

HILBERT'S FOURTH PROBLEM IN TWO DIMENSIONS I

J.C. ÁLVAREZ PAIVA

ABSTRACT. Hilbert's fourth problems asks *to construct and study the geometries in which the straight line segment is the shortest connection between two points*. In this paper the reader shall find an elementary introduction to the problem and its solutions in dimension two by Busemann, Pogorelov, and Ambartzumian. The relationship between integral geometry and inverse problems in variational calculus is emphasized.

CONTENTS

1. Introduction	1
2. Basic definitions and interpretations of the problem	2
3. Minkowski planes	6
4. Hilbert geometries	9
5. The Crofton formula	10
6. Busemann's construction of projective metrics	13
7. Ambartzumian's construction	14
8. Variational interpretation of Hilbert's fourth problem	15
9. Analytic solution of Hilbert's fourth problem	17
10. A glimpse ahead	20
References	21

1. INTRODUCTION

Hilbert's list of problems, read at the International Congress of Mathematicians in 1900, is perhaps one of the most influential documents in the history of mathematics. The twenty-three problems in this list have been the subject of numerous investigations for the last hundred years and continue to yield much beautiful mathematics. Even when one of Hilbert's problems has been solved in its original formulation, its variations and the developments arising from its solution continue to pique the curiosity of mathematicians. One example is Hilbert's third problem on the decomposition of polyhedra (see chapter seven in [1]). This problem was solved by M. Dehn ([23]) just two years after Hilbert's address, but the concepts he introduced have evolved in different directions. For instance, the theory of

Key words and phrases. Integral geometry, Crofton formula, Hilbert's fourth problem, Minkowski spaces, cosine transform.

valuations, a central part of modern convex geometry, is a direct descendent of Dehn's solution.

In posing his problems, Hilbert did not shy away from vague statements which would be subject to interpretation. Problem six on the mathematical treatment of the axioms of physics is perhaps the prime example of this. Hilbert's fourth problem, the subject of this note, is another.

In its original formulation, Hilbert's fourth problem asks *to construct and study the geometries in which the straight line segment is the shortest connection between two points*. The original wording (see [25]) makes one think the problem is part of Hilbert's project to study the foundations of geometry. However, the different modern approaches make it clear that the problem is at the basis of integral geometry, inverse problems in the calculus of variations, and Finsler geometry.

In this paper the reader will find various approaches to the solution of Hilbert's fourth problem in two dimensions. We will start with the examples given by Minkowski and Hilbert of metrics where the shortest connection between two points is a line segment, and proceed to the constructions of Busemann and Ambartzumian which together give the most elegant and natural solution to the problem. Later, we will cover Pogorelov's approach using Hamel's equations and the cosine transform. A second paper shall treat the subject from the symplectic viewpoint.

An effort has been made to make the paper partly accessible to undergraduate students and mathematically minded readers from other disciplines. The material is presented as a collection of exercises some of which point at extensions or digressions of the ideas which led to the solution of Hilbert's fourth problem in two dimensions. From the author's viewpoint, Hilbert's fourth problem is simply an axis around which revolve two fruitful lines of research: the relation between integral geometry and variational calculus and the relation between metric and symplectic geometry. The advanced reader is asked to keep these two themes in mind as he/she reads the paper.

For more information on Hilbert's problems the reader is referred to the two volumes "Mathematical developments arising from Hilbert problems", Proceedings of Symposia in Pure Math. Vol. XXVII Part 1, F. Browder (Ed.), AMS Rhode Island, 1974, and to the recently published "The Honors Class: Hilbert's problems and their solvers" by Benjamin Yandell, A K Peters, Ltd, 2001.

2. BASIC DEFINITIONS AND INTERPRETATIONS OF THE PROBLEM

Most interpretations of Hilbert's fourth problem revolve around the notion of *distance* or *metric*. *To find all the possible definitions of distance on the plane for which the shortest path between two given points is the line segment that joins them* is a clear and attractive formulation. Of course, our usual notion of distance in two dimensions, borrowed from Euclidean geometry, satisfies this property and we may wonder if there is any other that does. In

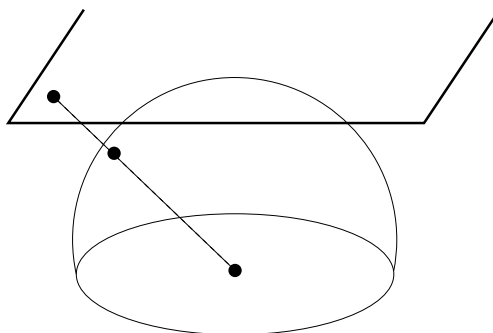


Fig. 1. Charting the northern hemisphere.

fact, all the three classical geometries, Euclidean, elliptic, and hyperbolic, define distances on the plane or portions of it for which the shortest path between two points is the line segment that joins them. Hyperbolic geometry and the generalization that led Hilbert to pose the fourth problem will be treated in the next section, but the elliptic case — the geometry of the sphere — is easy to explain to the proverbial “first man on the street”:

Suppose you’re a very small being living in the northern hemisphere of a perfectly round planet. At the last planetary convention, your nation’s geographers proposed the incredibly impractical idea of charting an infinite map of the surface of the northern hemisphere by drawing above the north pole a virtual plane parallel to the equator. For every point in the surface of the planet, they would compute the line from the center of the planet to the point and then compute where this line cut the plane. That would give them a point on the plane for every point on the surface. If we define the distance between two points in the map as the distance of the corresponding points on the surface of the planet, we will get a new sort of distance very different from the Euclidean, but for which the shortest curve between two points is still the straight line that joins them.

In order to tackle Hilbert’s fourth problem we need to understand precisely the notion of distance. We have seen that the distance between two points in the plane can be defined in different ways, but it would be hard for us to accept that the distance between two points be a negative number, or that two distinct points be at zero distance from each other. Mathematicians abstracted such “common sense” properties of distances and came up with the following three axioms:

- (1) The distance between any two points is greater than or equal to zero, and it is equal to zero if and only if the points coincide.
- (2) The distance between a point x and a point y is the same as the distance between y and x .
- (3) If x , y , and z are three points, then the distance between x and z is not greater than the distance between x and y plus the distance between y and z .

