III. Spaces with special properties

We have seen that one can derive a relatively sizable amount of information simply from the axioms for metric and topological spaces. However, it should not be surprising that oned needs to impose further conditions on spaces in order to prove more substantial results, including abstract versions of the Maximum Value Theorem and Intermediate Value Theorem from single variable calculus. It turns out that these two results relies on a separate basic properties of the open subsets in a closed interval. The underlying concepts are known as *compactness* and *connectedness*, and they are treated in this unit.

In calculus it is also important to know that certain infinite series have meaningful sums, and indeed one reasons that mathematicians tightened their standards of logical rigor in the nineteenth century was to analyze the validity of certain strange and unanticipated results that arose from casual manipulations with infinite series; some of the results were justified, but others were not (this was expected because some of the formulas contradicted each other). One abstract version of the basic condition guaranteeing convergence of reasonable infinite series is called *completeness*, and it is also discussed in this unit along with some important geometrical an analytical implications (however, the applications to analysis go far beyond the scope of this course).

III.1 : Compact spaces – I

(Munkres, §§ 26, 27)

One of the most fundamental properties of continuous functions on closed intervals is that they have maximum and minimum values. In contrast, a continuous function on an open or half open interval does not necessarily have this property. In most real variables courses, the existence of maximum and minimum values is established with the help of the Heine-Borel-Lebesgue Theorem (sometimes the third name is dropped when referring to this result, and sometimes the first name is dropped). The conclusion of this result is so important that it has become incorporated into a definition. However, before proceeding to the main result we need a preliminary concept.

Definition. If X is a topological space and $\mathcal{U} = \{U_\alpha\}$ is a family of open subsets of X, we say that $\mathcal{U}$ is an *open covering* of X if $\cup_\alpha\, U_\alpha = X$. A subfamily $\mathcal{V} \subset \mathcal{U}$ is said to be a *subcovering* if $\cup_\beta\, V_\beta = X$, where $\mathcal{V} = \{V_\beta\}$.

Definition. A topological space is said to be *compact* if every open covering has a finite subcovering.

The main point of the Heine-Borel-Lebesgue Theorem is that closed intervals in the real line are compact. An abstract version of this result is established as Theorem 27.1 on pages 172–173 of Munkres (see Corollary 27.2 on the second of these pages for the case of interest to us here). In fact, one has the following characterization of compact subsets of the real line.

Characterization of compact subsets. *A subset K of the real line is compact if and only if it is closed and bounded.*

Several portions of the proof are true under much more general conditions, so we shall establish these first.

THEOREM. *If A is a compact subset of a Hausdorff space, then A is closed in X.*

Proof. We use the Hausdorff Separation Property to show that $X - A$ is open.

Let $y \in X - A$, then for each $a \in A$ we have $y \neq a$, and therefore by the Hausdorff Separation Property there are open sets $U_{(a,y)}$ and $V_{(a,y)}$ (in X) containing y and a respectively such that $U_{(a,y)}$ and $V_{(a,y)}$ are disjoint. The family of subsets $\{A \cap V_{(a,y)}\}$ is an open covering of $A$and thus has a finite subcovering

$$A \cap V_{(a_1,y)} \ , \ \cdots \ , \ A \cap V_{(a_k,y)}.$$

By construction we have $A \subset V_{(a_1,y)} \cup \cdots \cup V_{(a_k,y)}$, and thus if we take

$$U_y \ = \ U_{(a_1,y)} \cap \cdots \cap U_{(a_k,y)}$$

then U_y is an open subset containing y and $U_y \cap A = \emptyset$. A (by now) familiar argument shows that $X - A = \cup_y \, U_y$ and hence that $X - A$ is open.∎

PROPOSITION. *If A is a closed subset of a compact topological space X, then A is compact.*

Proof. Let $\mathcal{U} = \{U_\alpha\}$ be an open covering of A; choose open sets V_α in X so that $U_\alpha = A \cap V_\alpha$, and let $\mathcal{V}$ be the open covering of X given by the sets V_α together with $X - A$. By compactness of X there is a finite subcovering, which we may as well assume contains $X - A$ as well as open subsets $V_{\alpha(1)}, \ \cdots \ , V_{\alpha(k)}$. It then follows that the corresponding subsets $U_{\alpha(1)}, \ \cdots \ , U_{\alpha(k)}$ for a finite subcovering of A.∎

THEOREM. *If $f : X \to Y$ is continuous and X is compact, then its image $f(X)$ is also compact.*

Proof. Let $\{U_\alpha\}$ be an open covering of $f(X)$, and choose open subsets V_α in Y so that $U_\alpha = f(X) \cap V_\alpha$. Then the sets

$$W_\alpha = f^{-1}(U_\alpha) = f^{-1}(V_\alpha)$$

form an open covering of X, so there is a finite subcovering of X having the form $W_1, \cdots, W_k$. But

$$f(W_j) = f\left(f^{-1}(U_j)\right) = U_j$$

and therefore the sets $U_1, \cdots, U_k$ for a finite (open) subcovering of $f(X)$.∎

Proof of characterization of compact subsets of the real line. ($\Longrightarrow$) By the first result above a compact subset of the real line is closed. To see that it is bounded, consider the open covering given by the intersections of A with the open intervals $(-n, n)$ where n runs through the positive integers. If A were not bounded, this open covering would not have a finite subcovering, so A must be bounded as claimed.

($\Longleftarrow$) If A is bounded then A is a subset of some closed interval $[-M, M]$. Since A is closed in $\mathbf{R}$, it is also closed in the compact set $[-M, M]$, and therefore A is compact by the second of the results above.∎

With a little additional effort one can modify the proof of the Heine-Borel-Lebesgue Theorem to show that every box-shaped subset of $\mathbf{R}^k$ of the form

$$[a_1, b_1] \times \ \cdots \ \times [a_k, b_k]$$

is compact (see Theorem 2.40 on page 39 of Rudin's book); we shall also give an alternate proof later in the course. This in turn yields an extension of the characterization of compact subsets from $\mathbf{R}$ to $\mathbf{R}^k$ for all positive integers k.

COROLLARY. *A subset A of $\mathbf{R}^k$ is compact if and only if it is closed and bounded.*

Sketch of proof. ($\Longrightarrow$) The subset A is closed for the same reasons as before. If f_j denotes the restriction of the j^{th} coordinate function to A, then $f_j(A)$ is a compact and hence bounded subset of $\mathbf{R}$. If we choose $M > 0$ so that $\cup_j \, f_j(A) \subset [-M, M]$, then $a \in A \Rightarrow a = (a_1, \cdots a_k)$ where $|a_j| \leq M$ for all j. Hence A is bounded.

($\Longleftarrow$) If A is closed and bounded then for some $M > 0$ we know that A is a closed subset of the compact set

$$[-M, M] \times \ \cdots \ \times [-M, M]$$

and therefore A is compact.∎

The following consequence is a significant generalization of a fundamental result from calculus.

COROLLARY. *If X is compact and $f : X \to \mathbf{R}$ is continuous, then f attains maximum and minimum values on X.*

Proof. This reduces to showing the following: *If $A \subset \mathbf{R}$ is compact, so that it is closed and bounded, then both the least upper bound and greatest lower bound of A belong to A.* We shall only verify the statement regarding the least upper bound; the other statement follows by reversing the directions of all inequalities.

Let M be the least upper bound of A (which exists because A is bounded). The for every positive integer n we can find a point $a_n \in A$ such that

$$M - \frac{1}{n} < a_n \leq M$$

where the second inequality is true because M is an upper bound for A. It follows immediately that $M = \lim_{n \to \infty} a_n$, and since A is closed it follows that $M \in A$.

The Finite Intersection Property

There is a characterization of compactness in terms of closed subsets. Given a family $\mathcal{A} = \{F_\alpha\}$ of closed subsets of a topological space, we shall say that $\mathcal{A}$ has the *finite intersection property* if

$$F_{\alpha(1)} \cap \ \cdots \ \cap F_{\alpha(k)} \neq \emptyset$$

for all finite subcollections

$$\{ \ F_{\alpha(1)}, \ \cdots \ F_{\alpha(k)} \ \} \subset \mathcal{A} \ .$$

THEOREM. *A topological space X is compact if and only if for every family of closed subsets $\mathcal{A} = \{F_\alpha\}$ with the finite intersection property we have $\cap_\alpha F_\alpha \neq \emptyset$.*

A proof of this result and some further remarks appear on page 170 of Munkres.∎

Compactness and continuous mappings

We have already noted that continuous map that is 1–1 and onto is not necessarily a homeomorphism. However, if one puts suitable hypotheses on the domain or codomain it is sometimes possible to prove that a 1–1 onto continuous map is a homeomorphism without checking the continuity of the inverse directly. In particular, this holds for compact metric spaces.

PROPOSITION. *Suppose that $f : X \to Y$ is a continuous map from a compact topological space to a Hausdorff space. Then f is a closed mapping.*

Proof. Suppose that A is closed in X. Then A is compact, and therefore $f(A)$ is also compact in Y. But since Y is a Hausdorff space this implies that $f(A)$ is closed in Y.∎

COROLLARY. *If $f : X \to Y$ is a continuous and $1-1$ onto map from a compact topological space to a Hausdorff space, then f is a homeomorphism.*∎

Products and compactness

The following sort of question arises frequently in mathematics:

PROPERTIES OF PRODUCTS. *If X and Y are systems that have some property* **P** *and there is a reasonable notion of direct product $X \times Y$, does this product also have property* **P***?*

Here are some examples involving topological spaces for which there is a positive answer:

1. Suppose that X and Y are discrete spaces. Then $X \times Y$ is also discrete. (**Proof:** If $(x, y) \in X \times Y$ then $\{x\}$ is open in X and $\{y\}$ is open in Y. Therefore

$$\{(x, y)\} = \{x\} \times \{y\}$$

is open in $X \times Y$, and since x and y are arbitrary this means that every subset of $X \times Y$ is open.)

2. Suppose that X and Y are spaces in which one point subsets are closed. Then the same is true for $X \times Y$; the proof is analogous to the previous one.

3. If X and Y are finite, then the same is true for $X \times Y$.

4. If X and Y are homeomorphic to metric spaces, then the same is true for $X \times Y$. In fact, we have given three ways of constructing a metric on the product.

