METRIC SPACES

These notes introduce the concept of a *metric space*, which will be an essential notion throughout this course and in others that follow. Some of this material is contained in optional sections of the book, but I will assume none of that and start from scratch. Still, you should check the corresponding sections in the book for a possibly different point of view on a few things.

The main idea to have in mind is that a metric space is some kind of generalization of $\mathbb{R}$ in the sense that it is some kind of "space" which has a notion of "distance". Having such a "distance" function will allow us to phrase many concepts from real analysis—such as the notions of *convergence* and *continuity*—in a more general setting, which (somewhat) surprisingly makes many things actually easier to understand.

Metric Spaces

Definition 1. A *metric* on a set X is a function $d : X \times X \to \mathbb{R}$ such that

- $d(x, y) \geq 0$ for all $x, y \in X$; moreover, $d(x, y) = 0$ if and only if $x = y$,
- $d(x, y) = d(y, x)$ for all $x, y \in X$, and
- $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$.

A *metric space* is a set X together with a metric d on it, and we will use the notation (X, d) for a metric space. Often, if the metric d is clear from context, we will simply denote the metric space (X, d) by X itself.

Example 1. The set of real numbers $\mathbb{R}$ with the function $d(x, y) = |x - y|$ is a metric space. More generally, let $\mathbb{R}^n$ denote the Cartesian product of $\mathbb{R}$ with itself n times:

$$\mathbb{R}^n = \{(x_1, \ldots, x_n) \mid x_i \in \mathbb{R} \text{ for each } i\}.$$

The function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

is a metric on $\mathbb{R}^n$, called the *Euclidean metric*. When $n = 1, 2, 3$, this function gives precisely the usual notion of distance between points in these spaces. These will be the standard examples of metric spaces. In particular, whenever we talk about the metric spaces $\mathbb{R}^n$ without explicitly specifying the metrics, these are the ones we are talking about.

The observation above that the given metric on $\mathbb{R}^n$ gives the usual notion of distance is what is going on in general: a metric d on a set X should be viewed as giving a notion of "distance" between elements of X.

Indeed, thinking of d as a "distance function" makes the three properties given in the definition of a metric clear: the first says that the distance between two "points" is always larger than or equal to 0, and the only way the distance between two "points" can be zero is if the two "points" are actually the same; the second says that the "distance" from x to y is the same as the "distance" from y to x; and the third says that the distance from x to y is always less than or equal to the sum of the distances from x and y to some intermediate "point" z.

To encourage thinking about metric spaces in this way, we will often refer to the metric d as the *distance function*, and to the elements of X as *points* of X. The third property in the definition of

a metric will be called the *triangle inequality* since in the case of $\mathbb{R}^2$ it says exactly that the length of one side of a triangle is less than or equal to the sum of the lengths of the other two sides.

For many purposes, the example of $\mathbb{R}^2$ with the usual distance function is precisely the one you should have in mind when thinking about metric spaces in general. Indeed, all pictures we draw which are meant to illustrate some property of a general metric space will be a two-dimensional picture drawn on paper or on the board—i.e. a picture in $\mathbb{R}^2$. Keep in mind, however, that we do this only to have some intuition for how to think about metric spaces in general, but that anything we prove about metric spaces must be phrased solely in terms of the definition of a metric itself.

Example 2. As we said, the standard example of a metric space is $\mathbb{R}^n$, and $\mathbb{R}$, $\mathbb{R}^2$, and $\mathbb{R}^3$ in particular. However, we can put other metrics on these sets beyond the standard ones.

Define $d_1 : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ by

$$d_1((x_1, y_1), (x_2, y_2)) = \max\{|x_1 - x_2|, |y_1 - y_2|\}.$$

Then d_1 is a metric on $\mathbb{R}^2$ called the "box" metric. Let us show that this is actually a metric.

For any points $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$, $|x_1 - x_2| \geq 0$ and $|y_1 - y_2| \geq 0$, so the maximum $d_1((x_1, y_1), (x_2, y_2))$ of these is larger than or equal to zero as well. Also, $d_1((x_1, y_1), (x_2, y_2)) = 0$ if and only if both $|x_1 - x_2| = 0$ and $|y_1 - y_2| = 0$, which means that $x_1 = x_2$ and $y_1 = y_2$. Thus $d_1((x_1, y_1), (x_2, y_2)) = 0$ if and only if $(x_1, y_1) = (x_2, y_2)$, so d_1 satisfies the first requirement in the definition of a metric on $\mathbb{R}^2$.

For any $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$, since $|x_1 - x_2| = |x_2 - x_1|$ and $|y_1 - y_2| = |y_2 - y_1|$, the maximum of $|x_1 - x_2|$ and $|y_1 - y_2|$ is the same as the maximum of $|x_2 - x_1|$ and $|y_2 - y_1|$, so $d_1((x_1, y_1), (x_2, y_2)) = d_1((x_2, y_2), (x_1, y_1))$, which is the second requirement.

Finally, suppose $(x_1, y_1), (x_2, y_2), (x_3, y_3) \in \mathbb{R}^2$. Then

$$|x_1 - x_2| \leq |x_1 - x_3| + |x_3 - x_2| \text{ and } |y_1 - y_2| \leq |y_1 - y_3| + |y_3 - y_1|$$

by the triangle inequality for the absolute value function on $\mathbb{R}$. Also,

$$|x_1 - x_3| + |x_3 - x_2| \leq \max\{|x_1 - x_3|, |y_1 - y_3|\} + \max\{|x_3 - x_2|, |y_3 - y_2|\}$$

and similarly for $|y_1 - y_3| + |y_3 - y_2|$. Thus both $|x_1 - x_2|$ and $|y_1 - y_2|$ are smaller than or equal to $d_1((x_1, y_1), (x_3, y_3)) + d_1((x_3, y_3), (x_2, y_2))$, so their maximum $d_1((x_1, y_1), (x_2, y_2))$ is as well. This is the triangle inequality for d_1, so we conclude that d_1 is a metric on $\mathbb{R}^2$.

To get a feel for this metric, take $(1, 2), (0, 0) \in \mathbb{R}^2$. The usual distance (with respect to the standard metric on $\mathbb{R}^2$) between these is $\sqrt{1 + 2^2} = \sqrt{5}$. However, the "distance" between them with respect to the box metric d_1 is

$$d_1((1, 2), (0, 0)) = \max\{|1 - 0|, |2 - 0|\} = 2.$$

So, in particular, the distance from a point (x, y) to the origin $(0, 0)$ is the larger of $|x|$ and $|y|$. In general, the distance between two points with respect to this metric is the length of the longest side of the rectangle with one corner at the first point and the other corner at the second—this is where the name "box metric" comes from.

Using the same basic definition, only taking the maximum of more terms, the box metric can be generalized to give a metric on any $\mathbb{R}^n$.

Example 3. Here is another metric on $\mathbb{R}^2$. Define the function $d_2 : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ by

$$d_2((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + |y_1 - y_2|.$$

Then d_2 is a metric (you should verify this!) on $\mathbb{R}^2$ called the "taxicab" metric.

The name comes from the following picture. Say you were in a taxicab in a city where the streets were laid out in a grid like pattern, so that the cab could only move straight, to the left, right, or backwards, but not "diagonally". Then the distance between two points with respect to the taxicab metric is the distance the cab would have to travel to get from one point to the other.

As for the box metric, the taxicab metric can be generalized to $\mathbb{R}^n$ for any n.

The above two nonstandard metric spaces show that "distance" in this setting does not mean the usual notion of distance, but rather the "distance" as determined by the metric. Again, to emphasize, we think of this as a "distance" since it satisfies the same sorts of conditions (the ones given in the definition of a metric) that the usual notion of distance does.

Example 4. Let X be any set, and define the function $d : X \times X \to \mathbb{R}$ by

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y. \end{cases}$$

Then it is straightforward to check (do it!) that d is a metric on X, called the *discrete* metric. Here, the distance between any two distinct points is always 1.

Example 5. After the standard metric spaces $\mathbb{R}^n$, this example will perhaps be the most important. First, recall that a function $f : X \to \mathbb{R}$ from a set X to $\mathbb{R}$ is *bounded* if there is some $M \in \mathbb{R}$ such that $|f(x)| \leq M$ for all $x \in X$. In other words, this says that the set $\{f(x) \mid x \in X\}$ of values of f is a bounded subset of $\mathbb{R}$. Note that because of this, the set of values of a bounded function has a supremum as a consequence of the completeness axiom for $\mathbb{R}$.

Let $C_b([a, b])$ denote the space of real-valued, bounded functions on the closed interval $[a, b]$:

$$C_b([a, b]) := \{f : [a, b] \to \mathbb{R} \mid f \text{ is bounded}\}.$$

The function d given by

$$d(f, g) = \sup\{|f(x) - g(x)| \mid x \in [a, b]\}$$

defines a metric on $C_b([a, b])$ called the *supremum* (or *sup* for short) metric. Again, you should try to verify on your own that this is indeed a metric.

A few remarks are in order. First, if f and g are bounded, then so is $f-g$, so the supremum used in the definition of d actually exists in $\mathbb{R}$. Thus the definition of d makes sense. Now, here is what this definition is saying: take two bounded functions f and g, and look at the distance between $f(x)$ and $g(x)$ over all possible x—the distance between f and g is defined to be the "largest" such distance. (I put the word "largest" in quotation marks since the supremum defining the sup metric is not necessarily in the set $\{|f(x) - g(x)| \mid x \in [a, b]\}$ itself, but you should intuitively think of it as the largest such value.) Also, note that the same definition works for bounded functions on all of $\mathbb{R}$. The notation

$$\sup_{x \in [a,b]} |f(x) - g(x)|$$

for the supremum used in the definition of $d(f, g)$ is also commonly used.

Eventually, it will help to try to "visualize" what this metric "looks like"—we will talk about this in class as we go on. As a start, consider the functions $f(x) = \cos x$ and $g(x) = -\cos x$ on $\mathbb{R}$. Draw their graphs and ask yourself: what is the largest possible distance between the values of f and g at points $x \in \mathbb{R}$? From the picture it should be easy to see that the answer is 2, say when $x = 0$. Because of this, 2 is indeed the distance between f and g with respect to the sup metric; the sup metric will give us a way to tell how "close" or how "far apart" two functions are from each other in terms of their graphs.

The upshot is that everything we learn about metric spaces in general will be applicable to each example given above—in particular to the "space" $C_b([a,b])$ whose "points" are functions $f : [a,b] \to \mathbb{R}$.

Definition 2. Let (X,d) be a metric space. For any $r > 0$ and $x \in X$, the *r-ball* or *ball of radius r* around p in X is the set:

$$B_r(x) := \{y \in X \mid d(x,y) < r\}$$

of points of X whose "distance" from x is less than r.

The name "ball of radius r" comes from the fact that in the case of $\mathbb{R}^2$ with the Euclidean metric, the r-ball around a point indeed looks like a ball (or disk) of radius r centered at that point. In the case of $\mathbb{R}$, note that $B_r(x) = (x-r, x+r)$ is the interval of radius r around x.

Example 6. The ball $B_1((0,0))$ of radius 1 in $\mathbb{R}^2$ centered at the origin with respect to the box metric is the region enclosed by the square with corners $(-1,-1)$, $(-1,1)$, $(1,-1)$, and $(1,1)$, but not including the square itself.

With respect to the taxicab metric, $B_1((0,0))$ looks like the region enclosed by a diamond.

Example 7. Let X be a set with the discrete metric. For any $x \in X$, the ball of radius 1 around x is simply the set containing only x. Indeed, the points y in this ball are those satisfying the condition that $d(x,y) < 1$, and by the definition of the discrete metric this is only possible if $d(x,y) = 0$ in which case $y = x$. Hence $B_1(x) = \{x\}$, and more generally you can convince yourself that $B_r(x) = \{x\}$ for any $0 < r < 1$.

Since every $y \in X$ satisfies the condition that $d(x,y) \leq 1$, for any $r > 1$ every $y \in X$ satisfies $d(x,y) < r$, so the r-ball around x in this case is all of X.

Example 8. Let $f : [0,1] \to \mathbb{R}$ be the constant function 0: i.e. $f(x) = 0$ for all $x \in [0,1]$. Let us determine what the ball of radius 1 around f in $C_b([0,1])$ "looks like".

A function $g \in C_b([0,1])$ is in this ball exactly when

$$d(f,g) = \sup_{x \in [0,1]} |f(x) - g(x)| < 1.$$

Since $f(x) = 0$ for all x, this is the same as requiring that

$$\sup_{x \in [0,1]} |g(x)| < 1.$$

Thus the ball of raduis 1 around $f = 0$ consists of those functions such that $|g(x)| < 1$ for all $x \in [0,1]$. (Actually, this is not quite right, but we will skip this for now and come back to it later once we are more comfortable with this space.)

Here's the idea: draw a "tube" of radius 1 around the graph of the function $f = 0$, so this is just a "tube" of radius 1 around the x-axis. Then the ball of radius 1 around $f = 0$ (essentially) consists of those functions whose graphs lie completely within this tube. In this way, we can "picture" what this ball looks like.

Definition 3. Let (X,d) be a metric space. A *subspace* of X is a subset $Y \subseteq X$ of X with the metric obtained by restricting the one on X to Y. This just means that we take the same notion of distance as we have in X but only allow ourselves to plug in points of Y.

Example 9. We can view the set of rational numbers $\mathbb{Q}$ as a subpsace of $\mathbb{R}$ with the standard Euclidean metric $d(x, y) = |x - y|$. Note the following: the ball of radius 1 in $\mathbb{R}$ around $r \in \mathbb{Q}$ is the interval $(r - 1, r + 1)$, but the ball of radius 1 *in* $\mathbb{Q}$ around $r \in \mathbb{Q}$ is $(r - 1, r + 1) \cap \mathbb{Q}$; i.e. the set of rational numbers in in the interval $(r - 1, r + 1)$. The point is that the notion of "r-ball" depends on what the "big" space is: if we are working inside the metric space $\mathbb{Q}$, then such an r-ball only consists of rational numbers since those are the only such elements of our big space; working inside $\mathbb{R}$, there are more elements in the "big" space and so an r-ball in $\mathbb{R}$ will contain more numbers than an r-ball in $\mathbb{Q}$.

