

1 Metric spaces

Definition 1 A metric space consists of a pair (X, d) , where X is a set and $d : X \times X \rightarrow \mathbb{R}$ is a function, called the metric or distance function, such that the following hold for all $x, y, z \in X$

1. (Symmetry) $d(x, y) = d(y, x)$
2. (Positive Definiteness) $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$
3. (Triangle Inequality) $d(x, z) \leq d(x, y) + d(y, z)$

We will refer to $d(x, y)$ as the *distance* between x and y . The letter d is the default for metric spaces; often we will simply state “let X be a metric space” rather than specifying (X, d) . In addition, we will sometimes even use d to represent distances in different metric spaces in the same discussion; when there is danger of confusion we will be more careful, e.g. using d_X for the distance on X and d_Y for the distance on Y .

The definition of a “metric” captures the most important and basic elements of what a “distance” should be. The distance from x to y should be the same as that from y to x ; different points should be at positive distance from one another, but a point should be at distance 0 from itself, and travelling between two points via an arbitrary third point should not be shorter than the distance between the original two. The most familiar metric space is Euclidean space of dimension n , which we will denote by \mathbb{R}^n , with the standard formula for the distance:

$$d((x_1, \dots, x_n), (y_1, \dots, y_n)) = ((x_1 - y_1)^2 + \dots + (x_n - y_n)^2)^{\frac{1}{2}}. \quad (1)$$

Exercise 2 Let X be any set. For all $x, y \in X$, define $d(x, y) = 1$ if x and y are distinct (i.e. different), and $d(x, x) = 0$. It is easy to check that d is a metric, called the trivial or discrete metric. This metric is of little interest except as a very simple example and counterexample to certain statements. For example, there are no “midpoints” in a discrete metric space.

Recall that the Euclidean norm in \mathbb{R}^n is defined, for $v = (x_1, \dots, x_n)$ by

$$\|v\| := (x_1^2 + \dots + x_n^2)^{\frac{1}{2}}.$$

Note that for any $v, w \in \mathbb{R}^n$,

$$d(v, w) = \|v - w\|.$$

In elementary linear algebra one shows that the Euclidean norm satisfies the three properties in the following definition:

Definition 3 A norm on a real vector space V consists of a real-valued function $\|\cdot\|$ defined on V that satisfies the following properties for $v, w \in V$ and $t \in \mathbb{R}$:

1. $\|v\| \geq 0$ and $\|v\| = 0$ if and only if $v = 0$.
2. $\|tv\| = |t| \|v\|$.
3. $\|v + w\| \leq \|v\| + \|w\|$.

Proposition 4 *Let V be a vector space with norm $\|\cdot\|$. The function $d(v, w) := \|v - w\|$ defines a metric on V .*

Proof. Let $v, w \in V$. By definition of the norm, $d(v, w) = \|v - w\| \geq 0$ and

$$0 = d(v, w) \Leftrightarrow \|v - w\| = 0 \Leftrightarrow v - w = 0 \Leftrightarrow v = w.$$

This proves positive definiteness (Property 1 in the definition of the norm is also referred to as “positive definiteness”). Symmetry uses the second property above:

$$d(v, w) = \|v - w\| = |-1| \|w - v\| = \|w - v\| = d(w, v).$$

The triangle inequality follows from the third property of the norm (also called the “triangle inequality”):

$$\begin{aligned} d(v, w) &= \|v - w\| = \|(v - u) + (u - w)\| \\ &\leq \|(v - u)\| + \|(u - w)\| = d(v, u) + d(u, w). \end{aligned}$$

■

Corollary 5 *The function defined by Formula 1 defines a metric on \mathbb{R}^n .*

Remark 6 *The above proof only requires a weaker statement than the second property, namely that $\|-v\| = \|v\|$.*

The metric that comes from a norm has the following important property:

Definition 7 *A metric on a vector space V is said to be invariant if for any $u, v, w \in V$, $d(v, w) = d(v + u, w + u)$.*

Exercise 8 *Prove that the metric of a norm is invariant. If d is a metric from a norm, define a new invariant metric \tilde{d} on V by $\tilde{d}(v, w) = \min\{d(v, w), 1\}$ (prove it is a metric).*

The term “invariant” hence means that the distance is unchanged when you translate by a fixed vector u . Functions that preserve distance play an extremely important role in geometry.

Definition 9 *Let X and Y be metric spaces. A function $f : X \rightarrow Y$ is called an isometry if f is onto (also known as surjective) and for every $x, y \in X$, $d_Y(f(x), f(y)) = d_X(x, y)$.*

Exercise 10 Prove that an isometry is one-to-one (also known as injective), hence by definition a bijection. Show that the identity mapping $I_X : X \rightarrow X$ defined by $I_X(x) = x$ is an isometry. Prove that the composition of two isometries is an isometry and the inverse of an isometry is an isometry.

Exercise 11 Show that every bijection between trivial metric spaces is an isometry.

If X and Y are different spaces then the existence of an isometry between them simply means that they are, from the standpoint of metric spaces, actually the same space—just with different names. Every metric space X has at least one isometry from itself to itself, namely the identity function. Metric spaces that have many self-isometries are very important; they have a lot of geometric structure. For example, any distance that comes from a norm has many isometries:

Exercise 12 Show that if V is a vector space with norm then the function $L_u : V \rightarrow V$ defined by $L_u(v) := u + v$ is always an isometry. For which $t \in \mathbb{R}$ is the function $S_t(v) := tv$ an isometry?

It is a fact from linear algebra that every isometry of Euclidean space consists of an orthogonal linear mapping composed with a translation L_u . The orthogonal mappings provide rotations and changes in orientation while the translations “shift” the space in some direction. These isometries are very important in physics, for example.

