Math Circle for March 24 and 31, 2004

Distance and Metric Spaces by David Hartenstine

Measuring the distance between two objects is a basic mathematical activity, and provides a starting point for geometry. A familiar example of this is finding the (Euclidean) distance between two points in the xy-plane. The set of all points in the plane together with this method of computing the distance between any two points is an example of a *metric space*, which is a set S of objects together with a *metric*, a way of computing the distance between any two elements of S. There are many examples of metric spaces other than points in the plane; metrics can be defined on matrices, functions, sets of points and other mathematical objects.

Distance in the Plane

The (Euclidean) distance $d((x_1, y_1), (x_2, y_2))$ between two points (x_1, y_1) and (x_2, y_2) is the length of the line segment with endpoints (x_1, y_1) and (x_2, y_2). By drawing a right triangle with vertices at (x_1, y_1), (x_2, y_2) and (x_2, y_1), and using the Pythagorean theorem, this length is

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

By considering this distance function, we can observe some of its properties and then define a *metric*, or distance function, for situations other than points in the plane, based on these properties.

1. The distance between two points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ is always non-negative, $d(P_1, P_2) \geq 0$. Also, the only way $d(P_1, P_2)$ can be zero is if $x_1 = x_2$ and $y_1 = y_2$, meaning that $P_1 = P_2$. So for any two points P_1 and P_2, $d(P_1, P_2) \geq 0$ and $d(P_1, P_2) = 0$ if and only if $P_1 = P_2$.

2. The distance is symmetric, meaning that the distance from P_1 to P_2 is the same as the distance from P_2 to P_1, or $d(P_1, P_2) = d(P_2, P_1)$ for any two points P_1 and P_2.

3. A less obvious property of the distance function is the *triangle inequality*: For any points P_1. P_2 and P_3.

$$d(P_1, P_2) \leq d(P_1, P_3) + d(P_3, P_2).$$

The reason for calling this the triangle inequality comes from the geometric property that if P_1, P_2 and P_3 are the vertices of a triangle. the sum of the lengths of any two sides is larger than the length of the third side. Another way of thinking of the triangle inequality is that if you want to from P_1 to P_2, it is never shorter to first go to some point P_3 and then go to P_2 from P_3.

These three properties are considered to be the most important features of the distance function, and provide the definition of a metric space.

Metric Spaces

Let S be a set, and let d be a (real-valued) function defined on pairs of elements of S. S is called a *metric space (with metric d)* if:

1. $d(P_1, P_2) \geq 0$ for all P_1, $P_2 \in S$ and $d(P_1, P_2) = 0$ if and only if $P_1 = P_2$.

2. $d(P_1, P_2) = d(P_2, P_1)$ for all P_1, $P_2 \in S$.

3. $d(P_1, P_2) \leq d(P_1, P_3) + d(P_3, P_2)$ for all P_1, P_2, $P_3 \in S$.

Note that the set of all points in the xy-plane with the distance formula above is a metric space.

Once a set has a metric (a way to measure distance), other geometric objects can be defined. One example is a circle. In standard geometry, a circle of radius r centered at a point P_0 is the set of all points that are (or are less than) a distance r away from P_0. The metric space version of this is: If d is a metric on a set S, the *circle centered at P_0 of radius r* is the set of points $P \in S$ such that $d(P, P_0) = (\leq)r$. These "circles" might look very different from Euclidean circles.

First we'll take a look at some metrics on the number line and in the $xy-$plane. Each of these metrics gives a different way of measuring the distance between points. Later we'll look at examples of metric spaces for different underlying sets.

A simple example of a metric space is the number line with the metric

$$d(x, y) = |x - y|.$$

Since this distance is defined by absolute value, $d(x, y) \geq 0$ for all x and y, and the only way it can be zero is if $x = y$. Also, for any x and y. $d(x, y) = d(y, x)$. The only thing left to check in order to verify that this is a metric space is the triangle inequality. We need to show that for all x, y and z,

$$|x - y| \leq |x - z| + |z - y| \tag{1}$$

If $x = y$, then $|x - y| = 0$, and inequality (1) holds for any point z. If $x \neq y$, we can assume that $x < y$. There are then three possible locations for the third point z (drawing a picture of each of these possibilities might make things more clear). If $z < x$, $|y - z| > |x - y|$ and (1) holds. If z is in between x and y, then

$$|x - y| = |x - z| + |z - y|$$

which is OK for (1) to be true. Finally, if $z > y$, then $|x - z| > |x - y|$ and (1) holds for this case too. Therefore, the triangle inequality holds, and this is indeed a metric space.

Examples and Anti-Examples

For each of the following functions, determine if it defines a metric on the given set. If it is a metric, what do circles look like in that metric space? If it is not a metric, is it possible to make it into a metric by working with a subset of the given set?

