A Course in Metric Geometry

Dmitri Burago
Yuri Burago
Sergei Ivanov

Graduate Studies
in Mathematics

Volume 33

American Mathematical Society
Providence, Rhode Island

Contents

Preface

This book is not a research monograph or a reference book (although research interests of the authors influenced it a lot)—this is a textbook. Its structure is similar to that of a graduate course. A graduate course usually begins with a course description, and so do we.

Course description. The objective of this book is twofold. First of all, we wanted to give a detailed exposition of basic notions and techniques in the theory of length spaces, a theory which experienced a very fast development in the past few decades and penetrated into many other mathematical disciplines (such as Group Theory, Dynamical Systems, and Partial Differential Equations). However, we have a wider goal of giving an elementary introduction into a broad variety of the most geometrical topics in geometry—the ones related to the notion of distance. This is the reason why we included metric introductions to Riemannian and hyperbolic geometries. This book tends to work with "easy-to-touch" mathematical objects by means of "easy-to-visualize" methods. There is a remarkable book [**Gro3**], which gives a vast panorama of "geometrical mathematics from a metric viewpoint". Unfortunately, Gromov's book seems hardly accessible to graduate students and non-experts in geometry. One of the objectives of this book is to bridge the gap between students and researchers interested in metric geometry, and modern mathematical literature.

Prerequisite. It is minimal. We set a challenging goal of making the core part of the book accessible to first-year graduate students. Our expectations of the reader's background gradually grow as we move further in the book. We tried to introduce and illustrate most of new concepts and methods by using their simplest case and avoiding technicalities that take attention away from the gist of the matter. For instance, our introduction to Riemannian geometry begins with metrics on planar regions, and we even avoid the notion of a manifold. Of course, manifolds do show up in more advanced sections. Some exercises and remarks assume more mathematical background than the rest of our exposition; they are optional, and a reader unfamiliar with some notions can just ignore them. For instance, solid background in differential geometry of curves and surfaces in $\mathbb{R}^3$ is not a mandatory prerequisite for this book. However, we would hope that the reader possesses some knowledge of differential geometry, and from time to time we draw analogies from or suggest exercises based on it. We also make a special emphasis on motivations and visualizations. A reader not interested in them will be able to skip certain sections. The first chapter is a clinic in metric topology; we recommend that the reader with a reasonable idea of metric spaces just skip it and use it for reference: it may be boring to read it. The last chapters are more advanced and dry than the first four.

Figures. There are several figures in the book, which are added just to make it look nicer. If we included all necessary figures, there would be at least five of them for each page.

- It is a must that the reader systematically studying this book makes a figure for every proposition, theorem, and construction!

Exercises. Exercises form a vital part of our exposition. This does not mean that the reader should solve all the exercises; it is very individual. The difficulty of exercises varies from trivial to rather tricky, and their importance goes all the way up from funny examples to statements that are extensively used later in the book. This is often indicated in the text. It is a very helpful strategy to perceive *every* proposition and theorem as an exercise. You should try to prove each on your own, possibly after having a brief glance at our argument to get a hint. Just reading our proof is the last resort.

Optional material. Our exposition can be conditionally subdivided into two parts: core material and optional sections. Some sections and chapters are preceded by a brief plan, which can be used as a guide through them. It is usually a good idea to begin with a first reading, skipping all optional sections (and even the less important parts of the core ones). Of course, this approach often requires going back and looking for important notions that were accidentally missed. A first reading can give a general picture of the theory, helping to separate its core and give a good idea of its logic. Then the reader goes through the book again, transforming theoretical knowledge into the genuine one by filling it with all the details, digressions, examples and experience that makes knowledge practical.

About metric geometry. Whereas the borderlines between mathematical disciplines are very conditional, geometry historically began from very “down-to-earth” notions (even literally). However, for most of the last century it was a common belief that “geometry of manifolds” basically boiled down to “analysis on manifolds”. Geometric methods heavily relied on differential machinery, as it can be guessed even from the name “Differential geometry”. It is now understood that a tremendous part of geometry essentially belongs to metric geometry, and the differential apparatus can be used just to define some class of objects and extract the starting data to feed into the synthetic methods. This certainly cannot be applied to all geometric notions. Even the curvature tensor remains an obscure monster, and the geometric meaning of only some of its simplest appearances (such as the sectional curvature) are more or less understood. Many modern results involving more advanced structures still sound quite analytical. On the other hand, expelling analytical machinery from a certain sphere of definitions and arguments brought several major benefits. First of all, it enhanced mathematical understanding of classical objects (such as smooth Riemannian manifolds) both ideologically, and by concrete results. From a methodological viewpoint, it is important to understand what assumptions a particular result relies on; for instance, in this respect it is more satisfying to know that geometrical properties of positively curved manifolds are based on a certain inequality on distances between quadruples of points rather than on some properties of the curvature tensor. This is very similar to two ways of thinking about convex functions. One can say that a function is convex if its second derivative is nonnegative (notice that the definition already assumes that the function is smooth, leaving out such functions as $f(x) = |x|$). An alternative definition says that a function is convex if its epigraph (the set $\{(x, y) : y \geq f(x)\}$) is; the latter definition is equivalent to Jensen’s inequality $f(\alpha x + \beta y) \leq \alpha f(x) + \beta f(y)$ for all nonnegative α, β with $\alpha + \beta = 1$, and it is robust and does not rely on the notion of a limit. From this viewpoint, the condition $f'' \geq 0$ can be regarded as a convenient criterion for a smooth function to be convex.

As a more specific illustration of an advantage of this way of thinking, imagine that one wants to estimate a certain quantity over all metrics on a sphere. It is so tempting to study a metric for which the quantity attains its maximum, but alas this metric may fail exist within smooth metrics, or even metrics that induce the same topology. It turns out that it still may exist if we widen our search to a class of more general length spaces. Furthermore, mathematical topics whose study used to lie outside the range of noticeable applications of geometrical technique now turned out to be traditional objects of methods originally rooted in differential geometry. Combinatorial group theory can serve as a model example of this situation. By now the scope of the theory of length spaces has grown quite far from its cradle (which was a theory of convex surfaces), including most of classical Riemannian geometry and many areas beyond it. At the same time, geometry of length spaces perhaps remains one of the most "hands-on" mathematical techniques. This combination of reasons urged us to write this "beginners' course in geometry from a length structure viewpoint".

Acknowledgements. The authors enjoyed hospitality and excellent working conditions during their stays at various institutions, including the University of Strasbourg, ETH Zurich, and Cambridge University. These unforgettable visits were of tremendous help to the progress of this book. The authors' research, which had essential impact on the book, was partially supported by the NSF Foundation, the Sloan Research Fellowship, CRDF, RFBR, and Shapiro Fund at Penn State, whose help we gratefully acknowledge. The authors are grateful to many people for their help and encouragement. We want to especially thank M. Gromov for provoking us to write this book; S. Alexander, R. Bishop, and C. Croke for undertaking immense labor of thoroughly reading the manuscript—their numerous corrections, suggestions, and remarks were of invaluable help; S. Buyalo for many useful comments and suggestions for Chapter 9; K. Shemyak for preparing most of the figures; and finally a group of graduate students at Penn State who took a Math 597c course using our manuscript as the base text and corrected dozens of typos and small errors (though we are confident that twice as many of them are still left for the reader).

Chapter 2

Length Spaces

2.1. Length Structures

First we want to informally illustrate our main concept. Imagine that you ask a mathematician: "What is the distance between New York and Sydney?". Perhaps, you get the answer "about 8 thousand miles". It is formally correct and still absolutely useless: this is the length of a straight tunnel through the Earth. Analogously, every mountaineer knows that distance in mountains is a tricky thing: if you measure it by an optical device, you get the distance "as a crow flies". It may be relevant for a crow, while wingless creatures confined to the surface of the Earth (like us) have to take long detours with lots of ups and downs; see Figure 2.1.

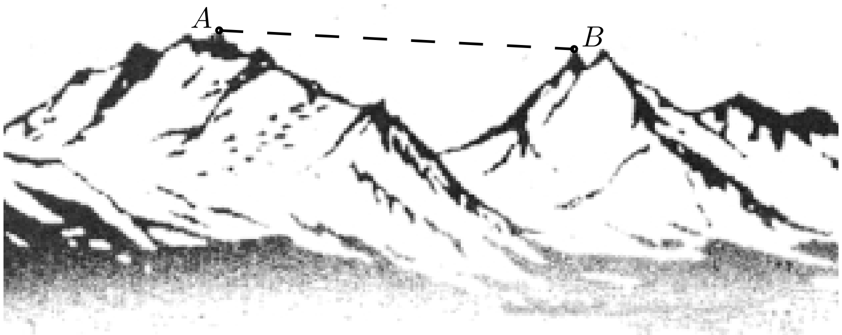

Figure 2.1: "A crow flies" along the segment AB; for a pedestrian it probably takes longer.

This little philosophical digression contains a very clear mathematical moral: in many cases, we have to begin with length of paths as the primary notion and only after that can we derive a distance function. Let us make this observation slightly more precise. For every two points on a surface in

Euclidean space (you may keep thinking of the surface of the Earth) we can measure Euclidean distance between the two points. What we do instead is we introduce a new distance which is measured along the shortest path between the two points. Generalizing this idea, one says that a distance function on a metric space is an intrinsic metric if the distance between two points can be realized by paths connecting the points (mathematically, it must be equal to the infimum of lengths of paths between the points—a shortest path may not exist).

If length of paths is our primary notion, one readily asks for its rigorous definition, where it may arise from and what are the properties of such structures. We will be occupied with these questions throughout this book.

2.1.1. Definition of length structures. Loosely speaking, a length structure consists of a *class of admissible paths* for which we can measure their length, and the *length* itself, which is a correspondence assigning a nonnegative number to every path from the class. Both the class and the correspondence have to possess several natural properties; in all reasonable examples (and in particular in all examples in this book) these requirements are automatically satisfied.

From now on we reserve the word *path* for maps of intervals: a path γ in a (topological) space X is a (continuous) map $\gamma : I \to X$ defined on an interval $I \subset \mathbb{R}$. By an *interval* we mean any connected subset of the real line; it may be open or closed, finite or infinite, and a single point is counted as an interval. Since a path is a map one can speak about its image, restrictions, etc.

A length structure on a topological space X is a class A of admissible paths, which is a subset of all continuous paths in X, together with a map $L : A \to \mathbb{R}_+ \cup \{\infty\}$; the map is called length of path. The class A has to satisfy the following assumptions:

(1) The class A is closed under restrictions: if $\gamma : [a,b] \to X$ is an admissible path and $a \le c \le d \le b$, then the restriction $\gamma_{|_{[c,d]}}$ of γ to $[c,d]$ is also admissible.

(2) A is closed under concatenations (products) of paths. Namely, if a path $\gamma : [a,b] \to X$ is such that its restrictions γ_1,γ_2 to $[a,c]$ and $[c,b]$ are both admissible paths, then so is γ. (Recall that γ is called the *product* or *concatenation* of γ_1 and γ_2, $\gamma = \gamma_1 \cdot \gamma_2$).

(3) A is closed under (at least) linear reparameterizations: for an admissible path $\gamma : [a,b] \to X$ and a homeomorphism $\varphi : [c,d] \to [a,b]$ of the form $\varphi(t) = \alpha t + \beta$, the composition $\gamma \circ \varphi(t) = \gamma(\varphi(t))$ is also an admissible path.

> **Remark 2.1.1.** Every natural class of paths comes with its own class of reparameterizations. For example, consider the class of all continuous paths and the class of homeomorphisms, the class of piecewise smooth paths and the class of diffeomorphisms. We only require that this class of reparameterizations includes all linear maps.

