distance from x to A by:

$$d(x, A) = \inf \{ d(x,a) : a \in A \}.$$

Clearly, $0 \leq d(x, A) < \infty$ for any x and any A, but it is not necessarily true that $d(x, A) > 0$ when $x \notin A$. For example, $d(x, \mathbb{Q}) = 0$ for any $x \in \mathbb{R}$.

Proposition: $d(x, A) = 0$ if and only if $x \in \overline{A}$.

Proof: $d(x, A) = 0$ if and only if there is a sequence of points $\{a_n\}_{n=1}^{\infty}$ in A such that $d(x, a_n) \to 0$. But this means that $a_n \to x$ and, hence, $x \in \overline{A}$.

Note that this proposition has given us another connection between limits in M and limits in $\mathbb{R}$. Loosely speaking, this proposition shows that 0 is a limit point of $\{d(x,a) : a \in A\}$ if and only if x is a limit point of A. We can get even more mileage out of this observation by checking that the map $x \to d(x, A)$ is actually continuous. For this it suffices to establish the following inequality:

Proposition: $|d(x, A) - d(y, A)| \leq d(x, y)$

Proof: $d(x,a) \leq d(x,y) + d(y, a)$ for any $a \in A$. But $d(x, A)$ is a lower bound for $d(x,a)$; hence $d(x, A) \leq d(x,y) + d(y,a)$. Now, by taking the infimum over $a \in A$, we get $d(x, A) \leq d(x,y) + d(y, A)$. Since the roles of x and y are interchangeable, we're done.

To appreciate what this has done for us, let's make two simple observations. First, if $f : M \to \mathbb{R}$ is a continuous function, then

the set $E = \{x \in M : f(x) = 0\}$ is closed. (Why?) Conversely, if E is a closed set in M, then E is the "zero set" of some continuous real-valued function on M; in particular, $E = \{x \in M : d(x, E) = 0\}$. Thus a set E is closed if and only if $E = f^{-1}(\{0\})$ for some continuous function $f : M \to \mathbb{R}$. Conclusion: If you know all of the closed (or open) sets in a metric space M, then you know all of the continuous real-valued functions on M. Conversely, if you know all of the continuous real-valued functions on M, then you know all of the closed (or open) sets in M.

Connected Sets

Our purpose in this section would be to understand the special nature of intervals in $\mathbb{R}$. This will allow us to prove and subsequently generalize the intermediate value theorem of calculus. The intermediate value theorem is the formal statement of the informal notion that the graph of a continuous function is "unbroken". The historical basis of the theorem is the concept of a function as measuring, over time, the relative position of an object moving along a straight line. Thus, if we track the position $y = f(x)$ of a moving object between time $x = a$ and some subsequent time $x = b$, we would expect the object to "visit" all of the positions y that are intermediate to $f(a)$ and $f(b)$. In short, the continuous image of the time interval $[a, b]$ should contain (at least)

the full interval of positions between $f(a)$ and $f(b)$.

The secret here is the intuitively obvious fact that no interval in $\mathbb{R}$ can be split into two relatively open parts. Let's prove this by "brute force" for the interval $[a,b]$ (we'll do the other cases shortly).

Suppose to the contrary that $[a,b] = A \cup B$, where A and B are nonempty, disjoint, relatively open sets in $[a,b]$. We are going to find a contradiction by examining the "border" between A and B. The trouble comes from the fact that A and B are necessarily also closed in $[a,b]$, since each is the complement of an open set: $A = [a,b] \setminus B$ and $B = [a,b] \setminus A$, and so each of A and B lays claim to the "border".

To get started, we might as well assume that $b \in B$, and so $(b-\epsilon, b] \subset B$, for some $\epsilon > 0$, since B is open. Now let $c = \sup A$. Clearly, $a \leq c \leq b$, but note that, since A and B are open in $[a,b]$, we actually have $a < c < b$. (Why?) Next, it follows from the definition of c that $(c-\epsilon, c) \cap A \neq \emptyset$ and $(c, c+\epsilon) \cap B \neq \emptyset$ for any $\epsilon > 0$; in fact $(c,b] \subset B$. That is, $c \in \overline{A}$ and $c \in \overline{B}$. But then, $c \in \overline{A} \cap \overline{B} = A \cap B = \emptyset$. This contradiction shows that no such splitting of $[a,b]$ into nonempty, disjoint, open sets is available.

Based on this observation, we say that a metric space M is disconnected (or not connected) if M can be split into the union of two nontrivial open sets, that is, if there are nonempty open sets A and B in M with $A \cap B = \emptyset$ and $A \cup B = M$.

The pair of open sets A and B is called a disconnection of M. We say that M is connected if no such disconnection can be found. Thus, for example, $[a,b]$ is connected.