1. Foi points P_1 and P_2 in the xy-plane, let

$$d(P_1, P_2) = \begin{cases} 0, & \text{if } P_1 = P_2 \\ 1, & \text{if } P_1 \neq P_2. \end{cases}$$

2. For x, y on the number line, let $d(x, y) = |x - 2y|$.

3. For points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ in the xy-plane, let

$$d(P_1, P_2) = \max\{|x_1 - x_2|, |y_1 - y_2|\}.$$

4. $d(x,y) = |x^2 - y^2|$ for x and y on the number line.

5. For $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ in the xy-plane. let
$$d(P_1, P_2) = |x_1 - x_2| + |y_1 - y_2|.$$

6. For x, y on the number line, let $d(x,y) = (x-y)^2$.

Answers: #1 is a metric and circles are either single points or the whole plane. #2 is not a metric because if $x \neq 0$, $d(x,x) \neq 0$. Note also that this distance is not symmetric: $d(x,y) \neq d(y,x)$. #3 is a metric and circles are squares with sides parallel to the $x-$ and $y-$ axes. #4 is not a metric, because for example $d(-1,1) = 0$; it is however a metric when the set is restricted to the interval $[0,\infty)$. #5 is a metric and circles are diamond-shaped (they are like the squares in #3, but rotated 45 degrees). #6 is not a metric, because the triangle inequality fails (take $x = -1$, $y = 1$ and $z = 0$).

Distance in Space and on Surfaces

The formula for the Euclidean distance between two points $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$ in 3-space is
$$d(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}.$$
This is the length of the line segment from P_1 to P_2, by using the Pythagorean theorem as in the plane. Any metric defined on points in 3-space can be used to define a distance between points. The metrics in the plane in the last section can be extended to 3-space.

The formula above, however, is not always the way we measure distance between points lying in a three dimensional space. For example, we can think of the Earth as being a subset of 3-space, and the surface of the Earth as being the surface of a sphere. When measuring the distance between points on the Earth's surface, say from Salt Lake City to Rome, we don't draw a line through the two points and measure the distance along those points. This distance would correspond to the distance traveled if a tunnel was dug in a straight line through the Earth, starting at Salt Lake and ending at Rome. This is not very practical, one usually does not travel between cities by tunneling from one to the other. Instead we measure the distance between two points P_1 and P_2 on the Earth by measuring the shortest path starting at P_1 and ending at P_2 that stays on the surface of the Earth.

Generally, if S is any surface, we can define the distance between two points P_1 and P_2 on S, by

$$d(P_1, P_2) = \min\ \text{length}(\gamma) \tag{2}$$

where γ is any path starting at P_1, ending at P_2, and always staying on the surface S ($\gamma \subset S$). Notice that when S is all of 3-space or is the plane, this is the same thing as what we get with the standard distance formulas.

Some Questions

1. Given two points on a sphere (like the Earth's surface), is there always a path of minimal length between the two points? If so, can you describe what it looks like? Can there be more than one shortest path between two given points?

2. Can you think of a set, say in 3-space or in the plane, where there is no path of minimal length?

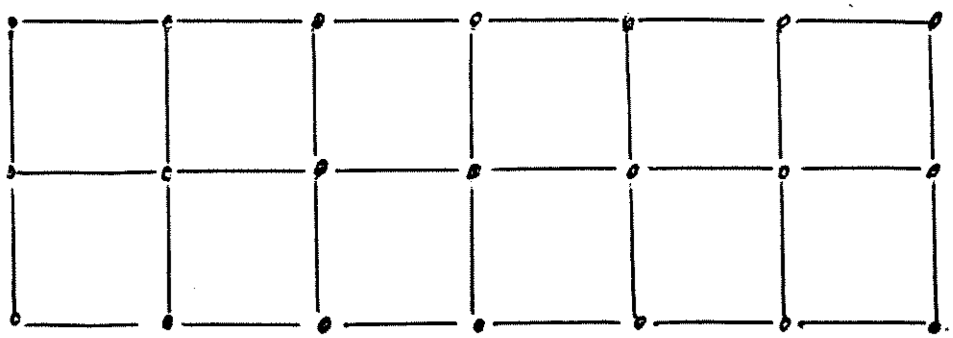
3. Is the distance defined above in (2) a metric for all surfaces? If not, what do we need to know about the surface to guarantee that it is a metric?

Answers: For #1, the (shorter) arc of the great circle through the two points is a path of minimal length. There are infinitely many minimal paths when the two points are antipodal, like the North and South Poles. For #2, if one point is deleted from the plane, there is no shortest path between any two points for which the line between them would pass through the missing point. The example of.#2 shows that this distance is not defined (and therefore not a metric) between two points when there is no minimal path. If the surface S is not connected, then there are no paths between points in different components, and again (2) is not a metric. When there exists a minimal path between any pair of points in S, (2) does define a metric.

