METRIC SPACES

KEITH CONRAD

1. Introduction

As calculus developed, eventually turning into analysis, concepts first explored on the real line (*e.g.*, a limit of a sequence of real numbers) eventually extended to other spaces (*e.g.*, a limit of a sequence of vectors or of functions), and in the early 20th century a general setting for analysis was formulated, called a metric space. It is a set on which a notion of distance between each pair of elements is defined, and in which notions from calculus in $\mathbf{R}$ (open and closed intervals, convergent sequences, continuous functions) can be studied. Many of the fundamental types of spaces used in analysis are metric spaces (*e.g.*, Hilbert spaces and Banach spaces), so metric spaces are one of the first abstractions that has to be mastered in order to learn analysis.

2. Metric spaces

In $\mathbf{R}$, the magnitude of a number x is its absolute value $|x|$ and the distance between two numbers x and y is the absolute value of their difference: $|x-y|$. In $\mathbf{R}^m$, the length of a vector $\mathbf{x} = (x_1, \ldots, x_m)$ is its norm $||\mathbf{x}|| = \sqrt{x_1^2 + \cdots + x_m^2}$ and the distance between two vectors $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_m)$ is the norm of their difference: $||\mathbf{x} - \mathbf{y}|| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_m - y_m)^2}$.

The distance between points is essential in defining limits, the central idea of calculus. There are limits of function values and limits of sequences. Focusing on the case of sequences (we will deal with limits and continuous functions in Section 8), we say a sequence $\{x_n\}$ of real numbers has limit x, and write $\lim_{n\to\infty} x_n = x$ or just $x_n \to x$, if for every $\varepsilon > 0$ there is an $N \geq 1$ (it is understood that $N = N_\varepsilon$ is an integer depending on ε) such that

$$n \geq N \Longrightarrow |x_n - x| < \varepsilon.$$

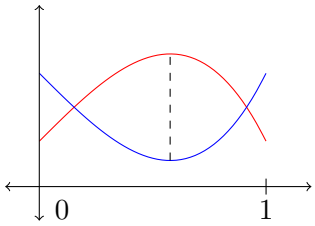
For a sequence $\{\mathbf{x}_n\}$ in $\mathbf{R}^m$, we write $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}$ or $\mathbf{x}_n \to \mathbf{x}$ for an $\mathbf{x} \in \mathbf{R}^m$ if for every $\varepsilon > 0$ there is an $N = N_\varepsilon$ such that

$$n \geq N \Longrightarrow ||\mathbf{x}_n - \mathbf{x}|| < \varepsilon.$$

Distances are useful not only between points in Euclidean space, but also between functions. For continuous functions $f, g\colon [0,1] \to \mathbf{R}$, here are two different ways of defining how far apart they are:

$$\max_{0\leq x\leq 1} |f(x) - g(x)|, \qquad \int_0^1 |f(x) - g(x)|\, dx. \tag{2.1}$$

What do these mean for the graphs of the functions below (in red and blue)?



The first formula in (2.1) is the length of the largest vertical line separating the graphs (the dashed line in the diagram), so saying f and g are close in this way means their graphs never get far apart from each other. The second formula is the area of the region over $[0, 1]$ that is enclosed by both graphs ("area between the curves"), so f and g are close in this way if, roughly speaking, the graphs can only be far apart over small regions (thereby not affecting the total area between the curves that much).

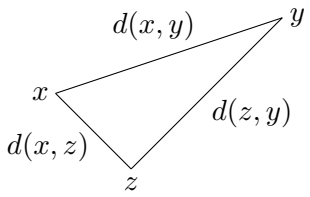
The desire to create a single framework for all the known settings where limit ideas are used inspired Maurice Fréchet in his 1906 PhD thesis [2] to make the following definition.

Definition 2.1. A *metric* on a set X is a function $d\colon X \times X \to \mathbf{R}$ satisfying the following three properties:

(*i*) $d(x, y) \geq 0$ for all x and y in X, with $d(x, y) = 0$ if and only if $x = y$,
(*ii*) $d(x, y) = d(y, x)$ for all x and y in X,
(*iii*) $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$.

A set X together with a choice of a metric d on it is called a *metric space* and is denoted (X, d), or just denoted X if the metric[1] is understood from context.

The third property in the definition of a metric is called the *triangle inequality* since it abstracts the fact that the length of one side of a triangle is at most the sum of the lengths of the other two sides (see figure below).



Example 2.2. On $\mathbf{R}^m$ the *Euclidean metric* is

$$d_E(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_m - y_m)^2},$$

This is the usual distance used in $\mathbf{R}^m$, and when we speak about $\mathbf{R}^m$ as a metric space without specifying a metric, it's the Euclidean metric that is intended.

To check d_E is a metric on $\mathbf{R}^m$, the first two conditions in the definition are obvious. The third condition is a consequence of the inequality $||\mathbf{x}+\mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ (replace $\mathbf{x}$ and $\mathbf{x} - \mathbf{z}$ and $\mathbf{y}$ with $\mathbf{z} - \mathbf{y}$), and to show this inequality holds we will write $||\mathbf{x}||$ in terms of

[1] In his thesis, Fréchet did not use the term "metric," but instead wrote écart [2, p. 30], which is French for "gap." He wrote (A, B) instead of $d(A, B)$.

the dot product: $||\mathbf{x}||^2 = x_1^2 + \cdots + x_m^2 = \mathbf{x} \cdot \mathbf{x}$, so

$$\begin{aligned} ||\mathbf{x}+\mathbf{y}||^2 &= (\mathbf{x}+\mathbf{y}) \cdot (\mathbf{x}+\mathbf{y}) \\ &= \mathbf{x}\cdot\mathbf{x} + \mathbf{x}\cdot\mathbf{y} + \mathbf{y}\cdot\mathbf{x} + \mathbf{y}\cdot\mathbf{y} \\ &= ||\mathbf{x}||^2 + 2\mathbf{x}\cdot\mathbf{y} + ||\mathbf{y}||^2 \\ &\leq ||\mathbf{x}||^2 + 2|\mathbf{x}\cdot\mathbf{y}| + ||\mathbf{y}||^2. \end{aligned}$$

The famous Cauchy–Schwarz inequality says $|\mathbf{x}\cdot\mathbf{y}| \leq ||\mathbf{x}||||\mathbf{y}||$, so

$$||\mathbf{x}+\mathbf{y}||^2 \leq ||\mathbf{x}||^2 + 2||\mathbf{x}||||\mathbf{y}|| + ||\mathbf{y}||^2 = (||\mathbf{x}|| + ||\mathbf{y}||)^2$$

and now take square roots.

A different metric on $\mathbf{R}^m$ is

$$d_\infty(\mathbf{x},\mathbf{y}) = \max_{1\leq i\leq m} |x_i - y_i|.$$

Again, the first two conditions of being a metric are clear, and to check the triangle inequality we use the fact that it is known for the absolute value. If $\max |x_i - y_i| = |x_k - y_k|$ for a particular k from 1 to m, then $d_\infty(\mathbf{x},\mathbf{y}) = |x_k - y_k|$, so

$$d_\infty(\mathbf{x},\mathbf{y}) \leq |x_k - z_k| + |z_k - y_k| \leq \max_{1\leq i\leq m} |x_i - z_i| + \max_{1\leq i\leq m} |z_i - y_i| = d_\infty(\mathbf{x},\mathbf{z}) + d_\infty(\mathbf{z},\mathbf{x}).$$

While the metrics d_E and d_∞ on $\mathbf{R}^m$ are different, they're not that different from each other since each is bounded by a constant multiple of the other one:

$$d_E(\mathbf{x},\mathbf{y}) \leq \sqrt{m}\, d_\infty(\mathbf{x},\mathbf{y}), \qquad d_\infty(\mathbf{x},\mathbf{y}) \leq d_E(\mathbf{x},\mathbf{y}). \tag{2.2}$$

Example 2.3. Let $C[0,1]$ be the space of all continuous functions $[0,1] \to \mathbf{R}$. Two metrics used on $C[0,1]$ are in (2.1):

$$d_\infty(f,g) = \max_{0\leq x\leq 1} |f(x) - g(x)|, \qquad d_1(f,g) = \int_0^1 |f(x) - g(x)|\, dx.$$

Checking these are metrics is left to the reader.[2] Notice for d_1 that the condition $d_1(f,g) = 0 \Longrightarrow f = g$ for being a metric uses continuity of the functions to know $\int_0^1 |f(x)-g(x)|\, dx = 0 \Longrightarrow |f(x) - g(x)| = 0$ for all $x \in [0,1]$.[3]

Unlike with the two metrics on $\mathbf{R}^m$ in Example 2.2, while we have $d_1(f,g) \leq d_\infty(f,g)$ there is no constant $A > 0$ that makes $d_\infty(f,g) \leq Ad_1(f,g)$ for all f and g. These metrics d_1 and d_∞ on $C[0,1]$ are quite different. In terminology we'll meet later, $C[0,1]$ is *complete* for d_∞ but not for d_1.

Example 2.4. If (X,d) is a metric space and Y is a subset of X, then Y with the metric $d|_Y$ that is d with its domain restricted to $Y \times Y$ is also a metric space (check!). For example, each subset of $\mathbf{R}$ is a metric space using $d(x,y) = |x-y|$ for x and y in the subset.

Example 2.5. Every set X can be given the *discrete metric*

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

[2]For d_∞ to make sense requires each continuous function on $[0,1]$ to have a maximum value. This is the Extreme Value Theorem, which we'll prove later as Theorem 8.2. It is convenient to have d_∞ available strictly for examples before then.

[3]For $p \geq 1$ the function $d_p(f,g) = \sqrt[p]{\int_0^1 |f(x) - g(x)|^p\, dx}$ is a metric on $C[0,1]$, and as $p \to \infty$, $d_p(f,g) \to \max_{0\leq x\leq 1} |f(x) - g(x)|$, which is why the metric d_∞ has the notation it does.

which sets all pairs of (distinct) points in X at distance 1 from each other. All three conditions for being a metric are easy to check.

Example 2.6. If d is a metric on X, then the functions $d'(x,y) = \min(d(x,y),1)$ and $d''(x,y) = d(x,y)/(1+d(x,y))$ are also metrics on X. The only part that requires careful checking is the triangle inequality, which is left to the reader. Note d' and d'' are both bounded: they never take a value larger than 1. These two metrics are similar to the original metric d when d has value at most 1:

$$d(x,y) \leq 1 \Longrightarrow d'(x,y) = d(x,y) \quad \text{and} \quad \frac{1}{2}d(x,y) \leq d''(x,y) \leq d(x,y). \tag{2.3}$$

If $d(x,y) > 1$ then d' and d'' change the distance between x and y in different ways: d' redefines the distance to be 1, while d'' makes the distance less than 1 in a smoother way.

3. Limit of a sequence in a metric space

Armed with a notion of distance, as codified in a choice of a metric, we can carry over the definition of the limit of a sequence from Euclidean space to metric spaces.

Definition 3.1. For a sequence x_n in a metric space (X,d), we say x_n *converges to* $x \in X$, and write $\lim_{n\to\infty} x_n = x$ or $x_n \to x$, if for every $\varepsilon > 0$ there is an $N = N_\varepsilon$ such that

$$n \geq N \Longrightarrow d(x_n, x) < \varepsilon.$$

If a sequence in (X,d) has a limit we say the sequence is *convergent*.

Example 3.2. For $x \in X$, the constant sequence $\{x, x, x, \ldots\}$ is convergent with limit x. Similarly, an eventually constant sequence ($x_n = x$ for all large n) is convergent. This is true no matter what metric is used.

Example 3.3. If d is the discrete metric on X then a convergent sequence must be eventually constant: if $d(x_n, x) < 1$ for large n then $x_n = x$ for large n.

These examples are boring. The reader should know many interesting examples of convergent sequences in $\mathbf{R}$ from calculus.

Saying $x_n \to x$ in a metric space (X,d) is the same as saying $d(x_n, x) \to 0$ in $\mathbf{R}$: convergence of a sequence to a specific value means the distance between the terms of the sequence and that value tends to 0.

To get used to the terminology, let's prove four intuitively reasonable theorems about convergent sequences (try drawing a picture for each one).

Theorem 3.4. *If a sequence $\{x_n\}$ in a metric space (X,d) converges then $d(x_n, x_{n+1}) \to 0$.*

Proof. Showing the numbers $d(x_n, x_{n+1})$ tend to 0 means we want to show for every $\varepsilon > 0$ that there is an N such that $n \geq N \Longrightarrow d(x_n, x_{n+1}) < \varepsilon$. (We don't need to say $|d(x_n, x_{n+1})| < \varepsilon$ since values of a metric are nonnegative.)

Say $\lim_{n\to\infty} x_n = x$. Using the triangle inequality,

$$d(x_n, x_{n+1}) \leq d(x_n, x) + d(x, x_{n+1}) = d(x_n, x) + d(x_{n+1}, x).$$

The two terms on the right get small when n is large, so $d(x_n, x_{n+1})$ gets small when n is large. To be precise, for $\varepsilon > 0$ also $\varepsilon/2 > 0$, so there's an $N \geq 1$ such that for all $m \geq N$ we have $d(x_m, x) < \varepsilon/2$. Therefore

$$n \geq N \Longrightarrow n+1 \geq N \Longrightarrow d(x_n, x_{n+1}) \leq d(x_n, x) + d(x_{n+1}, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

□

Remark 3.5. If we had stuck with ε rather than $\varepsilon/2$ all the way through the proof then we'd get 2ε at the end instead of ε, and then we'd have to say "Now go back and replace ε with $\varepsilon/2\ldots$" to get the desired conclusion. The idea of using $\varepsilon/2$ in place of ε in the middle in order to get a single ε at the end is called an $\varepsilon/2$ argument. This type of reasoning occurs all the time in analysis. Instead of $\varepsilon/2$ one might use $\varepsilon/3$, $\sqrt{\varepsilon}$, or $\varepsilon/2^n$ (the last one is good if we have a whole sequence of terms that need bounds whose sum is still less than ε).

Theorem 3.6. *Every subsequence of a convergent sequence in a metric space is also convergent, with the same limit.*

Proof. Let $x_n \to x$ in (X, d) and let $\{x_{n_i}\}$ be a subsequence of $\{x_n\}$. Then $n_1 < n_2 < \cdots$. Set $y_i = x_{n_i}$. We want to show $y_i \to x$.

For $\varepsilon > 0$ there is an N such that $n \geq N \implies d(x_n, x) < \varepsilon$. Since the integers n_i are increasing, we have $n_i \geq N$ if we go out far enough: there's an I such that $i \geq I \implies n_i \geq N \implies d(x_{n_i}, x) < \varepsilon$, so $d(y_i, x) < \varepsilon$. Thus $y_i \to x$. □

Theorem 3.7. *In a metric space (X, d), if two sequences $\{x_n\}$ and $\{x'_n\}$ converge to the same value then $d(x_n, x'_n) \to 0$.*

Proof. Suppose $x_n \to x$ and $x'_n \to x$. Then $d(x_n, x'_n) \leq d(x_n, x) + d(x, x'_n) = d(x_n, x) + d(x'_n, x)$ and the last two terms get small for large n. This suggests using an $\varepsilon/2$ argument.

For each $\varepsilon > 0$ there's an N_1 such that $n \geq N_1 \implies d(x_n, x) < \varepsilon/2$ and an N_2 such that $n \geq N_2 \implies d(x'_n, x) < \varepsilon/2$. Set $N = \max(N_1, N_2)$, so

$$n \geq N \implies d(x_n, x'_n) \leq d(x_n, x) + d(x'_n, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

□

The converse to Theorem 3.7 is false: sequences for which $d(x_n, x'_n) \to 0$ do not have to converge. After all, let $\{x_n\}$ be an arbitrary sequence and let $x'_n = x_n$ for all n, so $d(x_n, x'_n) = 0$ all the time.

The next result is a partial converse to Theorem 3.7.

Theorem 3.8. *In a metric space (X, d), if $x_n \to x$ and $\{x'_n\}$ is a sequence such that $d(x_n, x'_n) \to 0$ then $x'_n \to x$.*

Proof. This will be an $\varepsilon/2$ argument.

Pick $\varepsilon > 0$. We want to find an N such that $n \geq N \implies d(x'_n, x) < \varepsilon$.

There's an N_1 such that $n \geq N_1 \implies d(x_n, x) < \varepsilon/2$. Since the real numbers $d(x_n, x'_n)$ tend to 0, there's an N_2 such that $n \geq N_2 \implies d(x_n, x'_n) < \varepsilon/2$. Setting $N = \max(N_1, N_2)$, we have

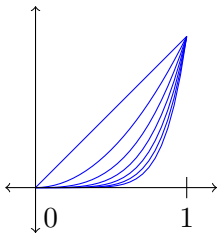
$$n \geq N \implies d(x'_n, x) \leq d(x'_n, x_n) + d(x_n, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

□

Example 3.9. On $\mathbf{R}^m$, because the metrics d_E and d_∞ are each bounded above by a constant multiple of the other (see (2.2)), we have $d_E(\mathbf{x}_n, \mathbf{x}) \to 0$ if and only if $d_\infty(\mathbf{x}_n, \mathbf{x}) \to 0$. Therefore convergence of sequences in $\mathbf{R}^m$ for both metrics means the same thing (with the same limits).

Example 3.10. In Example 2.6 we introduced two alternatives to a metric d that are both bounded metrics: $d'(x, y) = \min(d(x, y), 1)$ and $d''(x, y) = d(x, y)/(1 + d(x, y))$. Condition (2.3) shows d' and d'' have the *same* convergent sequences and limits as d.

Example 3.11. In $C[0,1]$ consider the sequence of functions x^n for $n \geq 1$, graphed below.



This sequence converges to 0 in the metric d_1 but not in the metric d_∞:

$$d_1(x^n, 0) = \int_0^1 |x^n|\, dx = \frac{1}{n+1} \to 0, \qquad d_\infty(x^n, 0) = \max_{0 \leq x \leq 1} |x^n| = 1.$$

In fact the sequence $\{x^n\}$ in $C[0,1]$ has no limit at all relative to the metric d_∞.