Notice that we could just as well have used closed sets in our definition. If a disconnection $A-B$ exists, then the disconnecting sets are also closed: $A = B^c$ and $B = A^c$. That is, A and B are clopen (simultaneously open and closed) sets. Conversely, if M contains a nontrivial clopen subset A (other than $\emptyset$ or M), then A and A^c are a disconnection for M. This gives us our first theorem:

Thm: M is connected if and only if M contains no nontrivial clopen sets.

Ex. (a) $\mathbb{R}$ is connected (why?)

(b) A discrete space containing two or more points is disconnected.

(c) The empty set $\emptyset$ and any one-point space are connected (by default).

(d) The Cantor set Δ is (very!) disconnected. Indeed, it follows from work done at home that for any $x, y \in \Delta$ with $x < y$ there is a $z \notin \Delta$ such that $x < z < y$. Thus, Δ is disconnected by the (relatively) open sets $A = [0, z) \cap \Delta$ and $B = (z, 1] \cap \Delta$.

Our terminology for connectedness is unavoidably fussy. After all, we have defined connectedness in terms of what it is not, namely, disconnected. To make matters worse, at least on the surface, part (d) of the example above and our proof that $[a,b]$ is

connected both suggest the frightening prospect of "relatively connected" as an altogether separate notion. Well, fear not! Connectedness is not a relative property for metric spaces. To see why, we will need to face the relative definition head-on.

A subset E of a metric space M is disconnected in E if there exists disjoint, nonempty, open (in E) sets U and V such that $E = U \cup V$. Now, it is immediate that this gives us a pair of open sets A and B in M such that $U = A \cap E$ and $V = B \cap E$. And so "unrelating" the relative definition, by writing it in terms of A and B, yields: $A \cap E \neq \emptyset$, $B \cap E \neq \emptyset$, $(A \cap E) \cap (B \cap E) = \emptyset$, and $E = (A \cap E) \cup (B \cap E)$, or $E \subset A \cup B$. This mess would be greatly simplified if we could take A and B to be disjoint in M. While this need not hold true in more general settings, luck is with us in a metric space.

Lemma: Let E be a subset of a metric space M. If U and V are disjoint open sets in E, then there are disjoint open sets A and B in M such that $U = A \cap E$ and $V = B \cap E$.

Proof: For each $x \in U$ there is an $\epsilon_x > 0$ such that $B_{\epsilon_x}^E(x) = E \cap B_{\epsilon_x}^M(x) \subset U$, because U is open in E. Likewise, for each $y \in V$ there is a $\delta_y > 0$ such that $B_{\delta_y}^E(y) = E \cap B_{\delta_y}^M(y) \subset V$. Since $U \cap V = \emptyset$, we also get $E \cap B_{\epsilon_x}^M(x) \cap B_{\delta_y}^M(y) = \emptyset$. We would like to get rid of the set E in this conclusion, and we can do so at a small price:

Claim. $B_{\epsilon_x/2}(x) \cap B_{\delta_y/2}(y) = \emptyset$ for every $x \in U$ and $y \in V$. (Just check)

Thus $A = \bigcup \{ B_{\epsilon_x/2}(x) : x \in U \}$ and $B = \bigcup \{ B_{\delta_y/2}(y) : y \in V \}$ work.

The conclusion to be drawn from this Lemma is that E is disconnected (in E) if and only if there exist disjoint, nonempty, open sets A and B in M such that $A \cap E \neq \emptyset$, $B \cap E \neq \emptyset$, and $E \subset A \cup B$. And it does not matter whether we take "open" to mean "open in E" or "open in M". That is, this statement reduces to the original definition in case $E = M$, and it gives the correct "relative" definition in any case (by taking $U = A \cap E$ and $V = B \cap E$). Thus, there is no harm in simply taking it as our new definition of a disconnected set, as opposed to a disconnected space. In other words, we have dodged a bullet! By adopting this harmless rewording of the definition of disconnected, and hence also rewording of the definition of connected, we have freed the concept from any apparent dependence on the relative metric. We would be foolish to do otherwise.

Henceforth, when considering a subset E of a given metric space M, we will call a pair of disjoint open sets A and B a disconnection of E if $A \cap E \neq \emptyset$, $B \cap E \neq \emptyset$, and $E \subset A \cup B$. And, of course, we will say that E is a connected set if no such disconnection of E can be found.

Let's put this new definition to use by giving another characterization of the intervals in $\mathbb{R}$.

Thm: A subset E of $\mathbb{R}$, containing more than one point, is connected if and only if, whenever $x, y \in E$ with $x < y$, we also have $[x, y] \subset E$. That is, the connected subsets of $\mathbb{R}$ (containing more than one point) are precisely the intervals.