Similarly, view $(-1, 1)$ as a subspace of $\mathbb{R}$. Then the ball of radius 2 in $(-1, 1)$ around 0 is $(-1, 1)$ itself. Yes, the ball of radius 2 in $\mathbb{R}$ around 0 is $(-2, 2)$, but these extra points do not exist in the "big" space $(-1, 1)$ we are working in.

Definition 4. A subset U of a metric space (X, d) is *bounded* if there exists a positive radius $r > 0$ and a point $x \in X$ such that $U \subseteq B_r(x)$.

In other words, a subset is bounded if there is a ball of a large enough radius which contains it; the point at which this ball centered does not really matter since if a subset is bounded, then around *any* point you can find a large enough ball which contains that subset. Indeed, suppose that U is bounded and contained in $B_r(x)$ for some $x \in X$ and $r > 0$. Then for any other point $x' \in X$, the ball of radius $r + d(x, x')$ around x' will also contain U; in particular, this ball contains $B_r(x)$ itself, which already contains U. Again, the point is that in the definition of bounded, neither the radius nor the center point really matter—what matters is that U is contained in some ball of *finite* radius.

Sequences

A *sequence* in a metric space X is an infinite list

$$x_1, x_2, x_3, \ldots$$

of points in X. (More formally, a sequence in X can be defined as a function $f : \mathbb{N} \to X$, but we will not need this point of view here.) Note that the terms in the sequence do not have to be distinct; in particular, for any $x \in X$ we can talk about the constant sequence

$$x, x, x, \ldots.$$

We will use the notation (x_n) for the sequence whose n-th term is x_n.

Definition 5. A sequence (x_n) in a metric space (X, d) is said to *converge* to $x \in X$ if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$d(x_n, x) < \epsilon \text{ for } n \geq N.$$

A sequence is *convergent* if it converges to something, and the point to which it converges is called the *limit* of the sequence; we will use the notation $(x_n) \to x$ to mean that the sequence (x_n) converges to x. The notation

$$\lim_{n \to \infty} x_n = x$$

to denote that x is the limit of the convergent sequence (x_n) is also common. A sequence which is not convergent is *divergent*.

Before looking at examples, we say a few words about what the definition above is actually saying. Intuitively, a sequence (x_n) converges to x if the terms of x_n are getting closer and closer to x as n increases. The given definition says this in a precise way.

First, consider any $\epsilon > 0$. The definition says that there is some index N so that starting at the index and beyond, the terms in the sequence are within a distance ϵ away from x. As ϵ gets smaller, maybe the index you have to start at to ensure this will get larger, but still starting at some index the terms in (x_n) are within that distance away from x.

Another way to say this is the following. Recall that the condition $d(x_n, x) < \epsilon$ says that x_n is in the ϵ-ball around x. Starting off the definition of convergence by saying "for every $\epsilon > 0$" says that we are going to consider an arbitrary ball around x. From this point of view, the definition says that given any ball around x, starting at some index the terms in the sequence are in that ball. Again the key is that ϵ here can get arbitrarily small: no matter how small a ball you start with around x, eventually the terms in (x_n) will be in there. This is precisely what it means to say that the terms in a sequence are getting closer and closer to the limit.

Example 10. In any metric space, a constant sequence $x, x, x, \ldots$ converges to x. (Show this!)

Example 11. This is a standard elementary example. Consider the sequence $(\frac{1}{n})$ in $\mathbb{R}$ with the standard metric. Intuitively the terms of this sequence are getting closer and closer to 0—let us prove that the sequence indeed converges to 0.

Let $\epsilon > 0$. We want to find an index N where for $n \geq N$ we have

$$\frac{1}{n} < \epsilon.$$

It is the Archimedian Property that tells us what N to choose: pick $N \in \mathbb{N}$ such that $\frac{1}{N} < \epsilon$. Then if $n \geq N$ we have

$$\frac{1}{n} \leq \frac{1}{N} < \epsilon.$$

We conclude that the sequence $(\frac{1}{n})$ converges to 0.

Check the book for other such examples of convergent sequences in $\mathbb{R}$. Let us now look at some nonstandard examples of metric spaces:

Example 12. Suppose that (X, d) is a discrete metric space. We claim that the only convergent sequences in X are those which are "eventually constant": i.e. those sequences (x_n) for which there is some element $x \in X$ and $N \in \mathbb{N}$ such that $x_n = x$ for $n \geq N$. In other words, a sequence is eventually constant if if terms starting at some index and beyond are the same.

It should not be hard to show that an eventually constant sequence converges—the proof is similar to the one showing that constant sequences always converge. Now suppose that (x_n) is a sequence in X which converges to $x \in X$. Applying the definition of convergence to $\epsilon = 1$, we know that there is some index N such that

$$d(x_n, x) < 1 \text{ for } n \geq N.$$

But according to the definition of the discrete metric on a set, the distance between two points is either 1 or 0, so the only way such a distance can be less than 1 is if it is actually 0. Thus the statement above becomes the statement that

$$d(x_n, x) = 0 \text{ for } n \geq N.$$

By the first property in the definition of a metric, $d(x_n, x) = 0$ if and only if $x_n = x$, so we find that $x_n = x$ for $n \geq N$, and thus a convergence sequence in a discrete space is eventually constant.

Example 13. Consider the sequence (f_n) in $C_b(\mathbb{R})$—the space of real-valued, bounded functions on $\mathbb{R}$ equipped with the sup metric—defined by

$$f_n(x) = \frac{1}{n}\sin x.$$

To clarify, since the "points" of $C_b(\mathbb{R})$ are actually bounded functions on $\mathbb{R}$, the terms in the sequence (f_n) are themselves such functions. We are defining the n-th term in this sequence to be the function f_n defined by the above expression. Note that each such f_n is indeed bounded since $\sin x$ is, so each f_n is in fact in $C_b(\mathbb{R})$.

We claim that (f_n) converges to 0, where 0 now denotes the constant zero function $0(x) = 0$. To see this, let $\epsilon > 0$. We want to find an index N where

$$d(f_n, 0) < \epsilon \text{ for } n \geq N.$$

According to the definition of the sup metric, this condition means we want

$$\sup_{x\in\mathbb{R}} |f_n(x) - 0(x)| = \sup_{x\in\mathbb{R}} \left|\frac{1}{n}\sin x\right| < \epsilon \text{ for } n \geq N.$$

Since the maximum of $|\sin x|$ is 1, it is not hard to see that

$$\sup_{x\in\mathbb{R}} \left|\frac{1}{n}\sin x\right| = \frac{1}{n}.$$

At this point we see again that the Archimedean Property will give us what we want: choose $N \in \mathbb{N}$ such that $\frac{1}{N} < \epsilon$. Then for $n \geq N$:

$$d(f_n, 0) = \sup_{x\in\mathbb{R}} |f_n(x) - 0(x)| = \sup_{x\in\mathbb{R}} \left|\frac{1}{n}\sin x\right| = \frac{1}{n} \leq \frac{1}{N} < \epsilon,$$

showing that (f_n) converges to 0 in $C_b(\mathbb{R})$.

To gain some more intuition for this, draw a picture of what the graphs of the functions f_n look like: these look like the graph of $\sin x$ only getting smaller as n increases because of the $\frac{1}{n}$ coefficient. As n gets larger and larger, it is clear that these graphs are approaching the x-axis, which we can view as the graph of the constant zero function.

In all the examples above, the sequences we considered only converged to one thing—this is not an accident. Note that this is something we have to prove since there is nothing in the definition of a convergent sequence which says that limit must be unique.

Proposition 1. *In any metric space, limits of convergent sequences are unique.*

Thoughts. We prove this the usual way we show something is unique: assume a convergent sequence (x_n) has two limits x and y and show that the two limits are the same. To show that $x = y$, we must think in an "analytic" way—in other words, how do we show two things are the same using inequalities? Note that in this case we cannot show that $x \leq y$ and $y \leq x$, since this only makes sense in $\mathbb{R}$; in a general metric space, it is not possible to compare elements directly using inequalities.

The only thing we have to work with in a general metric space is the metric d itself, so the question now becomes: how do we show two elements of a space are the same using only the metric? The answer is now clear: we show that $d(x, y) = 0$, and then the first definiing property of a metric

will tell us that $x = y$. Now again we have to think a bit: it may not be possible to show directly that $d(x, y) = 0$, so we have to find an "analytic" way to show this. But we know how to do this too: if we can show that $d(x, y) < \epsilon$ for *every* $\epsilon > 0$, it must be the case that $d(x, y) = 0$.

So, the end goal is now to show that if a convergent sequence (x_n) in a metric space has two limits x and y, then $d(x, y) < \epsilon$ for all $\epsilon > 0$. To do this, we find a way to bound $d(x, y)$ by things we know how to bound. In other words, here we use the triangle inequality to say that:

$$d(x, y) \leq d(x, x_n) + d(x_n, y)$$

for any n. The point is that now using the assumptions that (x_n) converges to x and (x_n) converges to y, we can bound these two latter terms; in particular, if we make each latter term smaller than $\frac{\epsilon}{2}$, then their sum will be smaller than ϵ—this is an instance of the so-called "$\frac{\epsilon}{2}$-trick", which we will be seeing a lot of. Let us now give the final proof.

Proof of Proposition. Suppose that (X, d) is a metric space and that the sequence (x_n) in X converges to x and y. Let $\epsilon > 0$. Since $(x_n) \to x$, there exists an index $N_1 \in \mathbb{N}$ such that

$$d(x_n, x) < \frac{\epsilon}{2} \text{ for } n \geq N_1.$$

Since $(x_n) \to y$, there exists $N_2 \in \mathbb{N}$ such that

$$d(x_n, y) < \frac{\epsilon}{2} \text{ for } n \geq N_2.$$

Let $N = \max\{N_1, N_2\}$. Then both inequalities above hold, and so the triangle inequality gives

$$d(x, y) \leq d(x, x_N) + d(x_N, y) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Thus for any $\epsilon > 0$, $d(x, y) < \epsilon$. This implies that $d(x, y) = 0$, which in turn implies that $x = y$. We conclude that the limit of a convergent sequence is unique. □

Here is another basic fact:

Proposition 2. *A convergent sequence in a metric space (X, d) is bounded, by which we mean that the set whose elements are the terms in the sequence is a bounded subset of X.*

Proof. Let (x_n) be a convergent sequence in X, and suppose that x is its limit. We want to show that the set $\{x_n \mid n \in \mathbb{N}\}$ is contained in some ball of finite radius. Since $(x_n) \to x$, there exists $N \in \mathbb{N}$ such that

$$d(x_n, x) < 1 \text{ for } n \geq N.$$

Thus the terms in the sequence for $n \geq N$ are contained in the ball of radius 1 around x.

Now we make the radius larger so that the resulting ball will contain the terms $x_1, x_2, \ldots, x_{N-1}$ as well. Set $r = 1 + \max\{d(x_1, x), \ldots, d(x_{N-1}, x), 1\}$. Since $r \geq 1$, r is positive, and we claim that the ball of radius r around x contains all the terms in the sequence (x_n). Indeed, as we said before, x_n is in this ball for $n \geq N$ since it was already contained in the smaller ball of radius 1 around x, and x_i for $i = 1, \ldots, N - 1$ is contained in this ball since x_i is contained in the possibly smaller ball of radius $1 + d(x_i, x)$ around x. We conclude that (x_n) is bounded. □

The special case that sequences of real numbers are bounded is proven in the book, but the above now shows that this holds in any metric space. The key is that we know that past some index the terms in the sequence are within a distance 1 away from x, so we only need to pick a large enough radius so that the terms before this are also in the ball of this radius; the way we did this—by constructing the radius using the maximum of a finite collection of positive numbers—is a common technique that we will see again.

Check the book for proofs of other properties of convergent sequences in $\mathbb{R}$, such as the fact that if (x_n) and (y_n) are sequences of real numbers converging to x and y respectively, then the sequence $(x_n + y_n)$ converges to $x + y$ and the sequence $(x_n y_n)$ converges to xy. Both facts use the "$\frac{\epsilon}{2}$-trick"; the second is tougher to prove, but illustrates some important techniques. You can also find a proof of this latter fact in my "Worked Examples" handout listed on the course website for Math 104 in Summer 2010, where I also say a bit about the motivation behind the proof.

For good measure, we should also give an example showing how to prove that a sequence is *divergent* using the definition of convergent itself:

Example 14. Consider the sequence (x_n) in $\mathbb{R}$ where $x_n = 2 + (-1)^n$. We claim that this sequence does not converge; to be precise, we claim that no real number satisfies the definition of what it means for (x_n) to converge to x.

To say that $x \in \mathbb{R}$ does not satisfy this definition means the following: there exists $\epsilon > 0$ such that for every $N \in \mathbb{N}$ there exists $n \geq N$ such that

$$d(x_n, x) \geq \epsilon.$$

Indeed, this is obtained by *negating* the definition of convergence: negating "for every $\epsilon > 0$" gives "there exists $\epsilon > 0$", negating "there exists $N \in \mathbb{N}$" gives "for every $N \in \mathbb{N}$", negating "for $n \geq N$" gives "there exists $n \geq N$", and finally negating "$d(x_n, x) < \epsilon$" gives "$d(x_n, x) \geq \epsilon$". Being able to negate statements—i.e. write down precisely what it means for a statement to be false—will be a useful thing to know how to do.

Let $x \in \mathbb{R}$. We claim that $\epsilon = 1$ satisfies the above requirement of what it means for (x_n) to *not* converge to x. Indeed, now let $N \in \mathbb{N}$. If $x < 2$, let n be an even integer larger than N. Then $x_n = 3$ and $d(x_n, x) > 1$. If $x \geq 2$, let n be an odd integer larger than N. Then $x_n = 1$ and $d(x_n, x) \geq 1$. Thus in either case, we have found a term in the sequence beyond the N-th one whose distance from x is at least 1, showing that x does not satisfy the definition of what it means for (x_n) to converge to x. Since $x \in \mathbb{R}$ was arbitrary, we conclude that (x_n) converges to no real number, so it diverges.