To prove $\{x^n\}$ has no limit in $(C[0,1], d_\infty)$, not just that the constant function 0 is not a limit, we seek a property that all convergent sequences satisfy and the sequence $\{x^n\}$ in $(C[0,1], d_\infty)$ does not satisfy. Theorem 3.4 tells us every convergent sequence in a metric space has the distance between consecutive terms tending to 0. Might $d_\infty(x^n, x^{n+1})$ not tend to 0? Well,

$$d_\infty(x^n, x^{n+1}) = \max_{0 \leq x \leq 1} |x^n - x^{n+1}| = \max_{0 \leq x \leq 1} (x^n - x^{n+1})$$

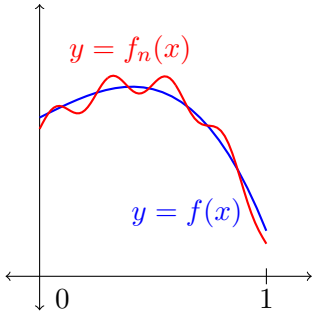
and by calculus $x^n - x^{n+1} = x^n(1-x)$ is is maximized on $[0,1]$ at $x = n/(n+1)$, where the value is $(n/(n+1))^n(1 - n/(n+1)) \sim (1/e)(1/(n+1)) \to 0$. Our idea failed to help.

Rather than looking at the distance between consecutive terms in a sequence, we can look at the distance between the nth and $(2n)$th terms. In the proof of Theorem 3.4 it wasn't so crucial that the terms from the sequence were consecutive. The exact same reasoning used there works with the nth and $(2n)$th terms, so if the sequence $\{x^n\}$ in $(C[0,1], d_\infty)$ has a limit then $d_\infty(x^n, x^{2n}) \to 0$ as $n \to \infty$. Since

$$d_\infty(x^n, x^{2n}) = \max_{0 \leq x \leq 1} |x^n - x^{2n}| = \max_{0 \leq x \leq 1} x^n(1 - x^n)$$

and the function $x^n(1-x^n)$ on $[0,1]$ has its maximum value at $x = 1/\sqrt[n]{2}$, where $x^n(1-x^n) = (1/2)(1/2) = 1/4$, which is independent of n, this proves $\{x^n\}$ has no limit in $(C[0,1], d_\infty)$.

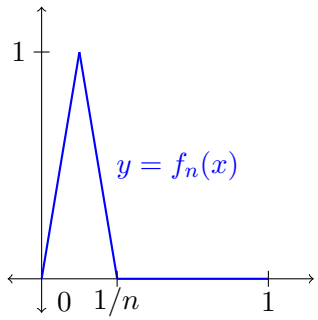
To know what a metric on a set X means, know what it means to say two points are close in that metric. For example, when we say $f_n \to f$ in $(C[0,1], d_\infty)$, it means the graph of f_n is getting close to f uniformly (at the same rate) over all of $[0,1]$ as $n \to \infty$. A picture of f (in blue) and an approximation f_n (in red) in the metric d_∞ is pictured below.



When $f_n \to f$ in $(C[0,1], d_\infty)$ we say $f_n \to f$ *uniformly* on $[0,1]$. This implies pointwise convergence: $f_n(a) \to f(a)$ for each $a \in [0,1]$ since

$$|f_n(a) - f(a)| \leq \max_{0 \leq x \leq 1} |f_n(x) - f(x)| = d_\infty(f_n, f) \to 0.$$

However the converse is false: if $f_n(a) \to f(a)$ for each $a \in [0,1]$ it does not mean $d_\infty(f_n, f) \to 0$. Pointwise convergence does not imply convergence can be controlled in the same way simultaneously on the whole domain $[0,1]$. Consider the functions $f_n(x)$ on $[0,1]$ graphed below which are an isoceles triangle of height 1 over $[0, 1/n]$ and 0 for $x \geq 1/n$. We have $f_n(0) = 0$ for all n, and for each $a \in (0,1]$ we have $f_n(a) = 0$ for large enough n, so $f_n(a) \to 0$ for each $a \in [0,1]$, but $d_\infty(f_n, 0) = 1$ so f_n does not get close to the function 0 as n gets large because every f_n has a peak of height 1.



Changing metrics from d_∞ to d_1, saying $f_n \to f$ in $(C[0,1], d_1)$ does not guarantee pointwise convergence. For example, $d_1(x^n, 0) \to 0$ and numerically $a^n \to 0$ for $0 \leq a < 1$, but *not* at $a = 1$.

4. Cauchy sequences and completeness

One aspect of infinite series that many students find hard to understand is how convergence tests really work: one may prove with the comparison test that $\sum_{k \geq 1} x^k/k^2$ converges for x in the interval $[-1, 1]$, but what does it converge to? The subtlety here is that we are saying something converges without identifying the limit in a concrete way. (Saying "the series is what the limit is" sounds more circular than explanatory.) Often we want to prove a sequence (of numbers or functions or shapes) has a limit even if we don't yet have a tidy name for the limiting object. How can convergence be detected before the limit is known?

A clue is in Theorem 3.4: if $x_n \to x$ in a metric space (X, d) then $d(x_n, x_{n+1}) \to 0$. The conclusion makes no reference to the original limit x. Unfortunately, this property that the terms become "consecutively close" is not good enough to characterize convergence in general metric spaces. We saw this in Example 3.11, where the sequence of power functions x^n in $(C[0,1], d_\infty)$ does not converge but $d_\infty(x^n, x^{n+1}) \sim (1/e)(1/(n+1)) \to 0$. A more basic example is the harmonic series, which diverges and its partial sums $H_n = 1 + 1/2 + \cdots + 1/n$ are consecutively close: $|H_n - H_{n+1}| = 1/(n+1) \to 0$.

By making a slight change in the proof of Theorem 3.4, we get a much stronger conclusion than consecutive closeness, and this stronger conclusion will be exactly what we need.

Theorem 4.1. *If $\{x_n\}$ is a convergent sequence in a metric space (X, d) then the terms of the sequence become "uniformly close": for every $\varepsilon > 0$ there is an $N \geq 1$ such that $m, n \geq N \Longrightarrow d(x_m, x_n) < \varepsilon$.*

Proof. We run through the proof of Theorem 3.4 and make a few changes. Letting $x = \lim_{n\to\infty} x_n$, the triangle inequality tells us for all m and n that

$$d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) = d(x_m, x) + d(x_n, x).$$

We now make an $\varepsilon/2$ argument. For every $\varepsilon > 0$ there's an $N \geq 1$ such that for all $n \geq N$ we have $d(x_n, x) < \varepsilon/2$. Therefore

$$m, n \geq N \Longrightarrow d(x_m, x_n) \leq d(x_m, x) + d(x_n, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

□

This concept of uniform closeness, which is a property of a sequence involving no direct reference to a hypothetical limit, is much more stringent than consecutive closeness. For example, if $H_n = 1 + 1/2 + \cdots + 1/n$ then the numbers H_n are consecutively close (that is, $|H_n - H_{n+1}| \to 0$) but they are not uniformly close. It can be shown, for instance, that $|H_n - H_{2n}| \to \log 2 \approx .693$.

When calculus was acquiring rigorous foundations in the 19th century, it was realized that uniform closeness in Theorem 4.1 captures the idea of convergence for sequences without mentioning a limit for the sequence. This property is not actually called uniform closeness, but is named in honor of Cauchy, who articulated and used it in the setting of infinite series.

Definition 4.2. A sequence $\{x_n\}$ in a metric space (X, d) is called a *Cauchy sequence* if for every $\varepsilon > 0$ there is an $N = N_\varepsilon$ such that for all $m, n \geq N$ we have $d(x_m, x_n) < \varepsilon$.

Theorem 4.3. *Every convergent sequence in a metric space is a Cauchy sequence.*

Proof. This is Theorem 4.1. □

Corollary 4.4. *If (X, d) is a metric space and Y is a subset of X given the induced metric $d|_Y$, then each sequence in Y that converges in X is a Cauchy sequence in $(Y, d|_Y)$.*

Proof. A sequence $\{y_n\}$ in Y that converges in X is Cauchy in X by Theorem 4.3. Since the metric d on X is the metric we are using on Y, the Cauchy property of $\{y_n\}$ in X can be viewed as the Cauchy property in Y. □

Example 4.5. Consider the interval $(0, \infty)$ as a metric space using the absolute value metric induced from $\mathbf{R}$. We have $1/n \to 0$ in $\mathbf{R}$, but the sequence $\{1/n\}$ has no limit in $(0, \infty)$ since $0 \notin (0, \infty)$. The sequence $\{1/n\}$ is a Cauchy sequence in $(0, \infty)$ by Corollary 4.4 but it is not a convergent sequence in $(0, \infty)$.

Example 4.6. On $\mathbf{R}^m$, the metrics d_E and d_∞ satisfy $d_\infty \leq d_E \leq \sqrt{m}\, d_\infty$, so a sequence in $\mathbf{R}^m$ is Cauchy with respect to one of these metrics if and only if it is Cauchy with respect to the other one.

Being a Cauchy sequence means if you go out far enough into the sequence then *all* the terms from some point onwards are as close together as you wish. While this property is much stronger than being a consecutively close sequence, a sequence whose terms get consecutively close *rapidly enough* is a Cauchy sequence. The next theorem says that being consecutively close at least at the rate of a geometric progression is rapid enough.

Theorem 4.7. *If $\{x_n\}$ is a sequence in a metric space (X, d) such that $d(x_n, x_{n+1}) \leq ar^n$ for all n, where $a > 0$ and $0 < r < 1$, then $\{x_n\}$ is a Cauchy sequence.*

Proof. For $1 \leq m < n$, a massive use of the triangle inequality tells us

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + \cdots + d(x_{n-1}, x_n) \\ &\leq ar^m + ar^{m+1} + \cdots + ar^{n-1} \\ &< \sum_{k=m}^{\infty} ar^k \\ &= \frac{ar^m}{1-r}. \end{aligned}$$

The bound $d(x_m, x_n) < ar^m/(1-r)$ is also true if $m = n$ since $d(x_m, x_m) = 0$.

Since $0 < r < 1$, the terms ar^n tend to 0 as $n \to \infty$. Now if we pick an $\varepsilon > 0$, choose N large enough that $r^N < (1-r)\varepsilon/a$. (Why $(1-r)\varepsilon/a$ and not ε? You'll soon see.) For $m, n \geq N$, without loss of generality $m \leq n$ so by our prior calculation $d(x_m, x_n) < ar^m/(1-r) \leq ar^N/(1-r)$, which is less than ε, so our sequence is Cauchy. □

More generally, if $d(x_n, x_{n+1}) \leq c_n$ and the infinite series $\sum_{n=1}^{\infty} c_n$ in $\mathbf{R}$ converges then $\{x_n\}$ is a Cauchy sequence.

Just as convergence is preserved when passing to a subsequence (Theorem 3.6), so is the Cauchy property.

Theorem 4.8. *Every subsequence of a Cauchy sequence in a metric space is also a Cauchy sequence.*

Proof. Adapt the proof of Theorem 3.6. Details are left to the reader. □

Theorem 3.8 also has an analogue for Cauchy sequences.

Theorem 4.9. *In a metric space (X, d), if $\{x_n\}$ is a Cauchy sequence and $\{x'_n\}$ is a sequence such that $d(x_n, x'_n) \to 0$ then $\{x'_n\}$ is a Cauchy sequence.*

Proof. Use $d(x'_m, x'_n) \leq d(x'_m, x_m) + d(x_m, x_n) + d(x_n, x'_n)$. and an $\varepsilon/3$ argument to adapt the proof of Theorem 3.8. □

A Cauchy sequence in a metric space X should be thought of as a sequence that wants to have a limit in X, but we saw in Example 4.5 that not all Cauchy sequences in X necessarily have a limit in X. If a metric space has a non-convergent Cauchy sequence then we should imagine the space has a point missing where the "ideal limit" for that non-convergent Cauchy sequence ought to be. A metric space in which all Cauchy sequences converge has no missing points, and such spaces have a special name.

Definition 4.10. A metric space (X, d) is called *complete* if every Cauchy sequence in X converges in X: if $\{x_n\}$ is Cauchy in X then there's an $x \in X$ such that $x_n \to x$.

Convergent sequences are always Cauchy (Theorem 4.3), and metric spaces are complete when the converse is true, so Cauchy = convergent for sequences in complete metric spaces. Analysis uses completeness of a metric since often the only way to construct a limit is to create a Cauchy sequence first and then pass to its limit.

Let's take a look at some complete and incomplete metric spaces.

Example 4.11. The metric space $\mathbf{R}$ with the absolute value metric $d(x, y) = |x - y|$ is complete. A proof of this uses fundamental properties of the real numbers like the existence of least upper bounds for nonempty bounded subsets of $\mathbf{R}$. See [3, Theorem 10.11].

Example 4.12. The closed intervals $[0,1]$ and $[0,\infty)$ with metric from $\mathbf{R}$ are complete. The open intervals $(0,1)$ and $(0,\infty)$ with metric from $\mathbf{R}$ are not complete: a sequence in the interval that tends to 0 is Cauchy but does not converge in the interval.

Example 4.13. The rational numbers $\mathbf{Q}$ with the absolute value metric $d(r,s) = |r-s|$ are not complete. To prove this, pick an irrational real number L (like $\sqrt{2}$ or π) and let r_n be the sequence of decimal approximations to L truncated at the $1/10^n$ place. Each finite decimal is rational, so $r_n \in \mathbf{Q}$. Since $r_n \to L$ in $\mathbf{R}$, $\{r_n\}$ is a Cauchy sequence in $\mathbf{Q}$ (Corollary 4.4). However, $\{r_n\}$ has no limit in $\mathbf{Q}$, so $\mathbf{Q}$ is not complete.

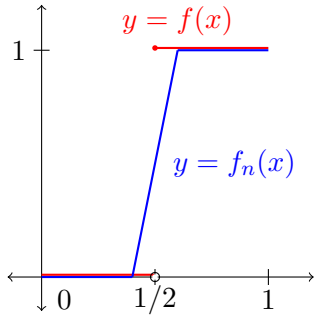
Example 4.14. The integers with the absolute value metric are complete: each Cauchy sequence in $\mathbf{Z}$ (as a metric space inside $\mathbf{R}$) is eventually constant. This is boring.

Example 4.15. A set X with the discrete metric on it (Example 2.5) is complete: if d is the discrete metric and $d(x_m, x_n) < 1$ then $x_m = x_n$, so every Cauchy sequence for the discrete metric is an eventually constant sequence, which clearly converges.

Example 4.16. The metric space $(C[0,1], d_\infty)$ is complete. A proof is in Appendix A.

Example 4.17. The metric space $(C[0,1], d_1)$ is not complete. To prove this we will follow the idea of Example 4.13 by writing down a discontinuous function that is a d_1-limit of continuous functions. Note the d_1-metric makes sense for piecewise continuous functions on $[0,1]$, since they are integrable.

In the picture below, let $f(x)$ (in red) be the discontinuous function that's 0 for $0 \leq x < 1/2$ and 1 for $1/2 \leq x \leq 1$. For $n \geq 2$ let $f_n(x)$ be a piecewise linear approximation (in blue) breaking from values 0 and 1 at $x = 1/2 \pm 1/n$, where it's linear in between.



The region below $f_n(x)$ and above the x-axis to the left of $x = 1/2$ is a right triangle with a base of width $1/2 - (1/2 - 1/n) = 1/n$ and height $1/2$, so

$$\int_0^1 |f_n(x) - f(x)|\,dx = 2\int_0^{1/2} |f_n(x)|\,dx = 2 \cdot \frac{1}{2}\frac{1}{n}\frac{1}{2} = \frac{1}{2n}.$$

Therefore

$$d_1(f_m, f_n) = \int_0^1 |f_m(x) - f_n(x)|\,dx \leq \int_0^1 |f_m(x) - f(x)|\,dx + \int_0^1 |f(x) - f_n(x)|\,dx = \frac{1}{2m} + \frac{1}{2n},$$

which tends to 0 as $m, n \to \infty$. Thus $\{f_n\}$ is Cauchy in $(C[0,1], d_1)$, but it has no limit in this metric space. If $f_n \to g$ in $(C[0,1], d_1)$ then $\int_0^1 |f(x) - g(x)|\,dx = 0$, so $\int_0^b |f(x) - g(x)|\,dx = 0$ and $\int_b^1 |f(x) - g(x)|\,dx = 0$ for all b in $(0,1)$. Using $0 < b < 1/2$ in the first integral and $1/2 < b < 1$ in the second integral, $\int_0^b |g(x)|\,dx = 0$ for $0 < b < 1/2$ and $\int_b^1 |1 - g(x)|\,dx = 0$

for $1/2 < b < 1$. Therefore $g(x) = 0$ for $0 \leq x < 1/2$ and $g(x) = 1$ for $1/2 < x \leq 1$, but no value can be assigned to $g(1/2)$ to make this continuous.

When a metric space is not complete, like $(\mathbf{Q}, |\cdot|)$ or $(C[0,1], d_1)$, we want to "fill in all the holes" to create a complete metric space containing the original one. To describe this larger space, we need one more concept.

Definition 4.18. A subset $S \subset X$ is called *dense* if every element of X is the limit of a sequence in S or equivalently every open ball around an element of X contains an element of S.

Example 4.19. The rational numbers are dense in $\mathbf{R}$ but the integers are not dense in $\mathbf{R}$.

Definition 4.20. A *completion* of a metric space (X, d) is a complete metric space $(\widehat{X}, \widehat{d})$ that contains X, the metric $\widehat{d}$ restricted to X is d, and X is a dense subset of $\widehat{X}$.

Example 4.21. The completion of $(\mathbf{Q}, |\cdot|)$ is $(\mathbf{R}, |\cdot|)$. If we think of $\mathbf{R}$ inside $\mathbf{R}^2$ as the x-axis then the completion of $\mathbf{Q}$ is not $\mathbf{R}^2$ even though $\mathbf{R}^2$ is complete because $\mathbf{Q}$ is *not dense* in $\mathbf{R}^2$. The idea of a completion is filling in the holes (missing limits of Cauchy sequences) and not adding anything more, which is why the completion of a metric space is required to have the original space as a dense subset.

Theorem 4.22. *Every metric space has a completion.*

Proof. See Appendix A. □

5. Open and closed subsets

In calculus, a lot of attention is paid to intervals: open intervals (a, b), closed intervals $[a, b]$, and half-open intervals $(a, b]$ and $[a, b)$. Allowing the endpoints to be $\pm\infty$ introduces the distinction between bounded and unbounded intervals. There are theorems about continuous or differentiable functions on intervals (Intermediate Value Theorem, Extreme Value Theorem, Mean Value Theorem), definite integrals are defined on intervals, and the set of numbers where a power series converges is an interval. What is so special about intervals?

It turns out that the key feature of intervals for one application may be quite different from the key feature needed for another application. Over the next few sections we will isolate several of these features and present the generalization of each one in metric spaces. This leads to many special types of subsets of a metric space, such as open balls, closed balls, compact subsets, and connected subsets. We will see that

- open intervals (a, b) with $a, b \in \mathbf{R}$ generalize to open balls and connected subsets,
- closed intervals $[a, b]$ with $a, b \in \mathbf{R}$ generalize to closed balls, compact subsets, and connected subsets.

