GEODESICS IN HYPERBOLIC SPACE AND NUMBER THEORY

YIANNIS N. PETRIDIS

Abstract. Geometry studies geodesics in various settings, in particular on hyperbolic surfaces. The distribution of geodesics on arithmetic surfaces gives information on the arithmetic of quadratic forms, an important branch of Number Theory.

1. Geodesics and the sphere

A geodesic curve is the path of a point in our space that is moving without friction and without external forces. A geodesic curve minimizes the distance between two points, at least when these are close enough to each other. While on our standard euclidean space the shortest distance between two points is given by the length of the line segment between them, the situation becomes more interesting and more complicated in general. For example the earth is (approximately) a sphere. When we fly from Frankfurt to Los Angeles, the plane goes over Greenland. This route at high latitude corresponds to the fact that the shortest path (geodesic) on the sphere is along a great circle, i.e., a circle centered at the center of the sphere (earth). Another example would be a trip from London (on the 0^o meridian) and Fiji in the Pacific Ocean with longitude 180^o. The geodesic between these two places goes over the North Pole. Such a geodesic is shown in Fig. 1. A direct flight to Fiji should follow this route. The geodesics starting at a point can go in whatever direction we choose (given by their vector of initial velocity) but they will all meet at the antipodal point. This is also shown in Fig. 1.

2. Hyperbolic space and its geodesics

Euclid's fifth postulate (or parallel axiom) can be stated as follows: Given any line and a point not on it, there exists one and only one line which passes through that point and never intersects the first line. In modern geometry the lines are the geodesics. Hyperbolic geometry is an example where Euclid's fifth postulate fails. It was created by Bolyai and Lobatchevsky. In hyperbolic geometry geodesics diverge from each other, i.e., the distance between two geodesics, which do not meet but get very close in one direction, increases exponentially in the other direction. The distance is measured on the curves meeting both geodesics perpendicularly. One simple model of hyperbolic geometry is the hyperbolic disc: the points in the interior of the circle of radius 1, given by the equation $x^2 + y^2 = 1$. The geodesics are the diameters (Fig. 2 a, b) and the circular arcs meeting the circle $x^2 + y^2 = 1$ perpendicularly (Fig. 2 c, d, e, f, g, h). The parallel axiom fails, see Fig. 2: from P there are two parallels g, h to f. The hyperbolic distance from the center to the

Date: March 22, 2006.

The author was partially supported by a Humboldt Foundation Research Fellowship, and NSF grant DMS 0401318.

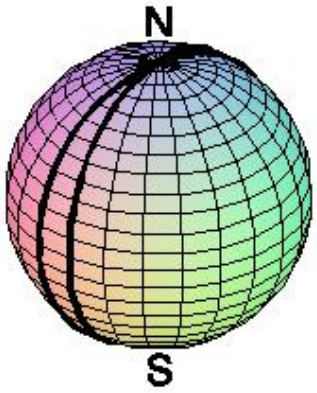

FIGURE 1. Two meridian geodesics on the sphere from the North Pole (N) to the South Pole (S) (antipodal points)

point (x, y) is $\log \frac{1+r}{1-r}$, where $r = \sqrt{x^2 + y^2}$ is the euclidean distance. As the point gets close to the circle of radius 1, the hyperbolic distance increases and becomes unbounded. Hyperbolic geometry is as real to mathematicians as euclidean. It could even be that the universe is a hyperbolic space, as analysis of the data on the cosmic microwave background (CMB) by Aurich, Lustig, Steiner, Then [ALST] has suggested.

3. Hyperbolic surfaces and their geodesics

Spaces modelled on the hyperbolic disc are called hyperbolic surfaces. They are important for many reasons: they are connected to the theory of automorphic forms, one of the main areas of research of the Max-Planck-Institut für Mathematik. They are among the easiest spaces (manifolds) that can be classified topologically, i.e., in terms of continuous bending into simpler spaces.