We emphasize that whenever we are asking whether or not a sequence converges, the "big" space we are working in is crucial. For example, consider the sequence (x_n) defined by $x_n = 2 - \frac{1}{n}$. As a sequence in $\mathbb{R}$, this converges to 2. However, since each $x_n \in (0, 2)$, we can also view this as a sequence in the subspace $(0, 2)$ of $\mathbb{R}$. In this subspace, however, (x_n) does *not* converge since the thing to which it should converge is not in the space $(0, 2)$. Thus, (x_n) is a convergent sequence in $\mathbb{R}$ but is a *divergent* sequence in $(0, 2)$. The moral is: limits of convergent sequences should actually exist in the space we are working in.

Let us give a final fact, which will be important later on. Note that the denseness of both $\mathbb{Q}$ and $\mathbb{R}\backslash\mathbb{Q}$ in $\mathbb{R}$ is crucial here—indeed, this proposition will be the idea behind the more general notion of "denseness" we will discuss later on:

Proposition 3. *Given any real number $x \in \mathbb{R}$, there exists a sequence of rationals converging to x and a sequence of irrationals converging to x.*

Proof. For each $n \in \mathbb{N}$, choose a rational number r_n such that

$$x - \frac{1}{n} < r_n < x$$

and an irrational number y_n such that

$$x - \frac{1}{n} < y_n < x.$$

(This is possible by the denseness of $\mathbb{Q}$ and $\mathbb{R}\backslash\mathbb{Q}$ in $\mathbb{R}$ respectively.) From this we get a sequence of rational numbers (r_n) and a sequence of irratinoal numbers (y_n) which we claim converge to x.

Indeed, let $\epsilon > 0$ and choose $N \in \mathbb{N}$ such that $\frac{1}{N} < \epsilon$. Then for $n \geq N$,

$$d(r_n, x) = |r_n - x| < \frac{1}{n} \leq \frac{1}{N} < \epsilon$$

and similarly $d(y_n, x) < \epsilon$. Thus $(r_n) \to x$ and $(y_n) \to x$ as claimed. □

To clarify a previous comment, note that if $x \in \mathbb{R}$ were irrational, the above proposition gives a sequence of rationals (r_n) converging to it. This sequence of rationals then would *not* converge in $\mathbb{Q}$ since the point to which they converge to in $\mathbb{R}$ is not in $\mathbb{Q}$. So, we would say that (r_n) is a convergent sequence in $\mathbb{R}$ but a divergent sequence in $\mathbb{Q}$.

Definition 6. A *subsequence* of a sequence (x_n) in a metric space X is a sequence (x_{n_k}) in X consisting of terms of the sequence (x_n) such that $n_k > n_{k'}$ if $k > k'$.

The final condition simply means that the terms in the subsequence occur in the same order as they did in the original sequence, so another way to state the above definition is the following: a subsequence of (x_n) is a sequence (x_{n_k}) of terms from (x_n) which occur in the same order as they did in (x_n). For example, given the sequence (n):

$$1, 2, 3, 4, \ldots$$

in $\mathbb{R}$, the sequence $(2n) = 2, 4, 6, \ldots$ is a subsequence, but the sequence $4, 2, 6, 8, \ldots$ is not since the first two terms do not occur in the same order as they did in (n).

The key property of a subsequence is the following:

Proposition 4. *Let (x_n) be a convergent sequence in a metric space X. Then any subsequence of (x_n) converges to the same limit as (x_n).*

Proof. Say that (x_n) converges to $x \in X$. Suppose that (x_{n_k}) is a subsequence of (x_n) and let $\epsilon > 0$. Since $(x_n) \to x$, there exists $N_1 \in \mathbb{N}$ such that

$$d(x_n, x) < \epsilon \text{ for } n \geq N_1.$$

Choose $N \in \mathbb{N}$ large enough so that $n_k \geq N_1$ for $k \geq N$; in other words, choose an index N for the subsequence large enough so that the N-term in the subsequence is beyond the N_1-th term in the original sequence. Then for $k \geq N$, $n_k \geq N_1$ so

$$d(x_{n_k}, x) < \epsilon.$$

We conclude that (x_{n_k}) converges to x. □

As an upshot, whenever we have a sequence in a metric space with two subsequences which converge to different things (or with a divergent subsequence), then the original sequence does not converge. This can be quite useful in practice: for instance, since the subsequence $(x_{2n} = 3)$ of the sequence $(x_n = 2 + (-1)^n)$ of a previous example converges to 3 and the subsequence $(x_{2n+1} = 1)$ converges to 1, the sequence (x_n) itself diverges. Note that this is much simpler than the way we previously proved this fact, where we used the precise statement of what it means for a sequence to diverge.

Cauchy Sequences

Definition 7. A sequence (x_n) in a metric space (X, d) is a *Cauchy sequence* if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$d(x_n, x_m) < \epsilon \text{ for all } n, m \geq N.$$

It is important to compare this with the definition of convergence for a sequence. In that case, we are talking about the sequence converging to some given point—in other words, to prove that a sequence converges using the definition of convergence we must already have a guess as to what the sequence should converge to. In a sense, we must already know what the limit is going to be.

The point here is that there is no such restriction in the definition of a Cauchy sequence: the definition makes no mention of any other point, only the terms of the sequence itself. Intuitively, the definition says that given some postive distance, eventually (i.e. past some index) the terms in the sequence will all be within that distance apart from each other, no matter how small a distance we started with.

Example 15. Recall that for any real number, there is a sequence of rational numbers converging to it—this is a consequence of the denseness of $\mathbb{Q}$ in $\mathbb{R}$. In particular, there is a sequence of rationals (r_n) converging to $\sqrt{2}$. To be specific, consider the decimal expansion of $\sqrt{2}$:

$$\sqrt{2} = 1.414\ldots$$

Define the sequence (r_n) by setting

$$r_1 = 1,\ r_2 = 1.4,\ r_3 = 1.41,\ r_4 = 1.414$$

and continuing in this manner, taking one more digit in the decimal expansion of $\sqrt{2}$ at a time. Note that each r_n is indeed rational since it has a finite decimal expansion, and that (r_n) then does converge to $\sqrt{2}$.

In fact, (r_n) is a Cauchy sequence. Indeed, let $\epsilon > 0$. Since $(r_n) \to \sqrt{2}$, there exists $N \in \mathbb{N}$ such that

$$|r_n - \sqrt{2}| < \frac{\epsilon}{2} \text{ for } n \geq N.$$

Then if $n, m \geq N$, we have

$$|r_n - r_m| \leq |r_n - \sqrt{2}| + |\sqrt{2} - r_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

(The first inequality is the triangle inequality.) This shows that (r_n) is a Cauchy sequence as claimed.

The above fact is not surprising—in fact, convergent sequences are *always* Cauchy, and the proof uses the same $\frac{\epsilon}{2}$-trick as the example:

Theorem 1. *Any convergent sequence in a metric space (X, d) is Cauchy.*

Proof. Suppose that (x_n) is a convergent sequence in X, and let x be its limit. Let $\epsilon > 0$ and choose $N \in \mathbb{N}$ so that

$$d(x_n, x) < \frac{\epsilon}{2} \text{ for } n \geq N.$$

If $n, m \geq N$, we have

$$d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

so (x_n) is Cauchy as claimed. □

Intuitively, since the terms of a convergent sequence are getting closer and closer to its limit, clearly the terms should be getting closer and closer to each other as well. This is what the above theorem says.

Now, we can ask about the converse: must a Cauchy sequence be convergent? After all, if the terms in a sequence are getting closer and closer to each other (i.e. the sequence is Cauchy), it is not crazy to think the terms are then getting closer and closer to some fixed point. The answer to this boils down to something we saw before: it depends on what space we are actually working in. For instance, the sequence of rationals (r_n) of the previous example converges to $\sqrt{2}$ in $\mathbb{R}$, but it does *not* converge in $\mathbb{Q}$ since the number it should converge to—namely $\sqrt{2}$—is not in $\mathbb{Q}$. This is then an example of a Cauchy sequence in $\mathbb{Q}$ (the definition of Cauchy only uses the terms of the sequence itself, so it does not really matter what space we are working in) which is not a convergent sequence in $\mathbb{Q}$.

The moral is: Cauchy sequences are precisely the sequences which "should" converge, only that the "point" to which a Cauchy sequence should converge to may not actually be in the given metric space.

Definition 8. A metric space (X, d) is said to be *complete* if every Cauchy sequence in X converges in X.

The "converges in X" part is the crucial point. So, $\mathbb{Q}$ is not complete, neither is $\mathbb{R}\backslash\mathbb{Q}$ (a sequence of irrationals converging to a rational number in $\mathbb{R}$ will be Cauchy but not convergent in $\mathbb{R}\backslash\mathbb{Q}$), and you should be able to convince yourselves that open intervals (a, b) are also not complete.

The standard example of a complete metric space is the following:

Theorem 2. *The set of real numbers $\mathbb{R}$ with the standard metric is complete. More generally, $\mathbb{R}^n$ with the standard metric is complete.*

A proof of this fact is given in the book. You can find another proof in the "Worked Examples" handout from my Summer 2010 course. After we talk about the notion of a "closed" subset of $\mathbb{R}$, we will be able to show that closed intervals $[a, b]$ are also complete. Another key example is the following:

Theorem 3. *The space $C_b([a, b])$ of bounded, real-valued functions on the interval $[a, b]$ is complete.*

The same is true if we replace $[a, b]$ by all of $\mathbb{R}$ itself. The proof of this theorem (which is not easy the first time you see it) uses the completeness of $\mathbb{R}$ in a crucial way: starting with a Cauchy sequence in the space of bounded functions, we first use the completeness of $\mathbb{R}$ to construct the function which should be the limit of the given Cauchy sequence, and then we show that it is indeed the limit. The proof is given in Section 25 of the book, only it is not phrased as a statement about completeness of the metric space $C_b([a, b])$. We will omit the proof here, and talk about it when we come to that part of the book.

Here is a final basic property of Cauchy sequences, analogous to a previously-mentioned property of convergent sequences:

Proposition 5. *Any Cauchy sequence in a metric space (X, d) is bounded.*

Proof. The proof is essentially the same as that for convergent sequences. Suppose (x_n) is a Cauchy sequence in X, and choose $N \in \mathbb{N}$ such that

$$d(x_n, x_m) < 1 \text{ for } n, m \geq N.$$

In particular then, x_n is in the ball of radius 1 around x_N for $n \geq N$. Thus, all terms in the sequence are in the ball of radius

$$r = 1 + \max\{d(x_1, x_N), \ldots, d(x_{N-1}, x_N), 1\} > 0$$

around x_N, so (x_n) is bounded as claimed; to emphasize, x_n for $n \geq N$ is in the smaller ball of radius 1 around x_N, and x_i for $i = 1, \ldots, N-1$ is in the possibly smaller ball of radius $1 + d(x_i, x_N)$ around x_N. □

As we have seen, metric spaces are not necessarily complete, and in such spaces Cauchy sequences do not necessarily converge—the main example being $\mathbb{Q}$ with the standard metric. In this example, however, there is a larger metric space—namely $\mathbb{R}$—in which every Cauchy sequence in $\mathbb{Q}$ *does* now converge. We can ask whether the same thing holds in general: given a noncomplete metric space X, does there always exist a complete metric space containing X as a subspace?

The answer is yes(!), and the "smallest" such complete space $\overline{X}$ is called the *completion* of X. For example, the completion of $\mathbb{Q}$ with respect to the standard metric indeed turns out to be $\mathbb{R}$ with the standard metric. We will not discuss this notion further here, nor say precisely what we mean by the "smallest" complete metric space containing a given noncomplete one. Roughly, given a noncomplete space X, we consider the set of all possible Cauchy sequences in X and *define* the completion $\overline{X}$ of X to essentially be this set of Cauchy sequences itself! To make this into a "space", we would then have to define what we mean by the distance between two Cauchy sequences, and show that with such a metric $\overline{X}$ is actually complete. To do everything carefully would take us quite a while, and for a first course in analysis the payoff is not so important. Still, you should be aware that this is an important construction in more advanced applications of analysis, although ones you probably would not see in an undergraduate course.

Open and Closed Sets

Now we begin talking about the *topology* of a metric space. "Topology" is a term which has a precise meaning in mathematics, but we will not discuss this in full generality in this course; for us, "topology" will simply mean "things having to do with open and closed sets". Here are the two main definitions:

Definition 9. A subset U of a metric space (X, d) is *open* in X if for any $x \in U$ there exists $r > 0$ such that $B_r(x) \subseteq U$. A subset $A \subseteq X$ is *closed* in X if whenever (a_n) is a sequence of points in A converging to some $a \in X$, then $a \in A$.

Let us think about these two definitions more carefully. The definition of open says the following: for any $x \in U$, we can find a ball around x which is fully contained in U. The idea is that if $x \in U$, then any point "close enough" to x is also in U. Note that we put no restriction on how large or small the radius r of the ball $B_r(x)$ around x has to be, simply that such a radius exists. Intuitively,

a set U is open if moving a short distance away from a point in U keeps you inside U; in other words, a set is open if it "surrounds" all of its points.

The definition of closed says this: if (a_n) is a convergent sequence in X whose terms happen to belong to A, then the point to which this sequence converges is also in A. In other words, A is closed under the process of taking limits; intuitively, a point which is "close" to A—in the sense that there are points of A arbitrarily close to it—is actually *in* A.

Example 16. As trivial examples, in any metric space (X, d), both the empty set $\emptyset$ and X itself are open *and* closed in X. Indeed, to say that $\emptyset$ is open in X means that around any $x \in \emptyset$ there should be a ball which is fully contained in $\emptyset$—but there is no such $x \in \emptyset$ on which to check this condition! Thus, the empty set trivially satisfies the definition of open, so it is open in X. On the other hand, for any $x \in X$, then *any* ball around x is contained in X simply because any ball around x by definition consists of elements from the "big" space X itself. So, the defining property of open is trivially satisfied by any $x \in X$, so X is open in itself.

Similarly, to say that $\emptyset$ is closed means that for any sequence of points in $\emptyset$ converging to something in X, the point the sequences converges to should be in $\emptyset$—but again there are no sequences in $\emptyset$ on which to check this condition, so $\emptyset$ is closed in X. If (x_n) is any convergent sequence in X, then the point it converges to is certainly in X simply because this is part of what it means to say that (x_n) is a "convergent sequence in X". Thus, X is closed in itself.