2 Groups

Definition 13 A group (G, \cdot, e) consists of a set G , a function $\cdot : G \times G \rightarrow G$ (called the product) and a distinguished element e (called the identity) satisfying the following axioms for all $x, y, z \in G$:

1. *Associative Law:* $(xy)z = x(yz)$
2. *Identity:* $ex = xe = x$
3. *Inverses:* for each $x \in G$ there exists an $x^{-1} \in G$ such that $xx^{-1} = x^{-1}x = e$.

Definition 14 Let G be a group. If G has $m < \infty$ elements, then the number m is called the order of G .

We will often simply refer to a group as G when there can be no confusion about the notation being used. A group G is called *abelian* (or commutative) if for all $x, y \in G$, $xy = yx$. We will often use sum notation for abelian groups: $(G, +, 0)$, where the inverse of an element x is denoted by $-x$.

The most basic examples of groups are the integers \mathbb{Z} , the rationals \mathbb{Q} , and the reals \mathbb{R} , with ordinary addition as the group operation. More generally, any

vector space is an abelian group with respect to the operation of $+$. In fact, a vector space may be defined as an abelian group with an additional operation of scalar multiplication satisfying additional axioms.

Definition 15 Let G be a group. A subgroup of G is a nonempty subset H of G such that for all $x, y \in H$, $x^{-1} \in H$ and $xy \in H$.

Exercise 16 Show that a subset H of a group G is a subgroup if and only if H is itself a group with respect to the group operations in G . In particular $e \in H$.

Example 17 \mathbb{Z} is a subgroup of \mathbb{Q} , which is in turn a subgroup of the abelian group $(\mathbb{R}, +, 0)$. Note that in each case \mathbb{Z} is a subgroup but not a vector subspace of the vector spaces \mathbb{Q} and \mathbb{R} . The set $2\mathbb{Z}$ of even integers is a subgroup of the integers.

Example 18 Letting \mathbb{R}^\times denote the set of non-zero real numbers, $(\mathbb{R}^\times, \times, 1)$ is a group. The set of non-zero rational numbers is a subgroup of \mathbb{R}^\times , but the set of non-zero integers is not (there are no multiplicative inverses). The set \mathbb{R}^+ consisting of positive real numbers is a subgroup of \mathbb{R}^\times .

Example 19 The set $M(n)$ of all $n \times n$ matrices forms an abelian group with respect to matrix addition.

Example 20 The set $GL(n, \mathbb{R})$ of all real $n \times n$ invertible matrices is a group with respect to matrix multiplication, where the identity I_n is the matrix with 1's on the diagonal and 0's elsewhere. Many important examples of groups are subgroups of $GL(n, \mathbb{R})$ for some n .

Example 21 Consider all matrices of the form $R_\theta := \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$, where θ is a real number. One can easily check that the set of all such matrices forms an abelian subgroup of $GL(2, \mathbb{R})$ with $R_{\theta_1} R_{\theta_2} = R_{\theta_1 + \theta_2}$. For the time being we will refer to this group as the "rotation group," denoted by C . The geometric motivation for this term is the fact that applying R_θ to any vector in the plane rotates that vector counterclockwise by an angle of θ .

Example 22 The two matrices $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$ together form a subgroup of C of order 2. More generally, the sets

$$\mathbb{Z}_n := \left\{ \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right], \left[\begin{array}{cc} \cos \frac{2\pi}{n} & -\sin \frac{2\pi}{n} \\ \sin \frac{2\pi}{n} & \cos \frac{2\pi}{n} \end{array} \right], \dots, \left[\begin{array}{cc} \cos \frac{2\pi(n-1)}{n} & -\sin \frac{2\pi(n-1)}{n} \\ \sin \frac{2\pi(n-1)}{n} & \cos \frac{2\pi(n-1)}{n} \end{array} \right] \right\}$$

form subgroups of the rotation group having n elements.

Example 23 Another way to realize the groups \mathbb{Z}_n is to consider the set of integers mod n using familiar modular arithmetic. Recall that the integers mod n consists of equivalence classes $\{\overline{0}, \overline{1}, \dots, \overline{n-1}\}$, where integers m_1 and m_2 are

equivalent if $m_1 - m_2$ is divisible by n . The practical way to determine \overline{m} is to subtract the largest multiple of n that is $\leq m$, i.e. \overline{m} is the remainder or “residue” that m differs from a multiple of n . For example, $\overline{47} = \overline{3} \pmod{4}$, since 44 is the largest multiple of 4 that is ≤ 47 . These equivalence classes are added mod n using the formula $\overline{m_1} + \overline{m_2} = \overline{m_1 + m_2}$ and it is easy to check that they form a group of order n . This group is “the same” as \mathbb{Z}_n in a sense we will now make precise.

Definition 24 Let G and H be groups. A function $\phi : G \rightarrow H$ is called a homomorphism if for all $x, y \in G$, $\phi(xy) = \phi(x)\phi(y)$. A homomorphism ϕ is called an isomorphism if ϕ is also a bijection.

We will discuss homomorphisms in more detail later, but for now note that an isomorphism plays the same role for groups that an isometry does for metric spaces. That is, isomorphic groups are essentially the same, except possibly in name.

Exercise 25 Show that the group \mathbb{Z}_n and the group defined in the preceding example are isomorphic.

Example 26 Let $I(X)$ denote the set of all isometries of a metric space X . We will show that $I(X)$ is a group with respect to the operation of composition of functions. It was shown in an earlier exercise that the composition of two isometries is an isometry, so isometries are closed with respect to this operation. The identity function is an identity in this group. In fact, if $g : X \rightarrow X$ is an isometry then for any $x \in X$, $I_X \circ g(x) = I_X(g(x)) = g(x)$ and so as functions $I_X \circ g = g$. A similar argument shows that $g \circ I_X = g$. By definition, $g \circ g^{-1} = g^{-1} \circ g = I_X$, and since it was shown earlier that the inverse of an isometry is an isometry, g^{-1} is in fact the inverse of g with respect to the group operation. The associative law follows quickly from the definition of composition of functions. Note that the isometry group of a metric space may consist of the identity function alone. In contrast, a normed vector space V is homogeneous in the sense that for every $v, w \in V$ there is an isometry g such that $g(v) = w$ (e.g. the translation by the vector $w - v$).