Examples of such classes include: all continuous paths; piecewise smooth paths (on a smooth manifold); broken lines in $\mathbb{R}^n$; see other examples below.

We require that L possesses the following properties:

(1) Length of paths is additive: $L(\gamma|_{[a,b]}) = L(\gamma|_{[a,c]}) + L(\gamma|_{[c,b]})$ for any $c \in [a, b]$.

(2) The length of a piece of a path continuously depends on the piece. More formally, for a path $\gamma : [a, b] \to X$ of finite length, denote by $L(\gamma, a, t)$ the length of the restriction of $\gamma : [a, b] \to X$ to the segment $[a, t]$. We require that $L(\gamma, a, \cdot)$ be a continuous function. (Observe that the previous property implies that $L(\gamma, a, a) = 0$.)

(3) The length is invariant under reparameterizations: $L(\gamma \circ \varphi) = L(\gamma)$ for a linear homeomorphism φ.

    (In fact, all reasonable length structures are invariant under arbitrary reparameterizations: $L(\gamma \circ \varphi) = L(\gamma)$ for any homeomorphism φ such that both γ and $\gamma \circ \varphi$ are admissible. However, it is not necessary to verify this in the beginning.)

(4) We require length structures to agree with the topology of X in the following sense: for a neighborhood U_x of a point x, the length of paths connecting x with points of the complement of U_x is separated from zero:

$$\inf\{L(\gamma) : \gamma(a) = x, \gamma(b) \in X \setminus U_x\} > 0.$$

There are several important types of length structures that will appear in this course. When the reader meets with these structures, it is advisable to come back to this definition and make sure that all of them belong to the same general scheme.

Notation. We will often use the notation $L(\gamma, a, b)$ introduced above. Namely, if $\gamma : I \to X$ is an (admissible) path and $[a, b] \subset I$, where $a \le b$, we will denote by $L(\gamma, a, b)$ the length of the restriction of γ to $[a, b]$, i.e., $L(\gamma, a, b) = L(\gamma|_{[a,b]})$. In addition, we define $L(\gamma, b, a) = -L(\gamma, a, b)$. This convention implies that $L(\gamma, a, b) = L(\gamma, a, c) + L(\gamma, c, b)$ for all $a, b, c \in I$ (verify this).

2.1.2. Length spaces. Once we have a length structure, we are ready to define a metric (a distance function) associated with the structure. We will always assume that the topological space X carrying the length structure is a Hausdorff space. For two points $x, y \in X$ we set the associated distance $d(x, y)$ between them to be the infimum of lengths of admissible paths connecting these points:

$$d_L(x, y) = \inf\{L(\gamma); \gamma : [a, b] \to X, \gamma \in A, \gamma(a) = x, \gamma(b) = y\}.$$

If it is clear from the context which length structure L gives rise to d_L, we usually drop L in the notation d_L.

Exercise 2.1.2. Verify that (X, d_L) is a metric space.

Note that d_L is not necessarily a finite metric. For instance, if X is a disconnected union of two components, no continuous path can go from one component to the other and therefore the distance between points of different components is infinite. On the other hand, there may be points such that continuous paths connecting them exist but all have infinite length. One says that two points $x, y \in X$ belong to the same accessibility component if they can be connected by a path of finite length.

Exercise 2.1.3. 1. Check that accessibility by paths of finite length is indeed an equivalence relation. Your argument should use additivity of length and the assumption that the concatenation of admissible paths is an admissible path.

2. Verify that accessibility components coincide with components of finiteness for d_L.

3. Verify that accessibility components coincide with both connectivity and path connectivity components of (X, d_L).

Did you notice that you have used the following fact?

Exercise 2.1.4. Prove that admissible paths of finite length are continuous with respect to (X, d_L).

This exercise deals with the topology determined by the metric d_L rather than the initial topology of the space X. There really are examples where these two topologies differ; such examples will appear later in this book.

Exercise 2.1.5. Prove that the topology determined by d_L can be only finer than that of X: any open set in X is open in (X, d_L) as well.

Definition 2.1.6. A metric that can be obtained as the distance function associated to a length structure is called an *intrinsic*, or *length, metric*. A metric space whose metric is intrinsic is called a *length space*.

Not every metric can arise as a length metric. Even if (X, d) is a length space and $A \subset X$, the restriction of d to A is not necessarily intrinsic. For example, consider a circle in the plane.

Moreover, not every metrizable topology can be induced by intrinsic metrics:

Exercise 2.1.7. 1. Prove that the set of rational numbers is not homeomorphic to a length space.

2. Prove that the union of the graph $\{(x, y) : y = \sin(1/x),\ x > 0\}$ and the y-axis (with its topology inherited from $\mathbb{R}^2$) is not homeomorphic to a length space.

There can be more delicate reasons why a topological space may be not homeomorphic to a length space:

Exercise 2.1.8. Consider the union of segments

$$\bigcup_{i=1}^{\infty} [(0,0), (\cos 1/i, \sin 1/i)] \cup [(0,0), (1,0)]$$

in the Euclidean plane, depicted in Figure 2.2.

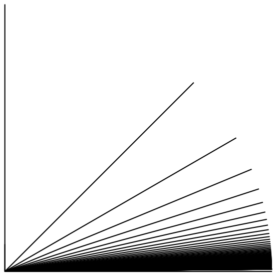

Figure 2.2: The space (with the topology inherited from $\mathbb{R}^2$) is not homeomorphic to a length space.

This set (resembling a fan made of segments) is a topological space with its topology inherited from the Euclidean plane (this is the topology of Euclidean distance restricted to the set). Prove that this topological space is not homeomorphic to a length space.

As a hint to the above exercises, consider the following more general one.

Exercise 2.1.9. Prove that a length space is locally path connected: every neighborhood of any point contains a smaller neighborhood which is path-connected.

One uses infimum instead of simple minimum when defining d_L since there may be no shortest path between two points. For instance, consider the Euclidean plane with an open segment removed; then for the endpoints of the segment, the shortest path does not exist: it is just removed. Still, its length can be approximated with a given precision by other paths connecting the points. Such situations rarely arise in "real-life" examples; in most cases they will be also prohibited by imposing completeness-compactness type assumptions. On the other hand, existence of shortest paths helps to avoid tedious and nonessential complications. For the simplicity of our exposition we will often restrict ourselves to complete length structures, which are defined as follows:

Definition 2.1.10. A length structure is said to be *complete* if for every two points x, y there exists an admissible path joining them whose length is equal to $d_L(x, y)$; in other words, a length structure is complete if there exists a shortest path between every two points.

Intrinsic metrics associated with complete length structures are said to be *strictly intrinsic*.

2.2. First Examples of Length Structures

To get better motivated, let us briefly meet with a few examples from the zoo of length structures and intrinsic metrics; we will not analyze them in this section, but we suggest that you keep them in mind and use them for testing further definitions and concepts. Notice that in many examples the space itself is a part of a Euclidean space, and there are two different ways of changing the usual Euclidean length structure: we change the class of admissible paths or change the notion of length of paths (or both).

Example 2.2.1 ("Driving in Manhattan")**.** The space here is the Euclidean plane, and length of paths is the same as usual. The only difference is that we restrict the class of admissible paths to broken lines with edges parallel to one of the coordinate axes. (A critically thinking reader should yell that paths are maps while broken lines are sets! This is absolutely true, and formally we mean the paths whose images are broken lines.) Can you draw a ball in the corresponding intrinsic metric? (It will not look round: you should get a diamond.)

Example 2.2.2 ("Metric on an island")**.** The space is a connected region in the Euclidean plane, and again length of paths is the same as usual. Admissible paths are all (piecewise smooth) paths contained in the region. If the region is convex, this length structure induces usual Euclidean distance. One may think of this region as an island, and the distance is measured by a creature who cannot swim. Drawing balls in intrinsic metrics arising this way may be quite fun; see Figure 2.3. Is this metric strictly intrinsic? What if we consider the closure of the region?

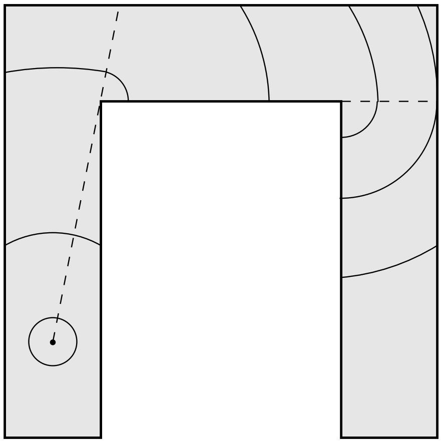

Figure 2.3: Metric balls in a nonconvex island.

The reader can generalize this example for a subspace of a space with length structure. Certainly, only sensible choices for a subspace lead to reasonable examples. For instance, restricting the Euclidean length structure in $\mathbb{R}^2$ to a circle leads to the angular metric. More generally, one obtains spherical geometry by restricting the usual Euclidean length structure to a round sphere. On the other hand, restricting the standard length structure of $\mathbb{R}$ to the set of rational points we obtain a space with each accessibility component consisting of just one point.

Example 2.2.3 (induced length structure)**.** The formal contents of this example is comprised in the following definition. Let $f : X \to Y$ be a continuous map from a topological space X to a space Y endowed with a length structure. One defines the *induced length structure* in X as follows. A path in X is admissible if its composition with f is admissible in Y. The length of an admissible path in X is set to the length of its composition with f with respect to the length structure in Y.

(In fact, this construction may not define a length structure in X because the new length function may fail to satisfy the fourth condition from section 2.1.1. We use the term "induced length structure" only if this is indeed a length structure.)

At first glance, the above definition may sound like a tautology. However, the properties of an induced metric may drastically differ from the properties of a metric we began with. For instance, the leading example of an induced metric when f is a *surface* (that is an immersion $f : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3$ of a two-dimensional region into $\mathbb{R}^3$) has served as the main motivating example in metric geometry for over a century. For a reader who is already familiar with Riemannian metrics, we mention that it is also true (though hard to believe and not easy to prove) that every Riemannian length structure on $\mathbb{R}^n$ can be induced by a map $f : \mathbb{R}^n \to \mathbb{R}^n$ (which makes lots of folds and is rarely smooth).

Example 2.2.4 ("Crossing a swamp": conformal length)**.** The space is the Euclidean plane, and admissible paths are all (piecewise smooth) paths. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a positively-valued continuous (or even L_∞) function. Define the length of a path $\gamma : [a, b] \to \mathbb{R}^2$ by

$$L(\gamma) = \int_a^b f(\gamma(t)) \cdot |\gamma'(t)| dt.$$

This length structure can be thought of as a weighted Euclidean distance. For instance, a traveler who measures the length (=time needed to cover) of a certain route would apparently assign big values to f in a territory that is difficult to traverse (for instance, a swamp or a mountain trail). From the mathematical viewpoint, this is the first example of a Riemannian length structure, which will be discussed further in Chapter 5; the word "conformal" in the title of this subsection reflects the fact that such types of Riemannian structures are called conformally flat.