Proof: One direction is easy: If there exist points $x < z < y$ such that

$x, y \in E$ but $z \notin E$, then $E \subset (-\infty, z) \cup (z, +\infty)$; that is, $A = (-\infty, z)$ and $B = (z, +\infty)$ is a disconnection of E.

For the other direction, suppose that E satisfies the condition that $[x,y] \subset E$ whenever $x, y \in E$ with $x < y$, but that E is disconnected. Then there are disjoint open sets A and B in $\mathbb{R}$ such that $A \cap E \neq \emptyset$, $B \cap E \neq \emptyset$, and $E \subset A \cup B$. Given points $a \in A \cap E$ and $b \in B \cap E$, we might as well assume that $a < b$ and hence that $[a,b] \subset E$. But now $[a,b] \subset E \subset A \cup B$; that is, A and B are a disconnection of the interval $[a,b]$. This contradicts the fact that $[a,b]$ is connected. Hence E is connected.

Finally, suppose that E satisfies $[x,y] \subset E$ whenever $x, y \in E$ with $x < y$. We want to prove that E is an interval. But it follows from this condition that E contains the open interval $(\inf E, \sup E)$, where we include the possibilities $\inf E = -\infty$ and $\sup E = \infty$. (Why?) Thus, E must be an interval; which particular type of interval depends on the disposition of $\inf E$ and $\sup E$ as finite, or not, and as elements, or not, of E.

We can now shed some light on the structure of open sets in $\mathbb{R}$. The proof of the theorem that characterizes open subsets of $\mathbb{R}$ shows that each nonempty open set $\mathcal{U}$ in $\mathbb{R}$ can be uniquely written as the union of connected subsets. Indeed, we wrote an open set in terms of "maximal" intervals I_x, and such intervals are actually maximal with respect to being connected subsets of $\mathcal{U}$ (i.e. no larger subset of $\mathcal{U}$ will be connected). At each $x \in \mathcal{U}$, we took I_x to be the union of all of the open subintervals in $\mathcal{U}$ that contain x. Thus, I_x is

both open and connected, and hence it is an open interval. The remainder of the proof shows that two such connected "components" of U are either identical or disjoint. There are at most countably many distinct I_x, the union of which must be all of U.

Given any set E, we call the maximal (with respect to containment) connected subsets of E the connected components of E. Essentially the same line of reasoning as above shows that every set can be written (uniquely) as the disjoint union of its connected components. A connected set, then, is a set with only one connected component (namely, itself).

We are more than ready to speak of continuous functions and connectedness. Our first result shows that the two-point discrete space is the canonical disconnected set.

Lemma: M is disconnected if and only if there exists a continuous map from M onto $\{0,1\}$ (the two-point discrete space).

Proof: If $f: M \to \{0,1\}$ is onto, then $A = f^{-1}(\{0\})$ and $B = f^{-1}(\{1\})$ are disjoint, nonempty, and satisfy $A \cup B = M$. If f is also continuous, then A and B are clopen sets and so form a disconnection of M.

Conversely, if A and B are a disconnection of M, then setting $f(a) = 0$ for $a \in A$ and $f(b) = 1$ for $b \in B$ defines a continuous map from M onto $\{0,1\}$. (Why?)

The lemma above is telling us that there is no continuous method of splitting a connected set M into two discrete "parcels". More generally, it follows that M is connected if and only if any continuous map from M into a discrete space is necessarily constant.

The lemma gives a nearly perfect replacement for the definition of disconnected. All of the notational difficulties that we faced earlier are now hidden in subtleties of language. For example, we have traded the cumbersome notation of relatively open sets for the tacit understanding that continuity may mean relative continuity. Most convenient. All of this hard work is beginning to pay off! In fact, we can now give a very short proof of that generalized intermediate value theorem we have been looking for:

Thm: Let $f:(M,d)\to(N,\rho)$ be continuous, and let E be a subset of M. If E is connected, then $f(E)$ is connected.

Proof: Suppose that $f(E)$ is not connected. Then there exists a continuous, onto map $g: f(E)\to\{0,1\}$. But this means that $g\circ f: E\to\{0,1\}$ is continuous and onto. That is, E is not connected.

To see that the above theorem is a generalization of the intermediate value theorem, we just need to bring our characterization of intervals of $\mathbb{R}$ as the only connected subsets of $\mathbb{R}$ back into the picture: The image of an interval under a nonconstant continuous function is again an interval.

Corollary: If I is an interval in $\mathbb{R}$ and if $f:I\to\mathbb{R}$ is a nonconstant continuous function, then $f(I)$ is an interval. In particular, if $a,b\in I$ with $f(a)\neq f(b)$, then f assumes every value between $f(a)$ and $f(b)$.