Any measurement that assigns to any two points a number and that satisfies the three properties above is deemed to be a possible way of measuring distances. It is sometimes convenient to admit that two points are at infinite distance from each other. The definition that you would find in a textbook runs as follows:

Definition 2.1. A metric space is a pair (X, d) , where X is a set and the *distance function*, or *metric*,

$$d : X \times X \longrightarrow \mathbb{R} \cup \{\infty\}$$

satisfies following properties:

- Positivity: $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$.
- Symmetry: $d(x, y) = d(y, x)$.
- Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

A very simple example of a metric space with a surprising connection to Hilbert's fourth problem is constructed as follows:

Take S to be the collection of all subsets of $\{1, \dots, n\}$. Some of the elements of S are the empty set, $\{1\}$, $\{2, 3\}$, and $\{1, \dots, n\}$. If x and y are in S define $d(x, y)$ as the number of elements in x which are not in y plus the number of elements of y that are not in x . For example,

$$d(\{1, 2\}, \{2, 5, 6\}) = 3.$$

Exercise 2.1. Show that (S, d) is a metric space. Moreover, S satisfies the following stronger version of the triangle inequality: if x_1, \dots, x_k are elements of S and b_1, \dots, b_k are integer numbers that add up to 1, then

$$\sum_{i,j=1}^k d(x_i, x_j) b_i b_j \leq 0.$$

For $k = 3$ the above property is just a fancy way to rewrite the triangle inequality. When a metric space satisfies this property for any positive integer k , it is called *hypermetric*. Remarkably, all the solutions to Hilbert's fourth problem in dimension two are hypermetric. This was first remarked by R. Alexander in [2].

Exercise 2.2. Define the distance between two points (x, y, z) and (x', y', z') in three-dimensional space as the biggest of the three numbers $|x - x'|$, $|y - y'|$, and $|z - z'|$. Show that this distance is not hypermetric.

Now that we understand the freedom we have in the choice of a distance function, it remains to make sense of the term "the shortest path" that appears in our metric interpretation of Hilbert's fourth problem. For that we must understand how to compute the length of a path in a metric space.

If all we know is to measure distances between pairs of points, a natural way to estimate the length of a path is to mark a number of points on it, measure the distances between consecutive points, and add them up.

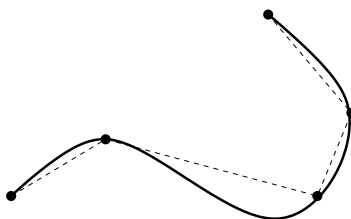


Fig. 2. Estimating the length of a path.

Of course, this is just an estimate. By taking the points more and more densely along the path, our intuition tells us that, in the limit, we will find its length. For many reasons, it is sometimes convenient not to evoke directly this passage to the limit and it is preferred to define the length of a path as follows:

Definition 2.2. Let (X, d) be a metric space and let $\gamma : [a, b] \rightarrow X$ be a continuous curve. The length of γ is defined as the supremum of

$$\left\{ \sum_{i=0}^{n-1} d(\gamma(t_i), \gamma(t_{i+1})) : a = t_0 < t_1 < \cdots < t_n = b \text{ is any partition of } [a, b] \right\}.$$

For almost every example one comes up with, the intuitive and the formal definition coincide.

We've seen that if we can measure distances, we can measure lengths of curves. However, if we can measure lengths, we may redefine the distance between two points as the infimum of the lengths of all curves joining them. Sometimes the redefined distance does not agree with the original. For example, think of a plane where you have taken out a circular obstacle (Fig. 3). The Euclidean distance between two points is the infimum of the lengths of the curves joining them only if the line segment that joins them does not pass through the interior of the obstacle.

Metric spaces where the distance between two points is the infimum of the lengths of all curves joining them are called *length spaces*. From now on we will be solely concerned with length spaces.

Definition 2.3. A continuous curve in a metric space is called a *segment* if the distance between its endpoints equals the length of the curve. A continuous curve is called a *geodesic* if it can be partitioned into segments.

Definition 2.4. A metric on the plane is said to be *projective* if straight lines are geodesics. A metric on the sphere is said to be projective if great circles are geodesics.

Hilbert's fourth problem can now be given a precise statement: *Construct and study all projective metrics on the plane.* This statement is usually extended to include not only the plane, but also the sphere, the projective plane, and convex domains in \mathbb{R}^2 .

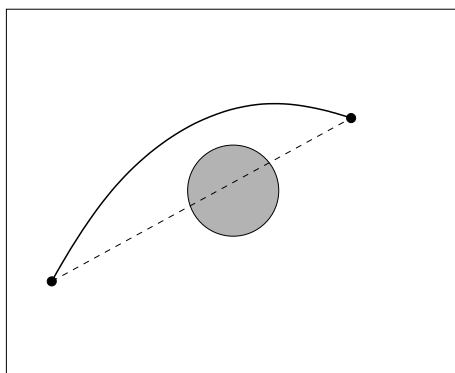


Fig. 3. Plane with an obstacle

Exercise 2.3. Extend the definition of projective metrics to the projective plane and try your hand at a definition of locally projective metrics on general surfaces. Hint: in this last case, your definition should be independent of coordinate changes.

We end this section with a very simple characterization of projective metrics on the plane.

Exercise 2.4. Show that a metric d on the plane is projective if and only if for any triple of collinear points x, y , and z , with y between x and z , we have that

$$d(x, z) = d(x, y) + d(y, z).$$

3. MINKOWSKI PLANES

In Hilbert's text, he mentions that Minkowski had already found many examples of projective metrics on the plane and that these metrics had the additional property that translating a line segment does not change its length. Hilbert was speaking about what we would now call *finite-dimensional normed spaces*, and which many people still call *Minkowski spaces* (see [32]).

Let us start, as Hilbert did, by remarking the importance of the invariance of distances under translations. Let us say that a metric, or distance, on the plane is *invariant under translations* if the distance between any two points x and y equals the distance between the points $x + v$ and $y + v$ for any vector v .

Exercise 3.1. Show that a metric on the plane that is invariant under translations is automatically projective.

In view of this exercise, constructing and studying distances that are invariant under translations is a logic first step in the solution of Hilbert's fourth problem.

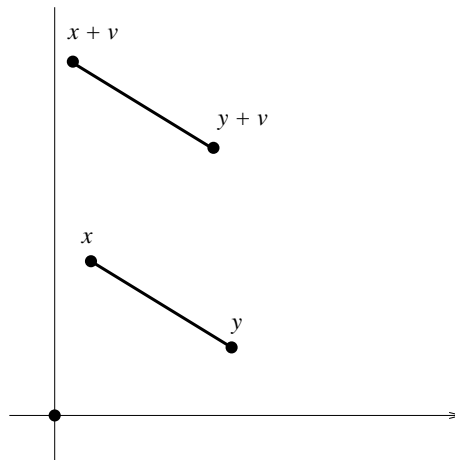


Fig. 4. Translations don't change distances.

Exercise 3.2. Assume d is a translation-invariant metric on \mathbb{R}^2 and define the *norm* of a vector x , denoted by $\|x\|$, as its distance from the origin. Show that the function

$$\|\cdot\| : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

satisfies the following properties:

- Positivity: $\|x\| \geq 0$ and $\|x\| = 0$ if and only if x is the origin.
- Homogeneity: if λ is any real number $\|\lambda x\| = |\lambda| \|x\|$.
- Convexity: $\|x + y\| \leq \|x\| + \|y\|$.