Example 17. Consider $\mathbb{R}$ with the standard metric. We claim that any open interval (a, b) is open in $\mathbb{R}$ and any closed interval $[a, b]$ is closed in $\mathbb{R}$. This is good, since otherwise using the terms "open" for open intervals and "closed" for closed intervals would conflict with the way we are now using these terms! Also, let's go ahead and assume that $a < b$ in the open interval case and $a \leq b$ in the closed interval case since otherwise these intervals would be empty.

To show that (a, b) is open, pick any $x \in (a, b)$. We want to show that there is some ball (which in the case of $\mathbb{R}$ with the standard metric means open interval) around x which is fully contained in (a, b). Drawing a picture of the interval (a, b) and the point x in it should easily let you see that such a ball exists, but let us makes this precise. Since $a < x < b$, both $x - a$ and $b - x$ are positive; set $r = \min\{x - a, b - x\}$, which is then also positive. We claim that the ball of radius r around x:

$$B_r(x) = \{y \in \mathbb{R} \mid d(x, y) < r\} = \{y \in \mathbb{R} \mid |x - y| < r\} = (x - r, x + r),$$

is fully contained in (a, b). Indeed, if $y \in (x - r, x + r)$, then

$$a \leq x - r < y < x + r \leq b$$

where the first inequality follows from the fact that $r \leq x - a$ and the last from the fact that $r \leq b - x$. Hence $y \in (a, b)$ for any $y \in B_r(x)$, so $B_r(x) \subseteq (a, b)$ and thus (a, b) is open in $\mathbb{R}$.

To see that $[a, b]$ is closed, take any convergent sequence (x_n) in $\mathbb{R}$ whose terms belong to $[a, b]$. Let $x \in \mathbb{R}$ be the limit of this sequence. Since $a \leq x_n \leq b$ for all n, Exercise 8.9 from our book shows that $a \leq x \leq b$. Thus $x \in [a, b]$ and hence $[a, b]$ is closed in $\mathbb{R}$.

Example 18. Now, we can ask the following: can (nonempty) open intervals (a, b) also be *closed* subsets of $\mathbb{R}$, and can (nonempty) closed intervals $[a, b]$ be *open* subsets of $\mathbb{R}$? It is easy to see that the answer to both of these is no—in particular, $\left(a + \frac{b-a}{n+1}\right)$ is a sequence in (a, b) which converges to a real number a not in (a, b), so (a, b) is not closed in $\mathbb{R}$, and there is no ball around $a \in [a, b]$ which is fully contained in $[a, b]$, so $[a, b]$ is not open in $\mathbb{R}$.

However, we note that (a, b) *is* closed in (a, b) itself and $[a, b]$ *is* open in $[a, b]$ itself. Indeed, a previous example showed that any metric space is both closed and open in itself. The point is that

the notion of something being open or closed is a *relative* one, meaning we only talk about a set being open or closed *in* something; for example, it does not make sense to ask whether or not the interval $(2, 3]$ is open or closed—it does, however, make sense to ask whether or not it is open or closed in $\mathbb{R}$. (It is neither.)

To clarify then, the sequence $(a + \frac{b-a}{n+1})$ used above to show that (a, b) is not closed in $\mathbb{R}$ does not work to show that (a, b) is not closed in (a, b) since it does not converge in the "big" space (a, b) we are working in. Similarly, thinking of $[a, b]$ as the "big" metric space, there is a ball around a which is fully contained in $[a, b]$ simply because a ball in $[a, b]$ can only consist of points from $[a, b]$; for example, the ball of radius 1 around 1 in $[1, 3]$ is $[1, 2)$, and does not contain any number less than 1 because the "big" space $[1, 3]$ does not.

Example 19. Consider the metric space $\mathbb{Q}$ with the standard metric and let S be the set of rational numbers between $-\sqrt{2}$ and $\sqrt{2}$:

$$S := \left(-\sqrt{2}, \sqrt{2}\right) \cap \mathbb{Q}.$$

We claim that S is both open and closed in $\mathbb{Q}$. (Note that S is neither open nor closed in $\mathbb{R}$; you should try to show this precisely.)

Indeed, suppose that (r_n) is a sequence of points in S converging to some $r \in \mathbb{Q}$. Since

$$-\sqrt{2} < r_n < \sqrt{2} \text{ for all } n,$$

the limit r satisfies $-\sqrt{2} \leq r \leq \sqrt{2}$. But since r is rational, we actually know that $-\sqrt{2} < r < \sqrt{2}$ and thus $r \in S$. This shows that S is closed in $\mathbb{Q}$. For any $r \in S$, the ball of radius $\epsilon = \min\{\sqrt{2} - r, r + \sqrt{2}\} > 0$ around r consists of those elements $x \in \mathbb{Q}$ such that $|x - r| < \epsilon$. Using the definition of ϵ it is not hard to show that this ball is contained in S itself. (Note that all we are doing is taking the smaller of the distances of r to the left endpoint $-\sqrt{2}$ and to the right endpoints $\sqrt{2}$ of S; the interval of this radius around r is then clearly smaller than the interval from $-\sqrt{2}$ to $\sqrt{2}$.) This shows that S is open in $\mathbb{Q}$.

We will see later that, besides the empty set and $\mathbb{R}$ itself, there are no other subsets of $\mathbb{R}$ which are both open and closed in $\mathbb{R}$—this is what it will mean to say that $\mathbb{R}$ is *connected*. We will speak about connectedness in more detail later, and indeed give a more geometric definition. The previous example shows that $\mathbb{Q}$ is not connected, it is *disconnected*.

The fact that open intervals are open in $\mathbb{R}$ is not surprising—it generalizes to other metric spaces in the following way:

Proposition 6. *Let (X, d) be a metric space. Then for any $x \in X$ and $r > 0$, the ball $B_r(x)$ of radius r around x is open in X.*

Proof. Let $y \in B_r(x)$. To show that $B_r(x)$ is open in X, we must show that there is some ball $B_s(y)$ in X of some radius $s > 0$ around y which is fully contained in $B_r(x)$. Indeed, we claim that the ball of radius $s = r - d(x, y)$ around y will work. (Draw a picture of what these balls look like in $\mathbb{R}^2$; $B_r(x)$ is a disk around x of radius r and $B_s(y)$ is a disk around $y \in B_r(x)$ of radius $s = r - d(x, y)$. From a well-drawn picture, it should be intuitively clear that the disk $B_s(y)$ is contained in the disk $B_r(x)$.)

First, since $y \in B_r(x)$, we know that $d(x, y) < r$ since this is the defining inequality for points in $B_r(x)$. Thus $s = r - d(x, y) > 0$. To show that $B_s(y) \subseteq B_r(x)$, let $p \in B_s(y)$. Then $d(p, y) < s$. We claim that $d(p, x) < r$, and if we can show this we are done since then $p \in B_r(x)$ as required. By the triangle inequality, we have

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

as needed. We conclude that $B_r(x)$ is open in X as claimed. □

The fact that closed intervals $[a, b]$ are closed in $\mathbb{R}$ generalizes to the fact that *closed* balls in a general metric space (X, d), which are subsets of the form

$$\{y \in X \mid d(x, y) \leq r\},$$

are always closed in X. Note that the difference between a closed ball and a usual ball (which we might now call an *open* ball) is that in the usual definition of ball the inequality $d(x, y) < r$ is strict whereas we allow a non-strict inequality in the definition of a closed ball.

Many other examples of open and closed sets in metric spaces can be constructed based on the following facts:

Theorem 4. *Let (X, d) be a metric space. The union of any collection open sets in X is open in X, and the intersection of finitely many open sets in X is open in X.*

Proof. Suppose that $\{U_\alpha\}$ is a collection of subsets of X with each U_α open in X, and let x be in their union $\bigcup U_\alpha$. Then there is some β such that $x \in U_\beta$. Since U_β is open in X, there is some ball $B_r(x)$ around x which is contained in U_β. Since U_β itself is contained in $\bigcup U_\alpha$, $B_r(x)$ is contained in $\bigcup U_\alpha$ as well and we thus conclude that $\bigcup U_\alpha$ is open in X.

Suppose now that $U_1, \ldots, U_n$ is a finite collection of open subsets of X, and let x be in their intersection $U_1 \cap \cdots \cap U_n$. Then $x \in U_i$ for each $i = 1, \ldots, n$. Since each U_i is open in X, there is a ball $B_{r_i}(x)$ of some radius $r_i > 0$ such that $B_{r_i}(x) \subseteq U_i$. Set

$$r = \min\{r_1, \ldots, r_n\},$$

which is positive since each r_i is. The ball $B_r(x)$ is then contained in each $B_{r_i}(x)$, which is in turn contained in U_i. Thus $B_r(x) \subseteq U_1 \cap \cdots \cap U_n$, showing that $U_1 \cap \cdots \cap U_n$ is open in X. □

Note that the intersection of infinitely many open sets is not necessarily open: for example, the intersection of all open intervals of the form $\left(-\frac{1}{n}, \frac{1}{n}\right)$ for $n \in \mathbb{N}$ is the set $\{0\}$ consisting only of zero, and is thus not open in $\mathbb{R}$ even though each such interval is open in $\mathbb{R}$. The proof of the theorem above does not work in this example since we can no longer define r as the minimum of a finite number of radii. Instead, we would have to use something like

$$r = \inf\{r_1, r_2, r_3, \ldots\},$$

which could now be zero even though each r_i is positive, and then the "ball" of radius $r = 0$ around x would actually be empty.

We have the following analog of the above theorem for closed sets:

Theorem 5. *Let (X, d) be a metric space. The intersection of any collection of closed sets in X is closed in X, and the union of finitely many closed sets in X is closed in X.*

You should try to prove this on your own for practice. Again, the finiteness in the second statement is important: for instance, the union of the infinite number of closed intervals of the form $\left[-1 + \frac{1}{n}, 1 - \frac{1}{n}\right]$ for $n \in \mathbb{N}$ is the interval $(-1, 1)$, which is not open in $\mathbb{R}$.

The above results suggest a deep connection between the notions of open and closed sets, and indeed we see that the two concepts are essentially the opposites of each other in the following way:

Theorem 6. *Let (X, d) be a metric space. A subset U of X is open in X if and only if the complement $X \backslash U$ is closed in X; a subset A of X is closed in X if and only if the complement $X \backslash A$ is open in X.*

Proof. Suppose that U is open in X and let (x_n) be a sequence of points in $X\backslash U$ converging to $x \in X$. To show that $X\backslash U$ is closed in X we must show that $x \in X\backslash U$. By way of contradiction, suppose that $x \in U$. Since U is open in X, there is a ball $B_\epsilon(x)$ of radius $\epsilon > 0$ around x contained in U. For this specific ϵ, there is some index N so that

$$d(x_n, x) < \epsilon \text{ for } n \geq N$$

since $(x_n) \to x$. But this is not possible since such x_n would be in $B_\epsilon(x) \subset U$ and $X\backslash U$ at the same time. Thus we must have $x \in X\backslash U$ and hence $X\backslash U$ is closed in X.

Conversely suppose that $X\backslash U$ is closed in X and let $x \in U$. To show that U is open in X we must show that there is some radius $r > 0$ so that $B_r(x) \subseteq U$. Suppose instead that no such radius existed. Then in particular, for each $n \in \mathbb{N}$ the ball of radius $\frac{1}{n}$ around x is not contained in U: $B_{\frac{1}{n}}(x) \nsubseteq U$. Thus for each n, there is some $x_n \in B_{\frac{1}{n}}(x)$ such that $x_n \notin U$. This gives a sequence (x_n) of points in $X\backslash U$ such that

$$d(x_n, x) < \frac{1}{n} \text{ for all } n,$$

which implies that $(x_n) \to x$. Since $X\backslash U$ is closed in X and each $x_n \in X\backslash U$, the limit x must then be in $X\backslash U$, contradicting the choice of $x \in U$. We conclude that there is a ball around x contained in U, so U is open in X as claimed.

The second part follows directly from the first using the fact that the complement of the complement of a set is that set itself: in other words, apply the first part to $U := X\backslash A$ and use the fact that $X\backslash(X\backslash A) = A$. □

So, the complements of open sets are closed and the complements of closed sets are open. This gives another way of seeing, for instance, that a closed interval $[a, b]$ is closed in $\mathbb{R}$: the complement of $[a, b]$ in $\mathbb{R}$ is $(-\infty, a) \cup (b, \infty)$, which is open since it is the union of open sets.

Example 20. Suppose that (X, d) is a discrete metric space. For any $x \in X$, we saw in Example **??** that the ball of radius 1 around x is the set containing only x itself: $B_1(x) = \{x\}$. This implies that for any $x \in X$, the set $\{x\}$ is open in X, and since the arbitrary union of open sets is open, it follows that *any* subset of a discrete space is open in that space. As a result of the previous theorem, any subset of a discrete space is also closed in that space.

Closure, Interior, Boundary, and Denseness

The conditions given in the definitions of open and closed sets can be singled out as follows:

Definition 10. Let S be a subset of a metric space X. A point $s \in S$ is said to be an *interior point* of S if there exists $r > 0$ such that $B_r(s) \subseteq S$. The set of all interior points of S is called the *interior* of S in X and is denoted by $\operatorname{int} S$. A point $x \in X$ is said to be a *limit point* of S if there exists a sequence (x_n) in S converging to x. The set of all limit points of S is called the *closure* of S in X and is denoted by $\overline{S}$.

Since by definition the interior of S only consists of points from S, it is always true that $\operatorname{int} S \subseteq S$. The requirement that $S \subseteq \operatorname{int} S$ (and hence $\operatorname{int} S = S$) is *precisely* what it means for S to be open. Similarly, for any $s \in S$, the constant sequence $s, s, s, s, \ldots$ is a sequence in S converging to s so $s \in \overline{S}$ and thus it is always true that $S \subseteq \overline{S}$. The requirement that $\overline{S} \subseteq S$ is *precisely* what it means for S to be closed. We summarize this:

Proposition 7. *A subset S of a metric space X is open in X if and only if $S \subseteq \operatorname{int} S$ (and hence $\operatorname{int} S = S$); S is closed in X if and only if $\overline{S} \subseteq S$ (and hence $\overline{S} = S$).*

Again, this is just restating the definitions of open and closed sets in terms of the notions of interior and limit points: S is open if every point in S is an interior point of S and S is closed if every limit point of S is in S.