Example 2.2.5 (Finslerian length)**.** Thinking of the previous example as a length structure for a traveler who assigns weights to different parts of his/her path, one notices that an important feature of real travel is not reflected here. Namely, the difficulty of traversing a region depends not only on the region itself but also on the direction of the route; for instance, choosing a direction in which most ravines are oriented might essentially simplify the trip. To incorporate this additional information, one introduces a function f in two variables and applies it to both γ and its velocity γ'. The expression for the length reads:

$$L(\gamma) = \int_a^b f(\gamma(t), \gamma'(t))\, dt.$$

(A physics-oriented reader recognizes that this structure can be interpreted as action.) In order for this expression to be invariant under bijective reparameterizations of paths, one has to require that f satisfies $f(x, kv) = |k| f(x, v)$ for all scalars k, points x and vectors v (check this as an exercise for change of variable in a definite integral). Usually a stronger requirement is imposed on f, namely for every point x the function $f(x, \cdot)$ must be a norm. A motivation for this will be explained in section 2.4.2. Length structures obtained from this type of constructions are called *Finslerian*, or *Finsler*.

Remark 2.2.6. A reader who seriously tests our definitions against needs of travelers and mountaineers will notice that some features are still missing.

Namely the fact that walking downhill may be easier than climbing uphill cannot be reflected in a length structure. Since the distance in a metric space must be symmetric, we had to require that the length is invariant under all changes of variable including the orientation-reversing ones like $t \mapsto -t$. One could modify the definitions to allow nonsymmetric length structures and metrics. Everything in this chapter can be adapted to such generalized settings; however, this would not make sense in the rest of the book.

Example 2.2.7 (A "cobweb" and a "notebook")**.** Begin with several disjoint segments and glue some of their endpoints together. Such a space may resemble a cobweb in Euclidean space. This space has a natural length structure. All continuous paths are admissible. The space is built out of segments, and we know how to measure the length of a path while it travels within one segment. Thus to find the length of a path we restrict it to (countably many) intervals such that the image of each interval is contained in one segment, and add the lengths of the restrictions. This is a first example of metric graphs, and the construction of its length structure is a particular case of gluing, discussed in detail in section 3.1.

Another example of the same type can be made out of several copies of a closed half-plane by attaching them together along their boundary lines. This is an example of a polyhedral length space. It can be visualized (and realized) in Euclidean spaces: it looks like an open book. Can you modify the definition of the length structure on a cobweb for this case? Caution: while we could disregard the part of a path spent in endpoints of segments (nodes of the cobweb) since they have zero length, this is not the case for the common edge of the half-planes.

2.3. Length Structures Induced by Metrics

2.3.1. Length of curves in metric spaces. Let us recall our motivating example from the very beginning of this chapter. We began with some distance function (Euclidean distance "as a crow flies") that was not satisfactory since there might be no paths realizing this distance. By saying this we already mean that we know how to measure length of paths, and this length is somehow derived from the Euclidean distance!

Indeed, some of the main examples of length structures are those induced by metric structures. For admissible paths one may use just all continuous paths; for some of them the length may be infinite. In some cases a better choice is the class of Lipschitz paths, that is, the class of maps $\gamma : [a,b] \to X$ such that $d_X(\gamma(t), \gamma(t')) \le C|t - t'|$, for all $t, t' \in [a,b]$; C is a positive constant.

How do we define the length of a path in Euclidean space? We approximate the path by broken lines and define the length as the limit of their lengths. For each of these broken lines, its vertices belong to (the image of) the path and they are well ordered with respect to the parameter of the path. Since all that we actually use are distances between neighboring vertices in this list, we can mimic this definition in a general metric space in the most straightforward way (compare also with 2.4.12).

Definition 2.3.1. Let (X, d) be a metric space and γ be a path in X, i.e., a continuous map $\gamma : [a, b] \to X$. Consider a *partition* Y of $[a, b]$), that is, a finite collection of points $Y = \{y_0, ..., y_N\}$ such that $a = y_0 \le y_1 \le y_2 \le \cdots \le y_N = b$. The supremum of the sums

$$\Sigma(Y) = \sum_{i=1}^{N} d(\gamma(y_{i-1}), \gamma(y_i)).$$

over all the partitions Y is called the *length* of γ (with respect to the metric d) and denoted $L_d(\gamma)$. A curve is said to be *rectifiable* if its length is finite.

The length structure induced by the metric d is defined as follows: all continuous paths (parameterized by closed intervals) are admissible, and the length is given by the function L_d.

This definitions can be formally applied to any metric space, but one gets sensible examples only by a wise choice of a metric space to begin with: for instance, if we start with a discrete space, there are no nonconstant continuous paths at all. If it is clear from the context which metric d induces the length L, we usually drop d in the notation L_d.

The usual "Euclidean" definition uses passing to a limit as the edges of broken lines approach zero. The following exercise shows that there is no difference here:

Exercise 2.3.2. Prove that $\Sigma(Y) \to L(\gamma)$ as $\max_i\{|y_i - y_{i+1}|\} \to 0$.

Exercise 2.3.3. Prove that the definition of length is compatible with the one used in differential geometry. Namely if $(V, |\cdot|)$ is a finite-dimensional normed vector space and $\gamma : [a, b] \to V$ is a differentiable map, then $L(\gamma) = \int_a^b |\gamma'(t)|\, dt$.

2.3.2. Properties of the induced length. All properties of length structures hold for this structure; this length is also *semi-continuous.* Let us verify (some of) them:

Proposition 2.3.4. *The length structure $L = L_d$ induced by a metric d possesses the following properties:*

(i) Generalized triangle inequality: $L(\gamma) \ge d(\gamma(a), \gamma(b))$.

(ii) Additivity: if $a < c < b$, then $L(\gamma, a, c) + L(\gamma, c, b) = L(\gamma)$. In particularly, $L(\gamma, a, c)$ is a nondecreasing function of c.

(iii) If γ is rectifiable, the function $L(\gamma|_{[c,d]}) = L(\gamma, c, d)$ is continuous in c and d.

(iv) L is a lower semi-continuous functional on the space of continuous maps of $[a, b]$ in X with respect to point-wise convergence, and hence with respect to the uniform (i.e., C^0-) topology. This means that if a sequence of rectifiable paths γ_i (with the same domain) is such that $\gamma_i(t)$ converges to $\gamma(t)$ (as $i \to \infty$, for every t in the domain), then $\liminf L(\gamma_i) \geq L(\gamma)$.

Proof. (i) Indeed, the triangle inequality implies that $\Sigma(Y) \geq d(\gamma(a), \gamma(b))$ for all Y's, and thus the inequality persists under passing to the limit.

(ii) First notice that if Y' is obtained from Y by adding one point, then $\Sigma(Y) \geq \Sigma(Y)$ (by the same triangle inequality). Adding c to a partition Y of $[a, b]$ and then splitting it into two partitions of $[a, c]$ and $[c, b]$ completes the argument.

We repeat again that the readers should try to consider such lemmas as exercises and try to prove them on their own. If reading the proof was needed, then drawing a figure with all notations is a must!

(iii) We prove the continuity of L in d, $a < d \leq b$, from the left (the other cases are analogous and are left to the reader). Take $\varepsilon > 0$ and consider a partition Y such that $L(\gamma) - \Sigma(Y) < \varepsilon$. One may suppose that $y_{j-1} < d = y_j$. Then

$$L(\gamma, y_{j-1}, d) - d(\gamma(y_{j-1}), \gamma(d)) < \varepsilon,$$

and the same inequality takes place for each c such that $y_{j-1} \leq c \leq d$.

(iv) Let paths γ_j converge pointwise to γ. Take $\varepsilon > 0$ and fix a partition Y for γ such that $L(\gamma) - \Sigma(Y) < \varepsilon$. Now consider the sums $\Sigma_j(Y)$ for paths γ_j corresponding to the same partition Y. Choose j to be so large that the inequality $d(\gamma_j(y_i), \gamma(y_i)) < \varepsilon$ holds for all $y_i \in Y$. Then

$$L(\gamma) \leq \Sigma(Y) + \varepsilon \leq \Sigma_j(Y) + \varepsilon + (N + 1)\varepsilon \leq L(\gamma_j) + (N + 2)\varepsilon.$$

Since ε is arbitrary, this implies (iv). □

Remark 2.3.5. In general, functional L is not continuous. A stairs-like example is shown in Figure 2.4.

2.3.3. Induced intrinsic metric.

A metric d induces a length structure. The latter, in its turn, gives rise to an intrinsic metric (on each component of accessibility by rectifiable paths). Thus we obtain a canonical construction of induced intrinsic metrics

$$(X, d) \to (X, \widehat{d})$$

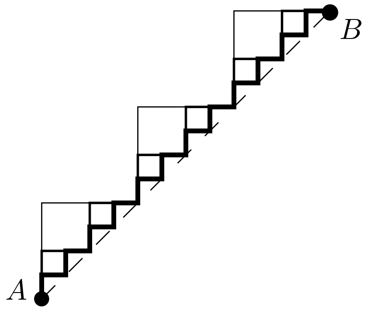

Figure 2.4: L is not continuous.

where $\widehat{d} = d_{L_d}$.

Exercise 2.3.6. Prove that the intrinsic metric induced by the restriction of Euclidean distance to the circle $x^2 + y^2 = 1$ is the angular metric.

Note that the topology of the induced intrinsic metric may be very poorly connected with the original topology of the space, as can be seen from the following examples-exercises:

Exercise 2.3.7. Find the induced intrinsic metric for the metric

$$d((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + \sqrt{|y_1 - y_2|}$$

on $\mathbb{R}^2$. What is the topology of the resulting length space?

Answer: A continuum of disjoint real lines, each with its standard metric.

Exercise 2.3.8. Consider the union of segments

$$U = \bigcup_{n=1}^{\infty} [(0,1), (1/n, 0)] \cup [(0,1), (0,0)] \subset \mathbb{R}^2.$$

The sequence of points $\{(1/n, 0)\}$ converges to $(0,0)$ in the topology inherited by U from $\mathbb{R}^2$. Since all pairwise distances between these points in the induced intrinsic metric are at least 2, this sequence diverges with respect to the intrinsic metric. Prove these statements.

Exercise 2.3.9. Connecting the point $(0,1) \in \mathbb{R}^2$ with all points of the standard Cantor set in the segment $[0,1] = [0,1] \times \{0\} \subset \mathbb{R}^2$, one obtains a compact connected set. Show that in the induced intrinsic metric this set is noncompact although it is still connected.

Exercise 2.3.10. Begin with a simple nonrectifiable curve in $\mathbb{R}^2 \subset \mathbb{R}^3$ and build a cone over its image by choosing a point (vertex of the cone) in $\mathbb{R}^3$ and connecting the vertex with every point in the image of the curve by a segment. The intrinsic distance in this cone induced by Euclidean distance in $\mathbb{R}^3$ is finite for every two points: one can go from one point to the vertex of the cone along a straight segment and then get to the other point along another segment. Show that removing the vertex of the cone makes it disconnected in the topology of induced intrinsic metric, while it is still connected in usual topology.

You can begin with a simple curve whose restriction to any nontrivial interval is nonrectifiable (prove that such a curve does exist). In the original topology this cone is still homeomorphic to a disc. Prove that in the induced intrinsic metric this cone is homeomorphic to the bouquet of a continuum of intervals (that is, the *disjoint* union of segments glued at one point).

Exercise 2.3.11. For two vectors $V, W \in \mathbb{R}^2$, set

$$d(V,W) = \big||V| - |W|\big| + \min(|V|,|W|) \cdot \sqrt{\measuredangle(V,W)},$$

where $\measuredangle(V,W)$ denotes the angle between V and W. Prove that

(i) the topology determined by d is the standard Euclidean one;

(ii) the induced intrinsic metric $\widehat{d}$ is

$$\widehat{d}(V,W) = \begin{cases} \big||V| - |W|\big| & \text{if } \measuredangle(V,W) = 0, \\ \big||V| + |W|\big| & \text{otherwise;} \end{cases}$$

(iii) $(\mathbb{R}^2, \widehat{d}\,)$ is homeomorphic to the bouquet of a continuum of rays.