5. If X and Y are Hausdorff spaces, then $X \times Y$ is also Hausdorff. (**Proof:** One way of doing this is to use the characterization of a Hausdorff space W in terms of the diagonal Δ_W in $W \times W$ being closed. Let **Shuff** be the "middle four shuffle map"

$$X \times X \times Y \times Y \longrightarrow X \times Y \times X \times Y$$

that sends (x_1, x_2, y_1, y_2) to (x_1, y_1, x_2, y_2). This map is continuous because its projections onto the four factors are continuous, and the same is true for the inverse map which sends (x_1, y_1, x_2, y_2) to (x_1, x_2, y_1, y_2). Since **Shuff** is a homeomorphism, it follows that

$$\Delta_{X \times Y} = \mathbf{Shuff}\big(\Delta_X \times \Delta_Y\big)$$

is closed in $(X \times Y) \times (X \times Y)$. This completes the argument.)

In contrast, here is one example where there is a negative answer:

6. If X and Y are homeomorphic to subsets of the real line, the product $X \times Y$ is not necessarily homeomorphic to a subset of the real line. An easy counterexample is given by taking $X = Y = \mathbf{R}$. We shall prove this when we discuss connectedness later in the course.

THEOREM. *The product of finitely many compact spaces is compact.*

Using the canonical homeomorphism

$$(X \times Y) \times Z \cong X \times Y \times Z$$

and finite induction we can reduce the proof to the case of a product of two compact spaces. The proof depends upon the following result which is also useful in other contexts.

TUBE LEMMA. *Let X and Y be topological spaces such that X is compact, let $y \in Y$, and let $\mathcal{W} = \{ W_\alpha \}$ be a family of open subsets of $X \times Y$ such that $X \times \{y\}$ is contained in $\cup_\alpha W_\alpha$. Then there is a finite open covering $\mathcal{U}(y) = \{ U_i \}$ of X and an open subset $V(y)$ of Y containing y such that each product set $U_i \times V(y)$ is contained in some W_α.*

Proof of Tube Lemma. First of all, we claim that $X \times \{y\}$ is homeomorphic to X and therefore is compact. To see this, consider the map $f : X \to X \times \{y\}$ defined by $h(x) = (x, y)$. The projections onto the factors are the identity and the constant map, and therefore h is continuous. Projection onto the X factor yields a continuous inverse to h. Maps of this form are often called *slice inclusions*.

For each $x \in X$ let $W(x)$ be an open subset in $\mathcal{W}$ such that $x \in W(x) \times \{y\}$. Let U_x and V_x be open subsets of X and Y respectively such that

$$(x, y) \in U_x \times V_x \subset W(x) \ .$$

Then $\mathcal{U} = \{ U_x \}$ is an open covering of X and hence there is a finite subcovering

$$\mathcal{U}(y) = \{ U_{x_1}, \ \cdots \ , \ U_{x_n} \} \ .$$

If $V(y) = \cap_i \ V_{x_i}$, it follows that

$$U_{x_i} \times V(y) \quad \subset \quad U_{x_i} \times V_{x_i} \quad \subset \quad W(x_i)$$

which proves the lemma.∎

A picture illustrating this proof is given in the files `tubelemma.`$*$ in the course directory for various formats $*$.

Proof of the Theorem. Let $\mathcal{W} = \{ W_\alpha \}$ be an open covering of $X \times Y$, and for each $y \in Y$ let $\mathcal{W}(y) \subset \mathcal{W}$ be a family that covers $X \times \{y\}$.

Given $y \in Y$, let $\mathcal{U}(y)$ and $V(y)$ be associated to $\mathcal{W}(y)$ as in the Tube Lemma. The sets $V(y)$ form an open covering of Y and therefore there is a finite subcovering $\{ V(y_1), \ \cdots \ , \ V(y_m) \}$. Then the finite family of sets

$$\mathcal{U} = \{ \ U_\beta \times V(y_j) \mid U_\beta \in \mathcal{U}(y_j) \ \}$$

is a finite open covering of $X \times Y$ and for each set in the family there is some $W_{\gamma(\beta,j)}$ in $\mathcal{W}$ such that

$$U_\beta \times V(y_j) \subset W_{\gamma(\beta,j)} \ .$$

The finite collection of sets $W_{\gamma(\beta,j)}$ is the desired finite subcovering of $\mathcal{W}$.∎

Compactness and infinite products

The preceding result on compactness of products extends to infinite products provided one assumes the Axiom of Choice (in fact, the statement of the theorem is equivalent to the latter). This was established by A. N. Tychonoff and is known as Tychonoff's Theorem. The result has fundamental applications in several mathematical contexts, and perhaps the most important involve functional analysis (*e.g.*, the Banach-Alaoglu Theorem; see pages 68–69 of Rudin, *Functional Analysis*, for a statement and proof). Proofs of Tychonoff's Theorem and a crucial preliminary result appear on pages 233–235 of Munkres.

Compact metric spaces

If a compact topological space is determined by a metric, then many additional statements can be made. For example, we have the following generalization of the boundedness property:

PROPOSITION. *If X is a compact metric space then there is a constant $K > 0$ such that $\mathbf{d}(x,y) \leq K$ for all $x, y \in X$.*

Proof. By Example 4 in the list of examples of continuous functions in the previous note, for a fixed $z \in X$ the function $f(x) = \mathbf{d}(x,z)$ is (uniformly) continuous. Let M be the maximum value of this function. Given two points $x, y \in X$ the triangle inequality now implies that

$$\mathbf{d}(x,y) \leq \mathbf{d}(x,z) + \mathbf{d}(y,z) \leq M + M = 2M$$

and therefore we can take $K = 2M$.∎

Another important and much deeper property of a closed interval in the real line is that every infinite sequence in the interval has a convergent subsequence (the Bolzano-Weierstrass Theorem). This property also holds for compact metric spaces.

THEOREM. *If X is a compact metric space, then every infinite sequence in X has a convergent subsequence.*

Later in this course we shall prove a converse to this result.

Proof. Let $\{a_n\}$ be an infinite sequence in X, and suppose it has no convergent subsequence. If the sequence takes only finitely many values, then at least one of them occurs infinitely many times, and thus one can find a convergent subsequence, so we may as well assume that the sequence takes infinitely many distinct values. Let $A \subset X$ be the set of all these values.

We claim that $\mathbf{L}(A) = \emptyset$; suppose that $b \in \mathbf{L}(A)$. One can then recursively construct a subsequence that converges to b as follows. Suppose that the first r terms of the subsequence $a_{n(k)}$ have been defined so that $\mathbf{d}(b, a_{n(j)}) < \frac{1}{k}$. Let U_{r+1} be the set containing b obtained by taking the open disk $N_{1/(r+1)}(b)$ and removing all elements a_ℓ of the original sequence for $\ell \leq n(k)$ that are not equal to b. Since one point (and hence finite) subsets of a metric space are closed, it follows that U_{r+1} is open. Therefore, by the definition of a limit point there is some $a \in A$ such that $a \neq b$ and $a \in U_{r+1}$. By construction this point has the form a_m for some $m > n(r)$, and we set $m = n(r+1)$. This yields a subsequence whose limit is b.

Since $\mathbf{L}(\mathbf{A})$ is empty it follows that it is contained in A and therefore A is closed. Since X is compact, so is A.

We shall obtain a contradiction by showing that the infinite set A is not compact. If $a \in A$, then since $a \notin \mathbf{L}(A)$ we can find an open subset U_a in X such that $A \cap U_a = \{a\}$. It follows that every one point subset of A is open in the subspace topology and hence that every subset is open in the subspace topology. Since A is infinite, it follows that the open covering of A by one point subsets does not have a finite subcovering, which shows that A is not compact.

The contradiction means that our original assumption — the existence of an infinite sequence in X with no convergent subsequence — is incorrect, and therefore it follows that every infinite sequence in X has a convergent subsequence.■

The next result plays an important role in several analytic and geometric considerations. In particular, we shall use it to show that a continuous map from a compact metric space to another metric space is uniformly continuous.

LEBESGUE'S COVERING LEMMA. *Let X be a compact metric space, and let $\mathcal{U}$ be an open covering of X. Then there is a number $\eta > 0$ such that for every pair of points $x, y \in X$ such that $\mathbf{d}(x, y) < \eta$ there is an open set V in $\mathcal{U}$ such that $x, y \in V$.*

Proof. For each $p \in X$ there is an $\varepsilon(x) > 0$ such that $N_{2\varepsilon(x)}(x)$ is contained in some element of $\mathcal{U}$. Let $W_x = N_{\varepsilon(x)}(x)$.

The family $\mathcal{W} = \{W_x\}$ is an open covering of X, so there is a finite subcovering of the form

$$\{W_{x_1},\ \cdots\ ,\ W_{x_k}\}\ .$$

Let $\varepsilon_j > 0$ be the positive number associated to x_j, and let η be the minimum of the positive numbers $\varepsilon_1,\ \cdots\ , \varepsilon_k$.

Suppose now that $\mathbf{d}(x, y) < \eta$. Choose i so that $x \in W_{x_i}$. Then by the triangle inequality we have

$$\mathbf{d}(y, x_i) \leq \mathbf{d}(y, x) + \mathbf{d}(x, x_i) < \eta + \varepsilon_i \leq \varepsilon_i + \varepsilon_i = 2\varepsilon_i$$

which shows that $y \in N_{2\varepsilon(x_i)}(x_i)$. The latter set is contained in some set V from the family $\mathcal{U}$ and by construction it also contains x.■

A number η satisfying the conditions of the conclusion of the preceding result is called a *Lebesgue number* for the open covering. It is easy to see this result fails for noncompact metric spaces. For example, consider the open covering of the set of the open unit interval $(0, 1)$ given by the open subintervals

$$\left(\frac{1}{2^{k+1}}, \frac{1}{2^{k-1}}\right)$$

where k runs through the positive integers.

The uniform continuity property is an immediate consequence of the Lebesgue Covering Lemma.

THEOREM. *Let X and Y be metric spaces where X is compact, and let $f : X \to Y$ be continuous. Then f is uniformly continuous.*

Proof. Let $\varepsilon > 0$ be arbitrary, and for each $y \in Y$ consider the open set $N_{\varepsilon/2}(y) \subset Y$). By continuity the sets $f^{-1}\left(N_{\varepsilon/2}(y)\right)$ form an open covering of X, and by the compactness of X this open covering has a Lebesgue number η. Suppose now that $u, v \in X$ satisfy $\mathbf{d}(u, v) < \eta$. Then

there is some $y \in Y$ such that $u, v \in f^{-1}\left(N_{\varepsilon/2}(y)\right)$. It follows that $f(u), f(v) \in N_{\varepsilon/2}(y)$, and by the triangle inequality we have

$$\mathbf{d}\left(f(u), f(v)\right) \leq \mathbf{d}\left(f(u), y\right) + \mathbf{d}\left(y, f(u)\right) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

so that

$$\mathbf{d}(u, v) < \eta \Longrightarrow \mathbf{d}\left(f(u), f(v)\right) < \varepsilon$$

for all $u, v \in X$.■

III.2 : Complete metric spaces

(Munkres, §§43, 45)

Infinite series play an extremely important role in the theory and applications of the real number system; this is particularly apparent in the computational view of real numbers in terms of infinite decimal expansions and the use of power series to work with large families of functions in calculus. However, some care is needed in working with infinite series to ensure the reliability of any calculations done with them; in particular, it is necessary to know whether or not a series actually produces a meaningful real number (in other words, it *converges*). Many of the important criteria for convergence of infinite series in calculus rely on the property of real numbers known as completeness. The definition of this concept requires an important preliminary notion.