Example 21. Consider $\mathbb{R}^2$ with the standard metric and let D be any disk of radius $r > 0$ centered at $p \in \mathbb{R}^2$; D either contains no piece of its boundary circle (in which case D is an open disk), it contains all of its boundary circle (in which case D is a closed disk), or it contains only part of its boundary circle (in which case D is neither open nor closed in $\mathbb{R}^2$). Then regardless of whether D is open, closed, or neither, the interior of D in $\mathbb{R}^2$ is the open disk of radius r centered at p and the closure of D in $\mathbb{R}^2$ is the closed disk of radius r centered at p. (Draw some pictures to convince yourselves that, at least intuitively, this is right!)

In the previous example, the interior of a disk D in $\mathbb{R}^2$ turned out to open in $\mathbb{R}^2$ and indeed was the largest open subset of $\mathbb{R}^2$ contained in D, and the closure turned out to be closed in $\mathbb{R}^2$ and moreover was the smallest closed subset of $\mathbb{R}^2$ containing D. This generalizes as follows:

Proposition 8. *Let (X, d) be a metric space and let S be a subset of X. Then $\operatorname{int} S$ is open in X and $\overline{S}$ is closed in X.*

Proof of First Claim. Let $x \in \operatorname{int} S$. We want to show there is some ball $B_r(x)$ around x contained in $\operatorname{int} S$. Since x is an interior point of S, we know there exists $r > 0$ so that $B_r(x) \subseteq S$. We claim that any point of $B_r(x)$ is actually an interior point of S, so that $B_r(x) \subseteq \operatorname{int} S$, proving our claim.

So, let $y \in B_r(x)$. Since $B_r(x)$ is open in X, there is some radius $s > 0$ so that $B_s(y) \subseteq B_r(x)$. But since $B_r(x) \subseteq S$, the smaller ball $B_s(y)$ is also contained in S, so y is an interior point of S as required. We conclude that $\operatorname{int} S$ is open in X. □

The proof that $\overline{S}$ is closed in X for any $S \subseteq X$ will be left to the homework. The tricky thing you have to think about is the following: $\overline{S}$ is formed by taking S and throwing in all the limits of convergent sequences in S—we are now claiming that if we take a sequence in *this* new space which converges to some $x \in S$, that x itself will be in $\overline{S}$.

So, suppose that (x_n) is a sequence in $\overline{S}$, meaning that each term x_n in this sequence is itself the limit of some sequence in S. If (x_n) converges to some $x \in X$, we then have to show that x is itself the limit of some sequence in S—this is kind of hard to think about at first: we know that x is the limit of the sequence (x_n) in $\overline{S}$, but from this we have to construct in a sequence in S (not $\overline{S}$) converging to x. I'll outline how to do this on the homework.

Proposition 9. *Let (X, d) be a metric space and let S be a subset of X. Then $\operatorname{int} S$ is the largest open subset of X contained in S in the sense that if U is any other open subset of X contained in S, then $U \subseteq \operatorname{int} S$. Also, $\overline{S}$ is the smallest closed subset of X containing S in the sense that if A is any other closed subset of X containing S, then $\overline{S} \subseteq A$.*

We will not prove this since it will not be so important, but it gives a nice interpretation of the interior and closure of a set. If you think about it, the above can be rephrased as follows:

Proposition 10. *Let (X, d) be a metric space and let S be a subset of X. Then $\operatorname{int} S$ is the union of all open subsets of X contained in S and $\overline{S}$ is the intersection of all closed subsets of X containing S:*

$$\operatorname{int} S = \bigcup_{U \text{ open in } X \text{ such that } U \subseteq S} U \quad \text{and} \quad \overline{S} = \bigcap_{A \text{ closed in } X \text{ such that } S \subseteq A} A.$$

We point out that as with the notions of open and closed sets, the notions of interior and limit points are *relative* ones in the sense that they depend on what the "big" metric space is.

Example 22. Consider $\mathbb{Q}$ as a subspace of $\mathbb{R}$ with the standard metric. Then $\mathbb{Q}$ has empty interior in $\mathbb{R}$. Indeed, no rational number $r \in \mathbb{Q}$ can be an interior point of $\mathbb{Q}$ in $\mathbb{R}$ since *any* interval in $\mathbb{R}$ around r will contain an irrational number due to the denseness of $\mathbb{R}\backslash\mathbb{Q}$ in $\mathbb{R}$. On the other hand, we saw in Proposition **??** that, due to the denseness of $\mathbb{Q}$ in $\mathbb{R}$, given any $x \in \mathbb{R}$ there exists a sequence in $\mathbb{Q}$ converging to x. This means that the closure of $\mathbb{Q}$ in $\mathbb{R}$ is all of $\mathbb{R}$. Similarly, you can show by a similar argument that the set of irrational numbers $\mathbb{R}\backslash\mathbb{Q}$ has empty interior in $\mathbb{R}$ and that the closure of $\mathbb{R}\backslash\mathbb{Q}$ in $\mathbb{R}$ is $\mathbb{R}$.

Now, instead view $\mathbb{Q}$ as a subspace of itself with the standard metric. Since "open balls" are now allowed to consist only of rational numbers, *any* ball in $\mathbb{Q}$ around a rational number will be contained in $\mathbb{Q}$, so any point of $\mathbb{Q}$ is an interior point of $\mathbb{Q}$ in $\mathbb{Q}$. Thus the interior of $\mathbb{Q}$ as a subspace of itself is $\mathbb{Q}$. Similarly, one can check that the closure of $\mathbb{Q}$ in $\mathbb{Q}$ is $\mathbb{Q}$. Indeed, viewing any metric space X as a subset of itself, the interior of X in X is X and the closure of X in X is X.

Definition 11. A subset S of a metric space X is *dense* in X if its closure in X is all of X.

Unwinding this definition gives the following concrete characterization: S is dense in X if for any $x \in X$ there is a sequence of points in S converging to x. The idea is that if S is dense in X, then points of X can be "approximated" arbitrarily well by points in S.

Example 23. We have already used the term "dense" before when saying that $\mathbb{Q}$ is dense in $\mathbb{R}$ or that $\mathbb{R}\backslash\mathbb{Q}$ is dense in $\mathbb{R}$. As Proposition **??** shows, "dense" in the previous sense indeed means dense in this new sense; in other words, $\mathbb{Q}$ is dense in $\mathbb{R}$ since $\overline{\mathbb{Q}} = \mathbb{R}$ and similarly for the irrationals.

In fact, to make the connection between this new notion of denseness and the previous way in which we used "dense" clearer, we have the following:

Proposition 11. *A subset S of a metric space X is dense in X if and only if every nonempty open set in X contains an element of S.*

The proof is given in the "Worked Examples" handout from Summer 2010. The idea is that if S is dense in X, then you can find points of S "everywhere" in X.

Example 24. Let $C([a, b])$ denote the space of *continuous* real-valued functions on $[a, b]$. (Admittedly, we have not defined "continuous" yet, but you can back to this example after we have done so.) This is a metric space with the sup metric—we will see later that the sup metric makes sense in this setting since such functions are always bounded. Then the set of polynomial functions on $[a, b]$ is dense in $C([a, b])$. This is known as the "Weierstrass Approximation Theorem" and is Theorem 27.5 in the book; the statement is book is not phrased in terms of "denseness" but we will see that this is exactly what the precisely means.

Definition 12. Let S be a subset of a metric space X. A point $x \in X$ is a *boundary point* of S if for any $r > 0$, the ball $B_r(x)$ of radius r around x contains an element of S and an element of $X\backslash S$. The set of all boundary points of S is called the *boundary* of S and is denoted by ∂S.

So, to say that x is a boundary point of S means that there are points of S arbitrarily close to x and there are points of the complement of S arbitrarily close to x. Intuitively, this is what the "boundary" of S should "look" like, at least in the following example:

Example 25. Let S be any subset of $\mathbb{R}^2$. The boundary of S is then precisely what you normally visualize as the boundary of S drawn in the plane. For example, the boundary of a disk (open, closed, or neither) is precisely the boundary circle. In $\mathbb{R}$, the boundary of any interval (a, b), $[a, b]$, $[a, b)$, or $(a, b]$ is $\{a, b\}$, again what you normally think of as the boundary of an interval.

The notions of open and closed sets now become much more intuitive when we state what they mean in terms of boundaries:

Proposition 12. *Let S be a subset of a metric space X. Then S is open in X if and only if S contains no piece of its boundary: i.e. $S \cap \partial S = \emptyset$; S is closed in X if and only if it contains all of its boundary: i.e. $\partial S \subseteq S$.*

The proof of this really just involves unwinding some definitions. We first prove a lemma, saying that any point of S is either an interior point of S or a boundary point of S; note also that $\operatorname{int} S$ and ∂S have nothing in common:

Lemma 1. *Let S be a subset of a metric space X. Then $S \subseteq \operatorname{int} S \cup \partial S$.*

Proof. Let $s \in S$. If there exists $r > 0$ such that $B_r(s) \subseteq S$, then $s \in \operatorname{int} S$ so $s \in \operatorname{int} S \cup \partial S$ and we are done. Otherwise, for any $r > 0$ $B_r(s)$ is not contained in S. This means that there is some $p \in B_r(s)$ which is not in S, so $B_r(s)$ contains an element of S (namely s itself) and an element not in S (namely p). Thus $s \in \partial S$ so $s \in \operatorname{int} S \cup \partial S$ as required. □

Proof of Proposition. Suppose that S is open in X. Then for any $s \in S$, $s \in \operatorname{int} S$, so $s \notin \partial S$ since $\operatorname{int} S$ and ∂S are disjoint. Thus ∂S contains no element of S, meaning that $S \cap \partial S = \emptyset$ as required. Conversely, suppose that $S \cap \partial S = \emptyset$. If $s \in S$, by the lemma we must have $s \in \operatorname{int} S$ or $s \in \partial S$. But since S and ∂S have nothing in common, we must have $s \in \operatorname{int} S$. Thus any point of S is an interior point, meaning that S is open in X.

Now suppose that S is closed in X and let $p \in \partial S$. As problem 5 on homework 5 shows, $p \in \overline{S}$. Since S is closed, $\overline{S} = S$ and thus $p \in S$. Thus $\partial S \subseteq S$ as required. Conversely suppose that $\partial S \subseteq S$ and let $s \in \overline{S}$. If $s \in \operatorname{int} S$, then $s \in S$. Otherwise, $s \in \overline{S} \setminus \operatorname{int} S$ so $s \in \partial S$ by the same homework problem. Since S contains its boundary, we have $s \in S$. Thus $\overline{S} \subseteq S$, showing that S is closed in X. □

Note that this is precisely how we visualize open and closed sets in $\mathbb{R}^2$, as ones which contain no piece of their boundary curves (in the open case) or all of their boundary curves (in the closed case). In general, a set which isn't open nor closed will contain some but not all of its boundary.

The above also suggests a relation between the boundary of a set and its closure—indeed we have:

Proposition 13. *For any subset S of a metric space X, $\partial S \subseteq \overline{S}$; in other words, any boundary point of S is also a limit point of S.*

Proof. Let $x \in \partial S$. To show that $x \in \overline{S}$, we must construct a sequence in S converging to x. For each n, consider the ball $B_{\frac{1}{n}}(x)$ of radius $\frac{1}{n}$ around x. Since x is a boundary point of S, this balls contains an element of S—call it s_n. This gives rise to a sequence (s_n) of points in S with the property that

$$d(s_n, x) < \frac{1}{n} \text{ for all } n,$$

which implies that $(s_n) \to x$. Hence $x \in \overline{S}$ as required. □

In fact, using the same idea as above, we can come up with all sorts of relations between the notions of interior, closure, and boundary. For example: $\partial S = \overline{S} \cap \overline{X \setminus S}$, $\overline{S} = \operatorname{int} S \cup \partial S$, $\partial S = \overline{S} \setminus \operatorname{int} S$. Some of these will appear on the next homework, and working through all these definitions is good practice in general for learning how to work with abstract concepts.

Connectedness

As we have seen, $\mathbb{Q}$ has nontrivial subsets—such as $\{r \in \mathbb{Q} \mid -\sqrt{2} < r < \sqrt{2}\}$—which are both open and closed in $\mathbb{Q}$. Subsets of a metric space which are both open and closed are called *clopen*. (Yes, that really is the standard mathematical term for such subsets, no matter how amusing the term may be.) So, we are saying that $\mathbb{Q}$ has at least one (many in fact) nonempty, proper clopen subset. (Proper just means a subset which is not the whole space.) We will see that this is not true for $\mathbb{R}$: the only clopen subsets of $\mathbb{R}$ are the empty set and $\mathbb{R}$ itself. This difference between $\mathbb{Q}$ and $\mathbb{R}$ is important enough that we give it a name—in fact, we give a more intuitive definition first and then see how it relates to the existence of clopen subsets.

Definition 13. A metric space X is *disconnected* if there exist nonempty, disjoint open subsets $U, V \subseteq X$ such that $X = U \cup V$. We say that X is *connected* if it is not disconnected.

To say that U and V are disjoint means that they have nothing in common, i.e. $U \cap V = \emptyset$. So, the definition says that X is disconnected if it can be broken up into nonempty open subsets which have nothing in common, while X is connected if it *cannot* be broken up in this manner. Intuitively, a disconnected space is one which is made up of different "pieces", while a connected one consists of a single "piece".