One can consider the length $L_{\widehat{d}}$ induced by the new metric $\widehat{d}$, and this length in turn determines a "second-stage" intrinsic metric. The reader may imagine how much confusion between various lengths and metrics might arise. However, this is not the case, as the length induced by $\widehat{d}$ is the same as induced by d.

Proposition 2.3.12. *Let (X,d) be a metric space and $\widehat{d}$ be the intrinsic metric induced by d.*

(1) *If γ is a rectifiable curve in (X,d), then $L_{\widehat{d}}(\gamma) = L_d(\gamma)$.*

(2) *The intrinsic metric induced by $\widehat{d}$ coincides with $\widehat{d}$. In other words, inducing a length metric is an idempotent operation.*

Proof. The fact that the length of every curve in (X,d) is not less than the distance between its endpoints implies that $\widehat{d} \geq d$. It follows immediately that $L_{\widehat{d}}(\gamma) \geq L_d(\gamma)$. To prove the inverse inequality, let $[a,b]$ be the domain of γ and let $Y = \{y_i\}$ be an arbitrary partition of $[a,b]$. Observe that $\widehat{d}(\gamma(y_i), \gamma(y_{i+1})) \leq L_d(\gamma, y_i, y_{i+1})$ because the left-hand value is the infimum of lengths one of which is written on the right-hand side. Therefore

$$\Sigma_{\widehat{d}}(Y) = \sum \widehat{d}(\gamma(y_i), \gamma(y_{i+1})) \leq L_d(\gamma).$$

Since Y is an arbitrary partition, the inequality $L_{\widehat{d}}(\gamma) \leq L_d(\gamma)$ follows. This proves the first statement of the proposition. The second one is a trivial consequence. □

Remark 2.3.13. The assumption that the curve γ is rectifiable is essential, simply because otherwise it may fail to be continuous in $(X, \widehat{d})$. The set of continuous curves in $(X, \widehat{d})$ is generally a subset of the respective set for (X, d) but it contains all rectifiable curves. See Exercises 2.1.4 and 2.1.5.

2.4. Characterization of Intrinsic Metrics

A metric was said to be intrinsic if it can be obtained by a certain construction. In this chapter we discuss properties that distinguish intrinsic metrics among all metrics and criteria that tell whether a given metric is intrinsic or not.

2.4.1. Another definition of length spaces.

The second statement of Proposition 2.3.12 gives a "constructive" criterion to find out whether a given metric can be obtained as an induced intrinsic one. Namely a metric is induced if and only if it induces itself. The following proposition generalizes this for arbitrary intrinsic metrics.

Proposition 2.4.1. *Let (X, d) be a length space and $\widehat{d}$ be the intrinsic metric induced by d. Then $\widehat{d} = d$.*

Proof. Let L be the length function that defines d and L_d be the length induced by d. Observe that $L_d(\gamma) \leq L(\gamma)$ for any admissible curve γ of finite length (repeat the respective argument from the proof of Proposition 2.3.12). Obviously a smaller length function determines a smaller metric; thus $\widehat{d} \leq d$. On the other hand, we already know that $\widehat{d} \geq d$; hence $\widehat{d} = d$. □

Note that the equality $d = \widehat{d}$ automatically implies that d is an intrinsic metric—just because the induced metric $\widehat{d}$ is always intrinsic. Hence it can be considered as an alternative definition of the term "intrinsic metric". In other words, (X, d) is a length space if and only if for any points $x, y \in X$ and any $\varepsilon > 0$ there exists a curve γ connecting x and y such that $L_d(\gamma) < d(x, y) + \varepsilon$.

If one considers an intrinsic metric d and it does not matter *which* length structure determines it, Proposition 2.4.1 allows us to assume that the length structure is the one induced by d and therefore use all properties of induced length that we established in previous sections.

2.4.2. Recovering a length structure. Now we are ready to answer another natural question, namely: given an intrinsic metric space (X, d), how can we recover the initial length structure L. In fact, recovering the length structure is not possible without additional assumptions because the same intrinsic metric usually can be obtained from many different length structures.

Exercise 2.4.2. Give an example of a length structure on the plane for which all continuous curves are admissible, the resulting intrinsic metric is the standard Euclidean one, but lengths of some curves differ from their Euclidean lengths.

Proposition 2.4.1 shows a natural candidate for the length structure. Namely, the length structure L_d induced by the metric indeed gives rise to the original metric. We will therefore reformulate our question: under what assumptions is $L = L_d$? This kind of question is also important for the following reason: in many examples the initial length function L has many specific features that one may want to exploit. For example, the definition of Finslerian length as the integral of "speed" (see example in Section 2.2) is much more suitable for computing actual lengths than the general definition of induced length applied to the resulting Finslerian metric. If coincidence of two lengths is known, one can combine specific features of the initial function L with general properties of lengths induced by metrics.

Certainly, two length structures may be different just because they have different classes of admissible paths. Thus, we care only for their values on the paths admissible for L. Then we notice that, besides all properties from the definition of length structures, L_d possesses an additional property of lower semi-continuity (see Proposition 2.3.4). Indeed, lower semi-continuity turns out to be the key property here:

Theorem 2.4.3. *If L is a lower semi-continuous length structure, then L coincides with the length structure induced by its intrinsic metric $d = d_L$ on all curves admissible for L: $L(\gamma) = L_d(\gamma)$. As usual, the semi-continuity means that if a sequence of paths $\gamma_i(t)$ pointwise converges to $\gamma(t)$, then* $\liminf L(\gamma_i) \geq L(\gamma)$.

Proof. The inequality $L_d(\gamma) \leq L(\gamma)$ holds for any length structure—see Propositions 2.3.12 and 2.4.1. Let us prove the opposite inequality. By property 2 of length structure, the function $L(t) = L(\gamma_{|[a,t]})$ is uniformly continuous in $[a, b]$ for each rectifiable curve $\gamma : [a, b] \to X$. Hence for every $\varepsilon > 0$, there exists a partition $a = t_0 \leq t_1 \leq \cdots \leq t_{k+1} = b$ such that $d_L(\gamma(t_i), \gamma(t_{i+1})) < \varepsilon$ for every integer i between 0 and k. According to the definition of d_L, for each $i = 0, 1, \ldots, k$ there exists a curve σ_i : $[t_i, t_{i+1}] \to X$ with endpoints $\sigma_i(t_i) = \gamma(t_i)$, $\sigma_i(t_{i+1}) = \gamma(t_{i+1})$ such that

$L(\sigma_i) \leq d(\gamma(t_i), \gamma(t_{i+1})) + \varepsilon/k$. For the concatenation h_ε of the curves σ_i we have

$$L(h_\varepsilon) = \sum_{i=0}^{k} L(\sigma_i) \leq \sum_{i=0}^{k} d_L(\gamma(t_i), \gamma(t_{i+1})) + \varepsilon \leq L_d(\gamma) + \varepsilon.$$

From the triangle inequality one readily sees that $d(\gamma(t), h_\varepsilon(t)) \leq 3\varepsilon$ for every $t \in [a, b]$. It follows that $h_\varepsilon(t) \to \gamma(t)$ since the topology determined by d is finer than the initial one. Now the lower semi-continuity for L implies

$$L(\gamma) \leq \lim_{\varepsilon \to 0} \inf L(h_\varepsilon) \leq L_d(\gamma),$$

proving the theorem. □

Example 2.4.4. Let us come back to the example from Section 2.2 entitled "Finslerian length". We have chosen a function f in two vector-valued variables and measured the length of a path γ by the formula

$$L(\gamma) = \int_a^b f((\gamma(t)), \gamma'(t))dt.$$

Let us take a function f that does not depend on the first argument and such that $f(x,(1,0)) = f(x,(0,1)) = 1/10$ and $f(x,(1,1)) = 1$. (The fact that f is independent of the first argument intuitively means that the cost of out travel depends on the direction only, and it does not depend on a particular location.) This Finslerian length structure is not lower semi-continuous. To see this, consider a sequence of (stairs-like) broken lines with edges parallel to coordinate axes and approaching the segment $[(0,0),(1,1)]$. The length of this segment (with respect to the Finslerian structure) is $\sqrt{2}$, while the length of each of the broken lines is $1/5$. The reason why this has happen is that we made our travel in the diagonal direction "too inexpensive" compared with the coordinate directions.

Exercise 2.4.5. Show that the Finslerian length structure constructed by a function $f(x, v) = F(v)$ is lower semi-continuous if and only if $F(v)$ satisfies the subadditivity assumption from Definition 1.2.11 of norm: $F(v + w) \leq F(v) + F(w)$. Show that this is also true for general $f(x, v)$: the lower semi-continuity of the corresponding length structure is equivalent to the assumption that f satisfies the inequality $f(x, v + w) \leq f(x, v) + f(x, w)$. This is the reason why, in defining Finslerian structures, one usually requires $f(x, \cdot)$ to be a *norm* for each x.

Exercise 2.4.6. Let d be a Finslerian metric (that is, the metric associated with a Finslerian length structure) on a domain $D \subset \mathbb{R}^n$. Prove that the topology of d coincides with the standard Euclidean one.

2.4.3. Existence of midpoints.

Definition 2.4.7. A point $z \in X$ is called a *midpoint* between points x, y in a metric space (X, d) if $d(x,z) = d(z,y) = \frac{1}{2}d(x,y)$.

The following lemma formulates a necessary condition for a metric to be strictly intrinsic (this condition, under mild assumptions, also turns out to be sufficient).

Lemma 2.4.8. *If d is a strictly intrinsic metric, then for every two points x, y there exists a midpoint z.*

Proof. The length of a shortest path $\gamma : [a,b] \to X$ between x and y is $L(\gamma) = d(x,y)$. Denote $L(t) = L(\gamma_{[a,t]})$. Since $L(t)$ is continuous in t and $L(0) = 0$, there is a $c \in [a,b]$ such that $L(c) = \frac{1}{2}L(b)$. Now choosing $z = \gamma(c)$ and using the fact that the length of a path is not less than the distance between its endpoints, one immediately sees that $d(x,z) = d(y,z) = 1/2d(x,y)$. □

In other words, the previous lemma tells us that if d is strictly intrinsic, then the (closed) balls $\overline{B}_{d(x,y)/2}(x)$ and $\overline{B}_{d(x,y)/2}(y)$ have a nonempty intersection.

Exercise 2.4.9. Show that, for a strictly intrinsic metric d, if $r_1 + r_2 = d(x,y)$, then the balls $\overline{B}_{r_1}(x)$ and $\overline{B}_{r_2}(y)$ have a nonempty intersection.

For intrinsic metrics, an analogous lemma asserts:

Lemma 2.4.10. *If d is an intrinsic metric, then, given a positive ε, for every two points $x, y \in X$ there exists an ε-midpoint z, that is, a point z such that $|2d(x,z) - d(x,y)| \le \varepsilon$ and $|2d(y,z) - d(x,y)| \le \varepsilon$. In other words, if $2r > d(x,y)$, then the balls $B_r(x)$ and $B_r(y)$ have a nonempty intersection.*

Proof. Repeat the same arguments as in the proof of the previous lemma for a path γ connecting x and y, and such that $L(\gamma) - d(x,y) \le \varepsilon$. □

Exercise 2.4.11. Let d be an intrinsic metric. Show that, if $r_1 + r_2 > d(x,y)$, then the balls $B_{r_1}(x)$ and $B_{r_2}(y)$ have a nonempty intersection.