Exercise 3.3. A function on the plane satisfying the properties in exercise 3.2 is called a *norm*. Show that if $\|\cdot\|$ is a norm on the plane, then the function

$$d : \mathbb{R}^2 \times \mathbb{R}^2 \longrightarrow [0, \infty)$$

defined by $d(x, y) = \|x - y\|$ is a metric that is invariant under translations.

Definition 3.1. A pair (\mathbb{R}^2, d) where d is a metric that is invariant under translations shall be called a *Minkowski plane*.

In order to construct and understand Minkowski planes, we look at their unit discs. Since these are all the same up to translation, we can concentrate on the disc with center at the origin.

Definition 3.2. A subset K of the plane is said to be *convex* if every two points in K are joined by a line segment entirely contained in K . A convex set is said to be a *convex body* if it is compact and its interior is not empty.

Exercise 3.4. The *unit disc* of a Minkowski plane is the set of vectors whose distance from the origin does not exceed one. Show that the unit disc of a Minkowski plane is a convex body symmetric around the origin.

Moreover, any convex body that is symmetric about the origin is the unit disc of some Minkowski plane: simply define the norm of a nonzero vector

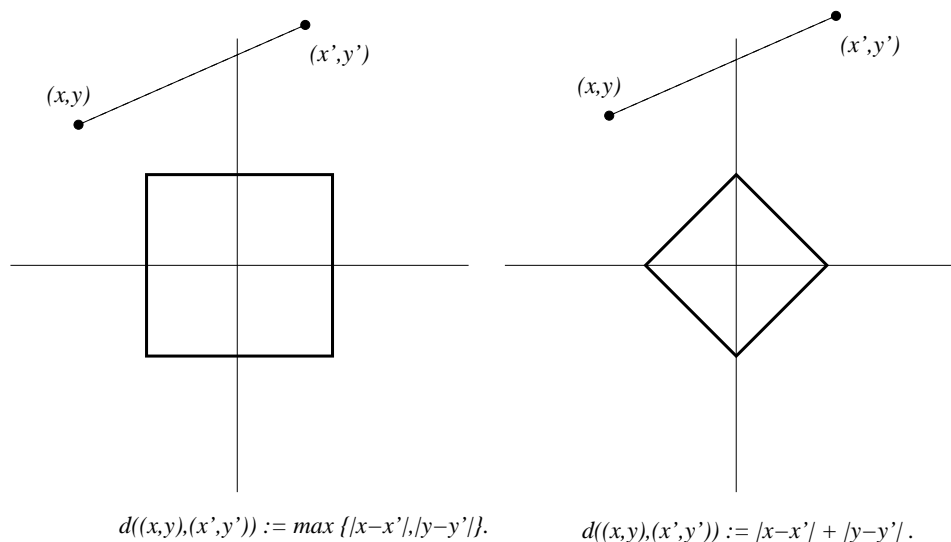


Fig. 5. Examples of Minkowski geometries.

x as the only positive number $\|x\|$ for which $x/\|x\|$ belongs to the boundary of the unit disc.

As a result of this simple remark, constructing Minkowski planes is very easy indeed: just draw a closed convex curve that is symmetric about the origin (e.g., a square, a rectangle, an ellipse, a rhombus) and we have a norm and hence a projective metric.

Another simple way of constructing and studying norms that is central to Pogorelov's approach to Hilbert's fourth problem is given in the following exercises:

Exercise 3.5. Given a piecewise continuous function $f : S^1 \rightarrow \mathbb{R}$, its *cosine transform* is defined as the function

$$L(v_1, v_2) := \int_0^{2\pi} |v_1 \cos(\theta) + v_2 \sin(\theta)| f(\theta) d\theta$$

- (1) Show that if f nonnegative and takes positive values in at least three distinct points, then L is a norm on the plane.
- (2) Show that if f is odd (i.e., $f(\theta + \pi) = -f(\theta)$), then its cosine transform is identically zero. Deduce from this that in order to study the range of the cosine transform we need only consider the case where f is even (i.e. $f(\theta + \pi) = f(\theta)$).
- (3) Assume that f is piecewise continuous and even. Show that in polar coordinates, (r, ϕ) , the cosine transform of f is given by the following formula:

$$L(r, \phi) = 2r \int_{\phi-\pi/2}^{\phi+\pi/2} \cos(\phi - \theta) f(\theta) d\theta.$$

(4) With f and L as in the previous item, show that

$$f(\phi + \pi/2) := \frac{1}{4r} \left(L(r, \phi) + \frac{\partial^2 L}{\partial \phi^2}(r, \phi) \right).$$

Exercise 3.6. Show that if f is a smooth positive function on the circle and L is its cosine transform in cartesian coordinates (v_1, v_2) , then the matrix formed by the second partial derivatives of L^2 is positive definite. Geometrically, this means that the unit circle of the Minkowski plane (\mathbb{R}^2, L) is smooth and has positive curvature everywhere.

A norm $L : \mathbb{R}^2 \rightarrow [0, \infty)$ that is smooth outside the origin and such that the matrix formed by the second partial derivatives of L^2 is positive definite is said to satisfy the *Legendre condition*. The previous exercise shows that the cosine transform of a smooth positive function on the circle is a such a norm, while item (4) on exercise 3.5 shows that every norm satisfying the Legendre condition is the cosine transform of a smooth positive function on the circle. Summarizing, we have the following result:

Theorem 3.1 (Blaschke, [17]). *The cosine transform establishes a bijection between the set of smooth, even, positive functions on the circle and norms on the plane that satisfy the Legendre condition.*

4. HILBERT GEOMETRIES

The reader has probably heard or read about the history of Euclid's parallel axiom, which states that *given a line and a point not belonging to it, there is a unique line passing through the point and not intersecting the original line*. For a long time geometers sought to show that this axiom — perhaps the less *self-evident* of Euclid's axioms — was a logical consequence of the others. In time it was shown that this axiom is independent of the others and that geometries in which it does not hold are not only possible, but natural. In fact, imagine a geometry in which the points range over the interior of the unit disc and the lines are the straight line segments joining pairs of points in the unit circle. Given a line and a point not belonging to it, there are infinitely many lines that pass through the point, but never intersect the original line (Fig. 6).

Such a geometry really exists and is none other than the hyperbolic geometry discovered in the nineteenth century independently by Gauss, Bolyai, and Lobachevsky. The way in which we shall present it here is due to Arthur Cayley and Felix Klein.