In this section we generalize open and closed intervals in $\mathbf{R}$ to open and closed balls of a metric space. Throughout, (X, d) is a metric space.

Definition 5.1. For $a \in X$ and $r \geq 0$, the *open ball* with center a and radius r is

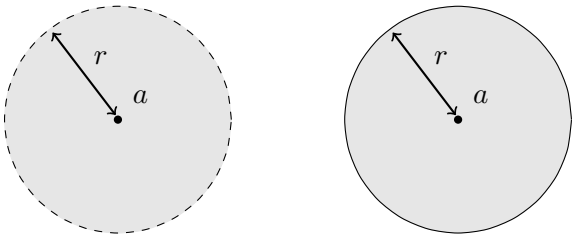
$$B(a, r) = \{x \in X : d(a, x) < r\}$$

and the *closed ball* with center a and radius r is

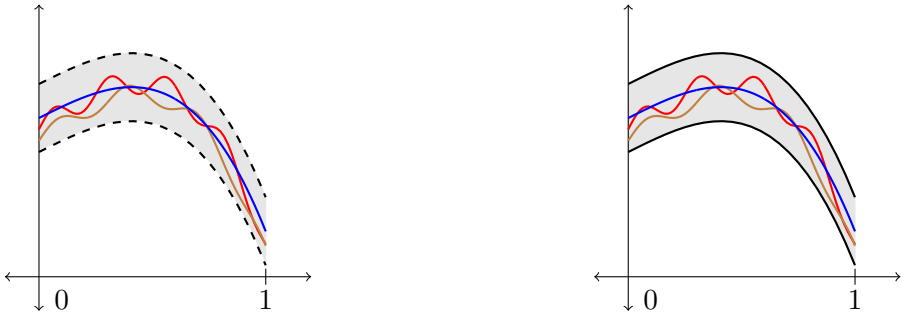
$$\overline{B}(a, r) = \{x \in X : d(a, x) \leq r\}.$$

When $r = 0$, $B(a,0) = \emptyset$ and $\overline{B}(a,0) = \{a\}$ by the axioms for a metric. Some writers only consider balls to have positive radius.

In a general metric space, the picture to have in mind of open and closed balls should be discs in the plane: the interior of a disc for open balls and the interior of a disc together with its boundary circle for closed balls, as shown below.



In specific metric spaces there are better pictures than these. In $\mathbf{R}$, an open ball $B(a,r)$ is the open interval $(a-r, a+r)$ and a closed ball $\overline{B}(a,r)$ is the closed interval $[a-r, a+r]$. Pictured below are the open and closed balls in $(C[0,1], d_\infty)$ of radius r around a function f (in blue), consisting of all functions whose graph deviates by less than r or at most r from the graph of f.



Definition 5.2. A subset of X is called *bounded* if it is contained in some ball $B(a,r)$. A subset that is not bounded is called *unbounded*.

Since $\overline{B}(a, r/2) \subset B(a,r) \subset \overline{B}(a,r)$, talking about a subset being contained in an open ball or a closed ball is the same thing: either case can be turned into the other by changing the radius if necessary. For example, the definition of a bounded subset does not change if we replace open balls in the definition with closed balls.

Theorem 5.3. *Every convergent sequence in a metric space is bounded.*

Proof. If $x_n \to x$ then there's an N such that $n \geq N \implies d(x_n, x) < 1$. Let

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

so $d(x_n, x) < r$ for all $n \geq 1$. Thus the whole sequence is contained in $B(x,r)$. □

Theorem 5.4. *Every Cauchy sequence in a metric space is bounded.*

Proof. Let $\{x_n\}$ be Cauchy. There's an N such that $m, n \geq N \implies d(x_m, x_n) < 1$. In particular, $d(x_N, x_n) < 1$ for $n \geq N$, so as in the proof of Theorem 5.3 (using x_N here in place of x there) an r can be found such that the whole sequence is in $B(x_N, r)$. □

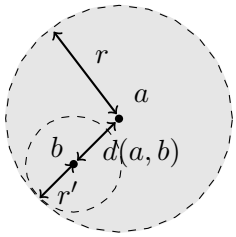
Since convergent sequences are Cauchy, Theorem 5.3 is a special case of Theorem 5.4.

Definition 5.5. A subset $U \subset X$ is called *open* if for each $x \in U$ there's an $r > 0$ such that $B(x, r) \subset U$. We also consider the empty subset of X to be an open subset.

The idea behind the concept of a subset U being open is that the property of being in U is stable under small perturbations: if we wiggle each element of U a little bit "in all directions" then we stay inside of U. Exactly how much we can wiggle a point in U without leaving U depends on how close we are to the "edge" of U.

Theorem 5.6. *Every open ball $B(a, r)$ in X is an open subset, and every open subset of X is a union of open balls in X.*

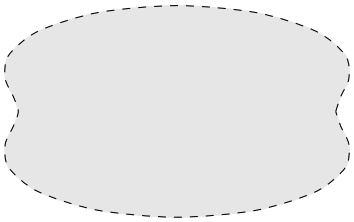
Proof. Since $B(a, 0) = \emptyset$ is open in X by definition, we can assume $r > 0$. Pick $b \in B(a, r)$. We want to find an $r' > 0$ such that $B(b, r') \subset B(a, r)$. The picture below suggests using $r' = r - d(a, b)$, which is positive since $d(a, b) < r$. A dashed circle with radius r' is drawn centered around b and just fits inside $B(a, r)$.



To show $B(b, r') \subset B(a, r)$, pick $x \in B(b, r')$. By the triangle inequality $d(x, a) \leq d(x, b) + d(b, a) < r' + d(a, b) = r$, so $x \in B(a, r)$.

If $U \subset X$ is open then each point in U is the center of an open ball that's inside U, by definition, so U is a union of open balls. □

A generic picture to have in mind for an open set in a metric space is a blob without its boundary, as in the picture below.



Theorem 5.7. *In a metric space, if $\{U_i\}$ is a collection of open subsets then $\bigcup_{i \in I} U_i$ is open. If $U_1, \ldots, U_n$ are finitely many open subsets then $U_1 \cap \cdots \cap U_n$ is open.*

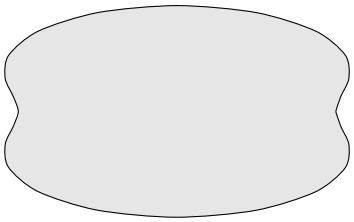
Proof. Each $x \in \bigcup_{i \in I} U_i$ is in some U_i, so there's an open ball $B(x, r)$ with $r > 0$ contained in that U_i. Thus every point of the union is in an open ball that's inside the union, so the union is an open subset.

For the second part, we may assume each U_i is nonempty, since otherwise the intersection is empty, and $\emptyset$ is open by definition. Pick $x \in U_1 \cap \cdots \cap U_n$. For each $j = 1, \ldots, n$ we have $B(x, r_j) \subset U_j$ for some $r_j > 0$, so $B(x, r) \subset U_1 \cap \cdots \cap U_n$ when $r = \min(r_1, \ldots, r_n) > 0$. □

It is false that the intersection of infinitely many open sets has to be open. For example, in $\mathbf{R}$ we can write $[0, 1] = \bigcap_{n \geq 1} (-1/n, 1 + 1/n)$ and $[0, 1]$ is not an open subset of $\mathbf{R}$.

Definition 5.8. A subset $C \subset X$ is called *closed* if for each sequence in C that has a limit in X, the limit is in C: if $c_n \in C$ and $c_n \to x \in X$ then $x \in C$. We also consider the empty subset of X to be a closed subset.

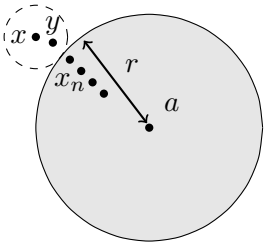
A subset being closed means limit operations don't take us outside the set. The following picture is a closed blob. It includes the boundary curve, so no sequence in the closed blob can have a limit outside the closed blob.



Theorem 5.9. *Every closed ball $\overline{B}(a,r)$ in X is a closed subset.*

Proof. We need to show for each sequence $\{x_n\}$ in $\overline{B}(a,r)$ with $x_n \to x \in X$ that $x \in \overline{B}(a,r)$. That is, if $x_n \to x$ and $d(a,x_n) \leq r$ for all n then $d(a,x) \leq r$.

Assume to the contrary that $d(a,x) > r$. We are going to prove from $x_n \to x$ that $d(a,x_n) > r$ for some n, which would be a contradiction since every x_n is in $\overline{B}(a,r)$.



Consider the open ball with center x and radius $r' = d(a,x) - r$. Then $r' > 0$. The picture above suggests that no element of $B(x,r')$ lies in $\overline{B}(a,r)$, and we can prove this using the triangle inequality:

$$y \in B(x,r') \Longrightarrow d(a,x) \leq d(a,y) + d(y,x) < d(a,y) + r' \Longrightarrow d(a,y) > d(a,x) - r' = r,$$

so $y \notin \overline{B}(a,r)$.

Since $x_n \to x$, some x_n is in $B(x,r')$, but we proved $B(x,r')$ is disjoint from $\overline{B}(a,r)$, so $x_n \notin \overline{B}(a,r)$ and that's a contradiction. Thus "$d(a,x) > r$" is false, so $d(a,x) \leq r$. □

Here is the "closed" counterpart of Theorem 5.7.

Theorem 5.10. *In a metric space, if $\{C_i\}$ is a collection of closed subsets then $\bigcap_{i\in I} C_i$ is closed. If $C_1, \ldots, C_n$ are finitely many closed subsets then $C_1 \cup \cdots \cup C_n$ is closed.*

Proof. We can assume $\bigcap_{i\in I} C_i \neq \emptyset$, since the empty set is closed by definition. Thus we can assume each C_i is nonempty.

Call the metric space X and let $\{c_n\}$ be a sequence in $\bigcap_{i\in I} C_i$ that converges to some $x \in X$. For each $i \in I$ we have $\{c_n\} \subset C_i$, so $x \in C_i$ since C_i is closed. Thus $x \in \bigcap_{i\in I} C_i$.

For the second part, we can assume at least one of $C_1, \ldots, C_n$ is nonempty, so the union is nonempty. Let $\{c_n\}$ be a sequence in $C_1 \cup \cdots \cup C_n$ with limit $x \in X$. We want to show x is in the union. A sequence has infinitely many terms, so one of $C_1, \ldots, C_n$ contains

infinitely many terms of the sequence. That is, a subsequence of $\{c_n\}$ lies entirely inside some C_j. Since each subsequence of a convergent sequence also converges with the same limit (Theorem 3.6), and C_j is closed, it follows that $x \in C_j \subset C_1 \cup \cdots \cup C_n$. □

It is false that the union of infinitely many closed subsets has to be closed. For example, in $\mathbf{R}$ we can write $(0,1) = \bigcup_{n\geq 1}[1/n, 1-1/n]$ and $(0,1)$ is not a closed subset of $\mathbf{R}$.

Theorem 5.11. *Every closed subset of a complete metric space is complete.*

Proof. Let C be a closed subset of X. Each Cauchy sequence in C is Cauchy in X, so it has a limit $x \in X$ by completeness of X. Since C is a closed subset of X, we must have $x \in C$. Thus all Cauchy sequences in C converge in C, so C is complete. □

Definition 5.12. A *limit point* of a subset $S \subset X$ is a point of X that is the limit of some sequence in S.

Definition 5.13. The *closure* of a subset $S \subset X$ is the union of S and its limit points in X. This set is denoted $\overline{S}$.

Example 5.14. For all $a \in \mathbf{R}^m$ and $r > 0$, $\overline{B(a,r)} = \overline{B}(a,r)$: the closure of every (nonemtpy) open ball in $\mathbf{R}^m$ is the closed ball with the same center and radius. In general metric spaces, this intuitively appealing property might be false.

Theorem 5.15. *For every subset $S \subset X$, its closure $\overline{S}$ is closed in X and $\overline{S}$ is the smallest closed subset of X containing S. In particular, S is closed if and only if $S = \overline{S}$.*

Proof. If $S \subset C \subset X$ and C is closed in X then each limit point of S lies in C, so $\overline{S} \subset C$. It remains to show $\overline{S}$ is closed.

Let $\{x_n\}$ be a sequence in $\overline{S}$ with limit $x \in X$. Since each x_n is a limit point of S, there's some $s_n \in S$ such that $d(x_n, s_n) < 1/n$. Then $x_n \to x$ and $d(x_n, s_n) \to 0$, so $s_n \to x$ by Theorem 3.8. Therefore $x \in \overline{S}$. □

Dense subsets were defined in Definition 4.18. They can also be described using closures.

Theorem 5.16. *A subset $S \subset X$ is dense if and only if $\overline{S} = X$.*

Proof. Saying S is dense means every element of X is a limit point of S, so $X \subset \overline{S}$. The reverse containment is true by definition. □

We have developed properties of open subsets and closed subsets separately. The following theorem shows they are in fact complementary concepts!

Theorem 5.17. *A subset of X is open if and only if its complement is closed.*

Proof. Pick an open subset U. To prove its complement $X - U$ is closed, we can assume $X - U \neq \emptyset$ since the empty set is closed by definition.

Let $\{x_n\}$ be a sequence in $X - U$ with limit $x \in X$. To prove $x \in X - U$, assume this is not true, so $x \in U$. Then there's some $r > 0$ such that $B(x,r) \subset U$. However, since $x_n \to x$ the ball $B(x,r)$ must contain some x_n, which is impossible since $x_n \in X - U$. Thus $x \notin U$.

Now pick a closed subset C. We want to prove its complement $X - C$ is open, and as before we can assume $X - C \neq \emptyset$ since the empty set is open by definition. Pick a point $x \in X - C$. We want to find an $r > 0$ such that $B(x,r) \subset X - C$. Assume there is no such r, so for every $r > 0$ the ball $B(x,r)$ contains an element of C. Using the sequence of radii $1/n$, for each $n \geq 1$ the ball $B(x, 1/n)$ contains some element, say x_n, of C. Then $d(x, x_n) < 1/n$ for all n, so $x_n \to x$. Since every x_n is in C and C is closed, we deduce that $x \in C$ too. That is a contradiction, so $X - C$ is open. □

Example 5.18. In $\mathbf{R}$, $(0,1)$ is open and its complement $(-\infty,0]\cup[1,\infty)$ is closed. The interval $[0,\infty)$ is closed and its complement $(-\infty,0)$ is open.

Theorem 5.17 is *not* saying every subset of a metric space is open or closed. Most subsets are neither. For example, in $\mathbf{R}$ the sets $[0,1)$ and $[0,1]\cup(2,3)$ are neither open nor closed.

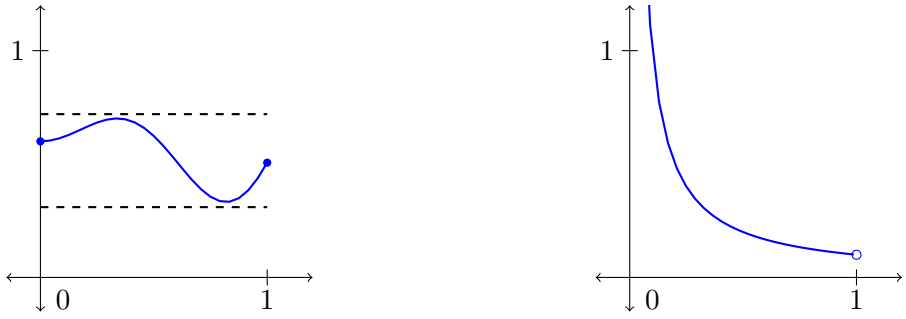
In light of Theorem 5.17, we now see that Theorems 5.7 and 5.10 are equivalent since complements exchange unions and intersections as well as open and closed subsets. For example, if C_i are closed subsets of X then their complements $U_i = X - C_i$ are open subsets and

$$X - \bigcap_{i\in I} C_i = \bigcup_{i\in I}(X - C_i) = \bigcup_{i\in I} U_i,$$

so from Theorem 5.6 saying $\bigcup_{i\in I} U_i$ is open, Theorem 5.17 implies that $\bigcap_{i\in I} C_i$ is closed.

6. Compact subsets

Closed bounded intervals are nice for functions: *every* continuous real-valued function on a closed bounded interval is bounded and has maximum and minimum values (see below on left), but $1/x$ on $(0,1)$ is unbounded above and has no minimum (see below on right).



What lies behind the better properties of continuous real-valued functions on closed bounded intervals turns out to be a property of the interval having nothing to do with functions: every sequence in a closed bounded interval has a subsequence converging in that interval. Even if a sequence itself does not converge, some subsequence does. In a general metric space X, subsets with this property get a special name.

Definition 6.1. A subset K of a metric space is called *compact* if every sequence in K has a subsequence that converges in K.

The notion of a compact set (in a metric space) was first defined by Fréchet. We will see in Section 8 some reasons why it is important. It is not an exaggeration to say compactness is one of the most important concepts in mathematics. Its initial applications were in analysis, but it is used in geometry, number theory, and even mathematical logic.

Here we will give some examples (and non-examples) and discuss some properties. We start with the fundamental example.

Theorem 6.2. *Every closed bounded interval $[a,b]$ in $\mathbf{R}$ is compact.*

Proof. Pick a sequence $\{x_n\}$ in $[a,b]$. All the terms of the sequence are within distance $b-a$ of each other. To extract a convergent subsequence, we use a repeated bisection method.

Break up $[a, b]$ into two halves: $[a, b] = [a, m] \cup [m, b]$ where $m = (a+b)/2$ is the midpoint. Infinitely many of the terms in the sequence $\{x_n\}$ have to be in $[a, m]$ or infinitely many have to be in $[m, b]$ (or maybe both happen). Focusing on a subinterval with infinitely many x_n in it, we get a subsequence denoted $x_n^{(1)}$ in which all terms are within $(b-a)/2$ of each other. Take that subinterval we chose and divide it into left and right halves (overlapping at the midpoint). Once again, infinitely many terms of $\{x_n^{(1)}\}$ have to be in one of the two halves, so by passing to the terms of the subsequence in such a half we get a new (refined) subsequence $\{x_n^{(2)}\}$ in which all the terms are within $(b-a)/4$ of each other.