We leave as an exercise the following corollary. In a sense it allows one to measure distances using (sufficiently fine) "dotted lines" between two points instead of measuring distances by means of connecting points by paths (compare also with Definition 2.3.1).

Corollary 2.4.12. *Given a positive ε and two points $x, y \in X$ in a space with a strictly intrinsic metric d, there exists a finite sequence of points $x_1 = x, x_2, \dots, x_k = y$ such that every two neighboring points in this* *sequence are ε-close (that is, $d(x_i, x_{i+1}) \leq \varepsilon$ for all $I = 1, \dots, k-1$) and $\sum_{i=1}^{k-1} d(x_i, x_{i+1}) = d(x, y)$.*

For intrinsic metrics the last formula in the corollary should be replaced by $\sum_{i=1}^{k-1} d(x_i, x_{i+1}) - d(x, y) \leq \varepsilon$.

Exercise 2.4.13. If x and y are two points in a length space (X, d) and $r < d(x, y)$, then $\operatorname{dist}(y, B_r(x)) = d(x, y) - r$. Prove this.

Exercise 2.4.14. Let X be a length space, Y a metric space, and let a map $f : X \to Y$ be locally Lipschitz with a Lipschitz constant C. Prove that f is Lipschitz with the same constant.

Exercise 2.4.15. Let (X, d) be a length space and A a connected open subset of X. Then d induces on A the (finite-valued) intrinsic metric d_A. Moreover each point $p \in A$ has a neighborhood $U \subset A$ such that for any points $p,\ q \in U$ we have $d(p, q) = d_A(p, q)$.

2.4.4. Complete intrinsic metrics. In many cases the converse of Lemma 2.4.8 is true: the existence of midpoints (resp. ε-midpoints) implies that a complete metric space is strictly intrinsic (resp. intrinsic). Thus we have a criterion that tells us whether a complete locally-compact metric space is a length space.

Theorem 2.4.16. *Let (X, d) be a complete metric space.*

1. If for every $x, y \in X$ there exists a midpoint, then d is strictly intrinsic.

2. If for every $x, y \in X$ and every positive ε there exists an ε-midpoint, then d is intrinsic.

This theorem has the following immediate corollary, which can be used as an alternative criterion for a metric to be intrinsic:

Corollary 2.4.17. *A complete metric space (X, d) is a length space iff, given a positive ε and two points $x, y \in X$, there exists a finite sequence of points $x_1 = x_1, x_2, \dots, x_k = y$ such that every two neighboring points in this sequence are ε-close (i.e., $d(x_i, x_{i+1}) \leq \varepsilon$ for all $i = 1, \dots, k-1$) and $\sum_{i=1}^{k-1} d(x_i, x_{i+1}) < d(x, y) + \varepsilon$.*

This corollary says that a metric is intrinsic if and only if, given two points and a positive ε, one can reach one of the points starting from the other one and hopping with jumps shorter than ε and with the total length of the jumps not exceeding the distance between the points plus ε.

Proof of Theorem 2.4.16. To prove that a metric is intrinsic we have to show that for any two points x and y there are paths connecting x and

y whose lengths approximate $d(x, y)$ with any given precision. In case of strictly intrinsic metric, there must be a path whose length is equal to $d(x, y)$. We proceed with the case of strictly intrinsic metrics; modifying this argument for the other case is left as an exercise.

We will construct a path $\gamma : [0, 1] \to X$ between x and y such that $\gamma(0) = x$, $\gamma(1) = y$ and $L(\gamma) = d(x, y)$. First we assign the values of γ for all dyadic rationals (rational numbers of the form $k/2^m$ for some natural numbers k, m). Then we extend this partially defined map by continuity; only this step of the argument will use the completeness of (X, d). Indeed, a path with the desirable properties must pass through a midpoint between x and y. Since such midpoints exist by the assumption of the theorem, choose such a midpoint and assign it to be the image $\gamma(1/2)$. Now we assign $\gamma(1/4)$ to be a midpoint between $x = \gamma(0)$ and $\gamma(1/2)$ and $\gamma(3/4)$ to be a midpoint between $\gamma(1/2)$ and $y = \gamma(1)$. Proceeding this way, we define γ for all dyadic rationals between 0 and 1.

According to our construction, for every two dyadic rationals t_i, t_j

$$d(\gamma(t), \gamma(t')) = |t - t'| \cdot d(x, y). \tag{2.1}$$

This equality implies that the map γ, defined on the set of dyadic rationals, is Lipschitz. Since X is complete and the set of dyadic rationals is dense in $[0, 1]$, this map can be extended to the entire interval $[0, 1]$ (cf. Proposition 1.5.9). Thus we obtained a path $\gamma : [0, 1] \to X$ connecting x and y. Then (2.1) implies that $L(\gamma) = d(x, y)$. □

To see how these results work, we suggest several exercises:

Exercise 2.4.18. Prove that the completion of a length space is a length space.

Exercise 2.4.19. Let X be a compact topological space and let $\{d_n\}_{n=1}^{\infty}$ be a sequence of intrinsic metrics on X that uniformly converge to a metric d (recall that metrics are functions on $X \times X$, so the notion of uniform convergence applies here). Prove that d is intrinsic too.

2.5. Shortest Paths

2.5.1. Curves and natural parameterizations. It is important to remember that, when speaking about paths, we do mean maps and not their images. Indeed, an image of a continuous path may fill a disc; two paths making different number of rounds around a circle are essentially different and in particular have different lengths. Still, changing the parameter by a strictly increasing change of variable means that we visit the same collection of points in the same order; in other words, we traverse the same "curve" in the same direction. One expects that such changes of variable do not change geometric properties.

This suggests the idea of a curve as a class of equivalent paths with respect to the following relation: paths $\gamma_1 : I_1 \to X$ and $\gamma_2 : I_2 \to X$ are equivalent if there exists a strictly increasing continuous map φ from I_1 onto I_2 such that $\gamma_1 = \gamma_2 \circ \varphi$. (Check that this indeed is an equivalence relation.) However such a definition would be too restrictive for our purposes. If a path is constant on some subinterval (i.e., it "stops for a while" at a point), so does any path obtained from this one by a change of variable. We want such a path to be equivalent to one that goes the same way except that it passes through the point without stopping. To achieve this, we allow nonstrictly monotone changes of variable. This is formalized by the following

Definition 2.5.1. An (*unparameterized*) *curve* is an equivalence class of the minimal equivalence relation satisfying the following: paths $\gamma_1 : I_1 \to X$ and $\gamma_2 : I_2 \to X$ are equivalent whenever there exists a nondecreasing continuous map φ from I_1 onto I_2 such that $\gamma_1 = \gamma_2 \circ \varphi$.

Paths (representatives of an equivalence class) are also called *parameterizations* of the curve and *re-parameterizations* of one another.

The term "curve" is used for both unparameterized curves and their parameterizations (i.e., paths). In most cases, "curve" is formally a synonym for "path". However, the former is more appropriate when parameterization-independent properties are considered.

Remark 2.5.2. Existence of a nonstrictly monotone change of variable is not an equivalence relation by itself (due to the lack of inverse changes of variable). This is why we consider the equivalence relation generated by it. In other words, two paths γ and $\bar{\gamma}$ are equivalent (represent the same curve) if and only if there exists a finite sequence of paths $\gamma_1, \gamma_2, \dots, \gamma_n$ such that $\gamma_1 = \gamma$, $\gamma_n = \bar{\gamma}$ and for every $i = 1, \dots, n-1$ either γ_i is obtained from γ_{i+1}, or γ_{i+1} is obtained from γ_i, by a nondecreasing change of variable. This description can be simplified by means of the following exercise.

Exercise 2.5.3. A path $\gamma : I \to X$ is said to be *never-locally-constant* if there exists no interval $[a, b] \subset I$ such that $a \neq b$ and the restriction of γ to $[a, b]$ is a constant map. Prove that

(a) Every curve admits a never-locally-constant parameterization.

Hint: Formally, for a path $\gamma : I \to X$ introduce an equivalence relation $\sim$ on I: $y \sim y'$ if and only if γ is constant on $[y, y']$. Show that the quotient $J = I/\sim$ is homeomorphic to an interval. Then observe that there is a unique map $\tilde{\gamma} : J \to X$ such that $\gamma = \tilde{\gamma} \circ \pi$ where π is the canonical projection I on J, and that $\tilde{\gamma}$ is never-locally-constant and continuous.

(Loosely speaking, J is obtained from I by cutting off all intervals where a path is constant and gluing together the ends of each of the resulting gaps.)

(b) Two paths $\gamma_1 : I_1 \to X$ and $\gamma_2 : I_2 \to X$ are equivalent if and only if there exist a path $\gamma : J \to X$ and changes of variable $\varphi_1 : I_1 \to J$ and $\varphi_2 : I_2$ $(i = 1, 2)$ such that $\gamma_i = \gamma \circ \varphi_i$.

Hint: Let γ be a never-locally-constant parameterization.

(c) Two paths $\gamma_1 : I_1 \to X$ and $\gamma_2 : I_2 \to X$ are equivalent if and only if there exist an interval J and changes of variable $\varphi_i : J \to I_i$ $(i = 1, 2)$ such that $\gamma_1 \circ \varphi_1 = \gamma_2 \circ \varphi_2$.

It is easy to see that all parameterizations of a curve have equal lengths (check this). We will denote a curve and its parameterization by the same letter.

Definition 2.5.4. A curve $\gamma : [a, b] \to X$ is called *simple* if the pre-image of every point is an interval.

This means that a simple path is allowed to stop at a point for a while, but having left a point it never comes back. Roughly speaking, the image of a simple path is a curve without "self-intersections". The following very easy exercise justifies the correctness of the definition:

Exercise 2.5.5. If one parameterization of a curve is simple, then so are all other parameterizations.

Exercise 2.5.6. If two simple curves have the same image, then they are equivalent up to a change of variable $t \mapsto -t$.

Our next goal is to choose our favorite parameterization for every curve. These parameterizations are analogous to motion with unit speed in physics or differential geometry:

Definition 2.5.7. A parameterization $\gamma : I \to X$ is *natural* if $L(\gamma, t, t') = t - t'$ for all $t, t' \in I$.

Remark 2.5.8. To verify that a parameterization γ is natural, it suffices to check that $L(\gamma, a, t) = t - a$ for a fixed a and all t. This follows from the formula $L(\gamma, t, t') = L(\gamma, a, t') - L(\gamma, a, t)$.

Other names for natural parameterization are *arc-length parameterization* and *parameterization by arc length.* In other words, a parameterization $\gamma(t)$ is natural when

$$\frac{d}{dt}L(\gamma, a, t) = 1,$$

and one also calls it a *unit speed parameterization.* More generally, one says that a parameterization γ is *of constant speed* v if $L(\gamma, t, t') = v(t - t')$ for all t, t'.

The following proposition tells us that every curve admits a natural parameterization.

Proposition 2.5.9. *Every rectifiable curve $\gamma : [a,b] \to X$ can be represented in the form $\gamma = \bar{\gamma} \circ \varphi$ where $\bar{\gamma} : [0, L(\gamma)] \to X$ is a natural parameterization and φ is a nondecreasing continuous map from $[a,b]$ onto $[0, L(\gamma)]$.*

Proof. The idea of construction of φ is trivial: let $\bar{\gamma}(\tau)$ be the point on (the image of) γ such that the length of the interval of γ between its origin and that point is equal to τ. This is formalized as follows. Define $\varphi(t) = L(\gamma, a, t)$ for all $t \in [a,b]$. Then the function φ is nondecreasing and continuous (by the continuity property of length). The set of its values is the interval $[0, L(\gamma)]$. Now for every $\tau \in [0, L(\gamma)]$ pick a $t \in [a,b]$ such that $\varphi(t) = \tau$ and define $\bar{\gamma}(\tau) = \gamma(t)$. This definition does not depend on the choice of t. Indeed, if $\varphi(t) = \varphi(t')$, then $\gamma(t) = \gamma(t')$ because $L(\gamma, t, t') = \varphi(t') - \varphi(t) = 0$.