Let D denote the open unit disc and let C denote its boundary. If x and y are two points in D , denote by a and b the points of intersection of C with the straight line passing through x and y with the provision that x lie between a and y and that y lie between x and b . Thus if we interchange x and y , we are forced to interchange a and b .

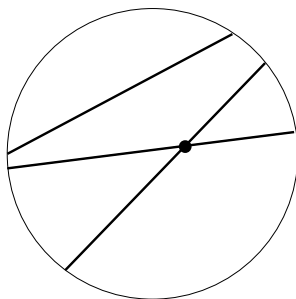


Fig. 6. Infinitely many parallel lines passing through the same point.

The *hyperbolic distance* between x and y is defined by the equation

$$d(x, y) := \frac{1}{2} \ln \left(\frac{\|y - a\| \|x - b\|}{\|x - a\| \|y - b\|} \right).$$

Exercise 4.1. Show that the function d satisfies the three following properties:

- (1) Positivity: $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$.
- (2) Symmetry: $d(x, y) = d(y, x)$.
- (3) If x, y , and z are three aligned points in D such that y lies between x and z . Show that $d(x, z) = d(x, y) + d(y, z)$.

The proof of the triangle inequality for d is somewhat more elaborate, but it only makes use of elementary projective geometry (see chapter 5 in [6]). Part (3) of exercise 4.1 shows that this metric is projective.

In [26], Hilbert shows that it is possible to replace the open unit disc D in the construction above by the interior of any convex body in the plane without changing the fact that d is a projective metric.

Hilbert geometries, as these metric spaces are called, are a natural generalization of hyperbolic geometry and furnish us with a second large class of examples of projective metrics. Recently, much progress has been made in the understanding of the Hilbert geometries (see, for example, [30] and [27]).

5. THE CROFTON FORMULA

The key to the solution of many problems in geometry is simply a change of viewpoint. So far we have concentrated on points and the distances between them. Hilbert's fourth problem asks for those metrics for which the geodesics are straight lines, so perhaps the best strategy is to replace the point for the straight line as the fundamental object of our investigations.

We shall take as departing point the following naive question: *what would Euclidean geometry look like if we take the straight line instead of the point as our basic geometric entity?*

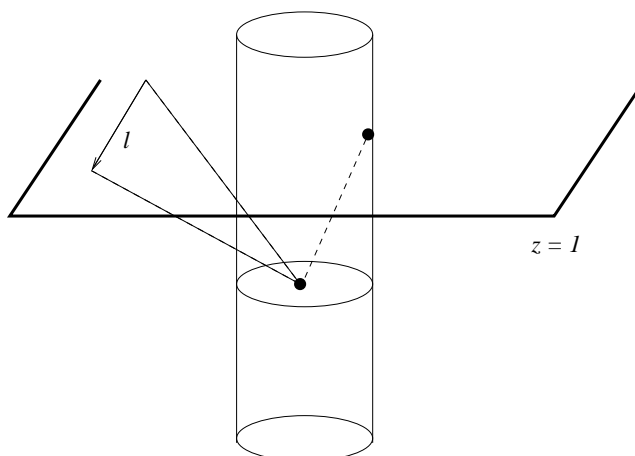


Fig. 7. The cylinder as the space of oriented lines.

The space of oriented lines. We may identify the space of *oriented* straight lines on the plane with the right circular cylinder $x^2 + y^2 = 1$ in three-dimensional space by means of the following construction:

An oriented line ℓ lying on the plane $z = 0$ in three-dimensional space is uniquely determined by the unit vector q defining its direction, and by the vector v representing the point on the line that is closest to the origin. The map $\ell \mapsto q + q \times v$ identifies the space of oriented lines on the plane $z = 0$ with the cylinder $x^2 + y^2 = 1$.

Another, more geometric, correspondence may be given if we consider not the plane $z = 0$, but the plane $z = 1$. In fact, any line ℓ on $z = 1$ is the intersection of this plane with some plane Π through the origin. Using the orientation on the line, we may determine which side of the plane Π is *up* and which is *down*: if you're walking head up along the line in the sense of its orientation, then the origin of \mathbb{R}^3 should be to your left. In order to identify ℓ with a point on the cylinder, simply trace an orthogonal ray *up* from Π at the origin and mark where it intersects the cylinder (Fig. 7).

Exercise 5.1. Discover the simple relationship between both identifications of the space of oriented lines with the right circular cylinder. *In the rest of the paper, we use the first of these identifications.*

The aim of the following exercises is to get an intuitive feeling for the correspondence between points on the cylinder $x^2 + y^2 = 1$, and henceforth denoted by \mathcal{C} , and oriented lines on the plane. This intuition is crucial for the understanding of the rest of the paper. Please take ten minutes to do them on your own and then come back to the paper.

Exercise 5.2. Draw the portion of the cylinder that corresponds to all oriented lines passing through

- (1) a point,
- (2) a line segment,

- (3) the unit disc centered at the origin,
- (4) two sides of a given triangle.

Exercise 5.3. The cylinder \mathcal{C} admits the parameterization

$$(p, \theta) \longmapsto (\cos(\theta), \sin(\theta), p),$$

where $0 \leq \theta < 2\pi$, and p is any real number. Draw the region in the (p, θ) -plane that corresponds to all oriented lines passing through

- (1) a point,
- (2) a line segment,
- (3) the unit disc centered at the origin,
- (4) two sides of a given triangle.

Euclidean transformations preserve areas. A rigid motion, or Euclidean transformation, on the plane is a transformation from the plane onto itself that preserves distances. It is not too hard to see that these transformations are composed of translations and rotations, and that they send oriented straight lines to oriented straight lines. We have then that Euclidean transformations of the plane *induce* a class of transformations on the space of oriented lines \mathcal{C} . Let us now understand this class of transformations and discover what geometric properties they may have.

Exercise 5.4. Parameterize each point in the cylinder \mathcal{C} by its height p and the angle, θ , that its projection makes with the positive x -axis (i.e., a point in \mathcal{C} has coordinates $(\cos(\theta), \sin(\theta), p)$).

- (1) Show that rotating the oriented line represented by (p, θ) counterclockwise by an angle ϕ around the origin results in the oriented line represented by $(p, \theta + \phi)$.
- (2) Show that translating the oriented line represented by (p, θ) by a vector (x, y) results in the oriented line represented by $(p + x \cos(\theta) + y \sin(\theta), \theta)$.

Exercise 5.5. Show that the area of the portion of the right circular cylinder $x^2 + y^2 = 1$ given by all points with coordinates $(\cos(\theta), \sin(\theta), p)$, where (p, θ) ranges over some region \mathcal{R} is just

$$\int_{\mathcal{R}} dpd\theta.$$

Conclude that the area of a region in the space of oriented lines is invariant under the action of Euclidean transformations.