Definition. Let X be a metric space. A sequence $\{ a_n \}$ in X is a *Cauchy sequence* if for every $\varepsilon > 0$ there is a positive integer N such that $m, n > M$ implies $\mathbf{d}(a_m, a_n) < \varepsilon$.

PROPOSITION. *Every convergent sequence is a Cauchy sequence.*

Proof. Suppose that $\lim_{n \to \infty} a_n = L$. Given a positive real number ε, choose M such that $n > M$ implies $\mathbf{d}(L, a_n) < \varepsilon/2$. The triangle inequality then implies that

$$\mathbf{d}(a_m, a_n) \leq \mathbf{d}(a_m, L) + \mathbf{d}(L, a_n) < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

and therefore $\{ a_n \}$ is a Cauchy sequence.∎

It is easy to find examples of Cauchy sequences in metric spaces that do not have limits. For example, take X to be the open interval $(0, 1)$ and consider the sequence $a_n = \frac{1}{n}$. Of course this sequence does have a limit if one expands X to the closed unit interval. This is a special case of the following basic property of the real numbers:

THEOREM. *Every Cauchy sequence in* $\mathbf{R}^k$ *converges for all* $k \geq 1$.∎

References for the proof of this fact include Theorem 43.2 on page 265 of Munkres and Theorem 3.11 on pages 53–54 of Rudin. The relevance of this theorem to infinite series is explained in a mathematically rigorous fashion on pages 58–78 of Rudin (see also Exercise 15 on page 81).

Definition. A metric space X is *complete* if every Cauchy sequence in X converges.

One of the main objective of this section is to show that **every** Cauchy sequence in a metric space X converges in some larger metric space Y containing X (isometrically) as a subspace.

Properties of complete metric spaces

Complete metric spaces behave like compact topological spaces in several respects. Of course, there are also some major differences; for example, the real numbers are complete but not compact, and the real numbers are also homeomorphic to the noncomplete subspace $(-1, 1)$, say by the map $f : (-1, 1) \to \mathbf{R}$ defined using the formula

$$f(x) = \frac{x}{1 - |x|}$$

but one can begin the analogies with the following result:

PROPOSITION. *A compact metric space is complete.*

Proof. Let $\{ a_n \}$ be a Cauchy sequence in X. By previous results we know this sequence has a convergent subsequence $\{ a_{n(k)} \}$. Let $\lim_{k\to\infty} a_{n(k)} = L$; we claim that $\lim_{n\to\infty} a_n = L$. Given $\varepsilon > 0$ choose M_1 so that $k > M_1$ implies $\mathbf{d}(L, a_{n(k)}) < \varepsilon/2$, and choose M_2 so that $m, n > M_2$ implies $\mathbf{d}(a_m, a_n) < \varepsilon/2$. Take M to be the larger of M_2 and $n(M_1)$. Then the triangle inequality implies that $\mathbf{d}(L, a_n) < \varepsilon$ if $n > M$ (we have used this sort of argument many times already; at this point the reader should try to fill in the details as an exercise).∎

One also has the following useful result about closed subsets and complete metric spaces:

PROPOSITION. *Let X be a metric space and let $A \subset X$. If A is complete in the subspace metric, then A is closed in X. Conversely, if A is closed in X and X is complete, then A is complete in the subspace metric.*

Proof. Suppose first that A is complete with respect to the subspace metric and that $\{ a_n \}$ is a sequence in A with a limit $L \in X$. By a previous result the sequence is a Cauchy sequence, and therefore it has a limit in A. Since the limit of a convergent sequence is unique, this limit in A must be the point L.

Now suppose that A is closed in X where X is complete. If $\{ a_n \}$ is a Cauchy sequence in A then the completeness of X implies that the sequence has a limit $L \in X$. Since A is a closed subspace, this means that $L \in A$, so that the Cauchy sequence in A has a limit in A. Therefore A is complete.∎

The following result on completeness and products reflects another similarity with compact topological spaces.

PROPOSITION. *If X and Y are complete metric spaces, then so is $X \times Y$ with respect to each of the three product metrics.*

Proof. Let $\mathbf{d}^{\langle q \rangle}$ be a product metric where $q = 1$, 2 or ∞. By construction each of the projection maps

$$\left(X \times Y, \mathbf{d}^{\langle q \rangle}\right) \to (X, \mathbf{d}_X)\,, \quad (Y, \mathbf{d}_Y)$$

sends points whose distance in $X \times Y$ is ε to points in X and Y with the same property.

It follows that if $\{ (x_n, y_n) \}$ is a Cauchy sequence in $X \times Y$ then $\{ x_n \}$ and $\{ y_n \}$ are Cauchy sequences in X and Y respectively. By the completeness of X and Y each of these sequences has a limit, and we shall call these limits x and y respectively. To complete the proof we need to show that

$$\lim_{n\to\infty} (x_n, y_n) = (x, y)$$

with respect to each of the three product metrics.

The previously established inequalities

$$\mathbf{d}^{\langle \infty \rangle} \leq \mathbf{d}^{\langle 2 \rangle} \leq \mathbf{d}^{\langle 1 \rangle}$$

show that it suffices to prove the limit statement for $\mathbf{d}^{\langle 1 \rangle}$. Given $\varepsilon > 0$, choose M so that $n > M$ implies that both $\mathbf{d}_X(x, x_n)$ and $\mathbf{d}_Y(y, y_n)$ are less than $\varepsilon/2$; then the standard arguments (supply them!) show that $n > M$ implies

$$\mathbf{d}^{\langle 1 \rangle}\Big((x, y)\,,\, (x_n, y_n) \Big) < \varepsilon$$

and this proves the statement(s) about limits.∎

Function spaces

The next result provides some important examples of complete metric spaces.

PROPOSITION. *If X is a set then $\mathbf{BF}(X)$ is a complete metric space with respect to the norm described previously; furthermore, if X is a topological space then $\mathbf{BC}(X)$ is a closed and hence complete subset of $\mathbf{BF}(X)$.*

Proof. We shall first prove that $\mathbf{BF}(X)$ is complete. If $\{ f_n \}$ is a Cauchy sequence in $\mathbf{BF}(X)$, then by definition we know that $\{ f_n(x) \}$ is a Cauchy sequence of real numbers for each $x \in X$. Since we know that $\mathbf{R}$ is complete, it follows that for each $x \in X$ there is a real number $f(x)$ to which $\{ f_n(x) \}$ converges. We need to prove two things. First, we need to show that f is bounded. Next we have to show that $\lim_{n\to\infty} f_n = f$ in $\mathbf{BF}(X)$.

To verify that $\{ f_n(x) \}$ is bounded, choose M so that $m, n > M$ implies $|f_m - f_n| < 1$. Then for all n and x we know that

$$|f(x)| \leq \max\{|f_i(x)|,\ i < M;\ |f_M(x)| + 1\}$$

and therefore we also have

$$|f(x)| \leq \max\{|f_i|,\ i < M;\ |f_M| + 1\}$$

so that the left hand side is bounded by the right hand side for all $x \in X$ and therefore $f \in \mathbf{BF}(X)$.

To show that $\{ f_n \}$ converges to f in $\mathbf{BF}(X)$, let $x \in X$ be arbitrary, and given $varepsilon > 0$ choose M so that $m, n > M$ implies $|f_m - f_n| < \varepsilon/2$. Since $\{ f_n(x) \}$ converges to $f(x)$, there is some $K_x > 0$ such that $m > K_x$ implies $|f_m(x) - f(x)| < \varepsilon/4$. Therefore if $n > M$ and $m > M + K_x$ we have

$$|f(x) - f_n(x)| \leq |f(x) - f_m(x)| + |f_m(x) - f_n(x)| < \frac{3\varepsilon}{4}$$

which means that

$$|f - f_n| = \sup\{|f(x) - f_n(x)|\} \leq \frac{3\varepsilon}{4} < \varepsilon$$

and thus that $\lim_{n\to\infty} f_n = f$ in $\mathbf{BF}(X)$.

We next need to show that f is continuous if each f_n is continuous. Given $\varepsilon > 0$ and $x \in X$ we shall find an open set U such that $x \in U$ and for all $y \in U$ we have $|f(y) - f(x)| < \varepsilon$.

First of all, choose M such that $n > M$ implies $|f_n - f| < \varepsilon/3$. Next choose an open set U containing x such that $y \in U$ implies $|f_{M+1}(y) - f_{M+1}(x)| < \varepsilon/3$. The triangle inequality for real numbers then implies that

$$|f(y) - f(x)| \leq |f(y) - f_{M+1}(y)| + |f_{M+1}(y) - f_{M+1}(x)| + |f(x) - f_{M+1}(x)| < \varepsilon$$

and therefore f is continuous at x; since x was arbitrary, this means that $f \in \mathbf{BC}(X)$.∎

The examples in the proposition are special cases of the following important mathematical structure:

Definition. If $(V, |...|)$ is a normed vector space, then V is said to be a *Banach space* if it is complete with respect to the associated metric.

Intersections of nested closed sets

We had previously noted that compact metric spaces are characterized by the fact that families of closed subspaces with the finite intersection property have nonempty intersections. An important special case involves *nested* sequences of closed subsets $\{ A_n \}$ that are nonempty and satisfy $A_{n+1} \subset A_n$ for all n. In this case compactness implies that the intersection $\cap_n A_n$ is nonempty. There is an analog of this property for compact metric spaces. Recall that the *diameter* of a (subset of a) metric space is given by

$$\text{diam } (A) = \sup_{u,v \in A} \mathbf{d}(u,v)$$

if the set of distances has an upper bound and by $+\infty$ if no such upper bound exists.