For a hyperbolic surface M some of the geodesics γ will come back to the point they start and fit in a smooth way. These are called closed geodesics. It ends up that there are finitely many closed geodesics of a given length (if any). Moreover, the lengths $l(\gamma)$ form an increasing set of numbers that can accumulate only at infinity. This allows us to count closed geodesics by

$$\pi(x) = \#\{\gamma, l(\gamma) \leq x\}.$$

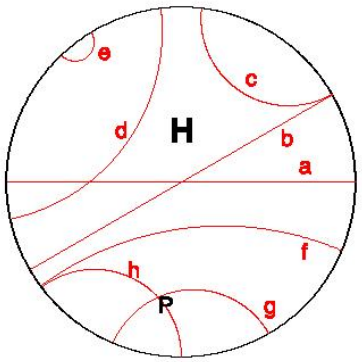

FIGURE 2. The hyperbolic disc and its geodesics

This function, which captures various aspects of the geometry of the surface, has been studied in dynamical systems and with methods of analysis. Huber [Hu] and Selberg, and in a more general case Margulis [M] proved that a good approximation of $\pi(x)$ is the function e^x/x. And this for every hyperbolic surface!

The Selberg trace formula is an important tool in understanding the function $\pi(x)$ and a research interest of many visitors at MPI and one of the scientific directors (D. Zagier). It is a generalization of the well-known fact that is taught in basic linear algebra courses. The trace of a symmetric matrix can be computed in two different ways: one way is to sum the diagonal entries, the other way is to sum the eigenvalues. While one may think that the second method is harder, it may provide valuable information, if, for instance, some information on the eigenvalues is easily obtainable. The Selberg trace formula can be considered as a generalization of the two methods of calculating the trace for infinite matrices: these give operators on Hilbert spaces. The lengths of the closed geodesics correspond to the diagonal entries and the eigenvalues are the eigenvalues of the Laplace-Beltrami operator. The asymptotic behavior of $\pi(x)$ is due to the fact that the smallest eigenvalue (basic frequency) is 0. The eigenvalues of the Laplace-Beltrami operator are the principal frequencies of M: they are the harmonics that the surface M will produce when used as a drum.

4. Arithmetic of quadratic forms and applications

If the hyperbolic surface has arithmetic nature, important number-theoretic consequences follow. The discussion starts with the simplest case of Pell's Equation. See [D, Chapter XII p. 341] and [He]. Proclus (410–485 A.D.) noted that the Pythagoreans made a geometric construction, which amounts to an algorithm for solving the diophantine equation

$$x^2 - 2y^2 = \pm 1.$$

Solving means to find all pairs of integers x and y that satisfy this equation. The method starts with the pair $(x_1, y_1) = (1, 1)$, which is the smallest solution for

$x^2 - 2y^2 = -1$. Given a solution (x_n, y_n), the numbers

$$x_{n+1} = 2y_n + x_n, \quad y_{n+1} = y_n + x_n$$

give a solution for the equation with the opposite sign. Theon of Smyrna (approx. 130 A.D.) stated this result and called the x_n diagonal numbers ($\delta\iota\alpha\mu\epsilon\tau\varrho\iota\kappa\omicron\iota$ $\alpha\varrho\iota\vartheta\mu\omicron\iota$) and the y_n side numbers ($\pi\lambda\varepsilon\upsilon\varrho\iota\kappa\omicron\iota$ $\alpha\varrho\iota\vartheta\mu\omicron\iota$).

The pairs constructed this way are $(x_2, y_2) = (3, 2)$, which is the smallest solution of $x^2 - 2y^2 = 1$, $(x_3, y_3) = (7, 5)$, $(x_4, y_4) = (17, 12)$, etc. Theon as a Neoplatonic and Plato's school were interested in this problem, because they knew that $2y^2$ with y integer cannot be the square of an integer (see Proclus' commentary to Euclid I. 47). This simply means that $\sqrt{2}$ cannot be expressed as a fraction with integer numerator and denominator (such numbers are called irrational). So Theon was looking for the closest possibility, i.e., that $2y^2$ differs from the square of a number x^2 just by one (± 1). In modern algebraic notation Theon's construction works, since

$$x_{n+1}^2 - 2y_{n+1}^2 = (2y_n + x_n)^2 - 2(y_n + x_n)^2 = -(x_n^2 - 2y_n^2).$$