Defining distance in terms of areas. So far our project of studying Euclidean geometry by taking the line instead of the point as basic geometric entity has led us to consider the space of oriented lines, to study the action of Euclidean transformations on this space, and to the remarkable fact that these transformations preserve areas. However, in order to gain full understanding, we must try to reconstruct the Euclidean distance on the plane by using the geometry on the space of lines.

Exercise 5.6. Use exercises 5.3 and 5.5 to show that the area in \mathcal{C} of the region representing the set of all oriented lines passing through a line segment equals four times its length.

Exercise 5.7. Use exercise 5.6 to show that the length of a polygonal curve γ in the plane is given by the following integral over the space of oriented lines \mathcal{C}

$$\frac{1}{4} \int_{\ell \in \mathcal{C}} \#(\ell \cap \gamma) dA,$$

where dA is the element of area on the cylinder.

By approximating rectifiable curves by polygonal ones, the previous exercise furnishes us with a proof of the *Crofton formula* in Euclidean geometry.

Theorem 5.1 (Crofton formula). *If $\gamma \subset \mathbb{R}^2$ is a rectifiable curve and dA is the standard area element on the cylinder \mathcal{C} , then*

$$\text{length}(\gamma) = \frac{1}{4} \int_{\ell \in \mathcal{C}} \#(\ell \cap \gamma) dA.$$

Exercise 5.8. Use the Crofton formula to give an easy proof of the following, apparently obvious, fact: if a convex body $K \subset \mathbb{R}^2$ contains a second convex body L , then the perimeter of K is greater than or equal to the perimeter of L with equality holding if and only if $K = L$.

6. BUSEMANN'S CONSTRUCTION OF PROJECTIVE METRICS

The simple remark that underlies the solution of Hilbert's fourth problem is that the Crofton formula implies that straight lines are geodesics. In fact, if γ is a rectifiable curve on the plane joining the points x and y , then any oriented line intersecting the line segment that joins them must intersect γ . Applying Crofton's formula, we see that the length of γ is greater than or equal to that of the line segment joining x and y . The independence of this argument on the special nature of the area element on \mathcal{C} suggests replacing dA by other ways of measuring areas on the space of oriented lines.

Any continuous area element on the space of oriented lines, \mathcal{C} , has the form $f dA$, where f is some positive, continuous function on \mathcal{C} . Let us agree to call the area, measured with $f dA$, of a region \mathcal{R} in \mathcal{C} the *f-area* of \mathcal{R} .

If f is a positive, continuous function of \mathcal{C} , *define* a metric on the plane by setting the distance between any two points as one fourth times the *f-area* of the set of all oriented lines that intersect the segment that joins them.

Exercise 6.1. Show that the construction above does indeed define a distance function on the plane and that this distance is necessarily projective. This is *Busemann's construction* of projective metrics.

The same arguments in your solution of exercise 5.7 show that Crofton's formula holds for the metrics defined by Busemann's construction. In fact, these metrics were defined so that Crofton's formula be a tautology.

Exercise 6.2. Let f be a positive, continuous function on the space of oriented lines that depends only on the direction of the oriented line. Show that Busemann's construction yields a translation invariant metric, and hence a norm, on the plane.

A metric d on the plane is said to be *periodic* if there exists two linearly independent vectors v and w such that translation by any vector of the form $mv + nw$, m, n any integers, preserves distances. In general, periodic metrics are not invariant under all translations, however we have the following exercise:

Exercise 6.3. Suppose a metric arising from Busemann's construction is periodic. Show that it is invariant under translations and, hence, comes from a norm on the plane.

7. AMBARTZUMIAN'S CONSTRUCTION

Busemann's construction shows that to any continuous way of measuring areas in the space of oriented lines we can associate a continuous way of measuring distances in the plane such that geodesics are straight lines. The natural question now is whether *all* projective metrics can be obtained from Busemann's construction. In [13], Ambartzumian gives a beautiful and simple construction that answers this question affirmatively.

We start by remarking a nearly obvious property of Busemann's construction.

Exercise 7.1. Let $f dA$ be a continuous element of area in the space of oriented lines and let d be the projective metric on the plane obtained from $f dA$ using Busemann's construction. Show that the f -area of the set of all oriented lines passing through both the xy and yz sides of a triangle xyz is equal to two times the quantity $d(x, y) + d(y, z) - d(x, z)$.

This simple remark is the basis of Ambartzumian's construction: if d is a continuous distance function on the plane that satisfies the projective condition $d(a, b) + d(b, c) = d(a, c)$ whenever b lies in the line segment joining a and c , we *define* the area of the set of all lines passing through both the xy and yz sides of a triangle xyz as two times the quantity $d(x, y) + d(y, z) - d(x, z)$. The projective condition guarantees that this set function is additive, and the family of sets on which it is defined is sufficiently large so that the set function admits a unique extension to a continuous measure on the set of oriented lines *provided* we specify that this measure be invariant under the map that changes the orientation of the lines. While the details in the proofs of these statements are beyond the scope of this article, the exercises below should give the reader a good sense of why they are true.

We start by describing the set of all oriented lines passing through two sides of a triangle in the plane as a subset of the cylinder. Of course, the conscientious reader has already done this exercise in section 5.

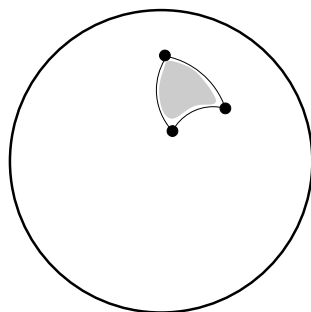


Fig. 8. A short triangle on the sphere.

Exercise 7.2. Define a *short triangle* on the unit sphere in three dimensional space as the interior of a spherical triangle in which the length of each side is less than π (see Fig. 8). In particular, a short triangle is disjoint from its image under the antipodal map. Show that a set in the cylinder $x^2 + y^2 = 1$ corresponds to the interior of the set of all oriented lines passing through two sides of a triangle in the plane if and only if it is the image of a short triangle and its antipode under the radial projection from the sphere to the cylinder. Use this description to see that any open set in the cylinder that is invariant under the antipodal map can be approximated by a disjoint union of such sets.

Exercise 7.3. Show that the measure of the set of all lines intersecting the segment xy equals four times the distance between x and y . Use this to explain why the projective condition on d is necessary for the additivity of the measure defined in Ambartzumian's construction.

8. VARIATIONAL INTERPRETATION OF HILBERT'S FOURTH PROBLEM

In the next two sections, we shall tackle Hilbert's fourth problem in a way that is apparently very different from the approach of Busemann and Ambartzumian. Our starting point is to consider Hilbert's fourth problem as an inverse problem in variational calculus: *construct and study all variational problems on the plane for which the extremals are straight lines*. In order to give a more precise statement, we must review some basic notions.