PROPOSITION. (Nested Intersection Property). *Let X be a complete metric space, and let $\{ A_n \}$ be a nested sequence of nonempty closed subsets of X such that* $\lim_{n \to \infty} \text{diam } (A_n) = 0$. *Then $\cap_n A_n$ consists of one point.*

A proof that the intersection is nonempty is given in Lemma 48.3 on page 297 of Munkres. Suppose that y and z lie in the intersection. Then $y, z \in A_n$ for all n, and therefore $\mathbf{d}(y,z) \leq$ diam (A_n) for all n. Since the right hand side goes to zero as $n \to \infty$, it follows that the left hand side is ≤ 0; since the left hand side is nonnegative by construction, it must be zero, so that $y = z$ and the intersection contains exactly one point.∎

Completions of metric spaces

We have already mentioned that one recurrent theme in theoretical mathematics is the desire to see whether empty spaces in mathematical structures can be filled in some sense, giving the real numbers as an example. In fact, one can view the real numbers as a system obtained from the rational numbers by filling in gaps so that every Cauchy sequence converges. In particular, the finite decimal fraction approximations to a real number form a set of rational numbers converging to the given real number, and as such they are Cauchy sequences that usually do not converge to rational points. Our objective here is to show that every metric space can be expanded to a larger one in which every Cauchy sequence converges. Often this is done by a brute force construction that starts with the set of all Cauchy sequences in the metric space (for example, see Munkres, Exercise 9, page 271). We shall construct this completion by a method that takes just about the same amount of work but also yields some illuminating insights of independent interest.

PROPOSITION. *If $(X, \mathbf{d})$ is a metric space, then there is a $1-1$ isometry from X into a Banach space.*

Proof. Before starting the proof we recall that a distance function satisfies

$$|\mathbf{d}(u,w) - \mathbf{d}(v,w)| \ \leq \ \mathbf{d}(u,v)$$

for all $u, v, w \in X$. This follows from two applications of the triangle inequality.

Let $\mathbf{BC}(x)$ be the Banach space of bounded continuous functions on X, and choose some point $a \in X$. Define a map $\varphi : X \to \mathbf{BC}(X)$ by the formula

$$[\varphi(x)](y) = \mathbf{d}(y,x) - \mathbf{d}(y,a) \ .$$

The right hand side is a continuous function of y because

$$\big|[\varphi(x)](y) - [\varphi(x)](z)\big| = \Big|(\mathbf{d}(y,x) - \mathbf{d}(y,a)) - (\mathbf{d}(z,x) - \mathbf{d}(z,a))\Big| \ \leq$$

$$\big|\mathbf{d}(y,x) - \mathbf{d}(y,x)\big| + \big|\mathbf{d}(z,a) - \mathbf{d}(y,a)\big| \ = \ 2\,\mathbf{d}(y,z)$$

for all $y, z \in X$, and it is a bounded function of y because

$$\Big|\mathbf{d}(y,x) - \mathbf{d}(y,a)\Big| \ \leq \ \mathbf{d}(x,a)$$

for all $y \in X$ if a and x are held fixed.

The estimates of the previous paragraph also show that

$$\big|\varphi(x) - \varphi(y)\big| \ \leq \ \mathbf{d}(x,y)$$

and in fact equality holds because

$$\big|[\varphi(x)](y) - [\varphi(y)](y)\big| \ = \ \Big|(\mathbf{d}(x,y) - \mathbf{d}(a,y)) - (\mathbf{d}(y,y) - \mathbf{d}(y,a))\Big| \ = \ \mathbf{d}(x,y)$$

so that φ is an isometry (hence uniformly continuous).∎

Definition. If X is a metric space, then a *completion* of X is a pair (f, Y) consisting of a complete metric space and an isometry $f : X \to Y$ such that $\overline{f(X)} = Y$.

The preceding result implies that completions exist because the closure of $\varphi(X)$ in $\mathbf{BC}(\mathbf{X})$ is a closed subset of a complete metric spaced and therefore is complete. We shall also prove that up to an isometry there is only one way of completing a metric space. Here is a formal statement:

UNIQUENESS THEOREM. *Let X be a metric space, and let (f, Y) and (g, Z) be completions of X. Then there is a unique $1-1$ onto isometry $h : Y \to Z$ such that $h \circ f = g$.*

This result is a consequence of the following more general statement:

THEOREM. *Let X be a metric space, let (f, Y) be a completion of X, let W be a complete metric space, and let $h : X \to W$ be a uniformly continuous function. Then there is a unique uniformly continuous function $H : Y \to W$ such that $H \circ f = h$. Furthermore, if h is an isometry then so is H.*

Proof. The basic idea of the proof is simple. Since $\overline{f(X)} = Y$ it follows that for each $y \in Y$ there is a sequence of points $\{\, x_n \,\}$ in X such that $\lim_{n\to\infty} f(x_n) = y$. The only way we can extend h is if we take $H(y) = \lim_{n\to\infty} h(x_n)$. We need to show this actually works. The first step is to verify that the definition of $H(y)$ makes sense (in particular, the sequence $\{\, h(x_n) \,\}$ actually converges) and does not depend upon the choice of sequence in $f(X)$ converging to Y. Next, we have to show that the function is uniformly continuous. Finally we have to show that H is an isometry if h is.

First step. How do we know that the sequence converges? The hypotheses on h and W suggest a couple of ideas. Since W is complete the sequence $\{\, h(x_n) \,\}$ will converge if it is a Cauchy

sequence, and thus one might hope that the uniform continuity of h and the convergence of the sequence $\{ f(x_n) \}$ imply the convergence of $\{ h(x_n) \}$. Thus it suffices to show that *a uniformly continuous map of metric spaces from X to Y takes Cauchy sequences in X to Cauchy sequences in Y.* To see this, let $h : Y \to W$ be a uniformly continuous map of metric spaces, and assume that $\{ y_n \}$ is a Cauchy sequence in Y. Given $\varepsilon > 0$ there is a $\delta > 0$ such that $\mathbf{d}(u, v) < \delta$ implies $\mathbf{d}(h(u), h(v)) < \varepsilon$. Since we have a Cauchy sequence it follows that there is an M such that $m, n > M$ implies $\mathbf{d}(y_m, y_n) < \delta$. It follows that $m, n > M$ implies $\mathbf{d}(h(y_m), h(y_n)) < \varepsilon$. — This implies that there is some point $w_0 \in W$ such that $w_0 = \lim_{n \to \infty} h(x_n)$.

Second step. If we have two sequences $\{ u_n \}$ and $\{ v_n \}$ in X such that

$$\lim_{n \to \infty} f(u_n) = y = \lim_{n \to \infty} f(v_n)$$

we need to show that

$$\lim_{n \to \infty} h(u_n) = \lim_{n \to \infty} h(v_n)$$

in order to know that H is well-defined. Denote the limits of these two sequences by α and β respectively; it will suffice to show that $\mathbf{d}(\alpha, \beta) < \varepsilon$ for all $\varepsilon > 0$.

Given $\varepsilon > 0$ one can find a positive integer M such that for all $n > M$ we have $\mathbf{d}(h(u_n), \alpha) < \varepsilon/3$ and $\mathbf{d}(h(v_n), \beta) < \varepsilon/3$. Now choose $\delta > 0$ so that $\mathbf{d}(s, t) < \delta$ implies $\mathbf{d}(h(s), h(t)) < \varepsilon/3$, and choose P so that $p \geq P$ implies $\mathbf{d}(f(u_p), y) < \delta/2$ and $\mathbf{d}(f(v_p), y) < \delta/2$; the latter imply that $\mathbf{d}(f(u_p), f(v_p)) < \delta$, so that $\mathbf{d}(f(u_p), f(v_p)) < \varepsilon/3$. If we choose $q > M + P$ then we have

$$\mathbf{d}(\alpha, \beta) \leq \mathbf{d}(h(u_q), \alpha) + \mathbf{d}(h(u_q), h(v_q)) + \mathbf{d}(h(v_q), \beta)$$

and since $q > M$ the right hand side is less than $\mathbf{d}(h(u_q), h(v_q)) + 2\varepsilon/3$. Finally, since $q > P$ we also know that $\mathbf{d}(h(u_q), h(v_q)) < \varepsilon/3$, and therefore $\mathbf{d}(\alpha, \beta) < \varepsilon$ as required.

Third step. We need to show that the function H is uniformly continuous. Since h is uniformly continuous, for each $\varepsilon > 0$ there is a $\delta > 0$ such that $\mathbf{d}(u, v) < \delta$ implies $\mathbf{d}(h(u), h(v)) < \varepsilon/3$. Given $a, b \in Y$, let $\{ u_n \}$ and $\{ v_n \}$ be sequences in X such that $\lim_{n \to \infty} f(u_n) = a$ and $\lim_{n \to \infty} f(v_n) = b$, and suppose that $\mathbf{d}(a, b) < \delta$. Choose M so large that $n > M$ implies

$$\mathbf{d}(u_n, a), \quad \mathbf{d}(v_n, b) \quad < \quad \frac{\delta - \mathbf{d}(a, b)}{2}$$

and also

$$\mathbf{d}\left(h(u_n), H(a)\right), \quad \mathbf{d}\left(h(v_n), H(b)\right) \quad < \quad \frac{\varepsilon}{3} \, .$$

We then have that $\mathbf{d}(u_n, v_n) < \delta$ and that

$$\mathbf{d}\left(H(a), H(b)\right) \leq \mathbf{d}\left(h(u_n), H(a)\right) + \mathbf{d}(h(u_n), h(v_n)) + \mathbf{d}\left(h(v_n), H(b)\right) < 3 \cdot \frac{\varepsilon}{3} = \varepsilon$$

which shows that H is uniformly continuous.

Fourth step. We need to show that H is an isometry if h is an isometry. This will follow from a more general fact:

Suppose that $\{ u_n \}$ and $\{ v_n \}$ are convergent sequences in a metric space E with $\lim_{n \to \infty} u_n = u$ and $\lim_{n \to \infty} v_n = v$. Then $\lim_{n \to \infty} \mathbf{d}(u_n, v_n) = \mathbf{d}(u, v)$.

Using this we may complete the fourth step as follows: Express $u, v \in Y$ as limits of sequences $\{ f(u_n) \}$ and $\{ f(u_n) \}$ respectively. Then the assertion implies that

$$\mathbf{d}(u,v) = \lim_{n\to\infty} \mathbf{d}(f(u_n), f(v_n)) = \lim_{n\to\infty} \mathbf{d}(u_n, v_n)$$

where the last equation holds because f is an isometry. On the other hand we also have

$$\mathbf{d}(H(u), H(v)) = \lim_{n\to\infty} \mathbf{d}(h(u_n), h(v_n)) = \lim_{n\to\infty} \mathbf{d}(u_n, v_n)$$

because h is also an isometry. Since limits of sequences are unique it follows that $\mathbf{d}(u,v) = \mathbf{d}(H(u), H(v))$ and therefore H is an isometry.