In modern algebraic number theory the pairs (x_n, y_n) are defined by the formula

$$x_n + y_n\sqrt{2} = (1 + \sqrt{2})^n.$$

More generally the expression

$$Q(x, y) = ax^2 + bxy + cy^2,$$

where the coefficients a, b, c are integers is a integral quadratic form and an integer N is represented by it, if we can find integer x and y such that $Q(x, y) = N$. For simplicity attention is restricted to the case when a, b, c are relatively prime numbers, i.e., the only integers dividing *all* three are ± 1. It is called indefinite, if $Q(x, y)$ represents both negative and positive numbers. Such a quadratic form has a discriminant $d = b^2 - 4ac > 0$, the same expression students meet in their school studies with relation to solving $ax^2 + bx + c = 0$. For a positive integer d we can have two different quadratic forms $Q(x, y)$ and $Q'(x, y)$ of discriminant d that represent exactly the same numbers. Such forms are 'identified'. Mathematicians say that they are equivalent. For example $Q(x, y) = x^2 - 2y^2$ is equivalent to $-2x^2 + y^2$, as obvious by interchanging the role of x and y. But forms can be equivalent in a more subtle way. The above two forms are also equivalent to $Q'(x, y) = x^2 + 2xy - y^2$. This follows by the simple calculations

$$(x + y)^2 - 2y^2 = x^2 + 2xy - y^2, (x - y)^2 + 2(x - y)y - y^2 = x^2 - 2y^2.$$

They show that $Q(x + y, y) = Q'(x, y)$ and $Q'(x - y, y) = Q(x, y)$ and consequently the two forms represent the same numbers. We encode the changes of variables showing the equivalence of Q with $-2x^2 + y^2$ and Q' in two matrices

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

giving us linear transformations in the xy-plane. These transformations are

$$S(x, y) = (-y, x), \quad T(x, y) = (x + y, y).$$

It can be proved that, given $d > 0$ there are only finitely many *nonequivalent* forms with discriminant d. Their number is called the class number and denoted by $h(d)$. The study of quadratic forms, their class numbers $h(d)$ and the numbers they represent is an active area of research. D. Hilbert outlines 23 important problems

to be studied in the twentieth century in the published version of his address at the International Congress of Mathematicians in Paris (Arkiv der Mathematik und Physik, 1901). Number 11 deals with quadratic forms with arbitrary algebraic number coefficients.

What the ancient Greeks could not possibly know is the close relationship of quadratic forms with geodesics in hyperbolic surfaces. This goes as follows: one searches for the smallest solution (x_d, y_d) of the equation

$$x^2 - dy^2 = 4.$$

This is a point on the hyperbola $x^2 - dy^2 = 4$. The fundamental unit is defined to be

$$\epsilon_d = \frac{x_d + y_d\sqrt{d}}{2}.$$

As noticed above, for $d = 2$, this is $3+2\sqrt{2}$. The lengths of the all closed geodesics in a certain arithmetic hyperbolic surface (associated with the matrices T and S) are identified with $2\log\epsilon_d$, with each length appearing $h(d)$ times. Here d is restricted to be nonsquare and leave residue 0 or 1, when devided by 4. This enabled Sarnak [S] to prove that (for the same d)

$$\sum_{\epsilon_d \le x} h(d) \sim \frac{x^2}{2\ln x}.$$

The $\sim$ symbol means that the quotient of the two expressions on its left and right is asymptotic to 1, as $x \to \infty$.

5. Refined distribution of the geodesics

The topological nature of M is captured by the number of holes of M. This is called the genus g. It is easier to understand the notion of homology on a surface of genus 1 as in Fig. 3. It looks like a doughnut and is called a torus. To every geodesic γ (or, generally, a closed loop) we can assign its homology $\phi(\gamma)$, which is a pair of integers describing how many times it wraps around the hole or the torus. The meridians traversed once wrap around the torus once, while the circles of fixed latitude can be pulled to shrink to the small circle wrapping around the hole. The more complicated loop in Fig. 3 wraps once around a meridian and once around the small inner circle, when we shrink it by pulling.