Let us start by specifying what we mean by a "variational problem on the plane". If $L : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ is a continuous function, then given a differentiable parameterized curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$ we may compute the integral

$$S(\gamma) := \int_a^b L(\gamma(t), \dot{\gamma}(t)) dt.$$

Given two points x and y on the plane, we consider the following problem: find the infimum of $S(\gamma)$, where γ ranges over all differentiable parameterized curves joining x and y . We are also interested in determining if this infimum is attained and, if so, in the curve or curves at which it is attained.

The integrand to keep in mind is the function

$$L_0(x_1, x_2, v_1, v_2) := \sqrt{v_1^2 + v_2^2},$$

which yields the variational problem of finding the shortest curve between two points in the Euclidean plane. This suggests taking an arbitrary function $L(x_1, x_2, v_1, v_2)$ and *defining* the length of a parameterized curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$ as the integral

$$S(\gamma) := \int_a^b L(\gamma(t), \dot{\gamma}(t)) dt.$$

Two basic properties of length that we would like our definition to satisfy are

- (1) The length of any differentiable curve γ is a nonnegative number and it is zero if and only if the curve is constant (i.e., it is a point).
- (2) The length of a curve does not depend on the way we traverse it. In other words, length should be invariant under changes of parameterizations.

Exercise 8.1. Let $L : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ be a continuous function and use it to define a “length” for differentiable parameterized curves in the manner described above.

- Show that property (1) holds if and only if $L(x_1, x_2, v_1, v_2)$ is nonnegative and it is zero if and only if v_1 and v_2 are both equal to zero.
- Show that property (2) holds if and only if L is homogeneous of order one in the last two coordinates: $L(x_1, x_2, \lambda v_1, \lambda v_2) = |\lambda|L(x_1, x_2, v_1, v_2)$.

The third requirement we shall make of our definition of length is more subtle: using lengths, we may define the distance between two points as the infimum of the lengths of all differentiable curves joining them. We may then use this definition of distance to redefine the lengths of curves as was done in section 2. The two ways of measuring lengths may not agree. Indeed, Busemann and Mayer proved in [22] that they agree if and only if the function L is convex in the last two variables:

$$L(x_1, x_2, v_1 + w_1, v_2 + w_2) \leq L(x_1, x_2, v_1, v_2) + L(x_1, x_2, w_1, w_2).$$

We may summarize all three requirements on the function L by saying that for every point (x_1, x_2) on the plane the function $(v_1, v_2) \mapsto L(x_1, x_2, v_1, v_2)$ is a norm.

Exercise 8.2. Let $L : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ be a continuous function such that for every point (x_1, x_2) on the plane the function $(v_1, v_2) \mapsto L(x_1, x_2, v_1, v_2)$ is a norm. Define the length of a differentiable parameterized curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$ as the integral

$$S(\gamma) := \int_a^b L(\gamma(t), \dot{\gamma}(t)) dt,$$

and define $d(x, y)$ as the infimum of the lengths of all differentiable parameterized curves joining the points x and y . Show that d is a distance.

Definition 8.1. A function $L : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ is called a *continuous Finsler metric* if it is continuous and if for every point (x_1, x_2) on the plane the function $(v_1, v_2) \mapsto L(x_1, x_2, v_1, v_2)$ is a norm.

Exercise 8.2 shows that a continuous Finsler metric defines a metric on the plane. By an abuse of notation we shall also call this metric (i.e., the distance function) a continuous Finsler metric. A possible interpretation of Hilbert's fourth problem is *to construct and study all continuous Finsler metrics on the plane for which straight lines are geodesics*. However, in order to apply the standard techniques of the calculus of variations in the study of the integrand L , we shall require the following regularity conditions:

- (1) The function L is smooth in $\mathbb{R}^2 \times (\mathbb{R}^2 \setminus (0, 0))$.
- (2) The matrix of second partial derivatives of L^2 with respect to the last two variables is positive definite at every point (x_1, x_2, v_1, v_2) with $(v_1, v_2) \neq (0, 0)$.

In the terminology of section three, we require that at each point (x_1, x_2) the norm $L(x_1, x_2, \cdot)$ satisfy the *Legendre condition*. This condition is quite standard in the calculus of variation and, among other things, it guarantees that all geodesics of the Finsler metric are smooth curves and that there is only one geodesic passing through each point in each direction.

Definition 8.2. A continuous Finsler metric L on the plane is said to be a (smooth) Finsler metric if it is smooth in $\mathbb{R}^2 \times (\mathbb{R}^2 \setminus (0, 0))$ and at each point (x_1, x_2) the norm $L(x_1, x_2, \cdot)$ satisfies the Legendre condition.

From now until the end of this paper, Hilbert's fourth problem will be interpreted as: *to construct and study all Finsler metrics on the plane whose geodesics are straight lines*. In [28], Pogorelov showed that the restriction to smooth Finsler metrics is not fundamental: any continuous projective Finsler metric on the plane can be uniformly approximated in each compact set by smooth projective Finsler metrics.

The reader interested in knowing more about Finsler metrics is referred to the books of Bao, Chern, and Shen ([15]), Álvarez and Durán ([8]), and to the survey article [7]. The last two works may be downloaded from the *Finsler Geometry Newsletter*, <http://www.math.poly.edu/research/finsler>, along with many other papers on the subject.

9. ANALYTIC SOLUTION OF HILBERT'S FOURTH PROBLEM

The first step in the analytic solution of Hilbert's fourth problem is the following simple result.

Theorem 9.1 (Hamel, [24]). *Let $L : \mathbb{R}^2 \times (\mathbb{R}^2 \setminus (0, 0)) \rightarrow \mathbb{R}$ be a smooth Finsler metric. Straight lines are geodesics for the metric defined by L if and only if L satisfies the following partial differential equation:*

$$\frac{\partial^2 L}{\partial x_1 \partial v_2} = \frac{\partial^2 L}{\partial x_2 \partial v_1}.$$

In order to prove this theorem we recall (or state) that if L is a smooth Finsler metric on the plane and $\gamma(t)$ is one of its geodesics, then the *Euler-Lagrange* equations

$$\begin{aligned}\frac{d}{dt} \frac{\partial L}{\partial v_1}(\gamma(t), \dot{\gamma}(t)) - \frac{\partial L}{\partial x_1}(\gamma(t), \dot{\gamma}(t)) &= 0 \\ \frac{d}{dt} \frac{\partial L}{\partial v_2}(\gamma(t), \dot{\gamma}(t)) - \frac{\partial L}{\partial x_2}(\gamma(t), \dot{\gamma}(t)) &= 0\end{aligned}$$

hold. These equations also hold for more general integrands than Finsler metrics (see, for example, Arnold's book [14] for a complete account of Lagrangian mechanics and the Euler-Lagrange equations), but that need not concern us at this point.