Repeating this process, we get for each $k \geq 0$ a subsequence $\{x_n^{(k)}\}$ with $n = 1, 2, 3, \ldots$ in which all the terms are within $(b-a)/2^k$ of each other (with $x_n^{(0)} = x_n$). The way we construct these subsequences makes them nested:

$$\{x_n\} = \{x_n^{(0)}\} \supset \{x_n^{(1)}\} \supset \{x_n^{(2)}\} \supset \{x_n^{(3)}\} \supset \cdots$$

Now set $y_n = x_n^{(n)}$ for $n \geq 1$. The sequence $\{y_n\}$ is a subsequence of $\{x_n\}$ and $|y_n - y_{n+1}| \leq (b-a)/2^n$. Since the sequence $\{y_n\}$ gets consecutively close at least as quickly as the geometric progression $(b-a)/2^n$, it is a Cauchy sequence by Theorem 4.7. By completeness of $\mathbf{R}$, the sequence y_n has a limit in $\mathbf{R}$, and this limit is in $[a, b]$ since closed intervals are closed subsets (Theorem 5.9). Thus $\{x_n\}$ has a subsequence that converges in $[a, b]$. □

Example 6.3. The open interval $(0, 1)$ is not compact in $\mathbf{R}$ since the sequence $1/2$, $1/3$, $1/4, \ldots, 1/n, \ldots$ does not have a convergent subsequence within $(0, 1)$: the sequence converges to 0, so all of its subsequences converge to 0, but $0 \notin (0, 1)$. In a similar way, no (nonempty) open interval in $\mathbf{R}$ is compact.[4]

Theorem 6.4. *Every closed bounded box $[a_1, b_1] \times \cdots \times [a_m, b_m]$ in $\mathbf{R}^m$ is compact.*

Proof. Let $\mathbf{x}_n = (x_{n1}, \ldots, x_{nm})$ be a sequence in the box. Look at the sequence of first components: $\{x_{n1}\}$ is a sequence in $[a_1, b_1]$, so by compactness of this interval (Theorem 6.2) there is a convergent subsequence $\{x_{n_i 1}\}$ where $n_1 < n_2 < \ldots$, with limit $y_1 \in [a_1, b_1]$.

Consider the subsequence $\mathbf{x}_{n_i} = (x_{n_i 1}, \ldots, x_{n_i m})$. The first components converge to y_1. Look now at the second components $x_{n_i 2}$: it is a sequence in $[a_2, b_2]$, so by compactness of this interval there is a convergent subsequence $x_{n_{i_j} 2}$ with limit $y_2 \in [a_2, b_2]$. The corresponding subsequence of first components $x_{n_{i_j} 1}$ still converges to y_1 since a subsequence of a convergent sequence has the same limit (Theorem 3.6).

Now the sub-subsequence $\mathbf{x}_{n_{i_j}}$ in the box has its first components converge to y_1 and its second components converge to y_2. Repeating this argument until we exhaust all the components, we will finally get a subsequence of $\{\mathbf{x}_n\}$ in which the kth components have a limit $y_k \in [a_k, b_k]$, so that subsequence converges to $(y_1, \ldots, y_m)$, which is in the box. □

Every bounded sequence in $\mathbf{R}^m$ lies in a closed bounded box, so Theorem 6.4 tells us that every bounded sequence in $\mathbf{R}^m$ has a convergent subsequence.[5]

Theorem 6.5. *Every closed and bounded subset of $\mathbf{R}^m$ is compact.*

[4] A student taking a real analysis course told me the instructor focused a lot on $[a, b]$ and the student didn't understand why there is a big fuss about distinguishing between $[a, b]$ and (a, b) since "they only differ in two points." Those two points make a *huge* difference, since it's why $[a, b]$ is compact and (a, b) is not.

[5] This property of bounded sequences in Euclidean space is called the Bolzano–Weierstrass theorem.

Proof. Let C be a closed and bounded subset of $\mathbf{R}^m$ and $\{c_n\}$ be a sequence in C. We want to show $\{c_n\}$ has a subsequence that converges in C.

Since C is bounded in $\mathbf{R}^m$ it lies in some open ball $B(\mathbf{a}, r)$, which in turn lies in the closed box $[a_1 - r, a_1 + r] \times \cdots \times [a_m - r, a_m + r]$. This box is compact (Theorem 6.4), so $\{c_n\}$ has a subsequence converging in this box, and the limit of this subsequence lies in C since C is closed. □

Example 6.6. Every closed ball in $\mathbf{R}^m$ is compact, since closed balls are closed subsets and are clearly bounded.

Theorem 6.7. *Every compact subset of a metric space is closed and bounded.*

Proof. Let K be a compact subset of the metric space X.

K is closed: Suppose $c_n \in K$ and $c_n \to x \in X$. We want to show $x \in K$. By compactness of K, there is a subsequence c_{n_i} with a limit $c \in K$. When a sequence converges, every subsequence converges to the same limit (Theorem 3.6), so $c_{n_i} \to x$. Thus $c = x$, so $x \in K$.

K is bounded: We will prove, contrapositively, that an unbounded subset S of a metric space is not compact. Pick $s_0 \in S$. Since S is unbounded, for each integer $n \geq 1$ there is an $s_n \in S$ such that $d(s_0, s_n) > n$. It should be intuitively clear that the sequence $\{s_n\}$ is not bounded. To prove this, we show every open ball in X can contain only finitely many s_n: if $s_n \in B(a, r)$, and then

$$n < d(s_0, s_n) \leq d(s_0, a) + d(a, s_n) < d(s_0, a) + r,$$

which is false for large enough n. Since no open ball contains infinitely many s_n, every subsequence of $\{s_n\}$ is unbounded. Therefore no subsequence of $\{s_n\}$ can converge, since convergent sequences are bounded (Theorem 5.3). □

Theorems 6.5 and 6.7 together give the following important characterization of compact subsets of Euclidean space.

Theorem 6.8. *A subset of $\mathbf{R}^m$ is compact if and only if it is closed and bounded.*[6]

Proof. By Theorem 6.7, every compact subset of a metric space is closed and bounded. By Theorem 6.5, every closed and bounded subset of $\mathbf{R}^m$ is compact. □

Theorem 6.8 is not true in general metric spaces: a closed and bounded subset of a metric space does *not* have to be compact. That is, the converse of Theorem 6.7 in some metric spaces is false.

Example 6.9. On $\mathbf{R}^m$ change the Euclidean metric d_E to one of the bounded metrics $\min(1, d_E)$ or $d_E/(1 + d_E)$. Convergent sequences and limits in $\mathbf{R}^m$ for these metrics are the same as for d_E (Example 3.10), so they define the same closed subsets (and, by taking complements, the same open subsets) of $\mathbf{R}^m$ as d_E does. All closed subsets of $\mathbf{R}^m$ are bounded in these metrics, but many closed subsets of $\mathbf{R}^m$ (like $\mathbf{R}^m$ itself) are not compact.

Here's a less weird counterexample to the converse of Theorem 6.7 (no strange metric).

Example 6.10. We will show in the complete metric space $(C[0, 1], d_\infty)$ that the closed unit ball $\overline{B}(0, 1)$ is not compact.[7] The sequence of functions x^n lies in this ball and we will

[6]This is called the Heine–Borel theorem.

[7]While compactness in $(C[0, 1], d_\infty)$ is not the same as being closed and bounded, there is a set of conditions in $(C[0, 1], d_\infty)$ useful for analysis that is equivalent to compactness. Google the Arzelà–Ascoli theorem.

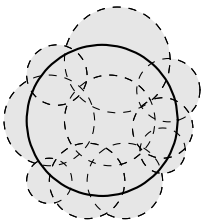
show it has no convergent subsequence. In Example 3.11 we showed this sequence is not convergent, but saying it has no convergent subsequence is much stronger.

Suppose a subsequence $\{x^{n_i}\}$ has a limit f in $(C[0,1], d_\infty)$. For each $a \in [0,1]$ we have

$$|a^{n_i} - f(a)| \leq \max_{0 \leq x \leq 1} |x^{n_i} - f(x)| = d_\infty(x^{n_i}, f),$$

and since $d_\infty(x^{n_i}, f) \to 0$ as $i \to \infty$ it follows that $a^{n_i} \to f(a)$ as $i \to \infty$. The exponents $n_1, n_2, \ldots$ are increasing, so if $0 \leq a < 1$ then $a^{n_i} \to 0$. Thus $f(a) = 0$. If $a = 1$ then $a^{n_i} = 1$ for all n_i, so $f(1) = 1$. But the function that's 0 on $[0,1)$ and 1 at 1 is not continuous.

The property of a subset S of a metric space X being compact can be described in a completely different way than with sequences and subsequences, using instead open coverings: an *open covering* of S is a collection of open subsets U_i in X such that $S \subset \bigcup_{i \in I} U_i$. Here is a picture of an open covering of a closed ball in the plane.



Theorem 6.11. *The following properties of a subset $K \subset X$ are equivalent.*

(1) *K is compact, i.e., every sequence in K has a subsequence converging in K.*
(2) *Every open covering of K in X has a finite subcovering: if $K \subset \bigcup_{i \in I} U_i$ where all U_i are open in X then there are $U_{i_1}, \ldots, U_{i_n}$ such that $K \subset U_{i_1} \cup \cdots \cup U_{i_n}$.*

Proof. (1) $\Longrightarrow$ (2):

Step 1: For each $r > 0$, K is contained in a finite union of balls of radius r centered at points in K.

We prove this by contradiction: assume K is not in a finite union of balls $B(x,r)$ for $x \in K$. Then starting with an $x_1 \in K$ we can build a sequence $\{x_n\}$ inductively: if $n \geq 2$ and we have $x_1, \ldots, x_n$ in K, by assumption K is not contained in $B(x_1, r) \cup \cdots \cup B(x_n, r)$, so there's some $x_{n+1} \in K$ that's not in that union. The sequence $\{x_n\}$ we have made has the property $d(x_{n+1}, x_k) \geq r$ for $k = 1, \ldots, n$ and all n, which is equivalent to saying $d(x_m, x_n) \geq r$ for all pairs of distinct integers m and n.

From (1), the sequence $\{x_n\}$ has a subsequence $\{x_{n_j}\}$ with a limit $x \in K$, so for all large enough n_j we have $d(x_{n_j}, x) < r/2$. Then for two different large n_j and $n_{j'}$ we get

$$d(x_{n_j}, x_{n_{j'}}) < d(x_{n_j}, x) + d(x, x_{n_{j'}}) < \frac{r}{2} + \frac{r}{2} = r,$$

and that contradicts the property at the end of the previous paragraph. This ends Step 1.

Step 2: Let $\{U_i\}_{i \in I}$ be an open covering of K in X. Then there is a number $r > 0$ such that for all $x \in K$ the ball $B(x,r)$ is in some U_i. (Stop at this point and check for $K = [0,1]$ and its open covering by the two intervals $(-.3, .75)$ and $(.5, 1.5)$ that this claim is true with $r = .25$.)

We will prove r exists by assuming it does not and getting a contradiction. If there is no such r then we can't use the numbers $1/n$ for $n = 1, 2, \ldots$ as r, so for each $n \geq 1$ there's an $x_n \in K$ such that $B(x_n, 1/n)$ is not in each U_i. We have built a sequence $\{x_n\}$ in K, so by (1) there is a subsequence $\{x_{n_j}\}$ converging to some $x \in K$. The point x has to be in some

member of the open covering, say $x \in U_i$. Since U_i is open, we have $B(x, 1/m) \subset U_i$ for some $m \geq 1$. Since $n_j \to \infty$ and $d(x_{n_j}, x) \to 0$, for suitably large n_j we have both $n_j > 2m$ and $d(x_{n_j}, x) < 1/(2m)$. For that n_j we have the implication

$$y \in B(x_{n_j}, 1/n_j) \Longrightarrow d(y, x) \leq d(y, x_{n_j}) + d(x_{n_j}, x) < \frac{1}{n_j} + \frac{1}{2m} < \frac{1}{2m} + \frac{1}{2m} = \frac{1}{m},$$

so $B(x_{n_j}, 1/n_j) \subset B(x, 1/m) \subset U_i$, but that is a contradiction since no $B(x_n, 1/n)$ lies inside a member of the open covering. Thus some r exists with the desired property.

Step 3: Every open covering $\{U_i\}$ of K in X has a finite subcovering containing K.

Using r as in Step 2, by Step 1 there are $x_1, \ldots, x_n \in K$ such that $K \subset \bigcup_{k=1}^n B(x_k, r)$. By the choice of r, each $B(x_k, r)$ is in some U_i, so K is contained in a finite union of members of the open covering. This completes the proof that (1) $\Rightarrow$ (2).

(2) $\Longrightarrow$ (1): Let $\{x_n\}$ be a sequence in K. We want to show (2) implies there is a convergent subsequence, or equivalently the sequence $\{x_n\}$ has a limit point in K. (This includes the possibility that a point occurs infinitely often in the sequence, making it a limit of a constant subsequence.) Assume there is no limit point: no point in K is the limit of a subsequence of $\{x_n\}$. Then for each $y \in K$ there must be an $r_y > 0$ such that the ball $B(y, r_y)$ contains only finitely many terms from $\{x_n\}$: if every ball centered at y had infinitely many terms from $\{x_n\}$ in it then we could build a subsequence tending to y by using radius 1, 1/2, 1/3, ...

The balls $B(y, r_y)$ for $y \in K$ are an open covering of K, so by (2) there is a finite subcovering: K is contained in a union of finitely many of these balls. Each of these balls has only finitely many terms from $\{x_n\}$ in it, so we'd get that the sequence $\{x_n\}$ has only finitely many terms, which is absurd. □

We call condition (2) in Theorem 6.11 the open covering criterion for compactness. It leads to a second proof that compact subsets of metric spaces are closed and bounded (Theorem 6.7).

Compact subsets are closed: (This will be similar to the proof that (2) $\Longrightarrow$ (1) above.) If $\{x_n\}$ is a sequence in K that converges to some $x \in X$ then we want to show $x \in K$. Every subsequence of $\{x_n\}$ also tends to x, so if x is not in K then every element of K is contained in an open ball that contains only finitely many terms of the sequence $\{x_n\}$. (If this were not true then some point in K would be the limit of a subsequence, which is impossible since all subsequences tend to x.) These open balls are an open covering of K, so by the open covering criterion for compactness we can extract a finite subcovering, but that implies the sequence has only finitely many terms, a contradiction.

Compact subsets are bounded: One open covering of K is $\bigcup_{x \in K} B(x, 1)$. By the open covering criterion for compactness there is a finite subcovering, so $K \subset \bigcup_{k=1}^n B(x_k, 1)$ for some finite set of points $x_1, \ldots, x_n$ in K. Thus K is in a finite union of balls, so it is bounded.

As an application of the open covering formulation of compactness, we show that all compact subsets of $\mathbf{R}$, no matter how complicated they may be, share a property with closed bounded intervals: they contain maximum and minimum elements.

Theorem 6.12. *For every nonempty compact subset K of $\mathbf{R}$ there are $a \in K$ and $b \in K$ such that all $x \in K$ satisfy $a \leq x \leq b$.*

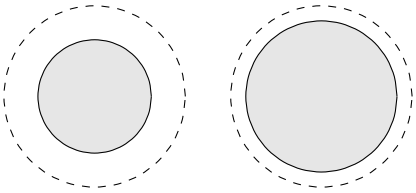
Proof. Suppose K does not have a maximum element. Then for each $x \in K$ there is $y > x$ in K, so $x \in (-\infty, y)$. Thus $K \subset \bigcup_{y \in K} (-\infty, y)$. This open covering of K has a finite

subcovering, say $K \subset (-\infty, y_1) \cup \cdots \cup (-\infty, y_n)$ with $y_i \in K$. But $\max(y_1, \ldots, y_n)$ is in K and is not in the finite subcovering, so we have a contradiction. The proof that K has a minimum element is similar. □

If we try covering compact sets with subsets slightly different from open subsets, there may not be a finite subcovering. Consider $[0, 1]$ covered by the intervals $[1/(n+1), 1/n]$ together with $\{0\}$, or by the intervals $[1/2 + 1/2^n, 1]$ for $n \geq 2$ together with $[0, 1/2]$.

7. Connected subsets

A connected subset of a metric space is a subset that is in "one piece." What does that mean? It's easier to say what it means *not* to be in one piece: the subset can be covered by two disjoint open sets. For example, two closed discs in the plane that don't overlap should not be considered to be one piece. We can surround them by two open discs that don't overlap, as shown below.



Definition 7.1. A subset S of a metric space X is called *connected* if whenever $S \subset U \cup V$ where U and V are disjoint open subsets of X, either $S \subset U$ or $S \subset V$. Equivalently, S is connected when it's impossible to have $S \subset U \cup V$ for disjoint open subsets with $S \cap U \neq \emptyset$ and $S \cap V \neq \emptyset$.

Example 7.2. The empty set and one-point subsets are connected. Verifying other examples requires work.

To say a metric space X is connected means that writing $X = U \cup V$ with U and V disjoint open subsets of X requires U or V to be empty. Since U and V are complements in X, saying U and V are open is the same as saying U is open and closed (or "clopen"), so X being connected is equivalent to saying there are no subsets of X that are both open and closed other than $\emptyset$ and X.

Theorem 7.3. *Every interval in* $\mathbf{R}$ *is connected.*

Proof. Step 1: Bounded open intervals are connected. We will show $(0, 1)$ is connected, but the same argument works with every open interval (a, b) where $a < b$ in $\mathbf{R}$.

Suppose $(0, 1) \subset U \cup V$ where U and V are disjoint open subsets of $\mathbf{R}$. Assume $(0, 1) \cap U$ and $(0, 1) \cap V$ are both nonempty. Setting $A = (0, 1) \cap U$ and $B = (0, 1) \cap V$, both A and B are open and we have $(0, 1) = A \cup B$ and $A \cap B = \emptyset$. Since A and B are nonempty by assumption, pick $a \in A$ and $b \in B$. Then $a \neq b$, so without loss of generality $a < b$.

Set $S = \{x \in A : x < b\}$, which is nonempty since $a \in S$. Since S is bounded above it has a least upper bound, say ℓ. Then $a \leq \ell \leq b$, so $\ell \in [a, b] \subset (0, 1) = A \cup B$.

If ℓ were in A, then $\ell \neq b$ so $\ell < b$. Since A is an open set in $\mathbf{R}$, some interval $(\ell - \varepsilon, \ell + \varepsilon)$ would lie in A. That means all numbers very close to ℓ on its right lie in A, but such numbers (taken close enough to ℓ) are less than b, and that contradicts ℓ being an upper bound on S. So $\ell \notin A$.

If ℓ were in B, which is also an open set in $\mathbf{R}$, then some interval around ℓ would lie entirely in B. However, since ℓ is the *least* upper bound of S there must be elements of S in every interval of the form $(\ell - \delta, \ell]$, and $S \subset A$, so we have a contradiction. Thus $\ell \notin B$.

Since ℓ is in neither A nor B, we have a final contradiction, so $(0,1)$ is connected.