We must now verify the general assertion about limits of distances. Consider the inequality

$$\bigl|\mathbf{d}(u_n, v_n) - \mathbf{d}(u,v)\bigr| \leq \bigl|\mathbf{d}(u_n, v_n) - \mathbf{d}(u, v_n)\bigr| + \bigl|\mathbf{d}(u, v_n) - \mathbf{d}(u,v)\bigr| \leq \mathbf{d}(u_n, u) + \mathbf{d}(v_n, v) \ .$$

If we choose M so that $n > M$ implies $\mathbf{d}(u_n, u),\ \mathbf{d}(v_n, v) < \varepsilon/2$ the inequalities imply that

$$\bigl|\mathbf{d}(u_n, v_n) - \mathbf{d}(u,v)\bigr| < \varepsilon$$

which proves the assertion about limits.∎

Proof of the Uniqueness Statement. Let X be a metric space, and let (f, Y) and (g, Z) be completions of X. By the preceding theorem there are unique isometries $G : Y \to Z$ and $F : Z \to X$ such that $G \circ f = g$ and $F \circ g = f$. These in turn imply that $F \circ G \circ f = f$ and $G \circ F \circ g = g$ for the maps $FG : Y \to Y$ and $GF : Z \to Z$. Since id_Y and id_Z also satisfy $\mathrm{id}_Y \circ f = f$ and $\mathrm{id}_Z \circ g = g$ it follows that $G \circ F = \mathrm{id}_Z$ and $F \circ G = \mathrm{id}_Y$, showing that the isometries F and G are 1–1 and onto. This completes the proof that any two completions of a given metric space are isometric by an isometry compatible with the isometric inclusions of X.∎

III.3 : Implications of completeness

(Munkres, § 48; Edwards, § III.1)

There are two fundamentally important properties of complete metric spaces that arise in numerous analytic and geometric contexts. One of these (Baire's Theorem) can be viewed informally as saying that a complete metric space cannot be decomposed into "thin" pieces. The other (the Banach Contraction Lemma) is a powerful method for finding solutions to various sorts of equations in a wide range of contexts. We shall give both topological and analytic examples in this course.

Nowhere dense and meager subspaces

Definition. A subset A of a topological space X is said to be *nowhere dense in* X if $\mathrm{Int}_X(\overline{A}) = \emptyset$.

The proofs of the following results are left to the reader as exercises.

PROPOSITION. *If A is a subset of X then A is nowhere dense in X if and only if $X - \overline{A}$ is dense in X.*∎

PROPOSITION. *Suppose that $A \subset B \subset C \subset X$ and that B is nowhere dense in C. Then A is nowhere dense in C and B is nowhere dense in X.*∎

The proof of this is left to the reader as an exercise.

PROPOSITION. *Suppose that A and B are nowhere dense subsets of X. Then $A \cup B$ is nowhere dense in X.*∎

It is particularly useful to understand when a one point subset of X is nowhere dense. Since one point subsets are closed, a one point subset $\{ x \}$ is nowhere dense in X if and only if it is **not** open; *i.e.*, the point x is not isolated in X.

Definition. A subset $B \subset X$ is said to be *meager* or *of the first (Baire) category* in X if it can be written as a countable union $B = \cup_n A_n$ where each A_n is closed and nowhere dense. A subset C is said to be *nonmeager* or *of the second (Baire) category* in X if it is not meager, and in this case the complement $X - C$ is said to be *residual* or *co-meager*.

Examples.

1. A closed nowhere dense subset is always meager.

2. The rationals are a meager, but definitely not nowhere dense, subset of the real numbers (since the rationals are dense in the reals).

3. Here is a more complicated but still important example. Let $\mathbf{R}^\infty$ denote the inner product space consisting of all sequences $(x_1, x_2, \cdots)$ such that $x_k = 0$ for all but finitely many k, with the inner product given by the convergent (in fact, finite) series

$$\langle x, y \rangle = \sum_j x_j y_j \ .$$

If B_n is the set of all points for which $x_j = 0$ for all $j > n$, then B_n is closed and nowhere dense in B_{n+1} and hence B_n is also closed and nowhere dense in X (why?). Since $\mathbf{R}^\infty = \cup_n B_n$ it follows that $\mathbf{R}^\infty$ is meager in itself.

Note. The concepts of first and second category were so named before category theory was invented; there is no connection between the two meanings of the word "category." In order to avoid confusion many authors have made conscious efforts to avoid terms like first and second category (compare the comments at the bottom of page 295 of Munkres), but this terminology is still very widely used by mathematicians and others. One way of dealing with this is to use the term *Baire category* when referring to concepts involving first and second category as defined above.

Baire spaces

Definition. A topological space X is said to be a *Baire space* if every open subset is nonmeager. Equivalent definitions are given on pages 295–296 of Munkres (in particular, see Lemma 48.1 on page 296).

The following result provides many examples of Baire spaces:

BAIRE CATEGORY THEOREM. *If X is a complete metric space, then X is a Baire space.*

A proof of this result for X itself is given in Theorem 48.2 on page 296 of Munkres (note that this proof uses Lemma 48.3 on the following page).∎

The Baire Category Theorem has many extremely important and useful consequences. We shall begin with one that only involves point set theory.

PROPOSITION. *If X is a countable compact metric space, then X has at least one isolated point.*

Proof. Let $X = \{ x_1, x_2, \ ... \}$, and for each i let $F_i = \{ x_i \}$. If X has no isolated points, then each F_i is closed and nowhere dense, and therefore X is meager in itself. By Baire's Category Theorem this is impossible.∎

COROLLARY. *If X is a compact metric space such that every point of X is a limit point, then X is uncountable.*∎

This is essentially the contrapositive of the theorem.

On pages 41–42 of Rudin's book this corollary is applied to show that the Cantor set is uncountable. One can prove that the Cantor set has the same cardinality as the real numbers by a different argument (see Problems 37–38 on page 46 of Royden, *Real Analysis, Third Edition*, as well as Exercise 6 on page 179 of Munkres).

In Section 49 of Munkres the Baire Category Theorem is used to prove the existence of continuous functions on the unit interval that are not differentiable at any point of the interval. The theorem also plays a crucial role in the foundations of the theory of Banach spaces and functional analysis; in particular, it is the key ingredient of the proofs of the Uniform Boundedness Principle and Open Mapping Theorem (*e.g.*, see Rudin's book on Functional Analysis for more information on this).

The Contraction Lemma

A special case of this result is established as Theorem 1.1 on pages 162–163 of Edwards, and later in that book there are assertions that the proof and conclusion can be generalized. Here is the explicit generalization.

CONTRACTION LEMMA. *Let X be a complete metric space, and let $T : X \to X$ be a map such that $\mathbf{d}(T(x), T(y)) \leq \alpha \cdot \mathbf{d}(x, y)$ for some fixed $\alpha \in (0, 1)$ and all $x, y \in X$ (in particular, T is uniformly continuous). Then there is a unique $z \in X$ such that $T(z) = z$ (in other words, a* **unique fixed point** *for T).*

To see the need for completeness, consider the open interval $(0, 1)$ and let T be multiplication by $\frac{1}{2}$.

Proof. The idea is beautifully simple. One starts with an arbitrary point $x \in X$ and considers the sequence of points $x,\ T(x),\ T^2(x), \cdots$. This sequence is shown to be a Cauchy sequence, and the limit z of this sequence turns out to be the unique fixed point.

More formally, we begin by noting that T has at most one fixed point. If $z, w \in X$ satisfy $T(z) = z$ and $T(w) = w$, then we have

$$0 \quad \leq \quad \mathbf{d}(z, w) \quad = \quad \mathbf{d}\left(T(z), T(w)\right) \quad \leq \quad \alpha \mathbf{d}(z, w)$$

and since $0 < \alpha < 1$ this can only happen if $\mathbf{d}(z, w) = 0$; *i.e.*, if $z = w$.

We now follow the idea described in the first paragraph of the proof. By induction on n we have

$$\mathbf{d}\left(T^n(x), T^{n+1}(x)\right) \quad \leq \quad \alpha^n \mathbf{d}\left(x, T(x)\right)$$

and therefore by the triangle inequality for $m > n$ we also have

$$\mathbf{d}\left(T^n(x), T^m(x)\right) \quad \leq \quad \sum_{i=n+1}^{m} \alpha^i \mathbf{d}\left(x, T(x)\right) \quad =$$

$$\frac{\alpha^{n+1}(1 - \alpha^{m-n})}{1 - \alpha} \mathbf{d}\left(x, T(x)\right) \quad \leq \quad \frac{\alpha^{n+1}}{1 - \alpha} \mathbf{d}\ \left(x, T(x)\right)$$

which implies that the sequence $\{\, T^n(x) \,\}$ is a Cauchy sequence. By the completeness of X there is a point z such that $z = \lim_{n \to \infty} T^n(x)$.

By Theorem 23.1 on page 130 of Munkres we have

$$T(z) \quad = \quad \lim_{n \to \infty} T\left(T^n(x)\right) \quad = \quad \lim_{n \to \infty} T^{n+1}(x)$$

and by a change of variable (specifically, take $k = n+1$) the right hand side is equal to $\lim_{k \to \infty} T^k(x)$, which by construction is z. Therefore we have $T(z) = z$.∎

Applications of the Contraction Lemma to solving equations in one and two real variables are discussed in Section III.1 on pages 160–172 of Edwards as well as an addendum to this section (`cubicroots.*` in the course directory), and further applications from Edwards will be discussed later. For the time being we shall merely use the Contraction Lemma to prove the basic existence and uniqueness theorem for solutions of first order differential equations in one variable.

PICARD SUCCESSIVE APPROXIMATION METHOD FOR THE SOLUTIONS OF DIFFERENTIAL EQUATIONS. *Let $F(x, y)$ be a real valued function of two variables on an open set U such that F has continuous partial derivatives on U. Then for each $(a, b) \in U$ there is a positive real number $\delta > 0$ such that there is a unique solution of the differential equation*

$$\frac{dy}{dx} = \mathbf{F}(x, y)$$

on the interval $(a-\delta, a+\delta)$ *satisfying the initial condition* $y(a) = b$.

Proof. To motivate the proof, note first that a function f is a solution of the differential equation with the given initial value condition if and only if

$$f(x) = b \; + \int_a^x \mathbf{F}\,(t, f(t))\; dt$$

where as usual the integral is zero if $x = a$, while if $x < a$ the integral from a to x is defined to be the negative of the integral from x to a.

The idea of the proof is to use the right hand side to define a map of bounded continuous functions and then to apply the Contraction Lemma. However, one needs to be a bit careful in order to specify exactly which sorts of functions form the space upon which the mapping is defined and in order to ensure that the map has the contraction property.