Example 27 Note that the group of all bijections of a set X is also a group with respect to composition. When X happens to be a metric space, $I(X)$ is a subgroup of the group of all bijections. When X is itself a group, the set of all isomorphisms from X to itself is a subgroup of the group of all bijections, since as it is easy to show, the composition of two isomorphisms and the inverse of an isomorphism are again isomorphisms.

Exercise 28 Consider all matrices of the form
$$\begin{bmatrix} 1 & x & y \\ 0 & 1 & z \\ 0 & 0 & 1 \end{bmatrix},$$
 where $x, y, z \in \mathbb{R}$.

Show that these matrices form a non-abelian subgroup of $GL(3, \mathbb{R})$, called the (3-dimensional real) Heisenberg group $H(\mathbb{R})$. Verify that if $x, y, z \in \mathbb{Z}$ then all

the entries of $\begin{bmatrix} 1 & x & y \\ 0 & 1 & z \\ 0 & 0 & 1 \end{bmatrix}^{-1}$ are again integers; the integral Heisenberg group $H(\mathbb{Z})$ is the subgroup of $H(\mathbb{R})$ with $x, y, z \in \mathbb{Z}$.

3 Curves in Metric Spaces

In elementary calculus one learns how to measure the length of a differentiable curve. Let $\alpha(t) = (x(t), y(t), z(t))$ be a curve in Euclidean space \mathbb{R}^3 . The velocity vector of the curve is $\alpha'(t) := (x'(t), y'(t), z'(t))$. This vector is represented as a vector based at the point $\alpha(t)$; it points in the direction of motion and its length $\|\alpha'(t)\|$ is defined to be the speed of motion along the curve. An infinitesimal distance is equal to the speed times an infinitesimal time: $\|\alpha'(t)\| dt$; to find the length of the curve on the interval $[a, b]$ we simply integrate this quantity: $L(\alpha) = \int_a^b \|\alpha'(t)\| dt$.

One way to alter the geometry of a Euclidean space is to change how we measure the lengths of tangent vectors at each point. For example, to obtain hyperbolic geometry, we modify the length formula by dividing by the y -coordinate of the point—that is, for $\alpha(t) = (x(t), y(t))$ we redefine

$$L(\alpha) := \int_a^b \frac{\|\alpha'(t)\|}{y(t)} dt = \int_a^b \frac{\|\alpha'\|}{y} dt$$

Thus the same curve “shifted up” becomes shorter, and it becomes more efficient to curve upwards when joining any two points. We will be more precise about this later, but for the time being it should seem reasonable that the shortest path between two points is generally not a straight line as it is in Euclidean space.

In metric spaces in general there is not even a way to define what differentiable means! We need another method of measuring lengths of curves that depends only on the distance function but gives us the same answer when we happen to apply it to a curve in Euclidean space with its usual metric. In fact, we even need a way to say what a curve is—and the key to this is continuity, not differentiability. In calculus one learns that a function f between Euclidean spaces is continuous at a point p if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $d(x, p) < \delta$ then $d(f(x), f(p)) < \varepsilon$. We can use the exact same definition for a function $f : X \rightarrow Y$ between metric spaces X and Y . Moreover, we say that f is continuous if f is continuous at every $p \in X$.

Exercise 29 Prove that every isometry is continuous.

Definition 30 Let X be a metric space. A curve in X is a continuous function $\alpha : I \rightarrow X$, where I is an interval in \mathbb{R} with the usual metric as a subset of \mathbb{R} .

Continuity is an extremely important concept, but for the present, intuitively speaking, a continuous curve is a mapping that carries the interval into the

metric space without breaking it. Continuity is not as stringent a requirement as might be supposed at first. For example, there are curves, discovered by Peano, that carry \mathbb{R} continuously *onto* the Euclidean plane. That is, continuous functions can actually raise the dimension of a space! Even so, there is a simple way to measure the length of any curve, although that length might be infinity.

Let $\alpha : [a, b] \rightarrow X$ be a curve and consider a partition $a = t_0 < t_1 < \cdots < t_n = b$ of $[a, b]$. The sum associated with this partition is $\sum_{i=1}^n d(\alpha(t_i), \alpha(t_{i-1}))$.

The simplest possible partition consists of the two points a and b , and the associated sum is simply $d(\alpha(a), \alpha(b))$. If we add a new point to the partition $a = t_0 < t_1 < t_2 = b$ then the associated sum is $d(\alpha(t_0), \alpha(t_1)) + d(\alpha(t_1), \alpha(t_2))$, which by the triangle inequality is greater than or equal to $d(\alpha(a), \alpha(b))$. In fact, any time we further subdivide a given partition the associated sum of the new partition is at least as large. Either the set of all associated sums of partitions is bounded above, or it is not. If it is not bounded above, we say that the curve has infinite length and write $L(\alpha) = \infty$. If it is bounded above, we define $L(\alpha)$ to be the supremum (or least upper bound) of all sums of partitions. Since these sums get larger as the partitions are subdivided, it is reasonable to say that $L(\alpha)$ is the “limit” of the sums of finer and finer partitions. Note that the length of a curve is greater than or equal to the distance between its endpoints, and in particular is nonnegative.

