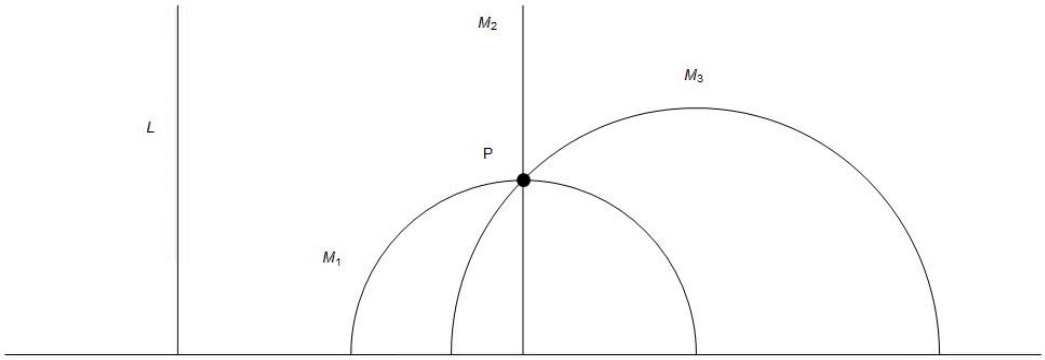

Figure 4.1 – hyperbolic parallels

For literal centuries, mathematicians had been trying to deduce the axiom of parallels from Euclid's other axioms. Instead all they managed to find were alternative, equivalent formulations for the parallel postulate. The above formulation is in fact due to Ludlam (1785) though it is usually attritbuted to Playfair; other formulations include:

- *parallel lines are everywhere equidistant.*
- *the sum of the angles of a triangle equals two right angles.*
- *given a triangle, we can construct a similar triangle of any area.*
- *Pythagoras' theorem.*
- *three non-collinear points always lie on a circle.*

In the nineteenth century certain mathematicians – notably Bolyai, Lobachevsky and Gauss – began to suspect that the parallel postulate was independent of the other axioms and proved alternative theory where more than one parallel existed. Such theory might still have contained inconsistencies, but this was shown not to be the case when Beltrami, Klein and Poincaré found models for the hyperbolic plane which showed the new geometry to be every bit as consistent as Euclidean geometry.

The elliptic plane (Example 3.32) is another example of non-Euclidean geometry, in which case there are no parallels to a line. The elliptic plane had previously been discounted as a non-Euclidean geometry as it did not seem to meet another axiom of Euclid that lines can be extended indefinitely. But if we permit lines to be extended repeatedly on to themselves then the elliptic plane is a valid non-Euclidean geometry.