Thus we have defined a map $\bar{\gamma} : [0, L(\gamma)] \to X$. The relation $\gamma = \bar{\gamma} \circ \varphi$ follows immediately from the definition. It remains to verify that $\bar{\gamma}$ is continuous and parameterized by arc length. For the former, let $\tau_1 = \varphi(t_1)$ and $\tau_2 = \varphi(t_2)$. Then $\bar{\gamma}(\tau_1)$ and $\bar{\gamma}(\tau_2)$ are the endpoints of the path $\gamma|_{[t_1,t_2]}$. The length of this path is $L(\gamma, t_1, t_2) = \varphi(t_2) - \varphi(t_1) = \tau_2 - \tau_1$. Since the distance between endpoints is no greater than the length, we obtain that $d(\bar{\gamma}(\tau_1), \bar{\gamma}(\tau_2)) \leq |\tau_1 - \tau_2|$. This means that $\bar{\gamma}$ is a nonexpanding and hence continuous map. Furthermore $\gamma|_{[t_1,t_2]}$ is a re-parameterization of $\bar{\gamma}|_{[\tau_1,\tau_2]}$ and hence $L(\bar{\gamma}, \tau_1, \tau_2) = L(\gamma, t_1, t_2) = \tau_2 - \tau_1$. Thus $\bar{\gamma}$ is a natural parameterization. □

Exercise 2.5.10. Prove that a natural parameterization of a curve is unique up to a translation $[a,b] \to [a+c, b+c] : t \mapsto t+c$ of the variable.

Exercise 2.5.11. Given $v > 0$, any rectifiable curve admits a parameterization with constant speed v.

The following exercise is a (trivial) corollary of Proposition 2.5.9:

Exercise 2.5.12. If a length space is homeomorphic to a segment, then it is isometric to a segment.

This corollary tells us a remarkable thing: one-dimensional intrinsic geometry is trivial since all intrinsic metrics on a line are locally indistinguishable! We will see that already two-dimensional surfaces are completely different in this respect. By the way, although there is an essentially unique intrinsic metric on $\mathbb{R}$, there are lots of different metrics, which are not intrinsic. For instance, consider $d(x,y) = \sqrt{|x-y|}$ (you may replace the square root by other concave functions to get more examples). The intrinsic distance induced by this metric is infinite everywhere.

2.5.2. Existence of shortest paths. The goal of this section is to prove that a complete locally compact length space is strictly intrinsic, that is, there is a shortest path between every two points.

We begin with the definition of uniform convergence for curves:

Definition 2.5.13. A sequence of curves uniformly converges to a curve γ if they admit parameterizations (with the same domain) that uniformly converge to a parameterization of γ.

The following theorem shows that the space of curves of uniformly bounded lengths in a compact space is compact with respect to the above convergence. This is a version of the Arzela–Ascoli Compactness Theorem in functional analysis.

Theorem 2.5.14 (Arzela–Ascoli Theorem)**.** *In a compact metric space, any sequence of curves with uniformly bounded lengths contains a uniformly converging subsequence.*

Proof. For each γ_i, there is a unique constant speed parameterization by the unit interval $[0,1]$. Uniform boundedness of the lengths of γ_i means that the speeds of these parameterizations are uniformly bounded. In its turn, this implies that for some $C < \infty$

$$d(\gamma_i(t), \gamma_i(t')) \le L(\gamma, t, t') \le C|t - t'| \tag{2.2}$$

for every integer I and all $t, t' \in [0,1]$.

Let $S = \{t_j\}$ be a countable dense subset of $[0,1]$. Using the Cantor diagonal process one can find a subsequence γ_{n_i} of $\{\gamma_i\}$ such that for each $j \in \mathbb{N}$ the sequence $\gamma_{n_i}(t_j)$ converges. We plan to show that the subsequence γ_{n_i} itself converges; to avoid double indices, we may assume (without loss of generality) that this subsequence is γ_i itself: for every t_j the limit $\lim_{i\to\infty} \gamma_i(t_j)$ exists.

To prove that the sequence $\{\gamma_i(t)\}$ converges for every $t \in [0,1]$, we will show that this is a Cauchy sequence.

Given $\varepsilon > 0$, choose $t_j \in S$ such that $|t - t_j| < \varepsilon$ and then $N \in \mathbb{N}$ such that $d(\gamma_i(t_j), \gamma_k(t_j)) < \varepsilon$ for all $i, k > N$. For these choices

$$d(\gamma_i(t), \gamma_k(t)) \le d(\gamma_i(t), \gamma_i(t_j)) + d(\gamma_i(t_j), \gamma_k(t_j)) + d(\gamma_k(t_j), \gamma_k(t)) \le 3\varepsilon.$$

This proves that $\gamma_i(t)$ is a Cauchy sequence and hence we can define $\gamma(t) = \lim_{j\to\infty} \gamma_j(t_i)$.

Passing to the limit in (2.2) we get

$$d(\gamma(t), \gamma(t')) \le C|t - t'|, \tag{2.3}$$

and thus γ is a continuous map.

Let us show that γ_i converges to γ uniformly. Given $\varepsilon > 0$, choose $N > \frac{C}{4\varepsilon}$ and let M be such that $d(\gamma(k/N), \gamma_i(k/N)) < \varepsilon/2$ for all $k = 0, 1, \ldots, N$ and all $i > M$. This choice is possible since γ_i converges to γ point-wise. Combining (2.2) with (2.3), for every $0 \le t \le 1$ and $k/N \le t \le (k+1)/N$ we have

$$d(\gamma(t), \gamma_i(k/N)) \le C|t - k/N| + \varepsilon/2 + C|t - k/N| \le \varepsilon$$

for all $i > M$. This concludes the proof. □

Although we already used the notion of shortest paths, this cornerstone notion deserves a formal definition:

Definition 2.5.15. A curve $\gamma : [a, b] \to X$ is a called a *shortest path* if its length is minimal among the curves with the same endpoints, in other words $L(\gamma_1) \ge L(\gamma)$ for any curve γ_1 connecting $\gamma(a)$ and $\gamma(b)$.

It is trivial that any interval of a shortest path is a shortest path (check this).

Remark 2.5.16. In a length space the above definition can be reformulated as follows: a curve $\gamma : [a, b] \to X$ is a shortest path if and only if its length is equal to the distance between its endpoints: $L(\gamma) = d(\gamma(a), \gamma(b))$. Shortest paths in length spaces are also called *distance minimizers.*

Shortest paths in length spaces possess some nice properties that do not hold in general metric spaces. One such property is the following

Proposition 2.5.17. *If shortest paths γ_i in a length space (X, d) converge to a path γ as $i \to \infty$, then γ is also a shortest path.*

Proof. Since the endpoints of γ_i converge to endpoints of γ and the length of each γ_i is equal to the distance between its points, we conclude that $L(\gamma_i) \to d(x, y)$, where x, y are the endpoints of γ. By the lower semi-continuity of length,

$$L(\gamma) \le \lim_{i\to\infty} L(\gamma_i) = d(x, y).$$

□

Exercise 2.5.18. Give an example showing that a limit of shortest paths in a metric space may fail to be a shortest path.

The example of $\mathbb{R}^2 \setminus \{0\}$ shows that there may be no shortest path between two points. On the other hand, there may be several different shortest paths between the same two points: for instance, consider two antipodal points in a sphere.

Convention. We use the notation $[x, y]$ to denote a shortest path between points x and y. This notation is convenient when either such shortest path is unique or it does not matter which of the shortest paths is considered. This notation is well compatible with our notation for segments in Euclidean space, since the latter are just shortest paths with respect to Euclidean distance.

Proposition 2.5.19. *Let (X, d) be a compact metric space and let $x, y \in X$ be points that can be connected by at least one rectifiable curve. Then there exists a shortest path between x and y.*

Proof. Let $L_{\inf}$ denote the infimum of lengths of rectifiable curves connecting x and y. Then there exists a sequence $\{\gamma_i\}$ of such curves with $L(\gamma_i) \to L_{\inf}$. According to Theorem 2.5.14, the sequence $\{\gamma_i\}$ contains a converging subsequence. Without loss of generality we may assume that $\{\gamma_i\}$ itself converges to a curve γ. Then γ has the same endpoints and, by the lower semi-continuity of length, $L(\gamma) \le \lim L(\gamma_i) = L_{\inf}$. Thus $L(\gamma) = L_{\inf}$. □

Corollary 2.5.20. *Let (X, d) be a boundedly compact metric space. Then for every two points $x, y \in X$ connected by a rectifiable curve there exists a shortest path between x and y.*

Proof. Let L be a length of some rectifiable curve connecting x and y. Observe that this curve, as well as any shorter curve with the same endpoints, is contained in the closed metric ball of radius L centered at x. So it is sufficient to prove the existence of a shortest path only inside this ball. Since the ball is compact, this follows from the previous proposition. □

Definition 2.5.21. A topological space X is called *locally compact* if every point of X has a pre-compact neighborhood.

Proposition 2.5.22. *If (X, d) is a complete locally compact length space, then every closed ball in X is compact (and hence X is boundedly compact).*

Note that in this proposition it is essential that X is a length space. For example, a space where all distances between points equal to 1 is locally compact and complete, but a closed unit ball in such a space is not compact unless the space has a finite cardinality.

Proof of Proposition 2.5.22. Let $x \in X$ be an arbitrary point. Observe that if $\bar{B}_r(x)$ is compact for some r, then $\bar{B}_\rho(x)$ is compact for any $\rho < r$. Define

$$R = \sup\{r > 0 : \bar{B}_r(x) \text{ is compact}\}.$$

Since x has a pre-compact neighborhood, we have $R > 0$. Suppose that $R < \infty$ and denote the ball $\bar{B}_R(x)$ by B.

First let us prove that B is compact. Since B is a closed set in a complete space, it suffices to prove that for any $\varepsilon > 0$ it contains a finite ε-net. We may assume that $\varepsilon < R$. Let B' denote the ball $\bar{B}_{R-\varepsilon/3}(x)$. This ball is compact and hence it contains a finite $(\varepsilon/3)$-net S. Let $y \in B$. Since X is a length space, we have $\operatorname{dist}(y, B') \le \varepsilon/3$. Therefore there exists a point $y' \in B'$ with $d(y, y') < \varepsilon/2$. On the other hand, $\operatorname{dist}(y', S) \le \varepsilon/2$; hence $\operatorname{dist}(y, S) < \varepsilon$. This means that S is an ε-net for B, and we have proven the compactness of B.

Every point $y \in B$ has a pre-compact neighborhood U_y. Pick a finite collection $\{U_y\}_{y \in Y}$ of such neighborhoods that cover B. Their union $U = \bigcup_{y \in Y} U_y$ is a pre-compact neighborhood of B. Using the compactness of B again, we can conclude that there exists a positive $\varepsilon > 0$ such that the ε-neighborhood of B is contained in U. Since X is a length space, the ε-neighborhood of B is the ball $B_{R+\varepsilon}(x)$, and its closure is $\bar{B}_{R+\varepsilon}(x)$. Therefore $\bar{B}_{R+\varepsilon}(x)$ is compact. This contradicts the choice of R.