Exercise 31 *Prove that the length (measured using partitions as above, not derivatives!) of a straight line segment $\alpha(t) = p + tv$ in Euclidean space is equal to the distance between its endpoints.*

Exercise 32 *Prove that if $\alpha : [a, b] \rightarrow X$ is a curve in a metric space then $L(\alpha) = 0$ if and only if α is a constant curve (that is, for some $p \in X$, $\alpha(t) = p$ for all t).*

There are several facts concerning lengths of curves that we will need. They may be proved using basic methods of analysis, but these arguments would take us too far from our subject of interest, given the limited time we have. Therefore we will simply state them here, for a curve $\alpha : [a, b] \rightarrow X$ in a metric space:

1. Length is additive. That is, if β is the restriction of α to $[a, c]$, and γ is the restriction of α to $[c, b]$ for some $a \leq c \leq b$, then $L(\alpha) = L(\beta) + L(\gamma)$. In particular $L(\beta), L(\gamma) \leq L(\alpha)$.
2. Length is independent of monotone reparameterization. That is, if $g : [c, d] \rightarrow [a, b]$ is a monotone continuous function (i.e. always increasing or always decreasing) then $L(\alpha) = L(\alpha \circ g)$.
3. If α has finite length L then α has a monotone reparameterization by arclength; that is, for some monotone increasing function $g : [0, L] \rightarrow [a, b]$, the curve $\beta := \alpha \circ g$ has the property that the length of β restricted to $[0, t]$ is t . Note that in some sense this curve has speed 1 because at time

t the length along the curve is exactly t . We will often assume that curves are unit parameterized.

The above properties should be familiar from calculus when applied to lengths of differentiable curves—in this case they follow from properties of the integral and are much easier to prove (given those properties of integrals, which are not so easy to prove!). It can be shown that if α is a differentiable curve in Euclidean space such that $\|\alpha'\|$ is an integrable function then $L(\alpha)$ as we have defined it is equal to $\int_a^b \|\alpha'\| dt$. Therefore our new definition is a true generalization of the definition from calculus.

Exercise 33 Define a curve in \mathbb{R} by $\alpha(t) := t$ on $[0, 1]$. Let $g(t) := t^2$ on $[-1, 1]$. Show that the reparameterization of α using g is twice as long as α . Explain what is happening to make this reparameterization change the length.

Exercise 34 Let d be the Euclidean metric on \mathbb{R}^n and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a positive continuous function. For any curve α with continuous derivatives, define the length of α by $L(\alpha) := \int_a^b f(\alpha(t)) \|\alpha'(t)\| dt$. Define the distance between two points to be the length of the shortest curve joining them (assume that such a shortest curve exists). Prove that this distance is a metric.

If α is defined on an interval I that is not closed and bounded, we may write I as an increasing union of closed bounded intervals $[a_i, b_i]$ and define the length of α to be the limit (possibly infinite) of the increasing sequence of lengths of α restricted to $[a_i, b_i]$. Of course it needs to be checked that this definition does not depend on the choice of the values a_i and b_i , but this follows from the above properties of length.

A curve α is said to be *parameterized proportional to arclength* if $\alpha(\frac{t}{k} - r)$ is arclength parameterized for some $k \neq 0$ and $r \in \mathbb{R}$. We refer to k as the speed of α ; if $k = 1$ we say α is unit parameterized. Thus α has constant speed k and starts at $t = rk$. The next exercise shows why this type of parameterization is useful:

Exercise 35 Let $\alpha : [a, b] \rightarrow X$ be a curve of finite length. Show that for any interval $[c, d]$, α has a monotone reparameterization that is proportional to arclength and defined on $[c, d]$.

Definition 36 A curve $\gamma : [a, b] \rightarrow X$ in a metric space is called a *geodesic* if γ is parameterized proportional to arclength and $L(\alpha) = d(\alpha(a), \alpha(b))$.

Exercise 37 Let γ be a unit geodesic in a metric space X .

1. Prove that the restriction of γ to any closed, bounded subinterval of its domain is also a unit geodesic.
2. Prove that γ is an isometry onto its image (that is, γ preserves distances although it may not be onto all of X).

Two points $x, y \in X$ are said to be “joined” by a curve $\alpha : [a, b] \rightarrow X$ if $x = \alpha(a)$ and $y = \alpha(b)$.

Definition 38 *A metric space X is called a geodesic space if every pair of points in X is joined by a geodesic.*

Euclidean space is a geodesic space, as was shown previously (the length of a line segment is equal to the distance between its endpoints, so a line segment is a geodesic). However, most metric spaces are not geodesic spaces. The following example is very important for many reasons:

Example 39 *Let $S^2 = \{v : \|v\| = 1\} \subset \mathbb{R}^3$. This space is called the unit 2-sphere; there are analogous definitions of unit n -spheres for nonnegative integers. We will first take the distance on S^2 obtained by restricting the usual Euclidean distance of \mathbb{R}^3 to S^2 . (In general, the restriction of a metric to a subset of a metric space is called the “subspace metric”.) That is, we measure the distance between points in S^2 just as we would when considering them as points in \mathbb{R}^3 . So the distance between any antipodal points is 2. On the other hand, it is well-known (but not so simple to prove!) that the shortest path between any antipodal points is a great circle, which has length $\pi > 2$. This shows that this particular metric on S^2 is not a geodesic metric. It is usually called the “subspace metric.”*

However, we can make S^2 into a geodesic space in a natural way. We can define the “intrinsic” metric on S^2 , by requiring that the distance between any two points be the length of the shortest path joining them. A proof similar to the proof of Exercise 34 shows that this in fact a metric. This metric describes the distances that would be experienced by a creature confined to live on the surface of the sphere. This metric defines what is known as “spherical geometry”, which is approximately the geometry of the Earth.

There is an important point that we will gloss over a bit concerning the above example. We really have two metric spaces, although the underlying set, S^2 is the same. Therefore we can consider the identity map I on S^2 as a function from the Euclidean metric to the intrinsic metric. It is an important fact that this function from one metric on S^2 to the other is continuous, as is its inverse. These two metrics are, in a sense, topologically the same.