Example 26. $\mathbb{Q}$ is disconnected since we can write $\mathbb{Q}$ as

$$\mathbb{Q} = \{r \in \mathbb{Q} \mid -\sqrt{2} < r < \sqrt{2}\} \cup \{r \in \mathbb{Q} \mid r < -\sqrt{2} \text{ or } \sqrt{2} < r\}.$$

Each of these sets is open in $\mathbb{Q}$, neither is empty, and they have nothing in common. Replacing $-\sqrt{2}$ and $\sqrt{2}$ by any irrational numbers will give other ways of breaking $\mathbb{Q}$ up into nonempty, disjoint open subsets. Taking similar subsets of $\mathbb{R}\backslash\mathbb{Q}$ with $-\sqrt{2}$ and $\sqrt{2}$ replaced by rational numbers will show that $\mathbb{R}\backslash\mathbb{Q}$ is disconnected.

The above example is hard to visualize, and it is not clear why we use the term "disconnected" to describe such spaces. The next example should hopefully make this clear:

Example 27. Let X be the union of any two disjoint open disks in $\mathbb{R}^2$. (Imagine that these two disks are far apart.) Then X is disconnected. Indeed, we have defined X precisely to be the union of nonempty, disjoint open subsets. Visually we see that X is disconnected because it is "broken" up into "pieces" which are separated from each other.

The above example also helps to illustrate why connectedness will be a useful concept: in a disconnected space, what is happening in one "piece" will have no effect on what is happening in any other "piece". Once we talk about differentiation, we will see an explicit reason as to why this is important.

Now, what kinds of spaces are connected? For us, the main examples of connected spaces will be intervals and $\mathbb{R}$ itself. The proof that $\mathbb{R}$ is connected is left to the homework, but it depends (at least if you prove it the way I suggest in the homework) on the fact that open intervals are connected:

Theorem 7. *Any open interval (a, b) with $a < b$ is connected.*

The condition that $a < b$ is just there to guarantee that (a, b) is not empty. This fact should be clear intuitively: draw an open interval and convince yourselves that it should not be possible to break it up into two smaller disjoint open intervals. The problem is that no matter how you try to do this, there will always be at least one point of (a, b) left out. Here is how we can make this precise.

Proof. Suppose that U and V are nonempty, disjoint open subsets of (a, b). We will show that $U \cup V$ cannot be all of (a, b), which will imply that (a, b) is connected.

Since (a, b) is bounded, so are U and V and thus the supremums of U and V both exist in $\mathbb{R}$. Since b is an upper bound of U and of V, we know that $\sup U \leq b$ and $\sup V \leq b$. These supremums cannot both be b, so without less of generality we assume that $x = \sup U < b$. Now, we first claim that $x \notin U$. If $x \in U$, since U is open there exists an interval $(x - \delta, x + \delta)$ with $\delta > 0$ contained in U. But then $x + \frac{\delta}{2}$ is an element of U larger than x, contradicting the fact that x is an upper bound of U.

We also claim that $x \notin V$. If $x \in V$, again there exists $\delta > 0$ so that $(x - \delta, x + \delta) \subseteq V$. Since x is the least upper bound of U and $x - \frac{\delta}{2} < x$, there exists $u \in U$ such that $x - \frac{\delta}{2} < u < x$. But then this u is in the interval $(x - \delta, x + \delta)$, which is contained in V, contradicting the fact that $U \cap V = \emptyset$. We conclude that $x \notin U \cup V$, so x is an element of (a, b) not contained in $U \cup V$, so $U \cup V$ cannot be all of (a, b) as required. □

It turns out that *any* interval, whether it be closed or half-open and half-closed, is connected; the same argument as above may not quite work as given, but slight modifications should work. As mentioned, $\mathbb{R}$ is also connected, but now the above proof completely fails since we may not be able to take $\sup U$ and $\sup V$ as we did. There may be a way to use the same kind of idea to show that $\mathbb{R}$ is connected, but the homework will give another way of proving this.

Compactness

We now come to one of the most important, and perhaps least intuitive, concepts in analysis: the notion of a *compact* metric space. The definition is simple enough to state, but it is not at all clear at first why this should be something we are interested in. Indeed, we won't really see why until we talk about the special properties that continuous functions on compact spaces have. To give a very vague idea, compact spaces are ones which are not too "large"; indeed, we will see that compact spaces are always bounded, so part of saying that something is not too "large" means that it cannot "extend to infinity". However, an open interval (a, b) in $\mathbb{R}$ is also bounded and yet will not be compact, so "large" in the above sense must mean more than simply being bounded. Again, the real key will come a bit later when we study continuous functions on compact spaces.

There are two possible definitions we can give for compactness: one which is easier to state but gives no hint at its deeper meaning, and another which is (psychologically) trickier to state but starts to get at the heart of the sense in which compact spaces are not too "large". We will start with the first definition (sequential compactness), give examples and prove some basic facts, and then we will give the second definition (covering compactness) and do the same. It is one of the miracles of analysis and topology that these two arbitrary-looking definitions are actually equivalent.

(Sequential) Compactness

Definition 14. A subset K of a metric space (X, d) is said to be *compact* if every sequence in K has a convergent subsequence in K.

Remark 1. Since a metric space can always be viewed as a subset of itself, it makes sense to ask whether or not a metric space itself is compact: (X, d) is *compact* if every sequence in X has a convergent subsequence. The point is that, unlike the definitions of open and closed, compactness is *not* a relative notion: we do not ask whether or not a space is compact in a larger space, only whether or not a space is itself compact. For instance, we will see shortly that closed intervals

$[a, b]$ are compact—yes, these are compact subsets of $\mathbb{R}$, but the point is that $[a, b]$ is compact when viewed as a metric space in its own right, regardless of the fact that it is a subset of $\mathbb{R}$.

So, according to this definition, a space is not too "large" if no matter what sample points you choose from the space, you can always find a subsequence among them which converges.

Example 28. Any finite subset $\{x_1, \ldots, x_n\}$ of a metric space X is compact. Indeed, pick any sequence in $\{x_1, \ldots, x_n\}$. Since there are only finitely many points in this subset but infinitely many terms in the sequence, at least one of the x_i must occur infinitely often in the sequence. The subsequence formed by picking out each occurrence of this term will then be the convergent subsequence in $\{x_1, \ldots, x_n\}$ we want.

The standard, and perhaps most important, example of a compact space is a closed interval $[a, b]$. Our book actually covers this, although it is not phrased in the language of compactness. Here is the relevant result:

Theorem 8 (Bolzano-Weierstrass, Theorem 11.5 in the book)**.** *Every bounded sequence in $\mathbb{R}$ has a convergent subsequence.*

Check the book for the proof, which should be relatively straightforward and is based on the fact (Theorem 11.3 in the book) that any sequence in $\mathbb{R}$ has a monotone subsequence. The compactness of closed intervals is an immediate consequence:

Corollary 1. *Any closed interval $[a, b]$ is compact.*

Proof. Let (x_n) be any sequence in $[a, b]$. Since $[a, b]$ is bounded, so is the sequence (x_n) so the Bolzano-Weierstrass Theorem implies that this has a subsequence (x_{n_k}) converging to some $x \in \mathbb{R}$. Since $[a, b]$ is closed, this limit x is actually in $[a, b]$, so (x_{n_k}) converges in $[a, b]$ and hence $[a, b]$ is compact. □

Note that the only important properties of $[a, b]$ which we actually used above were that it is bounded and closed in $\mathbb{R}$. Thus, the same proof works for any closed and bounded subset of $\mathbb{R}$; that is, any closed and bounded subset of $\mathbb{R}$ is also compact.

Example 29. $\mathbb{R}$ is not compact, and neither is a nonempty open interval (a, b). Indeed, the sequence $1, 2, 3, 4, 5, \ldots$ in $\mathbb{R}$ has no convergent subsequence, so $\mathbb{R}$ is not compact, and a sequence of points in (a, b) getting closer to a will not have a convergent subsequence in (a, b), so (a, b) is not compact.

Here are two properties which compact sets always have, which are already illustrated by the example of $[a, b]$ above.

Proposition 14. *A compact subset K of a metric space X is closed and bounded in X.*

Proof. First we show that K is closed in X. To this end, let (x_n) be a sequence in K converging to $x \in X$. We must show that $x \in K$. Since K is compact, (x_n) has a subsequence (x_{n_k}) converging to some $y \in K$. But since the original sequence converges to x, (x_{n_k}) also converges to x. Hence since limits are unique x and y must be the same, and since $y \in K$, $x \in K$ as claimed. We conclude that K is closed in X.

Now, to show that K is bounded, we establish the contrapositive: if K is not bounded, then K is not compact. We do this by constructing a sequence in K which no convergent subsequence. Fix $p \in K$. If K is not bounded, for any positive radius $r > 0$ we can find an element of K whose distance from p is larger than r. In particular, for each $n \in \mathbb{N}$, we can find $x_n \in K$ such that

$d(p, x_n) > n$. This gives a sequence (x_n) in K, and the condition that $d(p, x_n) > n$ for each n implies that every subsequence of (x_n) is unbounded and hence divergent. Thus (x_n) is a sequence of K with no convergent subsequence in K, so K is not compact. □

So, compact subsets of a metric space are always closed and bounded. Note that in the case of $\mathbb{R}$, as mentioned above, the converse is also true: a closed and bounded subset of $\mathbb{R}$ is compact. In fact, this is true more generally in $\mathbb{R}^n$:

Theorem 9 (Heine-Borel)**.** *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.*

This is not too hard to prove, given that we already know this for $n = 1$ and that a sequence of points in $\mathbb{R}^n$ converges if and only if each component sequence converges. We omit the proof here. The point is that compact subsets of $\mathbb{R}^n$—$\mathbb{R}$ and $\mathbb{R}^2$ in particular—are easy to describe: they are exactly the closed and bounded subsets.

The standard warning at this point is that while compact sets are always closed and bounded, it is not true that any closed and bounded subset of a general metric space will be compact. The fact that this is true for $\mathbb{R}^n$ is just one of the many amazing properties of $\mathbb{R}^n$.

Example 30. The subset $\{x \in \mathbb{Q} \mid -\sqrt{2} < x < \sqrt{2}\}$ of rational numbers between $-\sqrt{2}$ and $\sqrt{2}$ is closed and bounded but not compact. This subset is clearly bounded, and we have seen before that it is closed in $\mathbb{Q}$. However, take any sequence of rationals in this set converging to $\sqrt{2}$ in $\mathbb{R}$. Any subsequence would also have to converge to $\sqrt{2}$ in $\mathbb{R}$, and so no subsequence will converge *in* $\{x \in \mathbb{Q} \mid -\sqrt{2} < x < \sqrt{2}\}$. Thus, such a sequence of rationals would be a sequence in this set with no convergent subsequence, so this set is not compact.

(Covering) Compactness

We now give the second definition of compactness, which although more difficult to state and wrap your head around, is probably going to be the more intuitive one once you get the hang of it.

Definition 15. Let X be a metric space and $S \subseteq X$ a subspace. An *open cover* of S is a collection $\{U_\alpha\}$ of open subsets of X such that S is contained in their union—meaning that any element of S is in at least one of the U_α. A *subcover* of an open cover $\{U_\alpha\}$ is an open cover $\{V_\beta\}$ of S so that each V_β occurs in the collection $\{U_\alpha\}$. An open cover is *finite* if it contains finitely many sets.

Intuitively, an open cover of S is a collection of open subsets of X which all together "cover" all of S. A subcover is nothing but a collection of some (but not necessarily all) sets in the open cover which still "cover" all of S.

Definition 16. A subset K of a metric space (X, d) is said to be *compact* if every open cover of K has a finite subcover.

Thus, a compact set is one with the property that whenever we cover it by an infinite (even uncountable) number of open sets, there are actually a finite number of those open sets which still do the job. This is now getting closer to what we mean by saying that a compact set is one which is not too "large", in that questions dealing with an infinite number of open sets covering it can be reduced to one about a finite number of open sets covering it; i.e. for compact sets, questions about an infinite number of things can be reduced one about a finite number of things. We will see multiple examples of this.

The amazing (and completely surprising) fact is that this definition of compactness is equivalent to the previous one. In other words:

Theorem 10. *A subset K of a metric space X is sequentially compact (i.e. satisfies the first definition of compact) if and only if it is covering compact (i.e. satisfies the second definition of compact).*

This is a highly non-obvious and non-trivial theorem, and takes a lot of effort to prove. Actually, one direction (that covering compact implies sequentially compact) is not so hard to prove, but the other is a killer. We will omit the proof here since we will not need the techniques it uses, but I encourage you to look it up after this course is over. It really is quite astonishing how it all works out so nicely.

Example 31. Let us look at three examples considered previously. We have already seen that $\mathbb{R}$ is not compact according to the sequential definition of compact. It is easy to see that it also fails this new covering definition: in particular, the collection of open intervals of the form $(-n, n)$ for $n \in \mathbb{N}$ give an open cover of $\mathbb{R}$ with no finite subcover. Similarly, a nonempty open interval (a, b) is not compact according to this definition since the collection of open intervals of the form

$$\left(a + \frac{b-a}{n}, b - \frac{b-a}{n}\right) \text{ for } n \in \mathbb{N},$$

whose endpoints get closer to a and b as n increases, gives an open cover of (a, b) with no finite subcover.

Now, closed intervals $[a, b]$ do satisfy this new definition of compact. (Indeed, they'd better since we already know $[a, b]$ is compact according to the previous definition, and we have claimed that both definitions are equivalent.) Showing this is more difficult than showing $[a, b]$ is sequentially compact. The hard part is that given any open cover $\{U_\alpha\}$ of $[a, b]$ whatsoever, we have to somehow produce a finite number of the U_α which still cover $[a, b]$.

This is not at all easy to do directly, but it turns out there is an indirect approach. We proceed as follows. Let C denote the set of elements $x \in [a, b]$ with the property that there are finitely many of the U_α covering $[a, x]$:

$$C = \{x \in [a, b] \mid \text{there exist finitely many } U_\alpha \text{ covering } [a, x]\}.$$

Note that, in particular, $a \in C$ since there does exist a U_α covering $[a, a] = \{a\}$. Also, C is bounded since it is a subset of the bounded interval $[a, b]$. Thus C has a supremum, call it u, and since $C \subset [a, b]$, $a \leq u \leq b$. The proof now proceeds by showing that u is actually in C and that $u = b$. Combining these gives $b \in C$ which exactly says that finitely many of the U_α cover $[a, b]$, and hence that $[a, b]$ is compact. The full details can be found in the "Worked Examples" handout from Summer 2010, but this is not something you would be responsible for.