Step 2: Unbounded open intervals are connected. Consider the case of $(0,\infty)$. Suppose $(0,\infty) \subset U \cup V$ where U and V are disjoint open subsets of $\mathbf{R}$. The interval contains 1, and without loss of generality $1 \in U$. For all $m > 1$ we have $(0,m) \subset (0,\infty) \subset U \cup V$ and $(0,m) \cap U \neq \emptyset$, so by connectedness of $(0,m)$ we get $(0,m) \subset U$. Since this holds for all $m > 1$, we get $(0,\infty) \subset U$. The same argument works for every unbounded open interval with one endpoint in $\mathbf{R}$. The only open interval left is $\mathbf{R} = (-\infty,\infty)$. Write it as $(-\infty,2) \cup (0,\infty)$ and use each part separately (both contain 1) to see $\mathbf{R}$ is connected.

Step 3: Other intervals are connected. Let I be a non-open interval and $I \subset U \cup V$ for disjoint open U and V in $\mathbf{R}$. Let J be I without its finite endpoints, so J is an open interval and thus we know J is connected. Since $J \subset U \cup V$, either $J \subset U$ or $J \subset V$. Without loss of generality, $J \subset U$. If an endpoint of I were in V then a small open interval around that endpoint would be in V (since V is open in $\mathbf{R}$), but this is absurd since each open interval around an endpoint of I contains elements of J, which are all in U. Therefore finite endpoints of I are in U too, so $I \subset U$ and this proves I is connected. □

Theorem 7.4. *Every nonempty subset of* $\mathbf{R}$ *that is not an interval is not connected.*

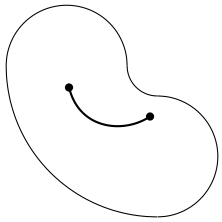
Proof. Say $S \subset \mathbf{R}$ is not an interval. Then S contains two points a and b, say with $a < b$, and does not contain some point c in between them. Let $U = (-\infty, c)$ and $V = (c, \infty)$. Then U and V are disjoint open subsets of $\mathbf{R}$, $S \subset U \cup V$, $a \in S \cap U$, and $b \in S \cap V$. □

Combining the last two theorems, the nonempty connected subsets of $\mathbf{R}$ are precisely the intervals. There is no simple characterization of connected subsets of $\mathbf{R}^m$ for $m > 1$. In practice nice subsets of $\mathbf{R}^m$ for $m > 1$ are proved to be connected by proving they have a stronger, more visually intuitive, property called being path-connected.

Definition 7.5. A subset S of a metric space X is called *path-connected* if, for every pair of points s and s' in S, there is a continuous function $p\colon [0,1] \to X$ such that $p(t) \in S$ for all t, $p(0) = s$, and $p(1) = s'$.

We call such a function p a path from s to s'. Since $q(t) = p(1-t)$ is also continuous with $q(0) = p(1) = s'$ and $q(1) = p(0) = s$, we can think of a path going in either direction, from s to s' or from s' to s.

The picture to have of a path-connected space is the inside of the blob below, where every pair of points can be linked by a path.

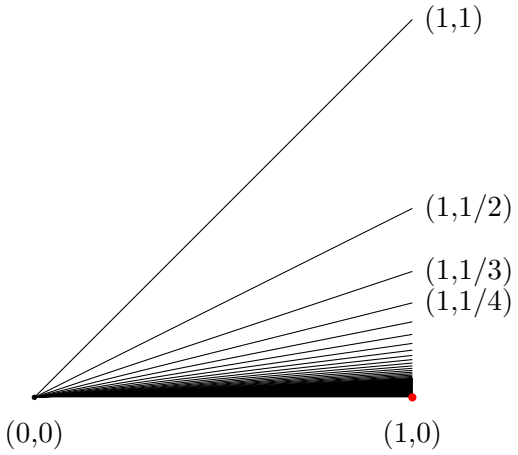


Theorem 7.6. *Every path-connected subset of a metric space is connected.*

Proving Theorem 7.6 requires information about continuous functions on metric spaces, the topic of Section 8, so we defer the proof until then (see after Example 8.12). Using

Theorem 7.6 we can "see" right away that nice regions in $\mathbf{R}^2$, $\mathbf{R}^3$, or $\mathbf{R}^m$ for $m > 3$ are connected because a path can be drawn within the region between every two of its points. For example, the surface of a sphere or a solid ball in $\mathbf{R}^3$ are path-connected and thus are connected. Theorem 7.6 does *not* imply $[0,1]$ is connected by being path connected since the proof of Theorem 7.6 uses the connectedness of $[0,1]$!

The converse of Theorem 7.6 in general is false: connectedness does not imply path-connectedness. An example is the "infinite broom" pictured below: it is the union of the closed line segments L_n from $(0,0)$ to $(1,1/n)$ as n runs over positive integers together with the (red) point $(1,0)$. The x-axis strictly between 0 and 1 is not part of the set. This is connected, but there is no path from $(1,0)$ to another point of the set. See [1] for a proof.



There is an important partial converse to Theorem 7.6: for *open subsets of* $\mathbf{R}^m$, being connected implies being path connected. The proof is omitted.

The concept most unlike being connected is being totally disconnected. A subset of a metric space is called *totally disconnected* if its only nonempty connected subsets are one-element subsets (a point is always connected). Examples of totally disconnected subsets of the metric space $\mathbf{R}$ include $\mathbf{Z}$, $\mathbf{Q}$, and fractals like the Cantor set. The p-adic integers and p-adic numbers, for a prime p, are important totally disconnected metric spaces in number theory. In $\mathbf{R}^m$ for $m > 1$ all open and closed balls are connected, which is a nice analogy with the one-dimensional case, but in a totally disconnected metric space no open or closed balls are connected (aside from closed balls of radius 0, *i.e.*, points).

8. Continuous functions between metric spaces

Up until now the only limits we have discussed in metric spaces were limits of sequences. Now we turn to limiting values of functions defined on a metric space, and specifically continuous functions on a metric space. This is where we will see the importance of connected sets and compact sets.

Specifically, we want to prove the following two properties of continuous functions.

Theorem 8.1 (Intermediate Value Theorem)**.** *Let I be an interval and $f\colon I \to \mathbf{R}$ be continuous. If $a < b$ in I and $f(a) \neq f(b)$ then for every y strictly between $f(a)$ and $f(b)$ there is $c \in (a,b)$ such that $f(c) = y$.*

Theorem 8.2 (Extreme Value Theorem)**.** *Every continuous real-valued function on a closed bounded interval has maximum and minimum values: if $f\colon [a,b] \to \mathbf{R}$ is continuous on a closed bounded interval then there are m and M such that (i) $m \leq f(x) \leq M$ for all x in $[a,b]$ and (ii) m and M are values of $f(x)$.*

The Extreme Value Theorem justifies the definition of the metric d_∞ on $C[0,1]$ back in Section 2 as a maximum value of a continuous function on $[0,1]$. In contrast to the Extreme Value Theorem, a continuous bounded function on an open interval doesn't have to have a maximum value, such as $1/x$ on $(1,2)$. Its values are bounded above by 1, but no value of the function is greater than all other values. The difference between closed bounded intervals and open intervals is that closed bounded intervals are compact, and we'll see that is what makes the Extreme Value Theorem work.

Before we prove theorems about continuous functions we have to define continuous functions. The definition of continuity for a real-valued function on an interval, usually called the (ε, δ)-definition, goes as follows. For a real number a and a real-valued function $f(x)$ defined on an interval containing a, we say $f(x)$ is *continuous at a* and write $\lim_{x\to a} f(x) = f(a)$ if for every $\varepsilon > 0$ there is a $\delta = \delta_{a,\varepsilon} > 0$ such that

$$|x - a| < \delta \Longrightarrow |f(x) - f(a)| < \varepsilon.$$

If we have a real-valued function $f\colon \mathbf{R}^m \to \mathbf{R}$, and $\mathbf{a} \in \mathbf{R}^m$, then we say $f(\mathbf{x})$ is *continuous at* $\mathbf{a}$ and write $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = f(\mathbf{a})$ if for every $\varepsilon > 0$ there is a $\delta = \delta_{\mathbf{a},\varepsilon} > 0$ such that

$$||\mathbf{x} - \mathbf{a}|| < \delta \Longrightarrow |f(\mathbf{x}) - f(\mathbf{a})| < \varepsilon.$$

Notice the different distances being used here, one on $\mathbf{R}^m$ (where the function is defined) and one on $\mathbf{R}$ (where the function takes its values).

Definition 8.3. A function $f\colon X \to Y$ between two metric spaces is called *continuous at* $a \in X$ if for every $\varepsilon > 0$ there is a $\delta = \delta_{a,\varepsilon} > 0$ such that

$$d_X(x,a) < \delta \Longrightarrow d_Y(f(x), f(a)) < \varepsilon.$$

If f is continuous at each point of X then we say f is *continuous on X*.

As a warm-up, let's show every metric is a continuous function on its metric space when we view it as a function of one of its variables, keeping the other one fixed.

Theorem 8.4. *For a metric space (X,d) and point $c \in X$, the function $f_c\colon X \to \mathbf{R}$ that is "distance to c", namely $f_c(x) = d(c,x)$, is continuous.*

Proof. Pick $a \in X$ and $\varepsilon > 0$. We need a $\delta > 0$ such that

$$d(x,a) < \delta \Longrightarrow |f_c(x) - f_c(a)| < \varepsilon.$$

The inequality on the right says $|d(c,x) - d(c,a)| < \varepsilon$.

Using the triangle inequality in two ways,

$$d(c,a) \leq d(c,x) + d(x,a) \quad \text{and} \quad d(c,x) \leq d(c,a) + d(a,x),$$

so

$$d(c,a) - d(c,x) \leq d(x,a) \quad \text{and} \quad d(c,x) - d(c,a) \leq d(a,x).$$

Thus $|d(c,x) - d(c,a)| \leq d(x,a)$. Therefore

$$d(x,a) < \varepsilon \Longrightarrow |d(c,x) - d(c,a)| \leq d(x,a) < \varepsilon$$

so we can use $\delta = \varepsilon$. □

Remark 8.5. By a two-way triangle inequality argument like the one used in this proof, show

$$|d(x,y) - d(x_n,y_n)| \leq d(x,x_n) + d(y,y_n)$$

for $x_n, x, y_n, y \in X$. Therefore if $x_n \to x$ and $y_n \to y$ in X then this inequality shows $d(x_n,y_n) \to d(x,y)$, an intuitively reasonable property.

Theorem 8.6. *For a metric space (X,d), the identity function $X \to X$ where $x \mapsto x$ is continuous.*

Proof. This is straightforward, using $\delta = \varepsilon$. □

Theorem 8.7. *Addition and multiplication, as functions $\mathbf{R}^2 \to \mathbf{R}$ given by $A(x,y) = x+y$ and $M(x,y) = xy$, are both continuous.*

Proof. First we prove continuity of addition. Pick $(a,b) \in \mathbf{R}^2$ and $\varepsilon > 0$. We need $\delta > 0$ such that

$$||(x,y) - (a,b)|| < \delta \Longrightarrow |A(x,y) - A(a,b)| < \varepsilon.$$

We will use $\delta = \varepsilon/2$.

If $||(x,y) - (a,b)|| < \varepsilon/2$ then $\sqrt{(x-a)^2 + (y-b)^2} < \varepsilon/2$, so $|x-a| < \varepsilon/2$ and $|y-b| < \varepsilon/2$. Then

$$|A(x,y) - A(a,b)| = |(x+y) - (a+b)| \leq |x-a| + |y-b| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

To prove multiplication is continuous we estimate $|M(x,y) - M(a,b)| = |xy - ab|$:

$$\begin{aligned} |xy - ab| &= |(x-a)y + (y-b)a| \\ &= |(x-a)(y-b) + (x-a)b + (y-b)a| \\ &\leq |x-a||y-b| + |x-a|b + |y-b|a. \end{aligned}$$

Thus if $|x-a| < \delta$ and $|y-b| < \delta$, then $|xy-ab| < \delta^2 + \delta b + \delta a = \delta(\delta + b + a)$. If

$$\delta \leq 1 \quad \text{and} \quad \delta \leq \frac{\varepsilon}{1+a+b} \tag{8.1}$$

then $\delta(\delta + b + a) \leq \delta(1 + b + a) \leq \varepsilon$. So pick $\delta = \min(1, \varepsilon/(1+a+b))$ to make δ satisfy the two inequalities in (8.1). Then

$$||(x,y) - (a,b)|| < \delta \Longrightarrow |x-a|, |y-b| < \delta$$

and our calculations above imply $|xy - ab| < \varepsilon$. □

In this proof notice that for continuity of multiplication our choice of δ depends not only on ε, but also on the point (a,b) where we are checking continuity. This is typical: in practice the choice for δ may depend on the point at which we are proving continuity. (In the definition of continuity at a, we wrote $\delta = \delta_{a,\varepsilon}$.) This did not happen for addition, where $\delta = \varepsilon/2$ everywhere. Such "independence of the point" is special; it will lead later to the concept of uniform continuity.

Theorem 8.8. *A composition of continuous functions on metric spaces is continuous: if $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous then the composite function $g \circ f\colon X \to Z$ is continuous.*

Proof. Pick $a \in X$ and $\varepsilon > 0$. We have $(g \circ f)(a) = g(f(a))$. By the definition of continuity of g at $f(a)$, there's an $\eta > 0$ (depending on $f(a)$ and ε) such that

$$d_Y(y, f(a)) < \eta \Longrightarrow d_Z(g(y), g(f(a))) < \varepsilon. \tag{8.2}$$

By the definition of continuity of f at a, there's a $\delta > 0$ (depending on η and a) such that

$$d_X(x, a) < \delta \Longrightarrow d_Y(f(x), f(a)) < \eta,$$

and by (8.2) that last inequality implies $d_Z(g(f(x)), g(f(a))) < \varepsilon$. □

Just as compactness has a formulation in terms of open sets rather than sequences (Theorem 6.11), using open coverings, continuity of a function on a metric space also has a formulation in terms of open sets rather than ε's and δ's. More precisely, continuity of a function can be expressed in terms of *inverse images* of open sets. For a function $f\colon X \to Y$ and a subset $S \subset Y$, the inverse image $f^{-1}(S)$ means $\{x \in X : f(x) \in S\}$. An inverse image of a function on a subset makes sense even if the function is not invertible. For example, if $f\colon \mathbf{R} \to \mathbf{R}$ by $f(x) = x^2$ then

$$f^{-1}((0,1)) = (-1,0) \cup (0,1) \qquad f^{-1}((-2,2)) = (-\sqrt{2}, \sqrt{2}),$$
$$f^{-1}((1,2)) = (-\sqrt{2}, -1) \cup (1, \sqrt{2}), \qquad f^{-1}((-1,0)) = \emptyset.$$

Inverse images of subsets under a function behave well for all set-theoretic operations: if $f\colon X \to Y$ is a function then for subsets S and T in Y,

$$f^{-1}(S \cap T) = f^{-1}(S) \cap f^{-1}(T), \qquad f^{-1}(S \cup T) = f^{-1}(S) \cup f^{-1}(T),$$
$$S \subset T \Longrightarrow f^{-1}(S) \subset f^{-1}(T), \qquad f^{-1}(S - T) = f^{-1}(S) - f^{-1}(T),$$

where $S - T = \{s \in S : s \notin T\}$. (For example, if $S = \{0,1\}$ and $T = \{1,2\}$ then $S - T = \{0\}$.) We will use these properties without comment below.[8] In particular, $f^{-1}(Y - S) = X - f^{-1}(S)$, so inverse images send complements to complements.

Theorem 8.9. *A function $f\colon X \to Y$ between two metric spaces is continuous if and only if the inverse image of every open set in Y is open in X: for all open U in Y, the set $f^{-1}(U) = \{x \in X : f(x) \in U\}$ is open in X.*

Proof. First suppose f fits the (ε, δ)-definition of continuity on X, so f is continuous at each element of X. For every open set $U \subset Y$ we want to show $f^{-1}(U)$ is open in X.

If $f^{-1}(U) = \emptyset$ then $f^{-1}(U)$ is open by our convention that the empty set is open, so suppose $f^{-1}(U) \neq \emptyset$. Pick $a \in f^{-1}(U)$, so $f(a) \in U$. Since U is open in Y there's some $\varepsilon > 0$ such that $B(f(a), \varepsilon) \subset U$. By the (ε, δ)-definition of continuity at a, there is a $\delta > 0$ such that $d_X(x, a) < \delta \Longrightarrow d_Y(f(x), f(a)) < \varepsilon$. This implication is saying $f(B(a, \delta)) \subset B(f(a), \varepsilon)$, so $B(a, \delta) \subset f^{-1}(B(f(a), \varepsilon)) \subset f^{-1}(U)$. This shows each a in $f^{-1}(U)$ is contained in an open ball that's contained in $f^{-1}(U)$, so $f^{-1}(U)$ is open in X.

Now we prove the converse. Suppose for all U open in Y we have $f^{-1}(U)$ open in X. For every $a \in X$ we will prove f is continuous at a. For each $\varepsilon > 0$, the open ball $B(f(a), \varepsilon)$ in Y is open, so using $B(f(a), \varepsilon)$ as U the inverse image $f^{-1}(B(f(a), \varepsilon))$ is open in X and this inverse image includes a. Thus there's a $\delta > 0$ such that $B(a, \delta) \subset f^{-1}(B(f(a), \varepsilon))$, and unwinding the notation this containment is saying that if $d_X(a, x) < \delta$ then $d_Y(f(a), f(x)) < \varepsilon$. That is exactly the (ε, δ)-definition of continuity of f at a. □

[8] Analogous formulas for images of subsets are true for unions and containments but false for intersections and complements: use $f(x) = x^2$ with $A = \{1,2\}$ and $B = \{1,-2\}$ to see $f(A \cap B) \neq f(A) \cap f(B)$ and $f(A - B) \neq f(A) - f(B)$.

It is *false* that continuous functions always send open sets to open sets. For example, the squaring function $\mathbf{R} \to \mathbf{R}$ sends $(-1, 1)$ to $[0, 1)$. Continuity aligns with inverse images of open sets, not images of open sets. C'est la vie.

Remark 8.10. Theorem 8.9 is an open-set formulation of continuity on a whole set, not at a particular point. There is an open-set formulation of continuity at a point: $f\colon X \to Y$ is continuous at a if and only if for all open $U \subset Y$ containing $f(a)$ there is an open $V \subset X$ containing a such that $f(V) \subset U$. Checking this matches the (ε, δ)-definition of continuity at a is left to the reader.