Thus the assumption $R < \infty$ is wrong; hence $R = \infty$. This means that all balls centered at x are compact. □

Combining Proposition 2.5.22 and the previous Corollary 2.5.20, we obtain the main result of this section.

Theorem 2.5.23. *Let (X, d) be a complete locally compact length space. Then this space is strictly intrinsic: for every $x, y \in X$ such that $d(x, y) < \infty$ there exists a shortest path γ connecting x and y, i.e., a curve $\gamma : [a, b] \to X$ such that $\gamma(a) = x$, $\gamma(b) = y$ and $L(\gamma) = d(x, y)$.*

The example of $\mathbb{R}^2 \setminus \{0\}$ shows that completeness in this theorem is essential. So is local compactness:

Exercise 2.5.24. Give an example of a complete length space (with finite metric) in which there is no shortest path between some points.

However, the condition that X is a length space in the above considerations can be omitted, at the expense of more complicated formulations:

Exercise 2.5.25. Let X be a complete locally compact metric space (not necessarily a length space). Prove that for every two points $x, y \in X$ that can be connected by a rectifiable curve, there exists a shortest path between x and y.

Hint: Prove that for every $R > 0$ the set of points that can be connected to x by a curve of length less than R is pre-compact. In other words, balls of induced length metric are pre-compact in the topology of the original metric (!). To prove this, modify the proof of Proposition 2.5.22.

Exercise 2.5.26. Is it true that a completion of a locally compact length space is locally compact?

2.5.3. Geodesics and the Hopf–Rinow Theorem.

Definition 2.5.27. Let X be a length space. A curve $\gamma\colon I \to X$ is called a *geodesic* if for every $t \in I$ there exists an interval J containing a neighborhood of t in I such that $\gamma|_J$ is a shortest path. In other words, a geodesic is a curve which is locally a distance minimizer (i.e., a shortest path).

The example of the sphere shows that there are geodesics that are not shortest paths: whereas every segment of a great circle on a sphere is a geodesic, it fails to be shortest as soon as it is longer than half of the equator. Although the spherical example suggests that a shortest path between two points should be unique at least locally, this is also not true in general. To see an example, consider the surface of a cube with its intrinsic metric induced by the Euclidean metric of its ambient space. We suggest that the reader shows that any neighborhood of a vertex contains points with multiple shortest paths between them. Another natural conjecture that turns out to be wrong is that a limit of geodesics is a geodesic; give a counterexample to this on the surface of the cube. We will see later that these phenomena are caused by nonsmoothness of the surface and that indeed such things never happen to smooth examples.

It is clear that a shortest path always admits a natural parameterization, and hence so does a geodesic.

Intuitively, a space is noncomplete if a point is missing. One may suspect that this absence of a point can be noticed by moving along a geodesic that would end at this point: the interval for which our motion is well-defined is not closed. The theorem below formalizes this observation.

Before formulating it, we generalize the notion of a shortest path so as to include paths defined on nonclosed intervals. Namely we say that a curve $\gamma\colon I \to X$ (where $I \subset \mathbb{R}$ is an interval and X is a metric space) is a shortest path, or a *minimal geodesic*, if its restriction to any interval $[a, b] \subset I$ is a shortest path in the sense of Definition 2.5.15.

Theorem 2.5.28 (Hopf–Rinow–Cohn-Vossen Theorem). *For a locally compact length space X, the following four assertions are equivalent:*

(i) *X is complete.*

(ii) *X is boundedly compact, i.e., every closed metric ball in X is compact.*

(iii) *Every geodesic $\gamma\colon [0, a) \to X$ can be extended to a continuous path $\overline{\gamma}\colon [0, a] \to X$.*

(iv) *There is a point $p \in X$ such that every shortest path $\gamma\colon [0,a) \to X$ with $\gamma(0) = p$ can be extended to a continuous path $\overline{\gamma}\colon [0,a] \to X$.*

The theorem generalizes the classical Hopf-Rinow theorem which originally was proved only in smooth situation, i.e., for Riemannian manifolds.

Remark 2.5.29. By Theorem 2.5.23, these conditions imply that every two points a, b can be connected by a shortest path.

Proof of the theorem. Implications (ii) $\Longrightarrow$ (i) $\Longrightarrow$ (iii) $\Longrightarrow$ (iv) are left as easy exercises. We will prove that (iv) implies (ii). The proof uses the same general scheme as the one of Proposition 2.5.22. The details are slightly more complicated because we cannot utilize completeness at this point. We can use only the property (iv) and this requires a more delicate argument.

Since X is locally compact, sufficiently small closed balls $\overline{B}_r(p)$ are compact. Reasoning by contradiction (that is, assuming that there are noncompact closed balls), define

$$R = \sup\{r : \overline{B}_r(p) \text{ is a compact set }\}$$

and assume that $R < \infty$. The argument consists of two steps.

1. First, we prove that the open ball $B_R(p)$ is pre-compact. To do this, it suffices to show that every sequence $\{x_i\}$ in this ball contains a converging subsequence (whose limit does not necessarily belong to this ball). Set $r_i = d(p, x_i)$. One may assume that $r_i \to R$ as $i \to \infty$; otherwise a subsequence of $\{x_i\}$ is contained in a ball $\overline{B}_r(p)$ for some $r < R$, and then there is a converging subsequence because this smaller ball is compact (by the choice of R).

Let $\gamma_i\colon [0, r_i] \to X$ be a (naturally parameterized) shortest path connection p to x_i. Such a shortest path exists because x_i belongs to a compact ball centered at p (see the proof of Corollary 2.5.20). We can choose a subsequence of $\{\gamma_i\}$ such that the restrictions of the paths to $[0, r_1]$ converge (along this subsequence). From this subsequence, we choose a further subsequence of paths whose restrictions to $[0, r_2]$ converge, and so on. Then the Cantor diagonal procedure (that is, picking the nth element from the nth subsequence for $n = 1, 2, \dots$) yields a sequence $\{\gamma_{i_n}\}$ such that for every $t \in [0, R)$ the sequence $\{\gamma_{i_n}(t)\}$ converges in X. (More precisely, points $\gamma_{i_n}(t)$ are well-defined for all large enough n and form a converging sequence.)

Define $\gamma(t) = \lim \gamma_{i_n}(t)$; then $\gamma\colon [0, R) \to X$ is a nonexpanding map and a shortest path (see Proposition 2.5.17). By the assertion (iv), there is a continuous extension $\overline{\gamma}\colon [0, R] \to X$. One easily sees (exercise!) that the points x_{i_n} (i.e., the endpoints of our converging curves $\{\gamma_{i_n}\}$) converge to $\overline{\gamma}(R)$.

2. Since the open ball $B_R(p)$ is pre-compact, the closed ball $\overline{B}_R(p)$ is compact. (Recall that a closed ball in a length space is the closure of the respective open ball.) Now we show that a ball $B_{R+\varepsilon}(p)$ is pre-compact for some $\varepsilon > 0$. Since X is locally compact, for every $x \in \overline{B}_R(p)$ there is an $r(x) > 0$ such that the ball $B_{r(x)}(x)$ is pre-compact. Choose a finite subcover $\overline{B}_{r(x_i)}(x_i)$ out of the cover of $\overline{B}_R(x)$ by these balls. The union of these balls is pre-compact and contains the ball $B_{R+\varepsilon}(p)$ for $\varepsilon = \min r_i > 0$. This contradicts the choice of R. □

2.6. Length and Hausdorff Measure

Recall that the length of a curve is independent of the parameterization. This fact suggests that the length of a simple curve can be recovered from its image in the space, i.e., the set of points it passes through. In this section we show that the length of a path actually equals the one-dimensional Hausdorff measure of the image. We assume that the normalization constant $C(1)$ for Hausdorff measure is 1 (for a definition and elementary properties of Hausdorff measure see Section 1.7).

Lemma 2.6.1. *If X is a connected metric space, then $\mu_1(X) \geq \operatorname{diam} X$.*

Proof. 1. A general observation: in the definition of Hausdorff measure we can restrict ourselves to coverings by *open* sets S_i. Indeed, an arbitrary covering $\{S_i\}$ can be replaced by an open covering $\{S_i'\}$ where

$$S_i' = U_{\delta/2^i}(S_i) := \{x \in X : \operatorname{dist}(x, S_i) < \delta/2^i\}$$

for a small $\delta > 0$. Then $\operatorname{diam} S_i' \leq \operatorname{diam} S_i + 2\delta/2^i$ and hence $w_1(\{S_i'\}) \leq w_1(\{S_i\}) + 2\delta$. Since δ is arbitrary, it follows that the measure can be approximated by 1-weights of open coverings. (*Exercise*: extend the argument to work for measures of all dimensions.)

2. Let X be a connected topological space and $\{S_i\}$ an open covering of X. Then for every two points $x, y \in X$ there exists a finite sequence $S_{i_1}, \dots, S_{i_n}$ of different sets such that $x \in S_{i_1}$, $y \in S_{i_n}$, and $S_{i_k} \cap S_{i_{k+1}} \neq \emptyset$ for all k, $1 \leq k \leq n-1$. To prove this, fix a point $x \in X$ and consider the set Y of all points $y \in X$ for which such a sequence exists. It is clear that for every set S_i one has either $S_i \subset Y$ or $S_i \subset X \setminus Y$. Therefore both Y and $X \setminus Y$ are open sets. Since X is connected, it follows that $Y = X$ and therefore any point $y \in Y$ is "accessible" by a sequence of sets as described above.

3. Let $\{S_i\}$ be an open covering of X, x and y two points of X, and $S_{i_1}, \dots, S_{i_n}$ a sequence from Step 2. For $k = 1, \dots, n-1$ pick a point $x_k \in S_{i_k} \cap S_{i_{k+1}}$, and define $x_0 = x$, $x_n = y$. Then $|x_{k-1}x_k| \leq \operatorname{diam} S_{i_k}$ for

all $k = 1, \dots, n$ because both x_{k-1} and x_k are in S_{i_k}. Hence

$$\sum \operatorname{diam} S_i \geq \sum_{k=1}^{n} \operatorname{diam} S_{i_k} \geq \sum_{k=1}^{n} |x_{k-1}x_k| \geq |x_0x_n| = |xy|.$$

Due to the observation made in step 1, it follows that $\mu_1(X) \geq |xy|$. Since x and y are arbitrary points, this means that $\mu_1(X) \geq \operatorname{diam} X$. □

Theorem 2.6.2. *Let X be a metric space, $\gamma\ [a,b] \to X$ a simple curve. Then $L(\gamma) = \mu_1(\gamma([a,b]))$.*

Proof. Let $S = \gamma([a,b])$ and $L = L(\gamma)$. Without loss of generality assume that γ is parameterized by arc length: $\gamma : [0,L] \to X$. Then for every natural number N, S is covered by N intervals of γ each of length L/N, namely by the sets $\gamma([i\frac{L}{N}, (i+1)\frac{L}{N}])$, $i = 0, 1, \dots, N-1$. The diameters of these sets are no greater than L/N. Hence the sum of the diameters is no greater than L, whereas the diameters themselves approach 0 as $N \to \infty$. This shows that $\mu_1(S) \leq L(\gamma)$.