Let us now return to a basic fact about compact sets we have previously seen, and reprove it from the covering point of view. As always, I suggest you draw a picture in $\mathbb{R}^2$ to illustrate what is going on in the proof!

Proposition 15. *A compact subset K of a metric space X is closed and bounded in X.*

Proof. First we show that K is closed in X. To do this, we show instead that the complement K^c of K in X is open in X. So, let $q \in K^c$. For each $p \in K$, let U_p be the ball of radius $d(p, q)/2$ around p and let V_p be the ball of radius $d(p, q)/2$ around q. Notice that these two balls are disjoint from each other. The collection $\{U_p \mid p \in K\}$ is then an open cover of K, so it has a finite subcover since K is compact; let

$$U_{p_1}, \ldots, U_{p_n}$$

be the sets in this finite subcover. We claim that $V := V_{p_1} \cap \cdots \cap V_{p_n}$ is an open set containing q and completely contained in K^c. Indeed, V is open since the intersection of finitely many open sets is open, and for any $p \in K$, $p \in U_{p_k}$ for some $k = 1, \ldots, n$, so $p \notin V_{p_k}$ and hence $p \notin V$—thus no element of K is in V, so $V \subset K^c$. Hence q is an interior point of K^c, and since $q \in K^c$ was arbitrary, every point in K^c is an interior point. This shows that K^c is open as required.

Now, to show that K is bounded, pick any $p \in K$ and consider the collection $\{B_r(p) \mid r > 0\}$ of all balls centered at p of any positive radius. This is an open cover of K since the element $q \in K$ is contained in the ball of radius $d(q, p) + 1$ around p. Since K is compact, this has a finite subcover—let $r_1, \ldots, r_n$ be the radii of the balls in this finite subcover, and set $r = \max\{r_1, \ldots, r_n\} > 0$. Then each of the balls $M_{r_i}(p)$ is contained in the ball $M_r(p)$, and since these balls cover K, it follows that K itself is contained in $M_r(p)$. Hence K is bounded. □

Granted, proving that a compact set is closed is not as straightforward using the covering characterization of compactness as it was using the sequence characterization. The proof illustrates what is meant by saying that the definition of covering compact allows one to pass from having to deal with an infinite number of things to instead deal with a finite number of things. More precisely, we constructed some sets U_p and V_p for each p in other space, and although there could be an absolutely huge number of such open sets, compactness of K allowed us to replace them but a finite number. Then when dealing with this finite number we can do things like take intersections. Similarly, in the proof that a compact set is bounded, we had an infinite number of radii to work with and compactness allowed us to replace them with a finite number of radii, after which taking their maximum makes sense.

In any event, I hope you agree that proving a compact set is bounded is much more intuitive and simpler using the covering definition. Indeed, this illustrates and important point: for certain applications, the sequence characterization of compactness is easier to apply, but for others the covering characterization works better. We will see the same idea pop up when we talk about properties of continuous functions on compact spaces—something which gets messy using the sequence definition becomes easier to visualize using the covering definition, and vice-versa.

Continuous Functions

We finally come to the notion of what it means for a function from one metric space to another to be *continuous*. Indeed, studying properties of continuous functions is one of the main reasons why we study metric spaces at all. In other words, although metric spaces are interesting in their own right, more important is understanding how different metric spaces are related to each other, and the notion of continuity allows us to do this. In particular, one of the punchlines in this final section of these notes is the so-called "Extreme Value Theorem", which says that a real-valued continuous function on a compact space always has a maximum and a minimum; as we will see in the rest of the course, this fact is one of the main reasons underlying why calculus works.

You've probably seen the so-called "ϵ-δ" definition of continuous in a previous calculus before. If you're like me, at the time this definition made no sense and seemed to come out of nowhere. Naively, a continuous function is one whose graph has no "jumps". However, this does not work so well when dealing with arbitrary metric spaces since it is not always easy to picture such spaces let alone the graphs of functions between them. A better intuition is the following: a continuous function is one with the property that points which are "close" together in the domain get sent to points which remain "close" together. This is exactly what goes wrong when a function "jumps": points which were close to begin with can be sent to points which are far apart. The definition

of continuous we will give precisely encodes this intuitive idea; we will then relate it to the ϵ-δ definition and also give a *third* equivalent characterization of continuity.

Continuity

Definition 17. A function $f : M \to N$ from a metric space (M, d_M) to a metric space (N, d_N) is said to be *continuous at* $p \in M$ if it has the property that whenever $(p_n) \to p$ in M, $(f(p_n)) \to f(p)$ in N. We say f is *continuous* if it is continuous at each $p \in M$.

Hopefully the notation used above is clear: M and N are both metric spaces, and we use M and N as subscripts when denoting their respective metrics. The given definition says exactly what we wanted "continuous" to mean: if points p_n are getting closer and closer to p, then the points $f(p_n)$ obtained after applying the function are getting closer and closer to $f(p)$.

Example 32. Suppose that M is a discrete metric space and let N be any metric space. We claim that any function $f : M \to N$ is continuous. Indeed, to check that f is continuous at $p \in M$, we must check whether or not it is true that $(f(p_n)) \to f(p)$ in N if $(p_n) \to p$ in M. However, we already know that the only convergent sequences in a discrete space are those which are eventually constant. So, if $(p_n) \to p$ in M, then $p_n = p$ for p_n past some index, and thus $f(p_n) = f(p)$ past that same index. Hence $(f(p_n))$ is eventually constant and thus converges to $f(p)$, showing that f is continuous at p. Since $p \in M$ was arbitrary, f is continuous as claimed.

More interestingly, I invite you to determine which functions *into* a discrete metric space are continuous: i.e. supposing that M is any metric space and N is discrete, which functions $f : M \to N$ are continuous? It is possible to give a concrete description of such functions.

Check Chapter 3 of the book for examples and properties of continuous functions from $\mathbb{R}$ (or subsets of $\mathbb{R}$) to $\mathbb{R}$. For example, the fact that $(x_n + y_n) \to x + y$ if $(x_n) \to x$ and $(y_n) \to y$ in $\mathbb{R}$ implies that the sum of continuous functions $\mathbb{R} \to \mathbb{R}$ is continuous, and similarly the fact that $(x_n y_n) \to xy$ if $(x_n) \to x$ and $(y_n) \to y$ in $\mathbb{R}$ implies that the product of continuous functions $\mathbb{R} \to \mathbb{R}$ is continuous.

The next example is so nice that I'm going to give it a special name—we will come across this same function as an interesting example of *integrability* later on.

Example 33. (*Santi's Favorite Example, Part I*) Recall that we can write any nonzero rational number r uniquely as a fraction $\frac{p}{q}$ with $p, q \in \mathbb{Z}$ having no common factors and $q > 0$. With this in mind, we define a function $f : [0, 1] \to \mathbb{R}$ as follows:

$$f(x) = \begin{cases} 0 & \text{if } x \in \mathbb{R} \backslash \mathbb{Q} \\ \dfrac{1}{q} & \text{if } x = \dfrac{p}{q} \in \mathbb{Q} \text{ where } p, q \in \mathbb{Z} \text{ have no common factors, } p \neq 0 \text{ and } q > 0 \\ 1 & \text{if } x = 0. \end{cases}$$

(We have to consider $x = 0$ separately since there is no unique way of writing 0 as a fraction in the required way since $\frac{0}{q} = 0$ for *any* $q > 0$. For the property of this function which we will soon state, the important thing is that $f(0) \neq 0$—we could just as easily have defined $f(0)$ to be π, 423040489, or any nonzero number.) For instance, $f(1) = 1$, $f\left(\frac{2}{4}\right) = \frac{1}{2}$, and $f\left(\frac{1}{\sqrt{2}}\right) = 0$. Then, as will be shown on a future homework, f is continuous at x if and only if x is irrational. This example is important since the graph of f in a sense indeed "jumps" at points throughout the interval $[0, 1]$, and yet it is still continuous at each irrational, showing that the notion of continuous is more subtle than simply "does not jump".

Now we show that the definition of continuous we have given is indeed equivalent to the ϵ-δ definition given in numerous calculus courses:

Theorem 11. *A function $f : M \to N$ from a metric space (M, d_M) to a metric space (N, d_N) is continuous at $p \in M$ if and only if for any $\epsilon > 0$ there exists $\delta > 0$ such that*

$$d_N(f(q), f(p)) < \epsilon \text{ whenever } d_M(q, p) < \delta.$$

Let us make clear what the above characterization is saying. Given any $\epsilon > 0$, say we want to end up within ϵ away from $f(p)$. Then the above says that there is some positive δ so that for any q within δ away from p, $f(q)$ is indeed within ϵ away from p. In other words: given any estimate as to how close we want to be to $f(p)$, it is possible to get close enough to p to ensure we end up within that estimate away from $f(p)$.

Proof. Suppose first that the second condition is satisfied and let $(p_n) \to p$ in M. We want to show that $(f(p_n)) \to f(p)$ in N. To this end, let $\epsilon > 0$. Then there exists $\delta > 0$ so that

$$\text{if } d_M(q, p) < \delta, \text{ then } d_N(f(q), f(p)) < \epsilon.$$

Pick K such that $d_M(p_n, p) < \delta$ for $n \geq K$, which exists since (p_n) converges to p. Then the condition which δ satisfies implies that $d_N(f(p_n), f(p)) < \epsilon$ for $n \geq K$, and hence $(f(p_n))$ converges to $f(p)$ as required.

To establish the converse, we instead prove the contrapositive. That is, suppose that there exists some $\epsilon > 0$ so that for any $\delta > 0$ there exists $q \in M$ such that

$$d_M(q, p) < \delta \text{ and } d_N(f(q), f(p)) \geq \epsilon.$$

We construct a sequence (p_n) converging to p in M for which $(f(p_n))$ does not converge to $f(p)$ in N. For each n, apply the above assumption to $\delta = \frac{1}{n}$ to pick $p_n \in M$ so that

$$d_M(p_n, p) < \frac{1}{n} \text{ and } d_N(f(p_n), f(p)) \geq \epsilon.$$

By construction, the terms in the resulting sequence (p_n) satisfy $d_M(p_n, p) < \frac{1}{n}$ for all n, implying that (p_n) converges to p in M. Also, since each $f(p_n)$ is bounded away form $f(p)$ by $\epsilon > 0$, the sequence $(f(p_n))$ does not converge to $f(p)$ as required. □

Let us restate the ϵ-δ characterization of continuous in the following way. Recall that the inequality $d_M(q, p) < \delta$ means precisely that $q \in B_\delta(p)$, and similarly $d_N(f(q), f(p)) < \epsilon$ means $f(q) \in B_\epsilon(f(p))$. Thus, the above theorem says that $f : M \to N$ is continuous at p if and only if for any ball $B_\epsilon(f(p))$ around $f(p)$ there is a ball $B_\delta(p)$ around p which is sent by f into the previous ball. This should now hopefully make clear what the ϵ-δ definition of continuity actually means.

Example 34. Let us show that the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ is continuous using the ϵ-δ definition. Fix $x \in \mathbb{R}$ and let $\epsilon > 0$. Set

$$\delta = \min\left\{1, \frac{\epsilon}{(1 + |x|)^2 + |x|(1 + |x|) + |x|^2}\right\},$$

which is positive since it is the minimum of two positive numbers. Suppose that $|y - x| < \delta$. In particular, $|y - x| < 1$ so

$$|y| - |x| \leq |y - x| < 1, \text{ and thus } |y| < 1 + |x|.$$

Then

$$\begin{aligned}|y^3 - x^3| &= |y-x||y^2+xy+x^2| \\ &\leq |y-x|(|y|^2+|x||y|+|x|^2) \\ &< \delta[(1+|x|)^2+|x|(1+|x|)+|x|^2] \\ &\leq \epsilon.\end{aligned}$$

Hence for the above choice of δ, $|y-x| < \delta$ implies $|f(y)-f(x)| < \epsilon$, so f is continuous at each $x \in \mathbb{R}$ and thus on $\mathbb{R}$.

See the "Worked Examples" handout from Summer 2010 for an explanation of where the motivation for the above choice of δ comes from and for a generalization of this proof showing that $f(x) = x^n$ is continuous for any n. The idea is one we've seen when dealing with sequences: we want to make $|f(y)-f(x)|$ smaller than ϵ, so we bound it by something we have some control over—in this case $|y-x|$.

Note something else about the previous example, which is indicative of a general fact: the δ we constructed depends on both ϵ and the point x we are checking continuity at. This makes sense, since the ϵ-δ definition of continuous says: first comes the point we are at, then comes ϵ, and *then* comes δ. Note that "y" in the above proof comes after the choice of δ: once δ is defined, we then look at y's satisfying $|y-x| < \delta$. Thus δ itself cannot depend on y, or in the general ϵ-δ definition of continuous at $p \in M$, δ cannot depend on q. But it is perfectly reasonable to expect that δ depends on other fixed quantities such as ϵ or the point we are checking continuity at.

In fact, continuous functions with the property that δ does not depend on the point we are at—i.e. the same δ works for *all* points—are special enough that we give them a special name:

Definition 18. A function $f : M \to N$ is *uniformly continuous* if for any $\epsilon > 0$ there exists $\delta > 0$ such that

$$d_N(f(q), f(p)) < \epsilon \text{ whenever } d_M(q,p) < \delta.$$

Let us again emphasize the difference between this and continuous: in the case of f being continuous at p, the p and $f(p)$ in the above inequalities are fixed and it is the q (and hence $f(q)$) which varies; here, it is both q *and* p which vary. Intuitively, a uniformly continuous function is one which is continuous "in the same way" near every point, which is why we use the term "uniformly" when describing such functions.

Example 35. The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ is not uniformly continuous. Looking back to the previous example, the point is that the δ we constructed satisfying the ϵ-δ definition of continuous gets smaller and smaller as x gets larger and larger (because of the $|x|$ terms in the denominator of the fraction used in defining δ), and so no such δ will work for all x at once. This is not a rigorous proof that this function is not uniformly continuous, but this basic idea can be made precise.

Notice that the above function has an unbounded derivative, which as we will later see is related to the fact that it is not uniformly continuous. We will soon see that continuous functions on compact spaces are *always* uniformly continuous, which is one of the special properties of compact metric spaces which I've been hyping.