Corollary 8.11. *A function $f\colon X \to Y$ between two metric spaces is continuous if and only if the inverse image of every closed set in Y is closed in X.*

Proof. Theorem 5.17 tells us that open and closed subsets are complementary to each other. For a closed subset C of Y, $U = Y - C$ is open and $f^{-1}(U) = f^{-1}(Y - C) = X - f^{-1}(C)$. Therefore $f^{-1}(C)$ is closed if and only if $f^{-1}(U)$ is open, so Theorem 8.9 tells us continuity of f is equivalent to f^{-1} sending closed subsets to closed subsets. □

Example 8.12. In $\mathbf{R}^3$, the sphere $S^2 = \{(x, y, z) \in \mathbf{R}^3 : x^2 + y^2 + z^2 = 1\}$ is closed. This can be proved directly using sequences in S^2 or by observing that $f(x, y, z) = x^2 + y^2 + z^2$ is a continuous function $\mathbf{R}^3 \to \mathbf{R}$ and $S^2 = f^{-1}(1)$ with the one-point set $\{1\}$ in $\mathbf{R}$ being closed.

To illustrate the open-set formulation of continuity of a function, we now prove Theorem 7.6: path-connected sets are connected. The proof will use connectedness of $[0, 1]$.

Proof. Let S be a path-connected subset of a metric space X. We will use paths in S to show that if S is not connected then $[0, 1]$ is not connected, which of course is a contradiction, so S has to be connected.

Suppose S is not connected, so we have $S \subset U \cup V$ where U and V are nonempty disjoint open subsets of X. Pick $s \in S \cap U$ and $s' \in S \cap V$. There is a path $p\colon [0, 1] \to S$ where $p(0) = s$ and $p(1) = s'$. The partition of S into $S \cap U$ and $S \cap V$ leads via this path to a partition of $[0, 1]$: $[0, 1] = p^{-1}(U) \cup p^{-1}(V)$. Set $A = p^{-1}(U)$ and $B = p^{-1}(V)$. Both are open subsets of $[0, 1]$ since p is continuous.[9]

Note $0 \in A$ and $1 \in B$, so A and B are nonempty. Obviously A and B are disjoint, since no point in $[0, 1]$ can have its p-value in both U and V. Thus the equation $[0, 1] = A \cup B$ exhibits $[0, 1]$ as a disjoint union of two nonempty open subsets of $[0, 1]$, which contradicts the connectedness of $[0, 1]$. □

Next we reprove Theorem 8.8 using open sets instead of ε's and δ's.

Proof. Let U be an arbitrary open set in Z. Then

$$\begin{aligned}(g \circ f)^{-1}(U) &= \{x \in X : (g \circ f)(x) \in U\} \\ &= \{x \in X : g(f(x)) \in U\} \\ &= \{x \in X : f(x) \in g^{-1}(U)\} \\ &= f^{-1}(g^{-1}(U)).\end{aligned}$$

By continuity of g, $g^{-1}(U)$ is open in Y, and by continuity of f, $f^{-1}(g^{-1}(U))$ is open in X. Thus $g \circ f$ is continuous. □

[9] In $[0, 1]$, intervals of the form $[0, \varepsilon)$ are open! This is the ball $B(0, \varepsilon)$ in the metric space $[0, 1]$, even though the same interval is not open in $\mathbf{R}$.

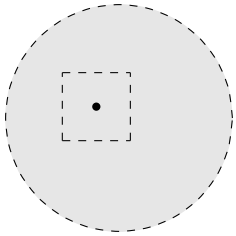
Think about why this proof of Theorem 8.8 and the first proof of Theorem 8.8 really are the same argument even though at first glance they might look different. Mathematicians consider the second proof, using open sets, to be more elegant.

As a substantial application of continuity being preserved under composition, we will prove polynomials with real coefficients are continuous. We'll need one lemma.

Lemma 8.13. *Let $f\colon \mathbf{R} \to \mathbf{R}$ and $g\colon \mathbf{R} \to \mathbf{R}$. If f and g are continuous then the function $\mathbf{R} \to \mathbf{R}^2$ given by $x \mapsto (f(x), g(x))$ is continuous.*

Proof. Set $F\colon \mathbf{R} \to \mathbf{R}^2$ by $F(x) = (f(x), g(x))$ and let U be an open set in $\mathbf{R}^2$. To prove $F^{-1}(U)$ is open in $\mathbf{R}$, we may assume $F^{-1}(U) \neq \emptyset$. Let $a \in F^{-1}(U)$, so $(f(a), g(a)) \in U$.

Around each point in U there is a small square, not just a small disc, centered at the point and contained in U. See the picture below.



Thus for some suitably small $\varepsilon > 0$ we have $(f(a)-\varepsilon, f(a)+\varepsilon) \times (g(a)-\varepsilon, g(a)+\varepsilon) \subset U$. By continuity of f and g, $V_f = f^{-1}((f(a)-\varepsilon, f(a)+\varepsilon))$ and $V_g = g^{-1}((g(a)-\varepsilon, g(a)+\varepsilon))$ are open in $\mathbf{R}$ and $a \in V_f \cap V_g \subset F^{-1}(U)$. Since $V_f \cap V_g$ is open, it contains an interval around a. Thus $F^{-1}(U)$ is open in $\mathbf{R}$. □

Theorem 8.14. *Every polynomial function in one variable with real coefficients is a continuous function $\mathbf{R} \to \mathbf{R}$.*

Proof. We will use Theorem 8.8 to get continuity of all polynomials from that of constant functions, $f(x) = x$, and addition and multiplication on $\mathbf{R}$. In particular, no ε's or δ's or open sets will appear. They were used in previous results that we will invoke.

First we prove by induction that x^n is continuous for each positive integer n. When $n = 1$ this is the identity function (Theorem 8.6). For $n \geq 2$ assume by induction that x^{n-1} is continuous. We can think of the function x^n as the composite $\mathbf{R} \to \mathbf{R}^2 \to \mathbf{R}$ where the first function is $x \mapsto (x, x^{n-1})$ and the second function is multiplication $(x, y) \mapsto xy$: their composite is $x \mapsto (x, x^{n-1}) \mapsto xx^{n-1} = x^n$. The first function is continuous by Lemma 8.13 and the second function is continuous by Theorem 8.7, so their composite is continuous.

Since x^n is continuous, a general monomial cx^n for $c \in \mathbf{R}$ can be regarded as a composite function $\mathbf{R} \to \mathbf{R}^2 \to \mathbf{R}$ where the first function is $x \mapsto (c, x^n)$ and the second function is multiplication $(x, y) \mapsto xy$. The first function is continuous by Lemma 8.13 since constant functions[10] and power functions are continuous. The second function is continuous by Theorem 8.7, so their composite is continuous.

Polynomials are finite sum of monomials, and we will prove they are continuous by induction on the number of monomials in the polynomial. The base case of monomials was proved above. A sum of two monomials, $ax^m + bx^n$, is a composite function $\mathbf{R} \to \mathbf{R}^2 \to \mathbf{R}$ where the first function is $x \mapsto (ax^m, bx^n)$ and the second function is addition $(x, y) \mapsto x+y$.

[10]Continuity of constant functions is easy to check.

Both of these functions are continuous by the base case, Lemma 8.13, and Theorem 8.7, so their composite is continuous. The general inductive step is left to the reader. □

Theorem 8.15. *Let $f\colon X \to Y$ be continuous. If $S \subset X$ is compact then $f(S)$ is compact in Y.*

Proof. We will give two proofs, one using the convergent subsequence description of compactness and the other using the open covering description of compactness.

First proof: Let $\{y_n\}$ be a sequence in $f(S)$, so we can write $y_n = f(x_n)$ for some $x_n \in S$. By compactness of S, the sequence $\{x_n\}$ in S has a convergent subsequence, say $x_{n_i} \to x \in S$. Then by continuity, $f(x_{n_i}) \to f(x)$, so $y_{n_i} \to f(x) \in f(S)$. We proved every sequence in $f(S)$ has a convergent subsequence, so $f(S)$ is compact.

Second proof: Let $\{U_i\}$ be an open covering of $f(S)$, so $f(S) \subset \bigcup_{i\in I} U_i$ in Y. Then $S \subset \bigcup_{i\in I} f^{-1}(U_i)$, and each $f^{-1}(U_i)$ is open in X, so $\{f^{-1}(U_i)\}$ is an open covering of S in X. By compactness of S there is a finite subcovering, say $S \subset f^{-1}(U_1) \cup \cdots \cup f^{-1}(U_n)$. Then $f(S) \subset U_1 \cup \cdots \cup U_n$, so every open covering of $f(S)$ has a finite subcovering. □

Theorem 8.16. *Let $f\colon X \to Y$ be continuous. If $S \subset X$ is connected then $f(S)$ is connected in Y.*

Proof. Suppose $f(S) \subset U \cup V$ where U and V are disjoint open subsets of Y. Then $S \subset f^{-1}(U \cup V) = f^{-1}(U) \cup f^{-1}(V)$. The inverse images $f^{-1}(U)$ and $f^{-1}(V)$ are open in X, and they are disjoint since U and V are disjoint (if $a \in f^{-1}(U) \cap f^{-1}(V)$ then $f(a) \in U \cap V = \emptyset$, a contradiction). Since S is connected, we have either $S \subset f^{-1}(U)$ or $S \subset f^{-1}(V)$, so $f(S) \subset U$ or $f(S) \subset V$. Thus $f(S)$ is connected. □

With Theorems 8.15 and 8.16 and our knowledge of connected and compact intervals in $\mathbf{R}$, we can now prove the Intermediate Value Theorem (Theorem 8.1) and Extreme Value Theorem (Theorem 8.2).

Proof. (Intermediate Value Theorem) Let I be an interval and $f\colon I \to \mathbf{R}$ be continuous. Suppose $a < b$ in I with $f(a) \neq f(b)$. The image $f([a,b])$ is connected in $\mathbf{R}$ by Theorem 8.16, so it must be an interval. An interval containing $f(a)$ and $f(b)$ contains all numbers between them, so every y strictly between $f(a)$ and $f(b)$ is $f(c)$ for some $c \in [a,b] \subset I$, and c is not a or b since y is not $f(a)$ or $f(b)$. □

Proof. (Extreme Value Theorem) Let $f\colon [a,b] \to \mathbf{R}$ be continuous. Since $[a,b]$ is compact and connected, $f([a,b])$ is compact and connected by Theorems 8.15 and 8.16. The connected subsets of $\mathbf{R}$ are intervals, and compact intervals are closed and bounded, so $f([a,b])$ must have the form $[m,M]$ for some $m, M \in \mathbf{R}$. The conclusions of the Extreme Value Theorem follow from this. □

Theorem 8.17. *If X is a compact metric space and $f\colon X \to \mathbf{R}$ is a continuous function with positive values everywhere then there is some $c > 0$ such that $f(x) \geq c$ for all $x \in X$, and we can take $c = f(x_0)$ for some $x_0 \in X$.*

Proof. The image $f(X)$ is compact in $\mathbf{R}$ by Theorem 8.15, so Theorem 6.12 implies there is an $x_0 \in X$ such that $f(x) \geq f(x_0)$ for all $x \in X$. Set $c = f(x_0) > 0$. □

Positive continuous functions on non-compact spaces need not have a positive lower bound. Consider $1/x$ as a function $(0,\infty) \to \mathbf{R}$. There is no $c > 0$ such that $1/x \geq c > 0$ for all $x > 0$.

Corollary 8.18. *If K is a compact subset of the metric space X then for each $x \in X$ the distance from x to the elements of K has a minimum: there is some $y_0 \in K$ such that $d(x,y) \geq d(x,y_0)$ for all $y \in K$.*

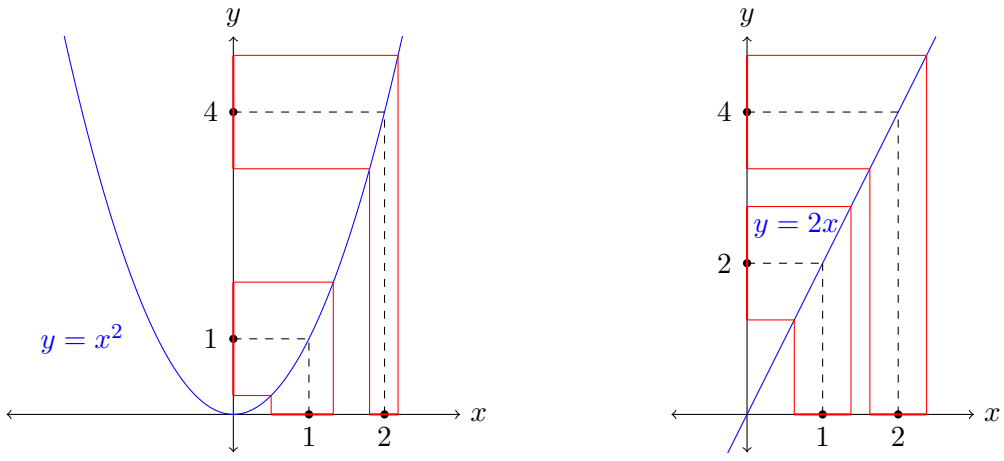
Proof. Let $f\colon K \to \mathbf{R}$ by $f(y) = d(x,y)$. This is continuous by Theorem 8.4 so its image $f(K)$ is compact in $\mathbf{R}$. By Theorem 6.12 (with K in that theorem being $f(K)$ here) some value $d(x,y_0)$ for $y_0 \in K$ is a minimum value of f. □

We now turn to the final topic about continuous functions that we'll discuss. It is an important refinement of continuity.

Definition 8.19. A function $f\colon X \to Y$ between two metric spaces is called *uniformly continuous* if for every $\varepsilon > 0$ there is a $\delta = \delta_\varepsilon > 0$ such that for all $a \in X$,

$$d_X(x,a) < \delta \Longrightarrow d_Y(f(x), f(a)) < \varepsilon.$$

Uniform continuity is a special kind of continuity. The difference is that uniform continuity says the value of δ can be chosen in terms of ε alone, independently of a choice of point a (so we can choose δ "uniformly in a," hence the name). Look at the graphs of $y = x^2$ and $y = 2x$ below. In each case we use $\varepsilon = .75$ and indicate in red the largest δ-intervals around $a = 1$ and $a = 2$ on the x-axis that make $|x - a| < \delta \Longrightarrow |f(x) - f(a)| < .75$. For $f(x) = x^2$ we need a shorter interval around $a = 2$ to keep the f-values within .75 of $f(2)$ than we need around $a = 1$ to keep f-values within .75 of $f(1)$, but for $f(x) = 2x$ intervals of the same radius around 1 and 2 lead to intervals of the same radius around $f(1)$ and $f(2)$. Linear functions on $\mathbf{R}$ are uniformly continuous while x^2 on $\mathbf{R}$ is not uniformly continuous.



It is instructive to compare the definitions of continuity and uniform continuity using quantifier notation. Continuity of a function $f\colon X \to Y$ says

$$\forall a \in X\, \forall \varepsilon > 0\, \exists\, \delta > 0 \text{ s.t. } \forall x \in X, d_X(x,a) < \delta \Longrightarrow d_Y(f(x), f(a)) < \varepsilon.$$

while uniform continuity of a function $f\colon X \to Y$ says

$$\forall \varepsilon > 0\, \exists\, \delta > 0 \text{ s.t. } \forall a \in X\, \forall x \in X, d_X(x,a) < \delta \Longrightarrow d_Y(f(x), f(a)) < \varepsilon.$$

The only difference is where "$\forall a \in X$" appears: in continuity it is at the start while in uniform continuity it is after δ. The order of quantifiers has a substantial impact on what something means. A simpler example: it's true that $\forall x \in \mathbf{R}$ s.t. $x > 0\, \exists\, y \in \mathbf{R}$ s.t. $y^2 = x$

(every positive real number has a real square root), but it's false that $\exists\, y \in \mathbf{R}$ s.t. $\forall x \in \mathbf{R}$ s.t. $x > 0$, $y^2 = x$ (there's a universal square root of all positive numbers?). Definitions in real analysis have many nested quantifiers, so it's understandable why even a mathematician of the stature of Cauchy, who introduced the (ε, δ)-definition[11] of continuity, confused it with uniform continuity in proofs.

Unlike continuity, uniform continuity can't be described as a property at individual points of the domain (why?). Also unlike continuity, there no way to convert uniform continuity into a statement about general open subsets.[12] Proofs about uniform continuity use (ε, δ)-language.

The last general comment we have about uniform continuity is that in its definition a and x play symmetric roles (they are quantified at the same time, unlike in the definition of continuity), so we can express uniform continuity in a way that puts a and x on an equal footing, which we therefore will write as x and x'. A function $f\colon X \to Y$ is called *uniformly continuous* if for every $\varepsilon > 0$ there is a $\delta = \delta_\varepsilon > 0$ such that for all $x, x' \in X$,

$$d_X(x, x') < \delta \Longrightarrow d_Y(f(x), f(x')) < \varepsilon.$$

This is the formulation of uniform continuity that we will use below. (Note: we are not saying all x and x' satisfy $d_X(x, x') < \delta$, but rather that they satisfy the total implication written above: if $d_X(x, x') < \delta$ then $d_Y(f(x), f(x')) < \varepsilon$.)

Example 8.20. We saw before that x^2 is not uniformly continuous on $\mathbf{R}$, since for a given ε the function values stay within ε of each other at larger numbers that are within δ of each other only if δ is made progressively smaller. We can't use the same δ as the points get too big. But if we restrict ourselves to a bounded domain then x^2 becomes uniformly continuous. For instance, on the interval $[-b, b]$ where $b > 0$,

$$|x - x'| < \delta \Longrightarrow |x^2 - x'^2| = |x - x'||x + x'| < \delta(|x| + |x'|) \leq \delta(2b),$$

so to make $|x^2 - x'^2| < \varepsilon$ when $|x - x'| < \delta$ in $[-b, b]$ we can use $\delta = \varepsilon/(2b)$.

Example 8.21. For $a > 0$ the function $1/x$ on $[a, \infty)$ is uniformly continuous:

$$|x - x'| < \delta \Longrightarrow \left|\frac{1}{x} - \frac{1}{x'}\right| = \frac{|x' - x|}{|x||x'|} < \frac{\delta}{a^2},$$

so we can make $|1/x - 1/x'| < \varepsilon$ if $|x - x'| < \delta$ in $[a, \infty)$ using $\delta = a^2\varepsilon$.

Example 8.22. For a metric space (X, d) and $c \in X$, the function $f_c(x) = d(c, x)$ is uniformly continuous on X. This comes from the proof of Theorem 8.4, where we could use $\delta = \varepsilon$ independently of the point x at which we were checking continuity of f_c.

Here is the fundamental property that gives rise to uniformly continuous functions.

Theorem 8.23. *A continuous function $f\colon X \to Y$ from a compact metric space to a metric space is uniformly continuous.*

This doesn't say continuous functions can't be uniformly continuous on non-compact metric spaces (see Example 8.21), only that they must be uniformly continuous on compact metric spaces.