Hyperbolic space is a geodesic space; as was discussed in the SG mini-course, the geodesics in hyperbolic space consist of circles having centers on the x -axis (which includes vertical lines, whose “centers” are at infinity on the x -axis). As with Euclidean space, every pair of points is joined by a unique geodesic. This particular property is somewhat rare among geodesic spaces. For example, in S^2 with its intrinsic metric every pair of antipodal points is joined by infinitely many geodesics.

4 More Metric Space Basics

Definition 40 Let x be a point in a metric space X , and $r > 0$. We define the ball centered at x of radius r to be $B(x, r) := \{y \in X : d(x, y) < r\}$.

For the metric space \mathbb{R} , we have

$$\begin{aligned} B(x, r) &= \{y \in X : d(x, y) < r\} \\ &= \{y \in X : |x - y| < r\} = \{y \in X : x - r < y < x + r\} \end{aligned}$$

which is simply the open interval $(x - r, x + r)$ of length $2r$ centered at x . (An interval doesn't "look" like a ball, but such preconceived images need to be abandoned—or replaced with other intuitive images.) Conversely, given any bounded open interval (a, b) in \mathbb{R} is a ball $B(x, r)$, where $x = \frac{a+b}{2}$ and $r = \frac{b-a}{2}$.

Exercise 41 Let x be a point in a metric space X , and suppose $r > r_0 > 0$. Show that $B(x, r_0) \subset B(x, r)$.

Exercise 42 Let x, y be distinct points in a metric space X , and let $r = d(x, y)$. Use the triangle inequality to prove by contradiction that $B(x, r/2)$ and $B(y, r/2)$ are disjoint.

Exercise 43 Let X be a non-empty set with the trivial metric, and let $x \in X$. Describe the following: $B(x, 2)$, $B(x, \frac{1}{2})$, $B(x, 1)$.

Definition 44 A subset A of a metric space X is said to be bounded if there exists some $x \in X$ and $r > 0$ such that $A \subset B(x, r)$.

How is this related to the notion of "bounded" from calculus? Recall that a subset A of \mathbb{R} is said to be bounded if it has an upper bound and a lower bound—that is, there exist numbers m and M such that for all $x \in A$, $m \leq x \leq M$. Taking $a < m$ and $b > M$ it follows that $A \subset (a, b)$. But we have already observed that open intervals in \mathbb{R} are balls, and hence if A is bounded in the "old" sense then A is bounded according to our new definition. Conversely, if A is contained in an open ball (a, b) then a is an upper bound for A and b is a lower bound for A , so A is bounded according to the old definition. This shows that our new definition is equivalent to the old one for subsets of \mathbb{R} , but our new definition makes sense in any metric space.

Exercise 45 Show that a nonempty set $A \subset X$ is bounded if and only if there exists some $r > 0$ such that for any $x, y \in A$, $d(x, y) < r$.

Definition 46 A subset A of a metric space X is called open if for every $x \in A$ there exists some $r > 0$ such that $B(x, r) \subset A$. A subset C of X is called closed if C^c is open.

Exercise 47 Prove that, for a metric space X , X itself and the empty set are subsets of X that are both open and closed.

In other words, unlike a door, a set can be both open and closed. As we will see later, sets can also be neither. We will frequently use the next lemma, the proof of which is an exercise.

Lemma 48 *A subset A of a metric space X is open if and only if for every $x \in A$ there exists some $\varepsilon > 0$ such that if $d(x, y) < \varepsilon$ then $y \in A$.*

Exercise 49 *Prove Lemma 48.*

Lemma 50 *If X is a metric space and $x \in X$ then $\{x\}$ is closed.*

Proof. We need to show that $\{x\}^c$ is open. If $y \in \{x\}^c$ then $y \neq x$. By the positive definiteness of the metric, $d(y, x) = r$ for some $r > 0$. But then $x \notin B(y, r)$, and hence $B(y, r) \subset \{x\}^c$. This proves that $\{x\}^c$ is open and so $\{x\}$ is closed. ■

A set containing a single point is often called a *singleton set*. Note that the number $r > 0$ in Lemma 48 may depend on x —that is, if y is closer to x then we need to use a smaller r . Similarly, it is not completely trivial that $B(x, r)$, which we have already named an “open ball”, is in fact open according to our definition. That is, every $y \in B(x, r)$ is contained in an open ball centered at x , namely $B(x, r)$, but is every $y \in B(x, r)$ contained in an open ball centered at y that is contained in $B(x, r)$? This is what we need to show in order to verify that $B(x, r)$ is open. We will check this now. Let $y \in B(x, r)$. By definition, $d(x, y) < r$. Let $\varepsilon := r - d(x, y) > 0$ and suppose that $d(z, y) < \varepsilon$. Then by the triangle inequality

$$d(z, x) \leq d(x, y) + d(y, z) < d(x, y) + (r - d(x, y)) = r,$$

and by definition $z \in B(x, r)$. That is, $B(y, \varepsilon) \subset B(x, r)$ and $B(x, r)$ is an open set. Therefore “open ball” is an appropriate name.

An indexing of a collection \mathcal{A} of sets “assigns” to each λ a unique set A_λ in the collection \mathcal{A} . For example, we can consider the collection $\{(-n, n)\}_{n \in \mathbb{N}}$ of all real intervals $(-n, n)$, where n is a natural number. Written more explicitly this is the set $\{(-1, 1), (-2, 2), (-3, 3), \dots\}$. In this example the indexing set is \mathbb{N} . However, we can consider sets indexed over any arbitrary set—e.g. $\{[-r, r]\}_{r \in (0, 1)}$, which consists of all closed intervals having endpoints $[-r, r]$, where $0 < r < 1$.