Choose $h, k > 0$ so that

$$S = [a-h, a+h] \times [b-k, b+k] \quad \subset \quad U$$

so that $\mathbf{F}$ and its (first) partial derivatives are bounded on S. Let L be an upper bound for $\mathbf{F}$. By the Mean Value Theorem we have that

$$\left|\mathbf{F}(x, y_1) - \mathbf{F}(x, y_2)\right| \leq \max_{(u,v)\in S} \left(\left| \frac{\partial \mathbf{F}}{\partial y}(u, v) \right| \cdot |y_1 - y_2| \right)$$

for all x, y_1, y_2; let A be the maximum value of the absolute value of the second partial derivative on S.

Choose $\delta > 0$ so that $\delta \leq h$, $L\,\delta < k$ and $A\,\delta < 1$. Define M to be the metric space of all bounded continuous functions g on $(a-\delta, a+\delta)$ for which $|g - b| \leq L\,\delta$, where as usual we identify a real number with the constant function whose value is that number.

For every metric space Z, every $z \in Z$ and every positive real number B, the set of points w with $\mathbf{d}(z, w) \leq B$ is closed (why?), and therefore M is a complete metric space. We need to show that the map

$$[T(g)](x) \quad = \quad b \; + \int_a^x \mathbf{F}\,(t, g(t))\; dt$$

is defined for all $g \in X$, it maps

$$X \subset \mathbf{BC}\,(\;(a-\delta, a+\delta)\;)$$

into itself, and it satisfies the hypothesis of the Contraction Lemma on M.

First of all, it follows immediately that $T(g)$ is continuous whenever g is continuous (fill in the details here). Next, by the boundedness of $\mathbf{F}$ on the closed solid rectangle S we have

$$|T(g) - b| \quad = \quad \left| \int_a^x \mathbf{F}\,(t, g(t))\; dt \right| \quad \leq \quad L \cdot \left| \int_a^x \, dt \right| \quad \leq \quad L\,\delta$$

so that $g \in M$ implies $T(g) \in M$.

Finally, let $g_1, g_2 \in X$ and consider $|T(g_1) - T(g_2)|$. By definition the latter is equal to the least upper bound of the numbers

$$\left| \int_a^x \left(\mathbf{F}\left(t, g_1(t)\right) - \left(t, g_2(t)\right)\right) \, dt \right| \quad \leq$$

$$\int_a^x \left| \mathbf{F}\left(t, g_1(t)\right) - \left(t, g_2(t)\right) \right| dt \quad \leq \quad A\,\delta \cdot |g_1 - g_2| \ .$$

Since $A\,\delta < 1$, all the hypotheses of the Contraction Lemma apply so that there is a unique fixed point, and as noted above this unique fixed point must be the (necessarily unique) solution of the original differential equation with the prescribed boundary condition.∎

Note. One can prove an existence theorem with the weaker hypothesis that $\mathbf{F}$ is continuous (compare Exercise 25 on pages 170–171 of Rudin's book), but uniqueness does not follow. For example, if $\mathbf{F}(x, y) = y^{1/2}$, then the zero function and $x^2/4$ are both solutions to the differential equation with initial condition $y(0) = 0$.

III.4 : Connected spaces

(Munkres, §§ 23, 24, 25)

One of the most basic results in single variable calculus is the *Intermediate Value Theorem*, which states that if a real valued function f is continuous on an interval $J \subset \mathbf{R}$ and $a, b \in J$ are points such that $f(a) \neq f(b)$, then for each real number y between $f(a)$ and $f(b)$ there is a real number x between a and b such that $f(x) = y$.

There are many other situations where one has such a conclusion, and it is useful to have a systematic understanding of when one has such intermediate value results. This requires one to find the abstract concept which underlies the Intermediate Value Theorem, and this notion is called *connectedness*.

Definition. A *separation* of topological space X is a pair of disjoint closed proper subspaces A and B whose union is X. A space is said to be *connected* if it does not have any separations. A space is said to be *disconnected* if it is not connected.

Of course we want intervals in the real line to be connected, but before addressing this point we give two equivalent formulations of the concept of separation.

PROPOSITION. *If X is a topological space, the the following are equivalent:*

(i) One can write $X = A \cup B$ where A and B are nonempty disjoint closed subsets.

(ii) One can write $X = A \cup B$ where A and B are nonempty disjoint open subsets.

(iii) There is a nonempty proper subset $A \subset X$ that is both open and closed (sometimes one says that such a subset is **clopen***).*

Proof. $(i) \Longrightarrow (ii)$ By construction we have $B = X - A$ and $A = X - B$, so that the subsets A and B **are also open in** X.

$(ii) \Longrightarrow (iii)$ The set A is nonempty and it is a proper subset because $B = X - A$ is nonempty. By hypothesis A is closed, and since $A = X - B$ we know that A is also open.

$(iii) \Longrightarrow (i)$ If $B = X - A$, then B is closed since A is open and B is open since A is closed. By hypothesis, A is nonempty, and since it is a proper subset we also know that B is nonempty. The conditions $A \cap B = \emptyset$ and $A \cup B = X$ follow immediately from the definition of B.∎

Connectedness and the real line

There are some immediate examples of spaces that are connected and spaces that are not connected. Every space with at most one point is connected because there are no nonempty proper subsets, and hence there are no subsets for which condition (iii) is meaningful. Every set with an indiscrete topology is connected because there are no nonempty proper subsets that are either open or closed. On the other hand, if a set S has at least two elements and is given the discrete topology, then every subset of S is open (hence by complementation every subset is also closed!), and therefore **every** nonempty proper subset of S will be open and closed.

In particular, the preceding discussion implies that *a topological space with the discrete topology is connected if and only if it contains at most one element.*

Without further discussion we proceed to the single most important family of examples of connected sets.

THEOREM. *Let A be a subset of $\mathbf{R}$ with at least two elements. Then A is connected if and only if for each $a, b \in A$ such that $a < b$ the entire interval $[a, b]$ is contained in A.*

Proof. ($\Longrightarrow$) Suppose that the conclusion is false, so that there is some c satisfying $a < c < b$ and $c \notin A$. Let

$$B = A \cap (c, \infty) = A \cap [c, \infty)$$

where the second equality holds because $c \notin A$. We know that $b \in B$ so that the latter is nonempty, and the two descriptions of B in the displayed formula show it is both open and closed in $\mathbf{R}$. Finally, since $a \in A - B$ we know that B is a proper subset, and therefore A is not connected. We have thus shown the contrapositive of what we wanted to prove (and this proves the latter).

($\Longleftarrow$) Suppose that C is a nonempty open and closed subset of A. Without loss of generality we may as well assume that $a \in C$; if this is false, then a lies in the nonempty open closed subset $A - C$ and we can go through the same argument reversing the roles of C and $A - C$ at each point.

By hypothesis we know that $[a, b] \subset A$. Since C is open in A there is a $\delta > 0$ such that

$$[a, a + \delta) \quad \subset \quad (a - \delta, a + \delta) \ \cap C \ \cap [a, b] \quad \subset \quad C$$

and thus the set

$$K = \{ \ y \in (a, b] \ \Big| \ [a, y] \quad \subset \quad C \ \}$$

is nonempty with an upper bound (specifically, b), Let y^* be the least upper bound of K; we claim that $y^* = b$. In order to do this we need to show that $y^* < b$ is impossible.

By the definition of least upper bound, for each $n > 0$ there is some $y_n \in K$ such that

$$y^* - \frac{1}{n} \quad < \quad y_n \quad \leq \quad y^*$$

and for this sequence we have $y^* = \lim_{n \to \infty} y_n$. Since each of the points of the sequence lies in C and the latter is closed, it follows that $y^* \in K$. Suppose now that $y^* < b$. On the other hand, since C is open there is some $\eta > 0$ such that $y^* + \eta < b$ and

$$(y^* - \eta, y^* + \eta) \quad \subset \quad C \ .$$

The latter in turn implies

$$\left[a, y^* + \frac{\eta}{2}\right] \quad \subset \quad C$$

which means that y^* is NOT an upper bound for K. This contradiction forces the conclusion $y^* = b$.

The preceding argument shows that *if C is a nonempty open and closed subset of A containing a point a and b is another point in A such that $a < b$, then $b \in C$.* — To prove that $C = A$ and hence that A is connected, it suffices to verify that the corresponding statement holds for all $b \in A$ such that $b < a$. The only way this could fail would be if $b \in A - C$. But in this case our argument would imply that $a \in A - C$, which is false. Therefore we have shown that a nonempty open and closed subset of A must be all of A if the latter has the intermediate point property described in the statement of the theorem.∎

COROLLARY. *The connected subsets of* $\mathbf{R}$ *are all given by the following list:*

(1) *Closed intervals* $[a, b]$ *where* $a, b \in \mathbf{R}$.

(2) *Open intervals* (a, b) *where* $a \in \mathbf{R} \cup \{-\infty\}$ *and* $b \in \mathbf{R} \cup \{+\infty\}$.

(3) *Half open intervals* $[a, b)$ *or* $(a, b]$ *where* $a \in \mathbf{R}$ *and* $b \in \mathbf{R} \cup \{+\infty\}$ *in the first case and* $a \in \mathbf{R} \cup \{-\infty\}$ *and* $b \in \mathbf{R}$ *in the second.*∎

It follows that the cardinality of the set of all connected subspaces of $\mathbf{R}$ is equal to the cardinality of $\mathbf{R}$ itself. The verifications of this statement and the corollary are left to the reader as exercises.

The proof of the Intermediate Value Theorem is essentially a combination of the characterization of connected subsets of the real line and the following abstract result, which establishes the conclusion of the Intermediate Value Theorem for arbitrary continuous real valued functions on connected topological spaces:

PROPOSITION. *If* $f : X \to Y$ *is continuous and* X *is connected, then* $f(X)$ *is also connected.*

Proof. Let $C \subset f(X)$ be a nonempty subset that is both open and closed. Then $C = U \cap f(X)$ where U is open in Y and $C = E \cap f(X)$ where E is closed in Y. We then have that

$$f^{-1}(C) = f^{-1}\big(U \cap f(X)\big) = f^{-1}(U)$$

is open in X and

$$f^{-1}(C) = f^{-1}\big(E \cap f(X)\big) = f^{-1}(E)$$

is closed in X. Since C is a nonempty subset of $f(X)$ it follows that $f^{-1}(C)$ is nonempty and therefore by the connectedness of X we have $f^{-1}(C) = X$. In particular, for all $x \in X$ this means that $f(x) \in C$, which in turn means that $C \supset f(X)$; by assumption the reverse inequality holds so that $C = f(X)$. Therefore we have shown that the only nonempty subset of $f(X)$ that is open and closed is $f(X)$ itself, which means that $f(X)$ is connected.∎

Finding connected (sub)sets

The Intermediate Value Theorem for connected spaces is a very powerful statement on the existence of solutions to equations of the form $y = f(x)$, and therefore it is important to recognize when it applies, particularly to subspaces of the plane or other Euclidean spaces. One expects that the theorem will apply to open and closed rectangles that are products of two open or closed intervals respectively. These facts will follow quickly from our abstract discussion below.