On the other hand, for a partition $a = t_0 \leq t_1 \leq \cdots \leq t_n = b$ of $[a,b]$, let $S_i = \gamma([t_i, t_{i+1}])$ for $i = 1, \dots, n-1$. The sets S_i are disjoint modulo a finite number of points $\gamma(t_i)$. Since the one-dimensional measure of a single point is obviously zero, one has $\mu_1(S) = \sum \mu_1(S_i)$. By Lemma 2.6.1, $\mu_1(S_i) \geq \operatorname{diam} S_i \geq |\gamma(t_i)\gamma(t_{i+1})|$, and hence

$$\mu_1(S) = \sum \mu_1(S_i) \geq \sum |\gamma(t_i)\gamma(t_{i+1})|$$

for any partition $\{t_i\}$. This implies that $\mu_1(S) \geq L(\gamma)$. □

Remark 2.6.3. If γ is not simple, the same argument shows that $L(\gamma) \geq \mu_1(\gamma([a,b]))$.

Exercise 2.6.4. Prove that, for any curve $\gamma : [a,b] \to X$,

$$\ell(\gamma) = \sum_{k \in \mathbb{N} \cup \{\infty\}} k \cdot \mu_1(\{x \in X : \#(\gamma^{-1}(x)) = k\})$$

where $\#$ denotes the cardinality and $\infty \cdot 0 = 0$.

Hint: The right-hand part is the integral of the function $x \mapsto \#(\gamma^{-1}(x))$ with respect to the measure μ_1. Combine Levy's limit theorem for integrals with inequalities "diam $\leq \mu_1 \leq \ell$" applied to small pieces of γ.

2.7. Length and Lipschitz Speed

This section is rather technical. We will generalize the formula "length equals the integral of speed" which is well known and trivial for smooth curves in $\mathbb{R}^n$ (cf. Exercise 2.3.3). The last theorem of this section requires the knowledge of Lebesgue integration, and we will use without proofs some facts from measure theory.

Definition 2.7.1. Let (X, d) be a metric space and $\gamma : I \to X$ a curve. The *speed* of γ at $t \in I$, denoted by $v_\gamma(t)$, is defined by

$$v_\gamma(t) := \lim_{\varepsilon \to 0} \frac{d(\gamma(t), \gamma(t+\varepsilon))}{|\varepsilon|}$$

if the limit exists.

Exercise 2.7.2. Let γ be a differentiable curve in $\mathbb{R}^n$ (more generally, in a normed vector space). Prove that $v_\gamma(t)$ exists for all t and $v_\gamma(t) = |\gamma'(t)|$.

Exercise 2.7.3. Let X be a metric space and $\gamma : [a, b] \to X$ a curve. Suppose that $v_\gamma(t)$ exists for all $t \in [a, b]$ and is continuous in t. Prove that

$$L(\gamma) = \int_a^b v_\gamma(t)\, dt$$

(compare Exercise 2.3.3).

Of course, the speed of a curve may not exist. However, it exists almost everywhere (i.e., except a set of zero measure) for a wide class of curves, namely for every Lipschitz curve. (Note that every rectifiable curve admits Lipschitz parameterization. For example, all natural parameterizations are Lipschitz with unit Lipschitz constant.) Moreover, the length of a Lipschitz curve equals the (Lebesgue) integral of its speed.

As a first step, we prove the following

Theorem 2.7.4. *Let (X, d) be a metric space, $\gamma : [a, b] \to X$ a rectifiable curve. Then for almost all $t \in [a, b]$ (i.e., for all t except a set of zero measure) the following holds: either*

$$\liminf_{\varepsilon, \varepsilon' \to 0+} \frac{L(\gamma|_{[t-\varepsilon, t+\varepsilon']})}{\varepsilon + \varepsilon'} = 0,$$

or

$$\lim_{\varepsilon, \varepsilon' \to 0+} \frac{d(\gamma(t-\varepsilon), \gamma(t+\varepsilon'))}{L(\gamma|_{[t-\varepsilon, t+\varepsilon']})} = 1.$$

One can let ε or ε' in the above formulas be zero. This gives the following

Corollary 2.7.5. *If γ is as in Theorem 2.7.4, then for almost all $t \in [a, b]$ either*

$$\liminf_{\varepsilon \to 0} \frac{L(\gamma|_{[t, t+\varepsilon]})}{|\varepsilon|} = 0,$$

or

$$\lim_{\varepsilon \to 0} \frac{d(\gamma(t), \gamma(t+\varepsilon))}{L(\gamma|_{[t, t+\varepsilon]})} = 1$$

(if $\varepsilon < 0$, the interval $[t, t+\varepsilon]$ in the denominator of the last formula should be interpreted as $[t+\varepsilon, t]$).

Proof of Theorem 2.7.4. Suppose the contrary. For every $\alpha > 0$ let Z_α denote the set of all $t \in [a, b]$ such that

$$\liminf_{\varepsilon,\varepsilon' \to 0+} \frac{L(\gamma|_{[t-\varepsilon,t+\varepsilon']})}{\varepsilon + \varepsilon'} > \alpha$$

and

$$\liminf_{\varepsilon,\varepsilon' \to 0+} \frac{d(\gamma(t-\varepsilon), \gamma(t+\varepsilon'))}{L(\gamma|_{[t-\varepsilon,t+\varepsilon']})} < 1 - \alpha.$$

Then $\mu_1(Z_\alpha) > 0$ for all sufficiently small α. Indeed, otherwise the set $Z_0 = \bigcup_{\alpha>0} Z_\alpha = \bigcup_{n\in\mathbb{N}} Z_{1/n}$ would have zero measure, and this is equivalent to the statement of the theorem. Fix an $\alpha > 0$ such that $\mu_1(Z_\alpha) > 0$. For brevity, we denote $Z = Z_\alpha$ and $\mu = \mu_1(Z)$. Choose ε_0 so small that for any partition $\{y_i\}_{i=1}^N$ ($a = y_0 \le y_1 \le \cdots \le y_N = b$) of $[a, b]$ such that $\max_i(y_i - y_{i-1}) < \varepsilon_0$, one has

$$L(\gamma) - \sum_{i=1}^N d(\gamma(y_{i-1}), \gamma(y_i)) < \mu\alpha^2/2.$$

Such an ε_0 exists by Exercise 2.3.2. Consider the set $\mathfrak{B}$ of all intervals of the form $[t-\varepsilon, t+\varepsilon']$ such that $t \in Z$, $\varepsilon + \varepsilon' < \varepsilon_0$, $L(\gamma|_{[t-\varepsilon,t+\varepsilon']}) > \alpha(\varepsilon + \varepsilon')$ and

$$d(\gamma(t-\varepsilon), \gamma(t+\varepsilon')) < (1-\alpha)L(\gamma|_{[t-\varepsilon,t+\varepsilon']}).$$

By the definition of $Z = Z_\alpha$, every point $t \in Z$ is contained in arbitrarily short elements of $\mathfrak{B}$. Applying Vitali's covering theorem (Theorem 1.7.14) we can extract from $\mathfrak{B}$ a countable collection $\{[t_i - \varepsilon_i, t_i + \varepsilon_i']\}_{i=1}^\infty$ of disjoint intervals that covers Z up to a set of zero measure. In particular,

$$\sum_{i=1}^\infty (\varepsilon_i + \varepsilon_i') = \mu_1\big(\bigcup [t_i - \varepsilon_i, t_i + \varepsilon_i']\big) \ge \mu_1(Z) = \mu.$$

Hence for a sufficiently large M,

$$\sum_{i=1}^M (\varepsilon_i + \varepsilon_i') > \mu/2.$$

Since the intervals $\{[t_i - \varepsilon_i, t_i + \varepsilon_i']\}_{i=1}^M$ are disjoint, they can be included in a partition $\{y_j\}_{j=1}^N$ all whose intervals are shorter than ε_0. We denote $L_j = L(\gamma|_{[y_{j-1},y_j]})$ and $d_j = d(\gamma(y_{j-1}), \gamma(y_j))$. By the choice of ε_0, we have

$$\sum_{j=1}^N (L_j - d_j) = L(\gamma) - \sum_{j=1}^N d_j < \mu\alpha^2/2.$$

In the left-hand sum above all terms are nonnegative and those for which $[y_{j-1}, y_j] \in \mathfrak{B}$ (i.e., $y_{j-1} = t_i - \varepsilon_i$ and $y_j = t_i + \varepsilon_i'$ for some i) satisfy

$$L_j - d_j > \alpha L_j > \alpha^2(y_{j-1} - y_j) = \alpha^2(\varepsilon_i + \varepsilon_i').$$

Therefore

$$\sum_{j=1}^{N}(L_j - d_j) \geq \alpha^2 \sum_{i=1}^{M}(\varepsilon_i + \varepsilon_i') > \mu\alpha^2/2.$$

This contradiction proves the theorem. □

Theorem 2.7.6. *Let X be a metric space; $\gamma : [a,b] \to X$ is a Lipschitz curve. Then the speed $v_\gamma(t)$ exists for almost all $t \in [a,b]$ and $L(\gamma) = \int_a^b v_\gamma(t)\,dt$ where $\int$ is the Lebesgue integral.*

Proof. We need the following fact ([**Fe**], Theorem 2.9.19): if $f : [a,b] \to \mathbb{R}$ is a Lipschitz function, then the derivative $f'(t)$ exists for almost all $t \in [a,b]$ and $\int_a^b f'(t)\,dt = f(b) - f(a)$. (*Remark*: the proof of this fact is based on the same ideas as the above proof of Theorem 2.7.4 though it is more complicated.)

Define $f(t) = L(\gamma|_{[a,t]})$ for $t \in [a,b]$. Then f is a Lipschitz function and hence is differentiable almost everywhere. We rewrite $f'(t)$ as follows:

$$f'(t) = \lim_{\varepsilon\to 0} \frac{L(\gamma|_{[t,t+\varepsilon]})}{|\varepsilon|} = \lim_{\varepsilon\to 0} \frac{L(\gamma|_{[t,t+\varepsilon]})}{d(\gamma(t),\gamma(t+\varepsilon))} \cdot \frac{d(\gamma(t),\gamma(t+\varepsilon))}{|\varepsilon|}.$$

By Corollary 2.7.5, for almost all $t \in [a,b]$ either $f'(t) = 0$ or the first term in the last product goes to 1 as $\varepsilon \to 0$. In the first case we have

$$v_\gamma(t) = \lim_{\varepsilon\to 0} \frac{d(\gamma(t),\gamma(t+\varepsilon))}{|\varepsilon|} = 0$$

because $d(\gamma(t),\gamma(t+\varepsilon)) \leq L(\gamma|_{[t,t+\varepsilon]})$. In the second case, it follows that

$$v_\gamma(t) = \lim_{\varepsilon\to 0+} \frac{d(\gamma(t),\gamma(t+\varepsilon))}{|\varepsilon|} = f'(t).$$

Thus $v_\gamma(t)$ exists and equals $f'(t)$ in both cases. The theorem follows. □

Exercise 2.7.7. Give an example of a nonconstant curve γ in $\mathbb{R}^2$ for which $v_\gamma = 0$ almost everywhere.

Exercise 2.7.8. Let X be a metric space and $\gamma : I \to X$ a curve. For a $t \in [a,b]$ define

$$\operatorname{dil}_t(\gamma) = \limsup_{\varepsilon\to 0+} \operatorname{dil}(\gamma|_{[t-\varepsilon,t+\varepsilon]}).$$

(1) Prove that $\operatorname{dil}_t(\gamma) \geq v_\gamma(t)$ whenever $v_\gamma(t)$ is defined.

(2) Prove that, if $v_\gamma(t)$ is defined for all t and is continuous in t, then $\operatorname{dil}_t(\gamma) = v_\gamma(t)$ for all t.

(3) Give an example where γ is Lipschitz but $\int_a^b \operatorname{dil}_t(\gamma) \neq L(\gamma)$.