Exercise 9.1. Prove Hamel's theorem using the Euler-Lagrange equations above and the following hints:

- (1) Use Euler's formula for homogeneous functions to replace the first partial derivatives of L with respect to x_1 and x_2 in the Euler-Lagrange equations by expressions in the second partial derivatives of L .
- (2) Use that lines $t \mapsto (x_1 + tv_1, x_2 + tv_2, v_1, v_2)$ are geodesics.

From exercise 3.5 we know that any norm $L(v_1, v_2)$ satisfying the Legendre condition can be written in a unique way as

$$L(v_1, v_2) = \int_0^{2\pi} |v_1 \cos(\theta) + v_2 \sin(\theta)| f(\theta) d\theta,$$

where f is a smooth, positive, and even function on the circle. Since for every fixed (x_1, x_2) a smooth Finsler metric $L(x_1, x_2, v_1, v_2)$ is a norm satisfying the Legendre condition, there is a unique function $f(x_1, x_2, \theta)$ that is positive and even in its last variable, and that satisfies

$$L(x_1, x_2, v_1, v_2) = \int_0^{2\pi} |v_1 \cos(\theta) + v_2 \sin(\theta)| f(x_1, x_2, \theta) d\theta.$$

It's not hard to see that f must be smooth as a function of its three variables.

Using this integral representation of Finsler metrics it is quite easy to solve Hamel's differential equation.

Theorem 9.2 (Pogorelov, [28]). *Let us use the cosine transform to represent a smooth Finsler metric on the plane, $L(x_1, x_2, v_1, v_2)$, as the integral*

$$L(x_1, x_2, v_1, v_2) = \int_0^{2\pi} |v_1 \cos(\theta) + v_2 \sin(\theta)| f(x_1, x_2, \theta) d\theta,$$

where f is a smooth positive function on $\mathbb{R}^2 \times S^1$ that is even in the last variable (i.e., $f(x_1, x_2, \theta + \pi) = f(x_1, x_2, \theta)$). The function L satisfies Hamel's differential equation if and only if there exists a smooth function $g : \mathbb{R} \times S^1 \rightarrow \mathbb{R}$ with $f(x_1, x_2, \theta) = g(x_1 \cos(\theta) + x_2 \sin(\theta), \theta)$.

Exercise 9.2. This exercise outlines the proof of theorem 9.2. The only idea in the proof is to get rid of the absolute values inside the integral by changing to polar coordinates in the velocities: $(x_1, x_2, v_1, v_2) \mapsto (x_1, x_2, r, \phi)$.

- (1) Show that in the variables (x_1, x_2, r, ϕ) Hamel's equations take the form

$$\sin(\phi) \frac{\partial^2 L}{\partial x_1 \partial r} + \frac{\cos(\phi)}{r} \frac{\partial^2 L}{\partial x_1 \partial \phi} = \cos(\phi) \frac{\partial^2 L}{\partial x_2 \partial r} - \frac{\sin(\phi)}{r} \frac{\partial^2 L}{\partial x_2 \partial \phi}.$$

- (2) Show that if $f(x_1, x_2, \theta)$ is a smooth function that is even in the last variable, then

$$L(x_1, x_2, r, \phi) = 2r \int_{\phi-\pi/2}^{\phi+\pi/2} \cos(\phi - \theta) f(x_1, x_2, \theta) d\theta$$

satisfies Hamel's equation if and only if the integral

$$\int_{\phi-\pi/2}^{\phi+\pi/2} \left(-\sin(\theta) \frac{\partial f}{\partial x_1} + \cos(\theta) \frac{\partial f}{\partial x_2} \right) d\theta$$

is zero for any value of ϕ .

- (3) Show that the last condition on the above item is met if and only if

$$-\sin(\theta) \frac{\partial f}{\partial x_1} + \cos(\theta) \frac{\partial f}{\partial x_2} = 0$$

and deduce Pogorelov's theorem from this fact.

To obtain a Finsler metric L on the plane whose geodesics are straight lines one just takes a function $g : \mathbb{R} \times S^1 \rightarrow \mathbb{R}$ that is smooth, positive, and even in the second variable, and plugs it into the formula

$$L(x_1, x_2, v_1, v_2) = \int_0^{2\pi} |v_1 \cos(\theta) + v_2 \sin(\theta)| g(x_1 \cos(\theta) + x_2 \sin(\theta), \theta) d\theta.$$

Exercise 9.3. Show that if $g(p, \theta) := 1 + p^2$, then the corresponding Finsler metric is

$$L(x_1, x_2, v_1, v_2) = \frac{1}{3\sqrt{v_1^2 + v_2^2}} [(3 + x_1^2 + x_2^2)(v_1^2 + v_2^2) + (x_1 v_1 + x_2 v_2)^2].$$

The reader may well ask about the relationship between this analytic approach and the geometric approach of Busemann and Ambartzumian. The answer is that they are the two sides of the same coin. If $g(p, \theta)$ is a smooth, positive function that is even in its last variable, then the Finsler metric associated to g by Pogorelov's construction is the same as the metric associated to the area element gdA in the space of oriented lines, \mathcal{C} , by Busemann's construction. For the details and a thorough study of the relationship between Crofton formulas and cosine transforms the reader is referred to the paper of Álvarez and Fernandes, [11].

10. A GLIMPSE AHEAD

It would be a pity if the reader of this article should conclude that Hilbert's fourth problem is completely solved and that there is nothing that he or she may contribute to the subject. Recent work by Álvarez ([3, 4, 5]), Alvarez and Fernandes ([10, 11]), Bryant ([18]), and Schneider ([29]) shows that the subject still has much to yield. In this last section I have collected some variations on Hilbert's fourth problem that may tempt those with a taste for simply stated geometric problems. I have followed the Russian tradition of formulating problems in their simplest nontrivial formulation and have included a short comment after each problem.

Problem 1 (Busemann, [20]). A nonsymmetric distance on a set X is a function $d : X \times X \rightarrow \mathbb{R} \cup \{\infty\}$ that satisfies

- Positivity: $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$.
- Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

Construct and study all nonsymmetric distances on the plane for which straight lines are geodesics.

Comment. Bryant constructed in [18] a class of nonsymmetric (Finsler) metrics on the two dimensional sphere for which the geodesics are great circles. His examples have constant flag curvature.

Problem 2 (Álvarez, [7]). Construct and study all metrics on the complex, quaternionic, and Cayley planes for which the geodesics agree with the geodesics for the standard homogeneous metrics in these spaces.

Comment. So far not a single nontrivial example of such a metric is known. However, Álvarez and Durán (see [9]) constructed a class of Finsler metrics on projective and quaternionic projective spaces for which projective lines are totally geodesic and all geodesics are geometric circles.