Finally, let us point out the following. Recall that the ϵ-δ definition of continuity can be restated in terms of balls: $f : M \to N$ is continuous at $p \in M$ if and only if for any ball $B_\epsilon(f(p))$ around $f(p)$ there is a ball $B_\delta(p)$ around p getting send into it. Saying that $B_\delta(p)$ gets sent into $B_\epsilon(f(p))$ means that $B_\delta(p)$ is contained in the pre-image $f^{-1}(B_\epsilon(f(p))$ of $B_\epsilon(f(p))$. Using this, we can recast the definition of continuous once more:

Theorem 12. *A function $f : M \to N$ is continuous if and only if $f^{-1}(U)$ is open in M for any open subset U of N.*

We will leave the full details of this to the reader, but the proof essentially amounts to unwinding the ϵ-δ characterization of continuity as mentioned in the previous paragraph. This is probably the shortest, but least intuitive, definition of continuous one could give. However, as you will see if you take a more advanced topology course at some point, this is the definition which best captures what continuity should *really* mean in general. Using the fact that the complement of an open set is closed and the complement of a closed set is open, it is not hard to also see that $f : M \to N$ is continuous if and only if $f^{-1}(A)$ is closed in M for any closed subset A of N.

Continuity and Compactness/Connectedness

As mentioned above, open and closed sets behave well under *pre-images* of continuous functions. On the other hand, we will see that compact and connected sets behave well under *images* of continuous functions. We start with connected sets:

Theorem 13. *The image of a connected space under a continuous function is connected: i.e. if $f : M \to N$ is continuous and M is connected, then $f(M)$ is connected.*

Proof. Suppose that $f(M) = U \cup V$ with U and V open in $f(M)$ and disjoint. To show that $f(M)$ is connected, we want to show that one of U or V must be empty. Now, by the third characterization of continuous, we have that $f^{-1}(U)$ and $f^{-1}(V)$ are open in M. Also, $M = f^{-1}(U) \cup f^{-1}(V)$ since everything in M has to map into $f(M) = U \cup V$, and $f^{-1}(U)$ and $f^{-1}(V)$ are disjoint since U and V are. Thus since M is connected one of these must be empty, say it is $f^{-1}(U)$. It follows that U itself is empty, so that $f(M)$ is connected as required. □

Notice how relatively straightforward this proof was, which essentially amounts to jumping back and forth between definitions. In the case of subsets of $\mathbb{R}$, this result says that the image of an interval under a continuous function is an interval; although this can be proven directly using only properties of $\mathbb{R}$, this type of result is an example of something which I think is much simpler to phrase in the setting of general metric spaces.

The following corollary is also some well-known property of continuous functions on $\mathbb{R}$ which you would have seen in a previous calculus course, but now we see that it is nothing but a consequence of the above general fact:

Corollary 2 (Intermediate Value Theorem)**.** *Any continuous function $f : M \to \mathbb{R}$ where M is connected has the intermediate value property: if $p, q \in M$ and $f(p) < f(q)$, then for any $z \in \mathbb{R}$ such that $f(p) < z < f(q)$ there exists $x \in M$ such that $f(x) = z$.*

In other words, if the image of f in $\mathbb{R}$ contains two points, then it contains the entire interval between those two points.

Proof. Let $p, q \in M$ and suppose that $f(p) < f(q)$. If there is $z \in \mathbb{R}$ such that $f(p) < z < f(q)$ and z is not in the image of f, then $f(M)$ is contained in the union of $(-\infty, z)$ and (z, ∞):

$$f(M) \subseteq (-\infty, z) \cup (z, \infty).$$

However, these two intervals are open, disjoint, and nonempty (in particular, $f(p)$ is in the first interval and $f(q)$ is in the second), which contradicts the fact that $f(M)$ is connected as a consequence of the previous theorem. Thus no such z exists. □

Now we move to the behavior of compact spaces under continuous functions:

Theorem 14. *The image of a compact space under a continuous function is compact: i.e. if $f : M \to N$ is continuous and M is compact, then $f(M)$ is compact.*

We give two proofs, based on the different but equivalent definitions of compactness.

Proof 1. Let $(f(p_n))$ be any sequence in the image $f(M)$. We want to show this has a convergent subsequence in $f(M)$. Since M is compact, the sequence (p_n) in M has a convergent subsequence, say (p_{n_k}). Since f is continuous, the image sequence $(f(p_{n_k}))$ converges to $f(p) \in f(M)$. Thus $(f(p_{n_k}))$ is a convergent subsequence in $f(M)$ of $(f(p_n))$, so $f(M)$ is compact. □

Proof 2. Let $\{U_\alpha\}$ be an open cover of $f(M)$. Since f is continuous, each preimage $f^{-1}(U_\alpha)$ is open in M. Since the U_α cover $f(M)$, it follows that the $f^{-1}(U_\alpha)$ cover M. Hence $\{f^{-1}(U_\alpha)\}$ is an open cover of M, so since M is compact, this has a finite subcover—say

$$\{f^{-1}(U_1), \ldots, f^{-1}(U_n)\}.$$

It follows that $\{U_1, \ldots, U_n\}$ is an open cover of $f(M)$, and this is then a finite subcover of the open cover $\{U_\alpha\}$ of $f(M)$. We conclude that $f(M)$ is compact. □

As in the proof that images of connected spaces are connected, note how relatively straightforward these proofs, which again amount to jumping back forth between definitions, are. This is also something I claim is simpler to appreciate in the general metric space setting.

The following corollary is what I claim is the single most important reason why we care about compact sets in this class at all. As with the intermediate value theorem, you may have seen a statement of this in a previous calculus course, but hopefully now you can appreciate how powerful it really is:

Corollary 3 (Extreme Value Theorem)**.** *Any continuous function $f : M \to \mathbb{R}$ with compact domain M achieves a maximum and a minimum.*

In particular, real-valued continuous functions on compact domains are always bounded, so we can talk about putting the sup metric on the space of continuous functions on $[a, b]$; this metric space will be denoted by $C([a, b])$ (note the absence of a subscript: C by itself means "continuous") and is a subspace of the space of all bounded functions on $[a, b]$. In this setting, Theorem 24.3 of the book (which we will soon get to) can be interpreted as saying that $C([a, b])$ is a *closed* subset of $C_b([a, b])$.

The above fact is the reason why being asked to find maximums and minimums of functions in a first semester calculus course makes sense: such maximums and minimums will exist for the types of functions considered in those courses.

Proof. Since M is compact, we know that $f(M)$ is compact. Now, a compact subset of $\mathbb{R}$ must be bounded, so $f(M) \subseteq \mathbb{R}$ has a supremum, say b. By an alternate characterization of supremums, there exists a sequence in $f(M)$ converging to b, so since compact subsets of $\mathbb{R}$ are also closed in $\mathbb{R}$, b must be in $f(M)$. Thus there exists $p \in M$ such that $f(p) = b$ is the maximum value of f, so f achieves a maximum. Similar reasoning applied to the infimum of $f(M)$ shows that f also achieves a minimum. □

Given this, we can finally give a better reason as to why compact spaces are not too "large": compact spaces are not too large in sense that real-valued continuous functions on them are always bounded and cannot blow-up to ∞. Clearly, this is not necessarily true for non-compact spaces; for instance, $f(x) = x$ is unbounded on $\mathbb{R}$, and $f(x) = \frac{1}{x}$ is unbounded on $(0, 1)$.

In a future analysis course dealing with more general forms in integration, you will see that this property of continuous functions on compact spaces is also essential for saying that continuous functions on compact spaces are always integrable; again, this is not true for non-compact spaces—for instance, the perfectly nice and continuous function $f(x) = x$ is not integrable on $\mathbb{R}$ precisely because it is not bounded.

Here is one final property of continuous functions on compact sets. We give a proof based on the open cover definition of compactness, and leave a sequence-based proof up to you. (The sequence-based proof is a lot simpler than the covering-based proof.)

Theorem 15. *A continuous function on a compact metric space is uniformly continuous.*

Proof. Suppose that $f : M \to N$ is a continuous function between metric spaces M and N and that M is compact. Let $\epsilon > 0$. Since f is continuous, for each $p \in M$ there exists $\delta_p > 0$ (which may depend on p) such that

$$d_M(q, p) < \delta_p \text{ implies } d_N(f(q), f(p)) < \frac{\epsilon}{2}.$$

We want to find a $\delta > 0$ satisfying this condition for any p, so a δ independent of p.

The collection $\{B_{\delta_p/2}(p)\}$, as p ranges over all points of M, is then an open cover of M. Since M is compact, this has a finite subcover—say

$$\{B_{\delta_{p_1}/2}(p_1), \ldots, B_{\delta_{p_n}/2}(p_n)\}.$$

Set $\delta = \min\{\delta_{p_1}/2, \ldots, \delta_{p_n}/2\}$. Note that since each $\delta_{p_i} > 0$ and there are only finitely many, the minimum of this set exists and is positive.

Suppose that q and q' are any two points of M such that

$$d_M(q, q') < \delta.$$

Since the radii $\delta_{p_1}/2, \ldots, \delta_{p_n}/2$ give balls which cover M, q' is in one of these balls—without loss of generality say that $q' \in B_{\delta_{p_1}/2}(p_1)$. Note that then

$$d_M(q, p_1) \leq d_M(q, q') + d_M(q', p_1) < \delta + \delta_{p_1}/2 \leq \delta_{p_1}/2 + \delta_{p_1}/2 \leq \delta_{p_1}.$$

Thus $d_M(q, p_1) < \delta_{p_1}$ and $d_M(q', p_1) < \delta_{p_1}/2 < \delta_{p_1}$, so by the choice of δ_{p_1} we have

$$d_N(f(q), f(q')) \leq d_N(f(q), f(p)) + d_N(f(q'), f(p)) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Hence we have that

$$d_M(q, q') < \delta \text{ implies } d_N(f(q), f(q')) < \epsilon,$$

so we conclude that f is uniformly continuous. □

Admittedly, this proof is not easy to follow at first, but the basic idea is one we have already seen when dealing with compact spaces: the covering definition of compact allows us to reduce dealing with an infinite number of things (in this case the radii $\delta_p/2$ of the balls $B_{\delta_p/2}(p)$) to dealing with a finite number of radii, and then we can do something like take their minimum.

Fixed Point Theorems

This final topic is not something we will focus on much, but we will see one amazing application of it at the end of course. For this reason alone, it is worth putting in these notes.

Definition 19. A function $f : M \to N$ is said to be a *contraction* if there exists a constant $K < 1$ such that

$$d_N(f(p), f(q)) \leq K d_M(p, q) \text{ for any } p, q \in M.$$

The term "contraction" should be clear: the definition says that f *shrinks* distances—in other words, starting with two distinct points p and q in M, applying f will give points in N which are closer to each other than p and q were. The requirement that distances are shrunk by at least a factor $K < 1$ is important; when $d_N(f(p), f(q)) < d_M(p, q)$ for all $p, q \in M$, we say that f is a *weak contraction*.

It should not be hard to see that any contraction is continuous, indeed on a future homework you will show that any function satisfying a condition similar to the one above but without the requirement that $K < 1$ is actually always uniformly continuous. (Such functions are called "Lipschitz", so a contraction is nothing but a Lipschitz function with "Lipschitz constant" K less than 1.)

An important question one could ask about functions in general is determining when they have *fixed points*, meaning a point such that $f(p) = p$. The main result here is that contractions on complete metric spaces always have fixed points, and those fixed points are unique:

Theorem 16 (Banach Contraction Principle, or Banach Fixed Point Theorem)**.** *A contraction $f : M \to M$ where M is complete always has a unique fixed point.*

The proof is very slick, and indeed gives a method for actually approximating what the fixed point is if we cared about such matters, which in the application we will give at the end of the course we do.

Proof. First we show uniqueness. Suppose that p and q are two fixed points of f. Then $f(p) = p$ and $f(q) = q$ so the contraction property gives

$$d(f, q) = d(f(p), f(q)) \leq K d(p, q).$$

Since $K < 1$, we must have $d(p, q) = 0$ so $p = q$ and a fixed point of f is unique.

Now we show that f has a fixed point, and for this we will use the note given at the end of Problem 4 of Homework 4. Let x be any point of M. Since f is a contraction, we have

$$d(f^2(x), f(x)) \leq K d(f(x), x)$$

where the notion f^n means f composed with itself n times, so for example $f^2(x) = f(f(x))$. Similarly, using the contraction property again, we have

$$d(f^3(x), f^2(x)) \leq K d(f^2(x), f(x)) \leq K^2 d(f(x), x).$$

In general, this same reasoning implies that

$$d(f^{n+1}(x), f^n(x)) \leq K^n d(f(x), x) \text{ for all } n.$$

According to the comment at the end of Problem 4 of Homework 4 (whose proof is similar to the solution to that problem), the sequence $(f^n(x))$ obtained by applying f over and over again to x is then Cauchy, and thus convergent since M is complete; say that this sequence converges to $p \in M$.

We claim that p is the fixed point we are looking for. Indeed, note that since f is continuous (as contractions always are), the sequence $(f^{n+1}(x))$ obtained by applying f to each term in the sequence $(f^n(x))$ converges to $f(p)$. But this former sequence is nothing but a subsequence of the latter, so it also converges to x. Since limits are unique, we must have $f(p) = p$, so p is a fixed points of f as claimed. □

Note how amazing this technique is: no matter which point x we start with, the sequence obtained by applying f over and over again will always converge, and indeed converge to the same point regardless of what x we started with. Finding fixed points of functions in general is hard, but at least here we have a way of approximating the fixed point—we simply start with any x and compute what we get as we apply f over and over again.

The completeness of M was essential above, as well as the fact that f was an honest contraction as opposed to merely a "weak contraction". When f is only a weak-contraction, we still have a fixed point theorem, but we have to put stronger conditions on M:

Theorem 17. *A weak contraction $f : M \to M$ with M compact has a unique fixed point.*

The proof is left to the reader, but it similar-in-spirit to the proof we have for contractions on complete spaces. You should also try to find an example of a weak contraction on a non-compact complete space with no fixed point.