PROPOSITION. *Suppose that* X *is a topological space, that* C *is an open and closed subset of* X*, and that* A *is a connected subset of* X*. Then either* $A \subset C$ *or* $A \subset X - C$.

Proof. The intersection $A \times C$ is an open and closed subset of A, so by connectedness of the latter it must either be all of A or empty. In the first case we have either $A \subset C$, and in the second we have $A \subset X - C$.∎

PROPOSITION. *Suppose that* X *is a topological space and that* A *and* B *are connected subsets of* X *with a nonempty intersection. Then* $A \cup B$ *is connected.*

Proof. Suppose that C is a nonempty open and closed subset of $A \cup B$. Without loss of generality we may as well assume that some point $x_0 \in A \subset B$ belongs to C; if instead we have $x_0 \in A \cup B - C$ then we can switch the roles of C and $A \cup B - C$ in the proof of the first case.

Since A is a connected subset of $A \cup B$ we must have $A \subset C$ or $A \subset A \cup B - C$. Since $x_0 \in C$ the latter is impossible, and hence $A \subset C$. If we interchange the roles of A and B in this argument we also conclude that $B \subset C$, so that

$$C \subset A \cup B \subset C$$

which implies that the two sets are equal and hence that $A \cup B$ must be connected.∎

Remark. If A and B are connected subsets of a topological space X it does **NOT** follow that $A \cap B$ is connected. Here is a counterexample when $X = \mathbf{R}^2$:

Let A and B be the semicircles in the unit circle (with equation $x^2 + y^2 = 1$) whose second coordinates are positive and negative respectively. Each subset is connected because the semicircles are the continuous images of the interval $[-1, 1]$ under the continuous mappings

$$\gamma_{\pm}(t) = \left(t, \pm \sqrt{1 - t^2}\right)$$

but the intersection is the pair of points with coordinates $(\pm 1, 0)$ and this set is not connected (it has the discrete topology).

Definition. Given a topological space X and $a, b \in X$, define a binary relation $\sim_{[\mathrm{CONN}]}$ by $a \sim_{[\mathrm{CONN}]} b$ (for $a, b \in X$) if and only if there is a connected subset of X that contains both a and b.

ELEMENTARY FACT. *The preceding binary relation is an equivalence relation, and its equivalence classes are called the* **(connected) components** *of* X.∎

PROPOSITION. *The connected components of X are maximal connected subsets of X.*

Proof. Let A be a connected component of X, and let C be a nonempty open and closed subset of A. Take a to be a point in $C \cap A$.

If $y \in A$, then by definition there is a connected subset A_y of X that contains both a and y. The definition of the equivalence relation implies that $A_y \subset A$. By an earlier result we know that either $A_y \subset C$ or $A_y \subset A - C$. The latter is impossible because $a \in A_y \cap C$, and therefore $A_y \subset C$ for all $y \in A$. Since $y \in A_y$ for all y, this means that $A \subset C$ so that $C = A$ and therefore A must be connected.

To verify maximality, suppose that B is a connected set such that $B \supset A$. Then for each $b \in B$ the set B itself is a connected subset of X that contains a and b, and therefore all points of B are in the same component as a, which is merely A. Therefore $B \subset A$ and A is maximal.∎

In a discrete space the connected components are just the one point subsets and as such they are open and closed. One can ask whether connected components in arbitrary spaces have similar properties. It turns out that these subsets are always closed but not necessarily open. The first of these will be an immediate consequence of the following result.

PROPOSITION. *If X is a topological space and $A \subset X$ is nonempty and connected, then its closure $\overline{A}$ is also connected.*

Since a component is a maximal connected subset, the preceding result shows that a component must be equal to its own closure and therefore must be closed.

Proof of proposition. Let C be a nonempty open and closed subset of $\overline{A}$ It follows that there is some point $y \in \overline{A}$ and since C is open it also follows that either $A \cap C \neq \emptyset$. Since A is connected it follows that $A \subset C$, and since C is closed in $\overline{A}$ it also follows that $\overline{A} \subset C$.∎

COROLLARY. *In the notation of the previous proposition, if $B \subset X$ satisfies $A \subset B \subset \overline{A}$, then B is connected.*

Proof. The proposition implies that Closure (A, B) is connected, and the latter is just the set $\overline{A} \cap B = B$.∎

Example. There are many examples to show that connected components are not necessarily open subsets. In particular, the rational numbers with the subspace topology inherited from the real numbers have this property. — Let $a \in \mathbf{Q}$ and consider the open sets $N^{\mathbf{Q}}_{\sqrt{2}/2n}(a)$. We claim these sets are both open and closed in $\mathbf{Q}$; openness is immediate, and they are closed because a rational number b lies in such a set if and only if $|b - a| \leq \sqrt{2}/2n$ because

$$a \ \pm \ \frac{\sqrt{2}}{2n}$$

is never a rational number. Therefore it follows that if A is the connected component of a in $\mathbf{Q}$ we must have that $A \subset N^{\mathbf{Q}}_{\sqrt{2}/2n}(a)$ for all $n > 0$. But this forces A to be equal to $\{a\}$, which is definitely not open in $\mathbf{Q}$ (each ε-neighborhood contains infinitely may other rational numbers).

Products and connectedness

The next result provides an important tool for recognizing connected subset in Euclidean spaces.

THEOREM. *If X and Y are connected spaces then so is their product $X \times Y$.*

Proof. The result is trivial if either X or Y is empty, so assume that both are nonempty. Let $(x_0, y_0) \in X \times Y$, and let C be the connected component of (x_0, y_0). We shall show that $C = X \times Y$.

For each $(u, v) \in X \times Y$ we have slice homeomorphisms from X and Y to $X \times \{v\}$ and $\{u\} \times Y$ respectively, and therefore the latter subspaces are all connected. This means that for each $(x, y) \in X \times Y$ the points (x, y) and (x_0, y) lie in the same connected component, and similarly the points (x_0, y) and (x_0, y_0) also lie in the same connected component. By the transitivity of the relation of belonging to the same component it follows that (x, y) and (x_0, y_0) also lie in the same connected component, and hence that $C = X \times Y$.∎

COMPLEMENT. *The same conclusion holds for arbitrary finite products.*

In fact, an arbitrary product of connected spaces is connected. An outline of the proof appears in Exercise 10 on page 152 of Munkres.

Proof. This follows from the theorem by induction and the canonical homeomorphism

$$\big(X_1 \times \ \cdots \ \times X_n\big) \times X_{n+1} \cong \prod_{i=1}^{n+1} X_i \ .∎$$

COROLLARY. *For each positive integer n the space $\mathbf{R}^n$ is connected, and for each sequence of closed intervals $[a_i, b_i]$ (where $1 \leq i \leq n$) the product $\prod_i [a_i, b_i]$ is connected.*∎

The results proven thus far have the following noteworthy consequence: *The cardinality of the set of connected subsets of $\mathbf{R}^2$ is the same as the cardinality of the set of all subsets of $\mathbf{R}^2$ (or*

equivalently the cardinality of the set of all subsets of $\mathbf{R}$*).* — To see this, begin by noting that the open rectangular region $(0,1)^2$ is connected by the theorem as is the closed rectangle $[0,1]^2$, and the closed rectangle is the closure of the open rectangle (it is easy to find infinite sequences in the open rectangle that converge to an arbitrary point of the closed rectangle; alternatively, one can use the general rule

$$\overline{A \times B} = \overline{A} \times \overline{B}$$

to show this). Given a subset S of $(0,1) \cong \mathbf{R}$. consider the set

$$C_S = [0,1) \times [0,1] \cup \{1\} \times S \ .$$

By the previous results on closures of connected sets it follows that each set C_S is connected and by construction $C_S \neq C_T$ if $S \neq T$. Therefore there are at least as many connected subsets of $\mathbf{R}^2$ as we have claimed. On the other hand, there are at most as many of these as there are subsets of $\mathbf{R}^2$, and therefore by the Schröder-Bernstein Theorem it follows that the cardinalities are the same.

The preceding yields an example for an assertion made earlier: $\mathbf{R}^2$ *is not homeomorphic to a subset of* $\mathbf{R}$. If A is an arbitrary subset of $\mathbf{R}$, the characterization of connected subsets of the real line shows that the cardinality of the set of connected subsets of A is at most the cardinality of the real numbers themselves, but we know that the cardinality of the set of connected subsets of $\mathbf{R}^2$ is greater than this.

Distinguishing homeomorphism types

Connectedness provides an effective means for showing that certain pairs of spaces are not homeomorphic to each other.

PROPOSITION. *No two sets in the following list are homeomorphic:*

(*i*) *The closed unit interval* $[0,1]$.
(*ii*) *The open unit interval* $(0,1)$.
(*iii*) *The half-open unit interval* $(0,1]$.
(*iv*) *The circle* $S^1 \subset \mathbf{R}^2$ *defined by the equation* $x^2 + y^2 = 1$.

Proof. The first and last sets are compact while the second and third are not, so it suffices to show that S^1 is not homeomorphic to $[0,1]$ and $(0,1)$ is not homeomorphic to $(0,1]$ (Would the result remain true if we added the half-open interval $[0,1)$ to the list? Why or why not?).

Our reasoning relies on the following observation: *If* $f : X \to Y$ *is a homeomorphism and* A *is a finite subset, then* $f|X - A$ *maps* $X - A$ *homeomorphically to* $Y - f(A)$*; in particular,* $X - A$ *is connected if and only if* $Y - f(A)$ *is connected.* If one removes two points from S^1 the resulting space is disconnected (supply the details!), but if one removes the endpoints of $[0,1]$ the resulting space is still connected. Therefore the observation shows that S^1 and $[0,1]$ cannot be homeomorphic.

Similarly, if one removes the endpoint from $[0,1)$ then the resulting space is connected, but if one removes any point from $(0,1)$ the resulting space is disconnected.∎

Further discussion along these lines yields complete topological characterizations of 1-dimensional objects like the unit circle or a closed interval. Textbook discussions of this appear in books by Hocking and Young (*Topology*, Section 2–5 on pages 52–55 with background material in the preceding section) and Christensen and Voxman (*Aspects of Topology* (First Edition), Section 9.A on pages 227–232 with accompanying exercises on page 251, and closely related material in Section 5.A on pages 127–128).

III.5 : Variants of connectedness

(Munkres, §§ 23, 24, 25)

If U is an open subset of some Euclidean space, then U is connected if and only if each pair of points in U can be joined by a broken line curve that lies entirely in U. This fact, which we shall prove at the end of the present section, reflects two important refinements of the concept of connectedness.