[11] Bolzano had developed similar ideas earlier, but his work was not widely read until after Cauchy.

[12] A partial translation into open sets is possible, but the family of all balls of a common radius has to be given a special status, called a "uniformity". We don't get into that here.

Proof. We prove the theorem in two ways, using the subsequence description of compactness and using the open covering description of compactness.

First proof (using subsequences): Our argument will be by contradiction, starting with the assumption that f is *not* uniformly continuous. Figuring out what it means not to be uniformly continuous is going to take up the bulk of the proof!

One way to say f is not uniformly continuous is to say it is not true that for every $\varepsilon > 0$ there is a $\delta > 0$ such that all x and x' in X satisfy $d_X(x, x') < \delta \Longrightarrow d_Y(f(x), f(x')) < \varepsilon$. But this is not really progress; adding "It is not true that" to the start of a sentence in order to negate it doesn't usually let us understand the *meaning* of the negation.

The negation of a statement of the form "for every $\varepsilon > 0$ there is a $\delta > 0$ such that something happens" is "there is an $\varepsilon > 0$ such that for all $\delta > 0$ the thing does not happen." Using quantifier notation, for a proposition P the negation of

$$\forall \varepsilon > 0 \, \exists \, \delta > 0 \text{ s.t. } P$$

is

$$\exists \, \varepsilon > 0 \text{ s.t. } \forall \delta > 0 \sim P,$$

where $\sim P$ is the negation of P. Negating turns $\forall$ into $\exists$ and *vice versa*.

For us, P is "all x and x' in X satisfy $d_X(x, x') < \delta \Longrightarrow d_Y(f(x), f(x')) < \varepsilon$." Its negation is "some x and x' in X satisfy $d_X(x, x') < \delta$ and $d_Y(f(x), f(x')) \geq \varepsilon$". (Observe "$< \varepsilon$" turned into "$\geq \varepsilon$" at the end, as a negation.) Putting everything together, the negation of f being uniformly continuous says:

There is $\varepsilon > 0$ s.t. for all $\delta > 0$, an x and x' in X satisfy $d_X(x, x') < \delta$ & $d_Y(f(x), f(x')) \geq \varepsilon$.

The numbers x and x' here may depend on δ (each δ needs an example of x and x').

Now that we can express what it means not to be uniformly continuous, assume f is not uniformly continuous. For the special ε that occurs, use $\delta = 1/n$ with $n = 1, 2, 3, \ldots$: for each n there are x_n and x_n' in X such that $d_X(x_n, x_n') < 1/n$ and $d_Y(f(x_n), f(x_n')) \geq \varepsilon$. We have built two sequences in X: $\{x_n\}$ and $\{x_n'\}$. By compactness of X, some subsequence $\{x_{n_i}\}$ of $\{x_n\}$ converges in X, say to x. Also $x_{n_i}' \to x$ by Theorem 3.8. By continuity of f, $f(x_{n_i}) \to f(x)$ and $f(x_{n_i}') \to f(x)$, so $d_Y(f(x_{n_i}), f(x_{n_i}')) \to 0$. This contradicts the inequality $d_Y(f(x_n), f(x_n')) \geq \varepsilon$ for all n, so we're done.

Second proof (using open coverings): This proof will not be by contradiction. Pick $\varepsilon > 0$. For each $x \in X$, continuity of f at x implies there is some $\delta_x > 0$ such that

$$d_X(x, x') < \delta_x \Longrightarrow d_Y(f(x), f(x')) < \frac{\varepsilon}{2}. \tag{8.3}$$

The balls $B(x, \delta_x)$ are an open covering of X, so by compactness there is a finite subcovering, say by $B(x_1, \delta_{x_1}), \ldots, B(x_n, \delta_{x_n})$. Set $\delta = \min(\delta_{x_1}/2, \ldots, \delta_{x_n}/2) > 0$.

Pick x and x' in X with $d_X(x, x') < \delta$. We have $x \in B(x_i, \delta_{x_i}/2)$ for some $i = 1, \ldots, n$. Then $d_X(x, x_i) < \delta_{x_i}/2 < \delta_{x_i}$, so $d_Y(f(x), f(x_i)) < \varepsilon/2$ by (8.3). Also

$$d_X(x', x_i) \leq d_X(x', x) + d_X(x, x_i) < \delta + \frac{\delta_{x_i}}{2} \leq \frac{\delta_{x_i}}{2} + \frac{\delta_{x_i}}{2} = \delta_{x_i},$$

so $d_Y(f(x'), f(x_i)) < \varepsilon/2$ by (8.3). Thus

$$d_Y(f(x), f(x')) \leq d_Y(f(x), f(x_i)) + d_Y(f(x_i), f(x')) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

□

To illustrate the better control that is built into uniform continuity compared to continuity, we will prove uniformly continuous functions can be extended from dense subsets to the whole space. This uses the following lemma about a property of uniformly continuous functions that is not generally true for continuous functions.

Lemma 8.24. *If $f\colon X \to Y$ is uniformly continuous and $\{x_n\}$ is a Cauchy sequence in X then $\{f(x_n)\}$ is a Cauchy sequence in Y.*

Proof. Pick $\varepsilon > 0$. By uniform continuity there is a $\delta > 0$ such that for all x and x' in X satisfying $d_X(x, x') < \delta$ we have $d_Y(f(x), f(x')) < \varepsilon$. Since $\{x_n\}$ is Cauchy in X, using the value of δ there is an $N \geq 1$ such that $m, n \geq N \implies d_X(x_m, x_n) < \delta$. Then uniform continuity implies $d_Y(f(x_m), f(x_n)) < \varepsilon$ for $m, n \geq N$, so $\{f(x_n)\}$ is Cauchy in Y. □

In contrast to this, a continuous function *might not* send Cauchy sequences to Cauchy sequences: $1/x$ maps $(0, \infty) \to \mathbf{R}$ and $\{1/n\}$ is Cauchy in $(0, \infty)$ but $\{1/(1/n)\} = \{n\}$ is not Cauchy in $\mathbf{R}$. There is nothing strange about this: continuous functions, by definition, send convergent sequences to convergent sequences, not Cauchy sequences to Cauchy sequences. In an incomplete metric space, such as $(0, \infty)$, Cauchy sequences need not be convergent.

Theorem 8.25. *Let X and Y be metric spaces, with Y complete. For a dense subset D of X, every uniformly continuous function $f\colon D \to Y$ uniquely extends to a continuous function $\widetilde{f}\colon X \to Y$, defined by $\widetilde{f}(x) = \lim_{n\to\infty} f(x_n)$, where $\{x_n\}$ is an arbitrary sequence in D that converges to x. Moreover, $\widetilde{f}$ is uniformly continuous.*

Proof. If f has some extension to a continuous function $\widetilde{f}$ on X then for each $x \in X$, $\widetilde{f}(x)$ has to equal $\lim_{n\to\infty} f(x_n)$ where $\{x_n\}$ is a sequence in D that converges to x (such a sequence in D exists for each x since D is dense in X). Therefore $\widetilde{f}$ on X is determined by f on D, so the uniqueness of a (uniformly) continuous extension of f from D to X is immediate. What we have to address is the existence of such an extension:

(1) For $x \in X$, if $x_n \to x$ where all x_n are in D, then why does the sequence $\{f(x_n)\}$ converge in Y?
(2) For $x \in X$, if $x_n \to x$ and $x'_n \to x$ where all x_n and x'_n are in D, then why does $\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} f(x'_n)$ in Y?
(3) Why does the function $\widetilde{f}\colon X \to Y$ defined by $\widetilde{f}(x) = \lim_{n\to\infty} f(x_n)$, if $x_n \to x$ in D, equal f on D and why is $\widetilde{f}$ uniformly continuous?

Proof of (1). For $x \in X$, let $\{x_n\}$ be a sequence in D such that $x_n \to x$. Then $\{x_n\}$ is a Cauchy sequence in D. By Lemma 8.24 for the uniformly continuous function $f\colon D \to Y$, $\{f(x_n)\}$ is a Cauchy sequence in Y, so this sequence has a limit since Y is complete.

Proof of (2). If $\{x_n\}$ and $\{x'_n\}$ are two sequences in D that both tend to x, why do the convergent sequences $\{f(x_n)\}$ and $\{f(x'_n)\}$ in Y have the same limit? It suffices to show $d_Y(f(x_n), f(x'_n)) \to 0$ (Theorem 3.8). That is, for every $\varepsilon > 0$ we want an N such that $n \geq N \implies d_Y(f(x_n), f(x'_n)) < \varepsilon$.

From uniform continuity of f on D, for each $\varepsilon > 0$ there is a $\delta > 0$ such that for all a and b in D, $d_X(a, b) < \delta \implies d_Y(f(a), f(b)) < \varepsilon$. Since $x_n \to x$ and $x'_n \to x$ we have $d_X(x_n, x'_n) \to 0$, so there's some N such that $n \geq N \implies d_X(x_n, x'_n) < \delta$. Therefore $n \geq N \implies d_Y(f(x_n), f(x'_n)) < \varepsilon$, which is what we wanted.

Proof of (3). By (1) and (2), the definition

$$\widetilde{f}(x) := \lim_{n\to\infty} f(x_n) \in Y$$

for $x \in X$ and a sequence $\{x_n\}$ in D with $x_n \to x$ is independent of the choice of $\{x_n\}$, so $\widetilde{f}$ is a genuine function from X to Y. In particular, if $x \in D$ then using $x_n = x$ for all n shows $\widetilde{f}(x) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} f(x) = f(x)$, so $\widetilde{f}$ restricts on D to f.

It remains to show $\widetilde{f}$ is uniformly continuous. This will be an $\varepsilon/3$ argument.

By uniform continuity of f on D, for every $\varepsilon > 0$ there is $\delta > 0$ such that

$$d_X(a,b) < \delta \text{ in } D \Longrightarrow d_Y(f(a), f(b)) < \frac{\varepsilon}{3}. \tag{8.4}$$

Suppose x and x' in X satisfy $d_X(x,x') < \delta$. There are sequences $\{x_n\}$ and $\{x'_n\}$ in D such that $x_n \to x$ and $x'_n \to x'$. Then

$$\begin{aligned} d_Y(\widetilde{f}(x), \widetilde{f}(x')) &\leq d_Y(\widetilde{f}(x), \widetilde{f}(x_n)) + d_Y(\widetilde{f}(x_n), \widetilde{f}(x'_n)) + d_Y(\widetilde{f}(x'_n), \widetilde{f}(x')) \\ &= d_Y(\widetilde{f}(x), f(x_n)) + d_Y(f(x_n), f(x'_n)) + d_Y(f(x'_n), \widetilde{f}(x')). \end{aligned} \tag{8.5}$$

Since, by definition, $\widetilde{f}(x)$ is the limit of $\{f(x_n)\}$ and $\widetilde{f}(x')$ is the limit of $\{f(x'_n)\}$, for large enough n both $d_Y(\widetilde{f}(x), f(x_n))$ and $d_Y(f(x'_n), \widetilde{f}(x'))$ are less than $\varepsilon/3$. Also for large enough n, $d_X(x_n, x'_n) < \delta$ since x_n and x'_n both tend to x, so $d_Y(f(x_n), f(x'_n)) < \varepsilon/3$ for large enough n by (8.4). Plugging this all into the inequality (8.5), for large enough n

$$d_Y(\widetilde{f}(x), \widetilde{f}(x')) \leq d_Y(\widetilde{f}(x), f(x_n)) + d_Y(f(x_n), f(x'_n)) + d_Y(f(x'_n), \widetilde{f}(x')) < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

□

Example 8.26. While $1/x$ is uniformly continuous as a function $[1,\infty) \to \mathbf{R}$ (Example 8.21 with $a = 1$), it is not uniformly continuous as a function $(0,1] \to \mathbf{R}$: if it were uniformly continuous on $(0,1]$, then it would extend to a continuous function on $[0,1]$, which of course it does not.

The reader should show $1/x$ is not uniformly continuous on $(0,1]$ directly from the definition of uniform continuity, without needing a result like Theorem 8.25. Use the negation of uniform continuity from the proof of Theorem 8.23.

Example 8.27. The function $f(x) = 1/(x^2 - 2)$ on $\mathbf{Q} \cap [1,2]$ is continuous, but it is not uniformly continuous: if it were uniformly continuous then the function would extend to a continuous function on $[1,2]$, which is clearly false due to a problem at $x = \sqrt{2}$.

Example 8.28. As a positive application of Theorem 8.23, for each $a > 1$ we use the theorem to construct[13] a^x for irrational $x > 0$ from the case of rational $x \geq 0$ and show a^x is continuous in x. To use Theorem 8.23 we restrict attention to bounded intervals of x.

We will take for granted as background the existence of positive rational powers. For each $r \in \mathbf{Q}$ with $r > 0$, written as the ratio of positive integers m/n, the number $a^r = a^{m/n}$ is the unique solution of $x^n = a^m$ in $(1,\infty)$. Its existence and uniqueness follow from the Intermediate Value Theorem since x^n is continuous, increasing, and unbounded as $x \to \infty$ with value 1 at $x = 1$.

Claim 1: For positive integers N, $1 < a^{1/N} < 1 + (a-1)/N$.

Proof: For $h > 0$, $(1+h)^N > 1 + Nh$ by the Binomial Theorem and $1 + Nh \geq a$ if $h \geq (a-1)/N$. Thus $h \geq (a-1)/N \Longrightarrow (1+h)^N > a$, so if $a^{1/N} = 1 + h$ then we see $a^{1/N} - 1 = h < (a-1)/N$, so $a^{1/N} < 1 + (a-1)/N$.

From Claim 1, $a^{1/N} \to 1$ as $N \to \infty$ in $\mathbf{Z}$.

[13]All we do could be adapted to $0 < a < 1$, but we stick to $a > 1$ to keep the notation simple.

Take for granted that $a^r a^s = a^{r+s}$ for rational $r, s \geq 0$. From this it follows that $0 \leq r < s \Longrightarrow a^r < a^s$: $s - r > 0 \Longrightarrow a^{s-r} > 1 \Longrightarrow a^s = a^r a^{s-r} > a^r$. Then $a^r \to 1$ as $r \to 0^+$ through rational numbers because it is true when $r = 1/N$ and a^r is increasing.

Claim 2: On each *bounded* interval $\mathbf{Q} \cap [0, M]$ for integers $M \geq 1$, the function $f(r) = a^r$ is uniformly continuous.

Proof: Pick a positive integer N and let $r, s \in \mathbf{Q} \cap [0, M]$ satisfy $|r - s| < 1/N$. Suppose $r \geq s$. Then

$$f(r) - f(s) = a^r - a^s = a^s(a^{r-s} - 1) \Longrightarrow |f(r) - f(s)| \leq a^M(a^{r-s} - 1).$$

From $r - s < 1/N$ we get $1 \leq a^{r-s} < a^{1/N} < 1 + (a-1)/N$ by Claim 1, so $0 \leq a^{r-s} - 1 < (a-1)/N$. Thus $|f(r) - f(s)| \leq \frac{a^M(a-1)}{N}$. If instead $r < s$ then we get the same bound by switching the roles of r and s. Thus

$$r, s \in \mathbf{Q} \cap [0, M] \text{ and } |r - s| < \frac{1}{N} \Longrightarrow |f(r) - f(s)| \leq \frac{a^M(a-1)}{N}.$$

This implies uniform continuity on $\mathbf{Q} \cap [0, M]$: for $\varepsilon > 0$ let $\delta = 1/N$ where N is chosen large enough that $a^M(a-1)/N < \varepsilon$.

By Claim 2 and Theorem 8.23, for each integer $M \geq 1$ the function $f(x) = a^x$ on $\mathbf{Q} \cap [0, M]$ uniquely extends to a uniformly continuous function on $[0, M]$, which we will also denoted by a^x. For example, $a^{\sqrt{2}}$ is the limit of a^r for rationals r tending to $\sqrt{2}$. The function a^x on $[0, M]$ for larger M has to agree with a^x on $[0, M]$ for smaller M since there's at most one continuous extension on each $[0, M]$. Therefore letting $M \to \infty$ we get a single continuous function a^x for all $x \geq 0$ that agrees with the originally defined a^x when the exponent x is rational.

EXERCISES.

(1) On $\mathbf{R}$, prove the function $d(x, y) = \sqrt{|x - y|}$ is a metric but the function $d(x, y) = (x - y)^2$ is not.
(2) Check the functions d_∞ and d_1 on $C[0, 1]$ in Example 2.3 are metrics.
(3) Check Example 2.4: if (X, d) is a metric space and Y is a nonemtpy subset of X then Y together with the restriction of d to Y (strictly speaking, to $Y \times Y$) is a metric space.
(4) Verify on each non-empty set that the discrete metric (Example 2.5) is a metric.
(5) Verify for a metric space (X, d) that the functions $d'(x, y) = \min(d(x, y), 1)$ and $d''(x, y) = d(x, y)/(1 + d(x, y))$ from Example 2.6 are both metrics.
(6) In the definition of a limit of a sequence, prove the limit is unique: if $x_n \to x$ and $x_n \to y$ then $x = y$. (This was implicitly assumed when we used limits, but it should be proved.)
(7) Prove Theorem 4.8.
(8) Let X be a metric space and Y be a subset, treated as a metric space using the metric induced from X. Show every open set in the metric space Y has the form $U \cap Y$ where U is an open set in X. (Hint: First treat the case of open balls in Y, and it may help to distinguish open balls in X or Y with the same center and radius, using notation like $B_X(y, r)$ and $B_Y(y, r)$.)
(9) Let (X, d) be a metric space. For $a \in X$ and $0 < r_1 < r_2$, show the annulus $\{x \in X : r_1 < d(a, x) < r_2\}$ is open and $\{x \in X : r_1 \leq d(a, x) \leq r_2\}$ is closed.