Given an indexed collection of sets $\{A_\lambda\}_{\lambda \in \Lambda}$, we define the *intersection* of the collection to be

$$\bigcap_{\lambda \in \Lambda} A_\lambda := \{x : x \in A_\lambda \text{ for all } \lambda \in \Lambda\}$$

and the *union* of the collection to be

$$\bigcup_{\lambda \in \Lambda} A_\lambda := \{x : x \in A_\lambda \text{ for some } \lambda \in \Lambda\}.$$

Example 51 We will prove that for sets A and $\{A_\lambda\}_{\lambda \in \Lambda}$ we have the following distributive law:

$$A \cap \left(\bigcup_{\lambda \in \Lambda} A_\lambda \right) = \bigcup_{\lambda \in \Lambda} (A \cap A_\lambda)$$

Now $x \in A \cap \left(\bigcup_{\lambda \in \Lambda} A_\lambda \right)$ if and only if

$$\begin{aligned} & x \in A \text{ and } x \in \bigcup_{\lambda \in \Lambda} A_\lambda \\ \Leftrightarrow & x \in A \text{ and } x \in A_\lambda \text{ for some } \lambda \in \Lambda \\ \Leftrightarrow & x \in A \cap A_\lambda \text{ for some } \lambda \in \Lambda \\ \Leftrightarrow & x \in \bigcup_{\lambda \in \Lambda} (A \cap A_\lambda). \end{aligned}$$

When $\Lambda = \mathbb{N}$ (or Λ is finite) we will often write, for example, $\bigcup_{i=1}^{\infty} A_i$ (or $\bigcap_{i=1}^m A_i$).

Exercise 52 Find the intersections and unions of the following collections

1. $\{(-n, n)\}_{n=1}^{\infty}$
2. $\{[-r, r]\}_{r \in (0,1)}$
3. $\{(0, \frac{1}{i}]\}_{i=1}^{\infty}$

De Morgan's laws have generalizations to general collections: it is not hard to prove that

$$A \setminus \left(\bigcup_{\lambda \in \Lambda} A_\lambda \right) = \bigcap_{\lambda \in \Lambda} (A \setminus A_\lambda)$$

and

$$A \setminus \left(\bigcap_{\lambda \in \Lambda} A_\lambda \right) = \left(A \setminus \bigcup_{\lambda \in \Lambda} A_\lambda \right).$$

Proposition 53 Let X be a metric space and $\{A_\lambda\}_{\lambda \in \Lambda}$ be a collection of open sets in X . Then $\bigcup_{\lambda \in \Lambda} A_\lambda$ is open in X .

Proof. Let $x \in \bigcup_{\lambda \in \Lambda} A_\lambda$. Then $x \in A_\lambda$ for some λ . Since A_λ is open there exists some $r > 0$ such that $B(x, r) \subset A_\lambda \subset \bigcup_{\lambda \in \Lambda} A_\lambda$. Hence $\bigcup_{\lambda \in \Lambda} A_\lambda$ is open. ■

Conversely, every open set A in X is a union of open sets—open balls. To see this, note that by definition, for each $x \in A$ there is some $r_x > 0$ and $A_x := B(x, r_x)$ such that $A_x \subset A$. Clearly $A = \bigcup_{x \in A} A_x$.

Exercise 54 Let X be a set with the trivial metric. Show that every subset of X is both open and closed. Hint: show that every subset $\{x\}$ with $x \in X$ is open.

Proposition 55 *If A_1, \dots, A_n are open sets in a metric space X then $\bigcap_{i=1}^n A_i$ is open.*

Proof. Let $x \in \bigcap_{i=1}^n A_i$. For each i there exists some $\varepsilon_i > 0$ such that if $d(x, y) < \varepsilon_i$ then $y \in A_i$. Then $\varepsilon := \min\{\varepsilon_1, \dots, \varepsilon_n\}$ is positive. If $d(x, y) < \varepsilon$ then $d(x, y) < \varepsilon_i$ for all i , and hence $y \in A_i$ for all i . By definition, $y \in \bigcap_{i=1}^n A_i$.

■

Example 56 *The above proof fails for infinitely many sets A_i , since infinite sets may not have minima—and even if we took the infimum of some infinite collection of epsilons, the infimum could well be 0. To see concretely that Proposition 55 is only valid for finitely many sets, consider the collection of open intervals $\{(-1/n, 1/n)\}_{n=1}^{\infty}$. The intersection of this collection is $\{0\}$, which we already know is closed. But it is also not open, because it is nonempty and does not contain any open interval at all!*

It follows from de Morgan's laws that the intersection of any collection of closed sets is closed, and the union of finitely many closed sets is closed.

Definition 57 *Let A be a subset of a metric space X . The closure \overline{A} of A is defined to be the set of all $x \in X$ such that for every $r > 0$, $B(x, r) \cap A \neq \emptyset$.*

Put another way, \overline{A} is the set of all points x in X such that there are points in A that are arbitrarily close to x . Note that certainly $A \subset \overline{A}$ since for every $r > 0$, if $x \in A$ then $x \in B(x, r) \cap A$. The next lemma has a satisfying ring to it. It says that closed sets are precisely those that are equal to their closures.

Lemma 58 *A is a closed subset of a metric space X if and only if $\overline{A} = A$.*

Proof. Suppose A is closed. Since $A \subset \overline{A}$ we need only show the opposite inclusion, which we will prove by contrapositive: Suppose $x \notin A$. Since A^c is open there exists an $r_0 > 0$ such that $B(x, r_0) \cap A = \emptyset$. But then by definition $x \notin \overline{A}$ and $\overline{A} \subset A$ is proved.