Locally connected spaces

Definition. A topological space X is said to be *locally connected* if for each $x \in X$ and each open set $U \subset X$ such that $x \in U$ there is a **connected** open set V such that $x \in V \subset U$.

Example. If U is open in some Euclidean space then U is locally connected. Suppose that $W \subset U$ is an open set and that $x \in W$. Choose $\delta > 0$ so that $N_\delta(x) \subset W$. To see that $N_\delta(x)$ is connected, given y in the latter consider the image J_y of the curve

$$\gamma(t) = x + t(y - x)$$

for $0 \leq t \leq 1$. The set J_y is a connected set containing x and y, and therefore y lies in the same connected component of $N_\delta(x)$ as x, and since y was arbitrary this implies that the set in question is connected.

If we take a disconnected open subet of some Euclidean space, we have an example of a locally connected space that is not connected. Finding examples of spaces that are connected but not locally connected takes more work. Before doing this we state a basic characterization of local connectedness that is established in Theorem 25.3 on page 161 of Munkres.

PROPOSITION. *A topological space X is locally connected if and only if for each open subset U, the components of U are open.*∎

Example. *A connected space that is not locally connected.* Let A be the graph of the function $f(x) = \sin(1/x)$ for $x > 0$, and let $B = \overline{A}$. It follows that B is the union of A with the set $\{0\} \times [-1, 1]$ and that B is connected because A is connected. We claim that B is not locally connected. Let W be the set of all points in B for which the second coordinate lines in $(-1, 1)$; then W is open and therefore it suffices to find a component of W that is not open. By construction W does not contain any points whose first coordinates have the form $2/m\pi$ where m is an odd positive integer. Therefore if C is the connected component in W containing the connected set $\{0\} \times (-1, 1)$, it follows that C cannot have any points whose first coordinates are $\geq 2/m\pi$ for all m and thus that the first components of points in C must be zero. In other words, we must have $C = \{0\} \times (-1, 1)$. This set is not open. In particular, for $\delta > 0$ the set $N_\delta([\textbf{origin}])$ contains points of W that are not in C — specifically all points of the form $(1/k\pi, 0)$ for all k sufficiently large.

The proof of the following result is left to the reader as an exercise:

PROPOSITION. *A finite product of locally connected spaces is locally connected.*

There is an analog for infinite products with a curious extra condition. Namely, an arbitrary product of locally connected spaces is locally connected if and only if all but at most finitely many

of the factors are also connected. To see the need for this condition, note that if $x \in \prod_\alpha X_\alpha$ and U is an open set containing x, then $p_\alpha(U) = X_\alpha$ for all but finitely many α (filling in the details and the proof of the original assertion are left to the reader as exercises).

Path or arcwise connectedness

The connectedness of the real interval leads to an important and useful criterion for recognizing connected spaces.

Definition. A topological space X is said to be *path connected* or *arcwise connected* if for each pair of points $x, y \in X$ there is a continuous function (or curve or path) $\gamma : [a, b] \to X$ such that $\gamma(a) = x$ and $\gamma(b) = y$.

PROPOSITION. *An arcwise connected space is connected.*

Proof. Let $x \in X$, and let C be the connected component of x in X. Given $y \in Y$, let γ be as in the definition, and let J_y be the image of γ. Then J_y is a connected set containing x and y, and consequently we must have $y \in C$. Since $y \in X$ was arbitrary, this means that $X = C$.

Example. The previous example B constructed from the graph of $f(x) = \sin(1/x)$ is a connected space that is not arcwise connected. — To see this, suppose that γ is a continuous curve in B defined on a closed interval $[a, b]$ that joins a point of the form $(0, y_0)$ to a point (x_1, y_1) with $x_1 > 0$. Since $\{0\} \times [-1, 1]$ is compact, it follows that the closed set

$$\gamma^{-1}\big(\{0\} \times [-1, 1]\big) \subset [a, b)$$

is a closed subset of $[a, b]$ and thus has a maximum point $c < b$. At least one of the open sets

$$V_- \;=\; B \cap (\mathbf{R} \times (-1, 1]\,)\,, \qquad V_+ \;=\; B \cap (\mathbf{R} \times [-1, 1)\,)$$

contains $\gamma(c)$ depending upon whether the last coordinate is ± 1 or neither so choose V_ε to be such an open subset. Then there is a $\delta > 0$ such that $\delta < b - c$ and $0 \leq \delta$ implies that $\gamma(c + t) \in V_\varepsilon$. It follows that for all but finitely many positive integers n there are points $t_n \in (0, \delta)$ such that the first coordinate of $\gamma(t_n)$ is equal to $2/n\pi$; specifically, pick any value $t_0 \in (0, \delta)$, so that the first coordinate η_0 of $\gamma(t_0)$ will be positive, and then notice that for all sufficiently large n we have

$$\frac{2}{n\pi} \;<\; t_0\ .$$

However, by construction V_- does not have any points whose first coordinates have the form $2/n\pi$ where n is an integer of the form $4k + 3$ (the values of x for which $\sin(1/x) = -1$), and V_+ does not have any points whose first coordinates have the form $2/n\pi$ where n is an integer of the form $4k + 1$ (the values of x for which $\sin(1/x) = +1$). Therefore it is not possible to construct a continuous curve in B joining a point with zero first coordinate to a point with positive first coordinate, and therefore the connected set B is not arcwise connected.

Analogies with connectedness

There is a concept of *path component* or *arc(wise) component* that is analogous to the concept of connected component.

Definition. Given a topological space X and $x, y \in X$, define a binary relation $\sim_{[\mathrm{ARC}]}$ by $x \sim_{[\mathrm{ARC}]} y$ (for $a, b \in X$) if and only if there is a continuous function $\gamma : [a, b] \to X$ such that $\gamma(a) = x$ and $\gamma(b) = y$. We often say that γ is a curve joining x and y if this condition holds.

It is obvious that this relation is reflexive (use the constant curve) and symmetric (consider the curve δ defined on $[-b, -a]$ with $\delta(t) = \gamma(-t)$. To see that the relation is transitive it is convenient to introduce a concept that arises frequently in mathematics.

Definition. Let X be a topological space, and suppose that $\gamma : [a, b] \to X$ and $\delta : [c, d] \to X$ are continuous curves such that $\gamma(b) = \delta(c)$. The *sum* or *concatenation* (stringing together) $\gamma + \delta$ is the continuous curve

$$\sigma : [a, b + d - c] \to X$$

defined by $\sigma(t) = \gamma(t)$ if $t \in [a, b]$ and $\sigma(t) = \delta(t + b - c)$ if $t \in [b, b + d - c]$. An illustration of this appears in the course directory as files of the form `concat.*`.

By construction the sum of curves is an associative operation whenever it is defined.

Transitivity of $\sim_{[\mathrm{ARC}]}$ follows immediately from this definition, for if γ joins x and y and δ joins y and z, then $\gamma + \delta$ joins x and z.

Definition. The equivalence classes of $\sim_{[\mathrm{ARC}]}$ are called *path components* or *arc components* of X. It follows that every arc component is arcwise connected, but the preceding example shows that arc components need not be closed or open in the ambient space X.

The statements of the next two results are parallel to those for connected spaces, but the proofs are entirely different.

PROPOSITION. *If X is an arcwise connected space and $f : X \to Y$ is continuous, then $f(X)$ is arcwise connected.*

Proof. Given $a, b \in f(X)$ write $a = f(c)$ and $b = f(d)$ for $c, d \in X$. Since X is arcwise connected there is a continuous curve γ in X joining c to d, and the composite $f \circ \gamma$ is a continuous curve in $f(X)$ joining a to b.∎

PROPOSITION. *A (finite) product of arcwise connected spaces is arcwise connected.*

In fact, the finiteness condition is completely unnecessary in the statement and proof of this result.

Proof. Let $X_1, \cdots, X_n$ be the spaces in question, let $u, v \in \prod_i X_i$ and for each j between 1 and n let u_j and v_j be the j^{th} coordinates of u and v respectively. Then for each j one can join u_j to v_j by a continuous curve γ_j. Suppose that γ_j is defined on the closed interval $[a_j, b_j]$ and let $L_j : [0, 1] \to [a_j, b_j]$ be the unique linear function that sends 0 to a_j and 1 to b_j. Then there is a unique continuous function $\alpha : [0, 1] \to \prod_i X_i$ whose projection onto the j^{th} coordinate is $\gamma_j \circ L_j$ for each j, and by construction α joins a to b.∎

There is also a corresponding notion of local path or arcwise connectedness (see Munkres, page 161), and as noted in Theorem 25.5 on that page (the proof continues to page 162), if a space is locally arcwise connected then its components and path components are identical.

Open subsets of Euclidean spaces

The discussion of local connectedness for open sets in Euclidean spaces actually proves that such sets are locally arcwise connected and hence their components, which are open, are the same

as their arc components. In particular, an open subset of a Euclidean space is connected if and only if it is arcwise connected, and the discussion at the beginning of this section asserts that one can choose the curves in question to be of a special type. In order to prove this we need to give a formal definition of a broken line curve.

Definition. A *closed line segment curve* in $\mathbf{R}^n$ is a continuous curve γ defined on $[0,1]$ by an equation of the form

$$\gamma(t) = a + t \cdot (b - a)$$

for some $a, b \in \mathbf{R}^n$. A *broken line curve* is a finite iterated concatenation of closed line segment curves.

PROPOSITION. *Let U be open in $\mathbf{R}^n$. Then U is connected if and only if every pair of points in U can be joined by a broken line curve that lies entirely in U.*

Proof. If the conclusion is true then U is arcwise connected. To prove the ($\Longrightarrow$) implication, define a binary relation $\sim$ by $u \sim v$ if and only if there is a broken line curve in U joining u to v. This is an equivalence relation, and in fact the equivalence classes are open subsets (if $x \in U$ and $N_\delta(x) \subset U$, then every point in $N_\delta(x)$ can be joined to x by a closed line segment curve). It follows that the union of any family of equivalence classes is also open, and in particular, if W is an equivalence class this means that $U - W$, which is the union of all the equivalence classes except W, is also open. The latter implies that W is closed, and since U is connected it follows that there can be only one equivalence class for the equivalence relation described above; this proves that each pair of points in U can be joined by a broken line curve in U.■

Given an open connected subset of $\mathbf{R}^n$ one can ask many different questions about the continuous curves joining two arbitrary points in U, including the following: *Given two points in U, can they be joined by a curve whose coordinate functions are infinitely differentiable? If so, can one find such a function such that the tangent vector at every point is nonzero?*

The answer to both questions is yes, but the proofs are more complicated than the ones given above. Such results can be established using techniques from an introductory graduate course on smooth manifolds.