(10) Show every closed subset of a compact metric space is a compact subset by two methods: using the convergent subsequence description of compactness and using the open covering description of compactness.
(11) Show every compact subset of a metric space is complete using the induced metric.
(12) Fill in the details from Remark 8.10: prove a function $f\colon X \to Y$ between two metric spaces is continuous at a point $a \in X$ (based on the (ε, δ)-definition) if and only if for every open $U \subset Y$ containing $f(a)$ there is an open $V \subset X$ containing a such that $f(V) \subset U$.
(13) Show each line in $\mathbf{R}^2$ is a closed subset in two ways: by the definition of a closed subset and by Corollary 8.11. (Hint for the second way: show each line has the form $f^{-1}(c)$ for some continuous function $f\colon \mathbf{R}^2 \to \mathbf{R}$ and $c \in \mathbf{R}$.)
(14) For every interval I in $\mathbf{R}$, construct a continuous function $f\colon (0,1) \to \mathbf{R}$ with image I. Therefore the connectedness of $(0,1)$ implies the connectedness of all other intervals by Theorem 8.16. (Hint: Start out with I being an interval (open, closed, half-open) having endpoints 0 and 1.)
(15) Prove the inequality in Remark 8.5.
(16) Let $f\colon X \to Y$ and $g\colon X \to Y$ be two continuous functions between the same metric spaces. Show $\{x \in X : f(x) = g(x)\}$ is a closed subset of X.
(17) Let K_1 and K_2 be disjoint compact subsets of a metric space. Prove there is $c > 0$ such that $d(x_1, x_2) \geq c$ for all $x_1 \in K_1$ and $x_2 \in K_2$. Give an example of two disjoint non-compact subsets of $\mathbf{R}^2$ for which the conclusion is false for those subsets (Hint: horizontal asymptote).
(18) Viewing $\mathbf{R}^\times$ as a metric space using the absolute value metric from $\mathbf{R}$, show inversion $\mathbf{R}^\times \to \mathbf{R}$ given by $f(x) = 1/x$ is a continuous function but is not uniformly continuous. Then show all rational functions with real coefficients are continuous on $\mathbf{R}$ away from the numbers where the denominator is 0.
(19) For functions $f\colon \mathbf{R} \to \mathbf{R}$ and $g\colon \mathbf{R} \to \mathbf{R}$, let $F\colon \mathbf{R} \to \mathbf{R}^2$ by $F(x) = (f(x), g(x))$. Lemma 8.13 says that if f and g are continuous then F is continuous. Prove the converse: if F is continuous then f and g are continuous.
(20) Show the function $1/x$ on $(0,1]$ is not uniformly continuous directly from the definition of uniform continuity.
(21) Is the function $f(x) = \sqrt{x}$ on $[0,\infty)$ uniformly continuous?
(22) Convergence of a function $f\colon [a,\infty) \to \mathbf{R}$ at ∞ is defined analogously to the case of sequences: say $f(x)$ has limit L as $x \to \infty$ if for every $\varepsilon > 0$ there is an N such that $x \geq N \Rightarrow |f(x) - L| < \varepsilon$.

For a continuous function $f\colon [a,\infty) \to \mathbf{R}$, prove that if $f(x)$ has a limit as $x \to \infty$ then f is uniformly continuous on $[a,\infty)$. (Hint: on each interval $[a,b]$ the function is uniformly continuous, so focus on large x.) Is the converse true?

Appendix A. Proofs about completeness

In Example 4.16 we stated the following theorem, which we will now prove.

Theorem A.1. *The metric space $(C[0,1], d_\infty)$ is complete.*

Proof. To show every Cauchy sequence $\{f_n\}$ in this space has a limit in this space falls into three steps:

Step 1: Create a candidate limit function f.

For each $a \in [0,1]$ the sequence of numbers $\{f_n(a)\}$ is Cauchy since

$$|f_m(a) - f_n(a)| \leq \max_{0 \leq x \leq 1} |f_m(x) - f_n(x)| = d_\infty(f_m, f_n)$$

and the value on the right is arbitrarily small for all large enough m and n. Therefore $\lim_{n\to\infty} f_n(a)$ exists by completeness of $\mathbf{R}$. Call the limit $f(a)$. We have defined a function $f\colon [0,1] \to \mathbf{R}$ using pointwise considerations.

Step 2: Show f is continuous.

This will be proved with an $\varepsilon/3$ argument.

Pick $a \in [0,1]$ and $\varepsilon > 0$. We need to find $\delta > 0$ such that $|x - a| < \delta \implies |f(x) - f(a)| < \varepsilon$.

Since $\{f_n\}$ is Cauchy, there is N such that $m, n \geq N \implies d_\infty(f_m, f_n) < \varepsilon/3$, so for all $x \in [0,1]$ and $m \geq N$ we have $|f_m(x) - f_N(x)| < \varepsilon/3$. Letting $m \to \infty$ in this inequality, $|f(x) - f_N(x)| \leq \varepsilon/3$ for all x in $[0,1]$. Thus for each x in $[0,1]$,

$$\begin{aligned} |f(x) - f(a)| &\leq |f(x) - f_n(x)| + |f_N(x) - f_N(a)| + |f_N(a) - f(a)| \\ &\leq \frac{\varepsilon}{3} + |f_N(x) - f_N(a)| + \frac{\varepsilon}{3} \\ &= \frac{2}{3}\varepsilon + |f_N(x) - f_N(a)|. \end{aligned}$$

Since f_N is continuous at a, there is $\delta > 0$ such that $|x-a| < \delta \implies |f_N(x)-f_N(a)| < \varepsilon/3$. Therefore

$$|x - a| < \delta \implies |f(x) - f(a)| \leq \frac{2}{3}\varepsilon + |f_N(x) - f_N(a)| < \varepsilon,$$

which proves f is continuous at each $a \in [0,1]$.

Step 3: Show $d_\infty(f_n, f) \to 0$.

We essentially repeat the beginning of Step 2. From $\{f_n\}$ being Cauchy there is an N such that $m, n \geq N \implies d_\infty(f_m, f_n) < \varepsilon/3$, so for all $x \in [0,1]$ and $m, n \geq N$ we have $|f_m(x) - f_n(x)| < \varepsilon/3$. Letting $m \to \infty$ here, we get $|f(x) - f_n(x)| \leq \varepsilon/3$ for all $x \in [0,1]$. Therefore $d_\infty(f, f_n) \leq \varepsilon/3 < \varepsilon$.

□

Next we will prove Theorem 4.22, which we restate.

Theorem A.2. *Every metric space has a completion.*

Proof. Let the metric space be (X, d). We seek a complete metric space $(\widehat{X}, \widehat{d})$ containing (X, d) as a dense subset. The purpose of the proof is to show *some* construction is possible, but in practice nobody thinks about a completion by the way it will be constructed here. In particular, analysts want to work with concrete complete metric spaces, not abstract completions. More comments about creating completions will be given after the proof.

Before building $\widehat{X}$, here are properties it must have if it exists.

(1) Each $\widehat{x} \in \widehat{X}$ is a limit of a sequence from X since X is dense in $\widehat{X}$: $\widehat{x} = \lim_{n\to\infty} x_n$ with $x_n \in X$, so $\{x_n\}$ is a Cauchy sequence in $\widehat{X}$ (Theorem 4.3) and thus is also a Cauchy sequence in X since $\widehat{d} = d$ on X.

(2) A second Cauchy sequence $\{x'_n\}$ in X converges to the same limit $\widehat{x}$ if and only if $\widehat{d}(x_n, x'_n) \to 0$ (Theorems 3.7 and 3.8), which is the same as $d(x_n, x'_n) \to 0$ since $\widehat{d} = d$ on X.

(3) The metric on $\widehat{X}$ can be expressed as a limit of values of the metric on X: if $\widehat{x}$ and $\widehat{y}$ are in $\widehat{X}$ and $x_n \to \widehat{x}$ and $y_n \to \widehat{y}$ with $x_n, y_n \in X$, then $d(x_n, y_n) \to \widehat{d}(\widehat{x}, \widehat{y})$ by Remark 8.5 (because $d(x_n, y_n) = \widehat{d}(x_n, y_n)$).

Since each element of $\widehat{X}$ is a limit of a Cauchy sequence in X, and two Cauchy sequences in X tend to the same value in $\widehat{X}$ exactly when their termwise distance tends to 0, this suggests an idea for how to define $\widehat{X}$: let it be the set of all Cauchy sequences in X, with sequences identified when the termwise distance goes to 0. This will turn out to work, and is a generalization of one of the constructions of the real numbers out of the rational numbers: a real number is an equivalence class of Cauchy sequences of rational numbers relative to the absolute value on $\mathbf{Q}$.

Let C be the set of all Cauchy sequences in X. Introduce a relation $\sim$ on C by

$$\{x_n\} \sim \{y_n\} \text{ if } d(x_n, y_n) \to 0.$$

It is left to the reader to check $\sim$ is an equivalence relation on C. Denote the equivalence class of $\{x_n\}$ in C as $[x_n]$, so

$$[x_n] = \{\{y_n\} \in C : d(x_n, y_n) \to 0\}.$$

We embed X into C by viewing each $x \in X$ as the constant Cauchy sequence $\{x, x, x, \ldots\}$. For different x and y in X, the constant sequences $(x, x, x, \ldots)$ and $(y, y, y, \ldots)$ are not equivalent since $d(x, y) > 0$.

Define $\widehat{X}$ to be the set of equivalence classes in C for the relation $\sim$. The motivation for this definition of $\widehat{X}$ is to have a set that has the first and second properties that we already observed a completion of (X, d) should have. To put a metric on $\widehat{X}$ we will get motivation from the third property listed above. It suggests the definition

$$\widehat{d}([x_n], [y_n]) = \lim_{n\to\infty} d(x_n, y_n).$$

There are now three things we need to verify:

(1) The function $\widehat{d}$ makes sense and is a metric on $\widehat{X}$.
(2) We can view X as a dense subset of $\widehat{X}$ and $\widehat{d}$ restricts on X to be d.
(3) The metric space $(\widehat{X}, \widehat{d})$ is complete.

(1) $\widehat{d}$ makes sense and is a metric on $\widehat{X}$.

First we show for Cauchy sequences $\{x_n\}$ and $\{y_n\}$ in X that the sequence $\{d(x_n, y_n)\}$ in $\mathbf{R}$ is Cauchy, so it has a limit by completeness of $\mathbf{R}$. (In the motivation to define $\widehat{d}$ we discussed how a metric behaves on convergent sequences, but we are no longer dealing with convergent sequences, only Cauchy sequences, so knowing the numbers $d(x_n, y_n)$ have a limit really requires a proof.) By the inequality in Remark 8.5,

$$|d(x_m, y_m) - d(x_n, y_n)| \leq d(x_m, x_n) + d(y_m, y_n).$$

From the Cauchy property in X both $d(x_m, x_n)$ and $d(y_m, y_n)$ become arbitrarily small for sufficiently large m and n, so the above inequality tells us that the sequence $\{d(x_n, y_n)\}$ is Cauchy in $\mathbf{R}$. Thus $\lim_{n\to\infty} d(x_n, y_n)$ exists.

Since $\widehat{X}$ is not the Cauchy sequences in X, but equivalence classes of them, to show $\widehat{d}$ makes sense we have to check its formula doesn't change if we replace Cauchy sequences by equivalent Cauchy sequences: if $\{x_n\} \sim \{x'_n\}$ and $\{y_n\} \sim \{y'_n\}$ in C, we need to show

$\lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(x'_n, y'_n)$. Once again we appeal to the inequality in Remark 8.5, writing it now in the form

$$|d(x_n, y_n) - d(x'_n, y'_n)| \leq d(x_n, x'_n) + d(y_n, y'_n).$$

The terms on the right side both tend to 0 as $n \to \infty$ by the definition of the relation $\sim$, so $d(x_n, y_n) - d(x'_n, y'_n) \to 0$ in $\mathbf{R}$. Thus the convergent sequences $\{d(x_n, y_n)\}$ and $\{d(x'_n, y'_n)\}$ in $\mathbf{R}$ have the same limit, and this finishes the proof that $\widehat{d}$ makes sense on $\widehat{X}$.

It is left to the reader to show $\widehat{d}$ is a metric on $\widehat{X}$. We raise one issue: the fact that

$$\widehat{d}([x_n], [y_n]) = 0 \Longrightarrow [x_n] = [y_n]$$

relies on the equivalence relation used to define $\widehat{X}$. This is why it's important for $\widehat{X}$ to be *equivalence classes* of Cauchy sequences in X and not the Cauchy sequences in X. If we had tried to work with the Cauchy sequences as the primary object, then wanting a metric on them would force us to introduce the equivalence relation $\sim$ anyway.

(2) X can be viewed as a dense subset of $\widehat{X}$ and $\widehat{d} = d$ on X.

Associate to $x \in X$ the equivalence class of the constant sequence where each term is x:

$$\overline{x} := [(x, x, x, \ldots)].$$

If $x \neq y$ in X then $\widehat{d}(\overline{x}, \overline{y}) = \lim_{n\to\infty} d(x, y) = d(x, y) > 0$. so the function $X \to \widehat{X}$ where $x \mapsto \overline{x}$ is injective. This is how we view X as a subset of $\widehat{X}$, and the calculation we just made shows $\widehat{d} = d$ on the copy of X inside $\widehat{X}$.

Why is the copy of X in $\widehat{X}$ a dense subset? Pick an element $[x_n]$ of $\widehat{X}$ and an $\varepsilon > 0$. We want to find an $x \in X$ such that $\overline{x}$ is within ε of $[x_n]$. We'll show that each element deep enough into the Cauchy sequence $\{x_n\}$ can be used as x. Since $\{x_n\}$ is Cauchy, there's an N such that $m, n \geq N \Longrightarrow d(x_m, x_n) < \varepsilon$. Setting $x = x_N$,

$$\widehat{d}(\overline{x}, [(x_1, x_2, x_3, \ldots)]) = \lim_{n\to\infty} d(x_N, x_n) \leq \varepsilon. \tag{A.1}$$

Anyone who is pedantic can replace ε with $\varepsilon/2$ to make the limit strictly less than ε.

By the same reasoning, for all $m \geq N$ we have $\widehat{d}(\overline{x_m}, [x_n]) \leq \varepsilon$, so $[(x_1, x_2, x_3, \ldots)]$ is the limit of $\{\overline{x_m}\}$ as $m \to \infty$. This shows that, in a loose sense, each Cauchy sequence in X has been turned into the limit of its own terms (but we have to work with equivalence classes of Cauchy sequences to make everything proper).

(3) $(\widehat{X}, \widehat{d})$ is complete.

Let $\widehat{x}_1, \widehat{x}_2, \widehat{x}_3, \ldots$ be a Cauchy sequence of elements of $\widehat{X}$. (This is a "Cauchy sequence of equivalence classes of Cauchy sequences from X," but don't think too hard about it that way.) By (2), for each $n \geq 1$ we can pick $y_n \in X$ such that $\widehat{d}(\widehat{x}_n, \overline{y_n}) \to 0$. (For example, we can make the distance at most $1/n$.) Then $\overline{y_1}, \overline{y_2}, \overline{y_3}, \ldots$ is Cauchy in $\widehat{X}$ by Theorem 4.9, so $y_1, y_2, y_3, \ldots$ is Cauchy in X since $d(y_m, y_n) = \widehat{d}(\overline{y_m}, \overline{y_n})$. Define $\widehat{y} = [y_n]$ in $\widehat{X}$. We will show $\widehat{x}_n \to \widehat{y}$.

Pick $\varepsilon > 0$. For all m and n,

$$\widehat{d}(\widehat{x}_n, \overline{y_m}) \leq \widehat{d}(\widehat{x}_n, \overline{y_n}) + \widehat{d}(\overline{y_n}, \overline{y_m}) = \widehat{d}(\widehat{x}_n, \overline{y_n}) + d(y_n, y_m).$$

For all sufficiently large m and n (depending on ε) we can make $\widehat{d}(\widehat{x}_n, \overline{y_n}) < \varepsilon/2$ by the definition of y_n and $d(y_n, y_m) < \varepsilon/2$ by the Cauchy property. Therefore $\widehat{d}(\widehat{x}_n, \overline{y_m}) < \varepsilon$ for

all sufficiently large m and n. Since $\overline{y_m} \to \widehat{y}$ by (2), by continuity of metrics (Theorem 8.4)

$$\lim_{m\to\infty} \widehat{d}(\widehat{x}_n, \overline{y_m}) = \widehat{d}(\widehat{x}_n, \widehat{y}).$$

For sufficiently large n, this limit is at most ε. That proves $\widehat{x}_n \to \widehat{y}$ in $\widehat{X}$. □

The method we just gave for constructing a completion is not the only possible construction. Let's describe a second one. For every metric space X, define $C_b(X)$ to be the set of all continuous bounded functions $X \to \mathbf{R}$. This is a metric space where the distance between two functions f and g is defined as $\sup_{x\in X} |f(x) - g(x)|$. (We use a supremum here since a continuous bounded function might not have a maximum; X could be non-compact.)

The space $C_b(X)$ with this metric is complete. A proof can be made by adapting the proof of Theorem 4.16, which is the special case $X = [0, 1]$. We can embed X into $C_b(X)$ by fixing a choice of $a \in X$ and, in terms of this choice, associating to each $y \in X$ the function $f_y \colon X \to \mathbf{R}$ given by $f_y(x) = d(y, x) - d(a, x)$. The function f_y is continuous by Theorem 8.4 and it is bounded since $|f_y(x)| \leq d(a, y)$ by the proof of Theorem 8.4.[14] It turns out that $\sup_{x\in X} |f_y(x) - f_z(x)| = d(y, z)$, so $y \mapsto f_y$ is an embedding of X into $C_b(X)$ and the metric on $C_b(X)$ restricts to the metric d on X under this embedding. Closures of subsets are closed (Theorem 5.15) and closed subsets of complete spaces are complete (Theorem 5.11), so the closure of the embedded copy of X in $C_b(X)$ is a completion of X.

Our first construction of a completion, using equivalence classes of Cauchy sequences, is a much more robust method than the one using $C_b(X)$. The reason is that often when X has some additional structure that interacts well with the metric (*e.g.*, being a normed vector space or a field), the method of completing using equivalence classes of Cauchy sequences shows the completion has the same additional structure automatically, whereas this is not clear by embedding of X into $C_b(X)$.

Sometimes no abstract construction of a completion shows a particular feature of X persists in its completion, so a more specific technique is needed to reveal that feature in the completion. For example, the metric space of functions $(C[0, 1], d_1)$ is incomplete (Example 4.17) and analysts want to think of elements in the completion as functions on $[0, 1]$. Part of a course on measure theory is devoted to explaining why the completion consists of certain real-valued functions on $[0, 1]$.

References

[1] K. Conrad, Spaces that are Connected but not Path-connected, https://kconrad.math.uconn.edu/blurbs/topology/connnotpathconn.pdf.

[2] M. Fréchet, "Sur quelques points du calcul fonctionnel," Rendiconti del Circolo Matematico di Palermo **22** (1906), pp. 1–72. Online at https://babel.hathitrust.org/cgi/pt?id=mdp.39015057352307;view=1up;seq=9.

[3] K. A. Ross, "Elementary Analysis: The Theory of Calculus," Springer-Verlag, New York, 1980.

[14] It is more natural to convert points of X into continuous functions on X by associating to each $y \in X$ the function $x \mapsto d(y, x)$, but this function is not bounded if the metric on X takes arbitrarily large values. Subtracting $d(a, x)$ creates a bounded function.