To prove the converse, suppose that $\overline{A} = A$ and let $x \in A^c = \overline{A}^c$. By definition of closure there exists some $r > 0$ such that $B(x, r) \subset \overline{A}^c = A^c$ and A^c is open; hence A is closed. ■

Remark 59 *Since A is always a subset of \overline{A} , a very useful strategy to prove that a set A is closed is to choose a point $x \in \overline{A}$ and show that $x \in A$.*

Exercise 60 *Let A be a subset of a metric space X .*

1. *Show that if C is any closed set in X containing A then $\overline{A} \subset C$.*
2. *Show that \overline{A} is the intersection of all closed sets containing A . In other words, the closure of A is the “smallest” closed set containing A .*

Exercise 61 *Let X be a trivial metric space with at least two points. Show that $\overline{B(x, 1)} \subsetneq C(x, 1)$; that is the closure of an open ball may not be equal to the closed ball of the same radius.*

5 Metric Groups

Definition 62 Let G be a group. A metric d on G is said to be left invariant if for every $x, y, z \in G$, $d(y, z) = d(xy, xz)$. Right invariant is defined similarly, and a metric is said to be bi-invariant if it is both left and right invariant. A group with a left-invariant metric such that the inversion function $x \mapsto x^{-1}$ is continuous is called a metric group.

If G is abelian then either left and right invariance implies bi-invariance and we simply say that d is invariant.

Proposition 63 If G has a left invariant metric d , then d is bi-invariant if and only if the inversion map is an isometry.

Proof. Suppose d is bi-invariant. Using the left and right invariance, multiplying by x^{-1} on the left and y^{-1} on the right, we have

$$d(x, y) = d(e, x^{-1}y) = d(y^{-1}, x^{-1}) = d(x^{-1}, y^{-1}).$$

On the other hand, if inversion is an isometry then

$$d(yx, zx) = d((yx)^{-1}, (zx)^{-1}) = d(x^{-1}y^{-1}, x^{-1}z^{-1})$$

and using left invariance, multiplying on the left by x , the last distance is equal to

$$d(y^{-1}, z^{-1}) = d(y, z).$$

■

Since isometries are continuous, we obtain:

Corollary 64 If G has a bi-invariant metric then G is a metric group.

Since a vector space is an abelian group with respect to the $+$ operation, the above definition generalizes to groups our earlier definition of invariant metric on a vector space. In particular, the Euclidean spaces are examples of metric groups.

Exercise 65 Show that any group with the trivial metric is a metric group.

Very important examples of metric groups come from what are known as finitely generated groups.

Definition 66 Let U be a subset of a group G . U is said to generate G if for all $x \in G$ there exist some $m \in \mathbb{Z}$ and $u_1, \dots, u_m \in U$ such that $x = u_1^{\pm 1} \cdot \dots \cdot u_m^{\pm 1}$. Here, for simplicity, we are using the following notation. $u^{\pm 1}$ means either $u = u^1$ or u^{-1} . If G is generated by a finite set then we say G is finitely generated.

In other words, a set U generates G if every element of G can be written as a (finite) product of elements of U and their inverses. To simplify notation we may also write $x = u_1 \cdot \dots \cdot u_m$, where $u_i \in U \cup U^{-1}$.

Definition 67 Let G be a group and U be a subset of G . The subgroup of G generated by U consists of all elements of G that are products of finitely many elements of $U \cup U^{-1}$.

Example 68 \mathbb{R} is generated by any interval $(0, \varepsilon)$. In fact, if $t \in \mathbb{R}$ then there are three cases. If $t > 0$ then for some n , $\frac{t}{n} \in (0, \varepsilon)$ and we can write $t = \frac{t}{n} + \dots + \frac{t}{n}$, where there are n terms in the sum. If $t < 0$ then in a similar way we can write $t = s + s + \dots + s$, where $s = -(\frac{-t}{n})$ and $\frac{-t}{n} \in (0, \varepsilon)$. Certainly $0 = a + (-a)$ for any $a \in (0, \varepsilon)$. Since \mathbb{R} is uncountably infinite, it cannot be finitely generated, or even generated by a countably infinite set:

Exercise 69 Prove that \mathbb{R} is not countably generated.

Example 70 Clearly \mathbb{Z} is generated by the set $\{1\}$ and hence is an infinite group that is finitely generated. Each of the groups \mathbb{Z}_n is also generated by a single element, $\bar{1}$. Groups generated by a single element are known as cyclic groups, and it is a theorem in abstract algebra that every cyclic group is isomorphic to the trivial group, \mathbb{Z}_n , or \mathbb{Z} .

Example 71 What about \mathbb{Q} ? This abelian group is countably infinite, so one cannot argue on the basis of cardinality alone that it is not finitely generated. Let's take a particular nonzero rational number $\frac{p}{q}$. What is the subgroup generated by $\frac{p}{q}$? This subgroup consists of all sums of $\frac{p}{q}$ with itself, which in turn is the cyclic group consisting of all fractions of the form $\frac{np}{q}$ where n is an integer. Certainly this subgroup does not contain $\frac{1}{r}$, when r and q are relatively prime. Note that the subgroup generated by $\frac{p}{q}$ is contained in the subgroup generated by $\frac{1}{q}$, so we need only consider generators of the latter form. Suppose that we have two generators $\frac{1}{p}$ and $\frac{1}{q}$. Then the subgroup generated by these generators consists of all fractions of the form $\frac{m}{p} + \frac{n}{q} = \frac{mq+np}{pq}$. If p and q have a common factor, we may cancel the factor in both the numerator and denominator of each such number, and hence we may assume that p and q are relatively prime. It is a fact from basic algebra that goes back to Euclid, that if p and q are relatively prime there are integers a and b such that $ap + bq = 1$. In other words, since we are assuming that p and q are relatively prime, the number $\frac{mq+np}{pq}$ is equal to $\frac{1}{pq}$ for some choice of integers m and n . From this discussion it follows that subgroup generated by $\frac{1}{p}$ and $\frac{1}{q}$ if p and q are relatively prime is the subgroup generated by $\frac{1}{pq}$. But this subgroup cannot contain $\frac{1}{r}$, when r is relatively prime to p and q . By similar arguments it follows that no finite collection of fractions can generate all of \mathbb{Q} .