More generally for a surface of genus g the homology $H_1(M, \mathbb{Z})$ can be represented by points $(n_1, n_2, \ldots, n_{2g})$ with integer coordinates. Now a refined problem can be posed: How are the lengths of geodesics distributed for geodesics having homology restricted in a given subset of $H_1(M, \mathbb{Z})$? For geodesics that all have the *same* homology the answer is due to Phillips and Sarnak [PS], who investigated the function

$$\pi(x, \beta) = \#\{\gamma \in \pi(x), \phi(\gamma) = \beta\}.$$

They proved that

$$\pi(x, \beta) \sim (g-1)^g \frac{e^x}{x^{g+1}}.$$

Note that this depends only on the topological invariant g and not on β! In recent work M. S. Risager and I investigated geodesics whose homology is restricted to belong to some set $A \subset H_1(M, \mathbb{Z})$. We define

$$\pi(x, a) = \#\{\gamma \in \pi(x), \phi(\gamma) \in A\}.$$

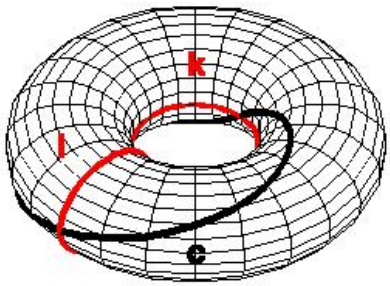

FIGURE 3. Torus with a homology basis k, l and a loop c

We measure the set A by its density $d(A)$, telling us what proportion of the lattice points are in A. We find that for lots of sets, including random sets A,

$$\pi(x, A) \sim d(A)\pi(x).$$

The geodesic flow on hyperbolic space is an example of a chaotic flow, an object of extreme interest in the theory of dynamical systems. Using dynamical systems Eskin, Margulis and Mozes [EMM] have studied the values of other interesting indefinite quadratic forms. The interplay between number theory and dynamical systems will undoubtedly produce new spectacular results in the future.

REFERENCES

[ALST] R. Aurich, S. Lustig, F. Steiner and H. Then: Indications about the Shape of the Universe from the Wilkinson Microwave Anisotropy Probe Data, Physics Review Letters 94 (2005) 021301.

[D] L. Dickson: History of the theory of numbers. Vol. II: Diophantine analysis. Chelsea Publishing Co., New York 1966 xxv+803 pp.

[EMM] A. Eskin, G. Margulis, S. Mozes: Quadratic forms of signature (2,2) and eigenvalue spacings on rectangular 2-tori. Annals of Mathematics (2) 161 (2005), no. 2, 679–725

[He] T. Heath: A history of Greek mathematics. Vol. I. From Thales to Euclid. Corrected reprint of the 1921 original. Dover Publications, Inc., New York, 1981. xv+446 pp. ISBN 0-486-24073-8 and Vol. II. From Aristarchus to Diophantus. Corrected reprint of the 1921 original. Dover Publications, Inc., New York, 1981. xi+586 pp. ISBN 0-486-24074-6

[Hu] H. Huber: Zur analytischen Theorie hyperbolischen Raumformen und Bewegungsgruppen. Mathematische Annalen 138 (1959) 1–26.

[M] G. Margulis: On some aspects of the theory of Anosov systems. With a survey by Richard Sharp: Periodic orbits of hyperbolic flows. Translated from the Russian by Valentina Vladimirovna Szulikowska. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2004. vi+139 pp. ISBN: 3-540-40121-0

[PS] R. Phillips, P. Sarnak: Geodesics in homology classes. Duke Mathematical Journal 55 (1987), no. 2, 287–297.

[S] P. Sarnak: Class numbers of indefinite binary quadratic forms. Journal of Number Theory 15 (1982), no. 2, 229–247.

The Graduate Center, Mathematics Ph.D. Program, 365 Fifth Avenue, Room 4208, New York, NY 10016-4309

Department of Mathematics and Computer Science, City University of New York, Lehman College, 250 Bedford Park Boulevard West Bronx, NY 10468-1589

E-mail address: petridis@mpim-bonn.mpg.de