Problem 3 (Bryant, [19]). Is there a Riemannian metric on the real projective space such that projective planes are minimal surfaces and which is not isometric to the standard?

Comment. If one allows the metric to be defined on an open subset of real projective space then there are Riemannian metrics not isometric to the standard for which projective planes are minimal. This has been studied by Bekkar and Bryant (see [16] and [19]). If we allow the metric to be Finsler, then Álvarez and Fernandes show in [11] that there is a correspondence between these metrics and volume forms on the Grassmannian of oriented 2-planes in \mathbb{R}^4 .

Problem 4. Do there exist non-Riemannian Finsler metrics on the complex projective plane for which all complex curves are minimal?

Comment. It can be shown that the examples of Álvarez and Durán (see [9]) of metrics in the complex projective plane whose geodesics are circles

satisfy the additional property that complex projective lines are minimal. It is not known whether conics or other complex curves are minimal as well.

REFERENCES

- [1] M. Aigner and G.M. Ziegler, "Proofs from the Book" Second Corrected Printing, Springer-Verlag, Berlin Heidelberg New York, 1999.
- [2] R. Alexander, *Zonoid theory and Hilbert's fourth problem*, *Geom. Dedicata* **28** (1988), no. 2, 199–211.
- [3] J.C. Álvarez Paiva, *Anti-self-dual symplectic forms and integral geometry*, in "Analysis, Geometry, Number Theory, The Mathematics of Leon Ehrenpreis.", 15 – 25, *Contemp. Math.*, 251, Amer. Math. Soc., Providence, RI, 2000.
- [4] J.C. Álvarez Paiva, *Contact topology, taut immersions, and Hilbert's fourth problem*, in "Differential and Symplectic Topology of Knots and Curves", 1–21, Amer. Math. Soc. Transl. Ser. 2, 190, Amer. Math. Soc., Providence, RI, 1999.
- [5] J.C. Álvarez Paiva, *Symplectic geometry and Hilbert's fourth problem*, preprint (2002).
- [6] J.C. Álvarez Paiva, "Interactive Course on Projective Geometry", http://www.math.poly.edu/courses/projective_geometry.
- [7] J.C. Álvarez Paiva, *Some problems on Finsler geometry*, preprint (2000).
- [8] J.C. Álvarez Paiva and C. Durán, "An Introduction to Finsler Geometry", *Notas de la Escuela Venezolana de Matemáticas*, 1998.
- [9] J.C. Álvarez Paiva and C. Durán, *Isometric submersions of Finsler manifolds*, to appear in *Proc. of the Amer. Math. Soc.*
- [10] J.C. Álvarez Paiva and E. Fernandes, *Crofton formulas in projective Finsler spaces*, *Electronic Research Announcements of the Amer. Math. Soc.* **4** (1998), 91–100.
- [11] J.C. Álvarez Paiva and E. Fernandes, *Crofton formulas and Gelfand transforms*, preprint 2000.
- [12] J.C. Álvarez Paiva, I.M. Gelfand, and M. Smirnov, *Crofton densities, symplectic geometry, and Hilbert's fourth problem*, in "Arnold-Gelfand Mathematical Seminars, Geometry and Singularity Theory", V.I. Arnold, I.M. Gelfand, M. Smirnov, and V.S. Retakh (eds.). Birkhauser, Boston, 1997, pp. 77–92.
- [13] R. Ambartzumian, *A note on pseudo-metrics on the plane*, *Z. Wahrsch. Verw. Gebiete* **37** (1976), 145 – 155.
- [14] V.I. Arnold, "Mathematical Methods of Classical Mechanics", *Graduate Texts in Mathematics*, Springer-Verlag, 1989.
- [15] D. Bao and S.S. Chern and Z. Shen, "An Introduction to Riemann-Finsler Geometry", *Graduate Texts in Mathematics*, Springer-Verlag, Berlin, 2000.
- [16] M. Bekkar, *Sur les métriques admettant les plans comme surfaces minimales*, *Proc. Amer. Math. Soc.* **124** (1996), 3077–3083.
- [17] W. Blaschke, "Kreis und Kugel", Chelsea, New York, 1955.
- [18] R.L. Bryant, *Projectively flat Finsler 2-spheres of constant curvature*, *Selecta Mathematica (New Series)* **3** (1997), 161–203.
- [19] R.L. Bryant, *On metrics in 3-space for which the planes are minimal*, preprint, 1995.
- [20] H. Busemann, *Problem IV: Desarguesian spaces* in *Mathematical developments arising from Hilbert problems*, *Proc. Sympos. Pure Math.*, Vol. XXVIII, Amer. Math. Soc., Providence, R. I., 1976.
- [21] H. Busemann, *Geometries in which the planes minimize area*, *Ann. Mat. Pura Appl.* (4) **55** (1961), 171–190.
- [22] H. Busemann and W. Mayer, *On the foundations of variational calculus*, *Trans. AMS* **49** (1941), 173–198.
- [23] M. Dehn, *Über den Rauminhalt*, *Math. Ann.* **55** (1902), 465–478.

- [24] Hamel, *Über die Geometrien, in denen die Geraden die kürzesten sind*, Math. Ann. **57** (1903), 231 – 264.
- [25] D. Hilbert, *Mathematical problems*, translation in “Mathematical developments arising from Hilbert problems”, Proceedings of Symposia in Pure Math. Vol. XXVII Part 1, F. Browder (Ed.), AMS Rhode Island, 1974.
- [26] D. Hilbert, “Foundations of Geometry”, Open Court Classics, Lasalle, Illinois, 1971.
- [27] A. Karlsson and G.A. Noskov, *The Hilbert metric and Gromov hyperbolicity*, preprint 2001.
- [28] A.V. Pogorelov, “Hilbert’s Fourth Problem”, Scripta Series in Mathematics, Winston and Sons, 1979.
- [29] R. Schneider, *Crofton formulas in hypermetric projective Finsler spaces*, Arch. Math. **77** (2001), 85–97.
- [30] E. Socié-Méthou, *Comportements asymptotiques et rigidités en géométries de Hilbert*, thèse, Université de Strasbourg, 2000.
- [31] Z.I. Szabo, *Hilbert’s fourth problem, I*, Adv. in Math. **59** (1986), 185–301.
- [32] A.C. Thompson, “Minkowski Geometry”, Encyclopedia of Math. and Its Applications, Vol. 63, Cambridge Univ. Press, Cambridge, 1996.

J.C. ÁLVAREZ PAIVA, POLYTECHNIC UNIVERSITY BROOKLYN, SIX METROTECH CENTER, BROOKLYN, NY 11201 USA.

E-mail address: jalvarez@duke.poly.edu