Finite groups are, of course, finitely generated.

Many finitely generated groups may be described in the following way. Let G have a finite generating set $\{g_1, \dots, g_n\}$. A *word* of length k in these generators is of the following form: $g_{i_1}^{\pm 1} g_{i_2}^{\pm 1} \cdots g_{i_k}^{\pm 1}$. A *relator* or *relation* in G is a word $g_{i_1}^{\pm 1} g_{i_2}^{\pm 1} \cdots g_{i_k}^{\pm 1}$ whose product is equal to the identity. To specify a group we may list its generators, say $U := \{g_1, \dots, g_n\}$. We may then provide a collection (usually finite) of *relators* r_1, \dots, r_k . (If the numbers of generators and relators are both finite, the group is generally referred to as *finitely presented*.) We will define an equivalence relation the set of all words $h_1 \cdots h_m$, where $h_i \in U \cup U^{-1}$. The equivalence relation is best illustrated by an example. Suppose that our generating set consists of $\{a, b\}$ and our set of relators is $\{aba, b^2\}$. Consider the word ab^2a^2ba . Since b^2 is a relator, we may remove it to reduce the word to aa^2ba . Using $aa^2 = a^3$, which is a relator, we see that we may now reduce the word again by removing a^3 to finish with ba . Since ba is obtained from ab^2a^2ba by finitely many steps, each of which eliminates a relator, we would like to declare these two words equivalent. But in order for our equivalence relation to be symmetric, we must be allowed to also *insert* a relator into a word at any place. In other words, we declare two words to be equivalent if one can be obtained from the other by a finite sequence of inserting or removing relators. We will denote the equivalence class of a word by putting square brackets around it; so the equivalence class of aba^2 is $[aba^2]$. The group operation is defined by concatenation: for example $[ab][ab^2] = [abab^2] = [b^2]$. Now b^2 is a relator in our example, so we could actually remove it to obtain the equivalence class of the “empty word”. This class clearly functions as identity, since concatenation with the empty word doesn’t change anything. Inverses are defined in the obvious way—for example $[ab]^{-1} = [b^{-1}a^{-1}]$.

Finitely presented groups occur in many area of mathematics and science. For example, the group \mathbb{Z}_n may be described as the group with a single generator, call it a , with the single relator a^n .

There is a very natural way to make a finitely generated group into a metric group. Suppose g_1, \dots, g_n are the generators of the group. Assign each generator and its inverse, $g_i^{\pm 1}$, a “length”, denoted by L_i . Every element of g may be expressed as some word in the generators: $g = g_{i_1}^{\pm 1} \cdots g_{i_k}^{\pm 1}$, and we define $\|g\|$ to be the minimum of $L_{i_1} + \cdots + L_{i_k}$, where the minimum is taken over all possible expressions of g in terms of the generators. Note that if one of the generators is redundant in the sense that it can be expressed using the others, then it is possible that its norm will be smaller than its initial length—but of course its length would have to be longer than the sum of the lengths of those other generators. In the same way that we defined the distance induced by a norm, we define $d(g, h) := \|g^{-1}h\|$. For any $k \in G$ we have

$$d(kg, kh) = \|g^{-1}k^{-1}kh\| = \|g^{-1}h\| = d(g, h)$$

so the metric is invariant.

Exercise 72 How would you define the distance using $\| \cdot \|$ in order to make the distance right invariant?

Exercise 73 Show that with the metric constructed above, the inversion mapping is continuous. Hint: If L_i is a shortest length of a generator, what is $B(g, L_i)$?

One standard construction of a metric group of this kind is to simply take the lengths of all generators to be 1. The study of finitely generated groups with this type of metric is a very deep and important area of mathematics known as geometric group theory. As the previous exercise shows, these groups themselves are topologically uninteresting, but there are many important relationships between the metric and algebraic structures, including what are known as structures at infinity, which may be topologically quite interesting.

There is a natural way to connect these metrics to the geodesic metrics we have already studied. Given a group G , the *Cayley graph* of the group is constructed in the following way. The vertices of the graph are the elements of G . Two vertices g, h are joined by an edge if $h = gg_i$, where g_i is one of the generators of the group. So the identity vertex e is joined to each of the generators and their inverses by an edge. If there are n generators then there are $2n$ edges from the identity vertex.

We assign each edge length 1. Note that (since we are assigning each generator length 1) $d(g, h) = \|g^{-1}h\| = \|g_i\| = 1$ when g and h are joined by a vertex. For any $g \in G$ we can write $g = g_{i_1}^{\pm 1} \dots g_{i_k}^{\pm 1}$. Note that there is an edge joining e and $g_{i_1}^{\pm 1}$, and an edge joining $g_{i_1}^{\pm 1}$ to $g_{i_1}^{\pm 1}g_{i_2}^{\pm 1}$, and so on. That is there is a path of edges from e to g having k edges of length 1. If we take the infimum of the lengths of all such paths we obtain precisely $\|g\|$. In other words, the distance on G , having generators all of length 1, is the length of the shortest edge path joining the two points in the Cayley graph. In fact, we can define a metric on the Cayley graph by declaring the distance between any two points to be the shortest path in the Cayley graph, where all edges are declared to have length 1. This is a geodesic metric, and the natural metric on G is obtained by considering G as the set of vertices in the Cayley graph and taking the induced metric from the Cayley graph. More generally one may carry out the same procedure when assigning different lengths to the generators, as long as one is a little careful about redundant generators.