Metrics on Graphs

Metrics are used to measure more than distance between points in the plane or in space or on surfaces. They can also be used to measure distance between points on graphs, the distance between two matrices or functions, or between two sets.

Example 1 Taxicab metric: Consider a square grid of points connected by vertical and horizontal lines, like that pictured below. We can define the distance between two of the points to be the minimal number of line segments that have to be covered when going from one point to the other. We can think of the lines as (two-way) streets and the points as intersections in a city like Salt Lake that is set up on a grid. Then the distance between P_1 and P_2 is the smallest number of city blocks that have to be driven past in order to get from P_1 to P_2.



1. Verify that this is a metric. Is this similar to any of the metrics we saw for points in the plane?
2. What if one-way streets are allowed, so that some of the edges can only be traveled in one direction? We can still define the distance between two points as the smallest number of blocks that must be passed to get from one to the other. Is this distance a metric? Why or why not?

Answer to #2: No. If P_1 and P_2 are adjacent intersections connected by a one-way street from P_1 to P_2, then $d(P_1, P_2) < d(P_2, P_1)$.

Example 2 Chess pieces: The game of chess is played on an 8×8 checkerboard. There are several different pieces involved in the game, and each is allowed to move according to different rules. For each piece, we can measure the distance between any two squares S_1 and S_2 on the board by defining it to be the smallest number of moves permitted for that piece that gets the piece from S_1 to S_2. Assume that nothing (like another piece in the way) prevents the piece from performing its moves. As a result of this assumption, this distance is not of much practical use from the point of view of the actual game of chess.

Let's focus on just four of the chesspieces. The bishop can move in any diagonal direction any number of squares until the edge of the board is reached (but in one particular move can only move in one direction). The rook (the one that looks like a castle) moves in a similar way to the bishop, but has to move in either a horizontal or vertical direction. The queen can do anything the rook or bishop can. The knight (horse) can move in either of the following ways: move 1 square in a vertical direction followed by 2 squares horizontally, or move two squares vertically, then 1 square horizontally. Note that there are eight possible moves for the knight unless the edge of the board makes some of them impossible.

1. Which pieces lead to metrics? You might want to convince yourself that all of the requirements for metrics are satisfied (note that this distance is similar to the distance used on surfaces above and the taxicab metric), because the distance is defined in terms of a minimum number of moves (and all moves are reversible), except for possibly the ability to measure the distance between any two squares. If a piece cannot get from S_1 to S_2, the distance is not defined.

2. For those pieces whose distances aren't metrics, is there a way to define the distance between squares that a piece can't move between in order to make it a metric?

3. For each of the distances (metric or not), what do the circles of radius 1, 2 and 3 look like?

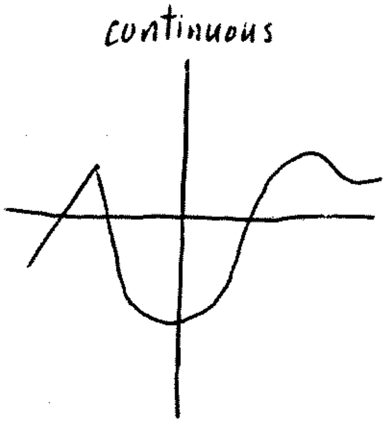
4. What is the maximum distance between any two squares in each metric? This number is called the *diameter* of the metric space.

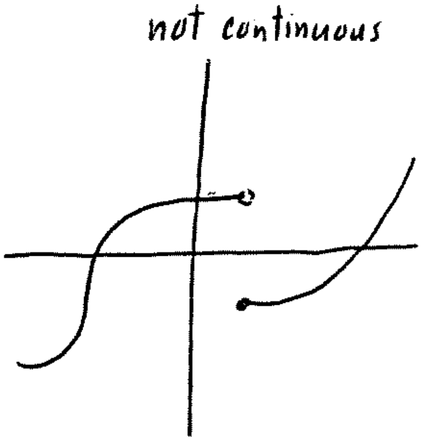
Answers: All of them except the bishop, which can't move from a black square to a red square, produce metrics. We can define a metric by the bishop by defining the "distance" between any red and black square to be some large number (larger than the maximum number of moves to get from any square to another square of the same color); the triangle inequality will then hold. Of course, the distance between red and black squares will not correspond to the number of moves necessary to move between the squares. In the rook metric, the circle of radius 1 is the vertical and horizontal lines of squares passing through the center (of the circle). Any circle of radius two is the whole board. In the bishop distance. the circle of radius one is the squares that lie on the diagonals through the center. and a circle of radius two is the set of all squares with the same color as the center. The circle of radius one for the queen is the set of all squares that lie on the diagonals or on the horizontal and vertical lines through the center. a circle with radius two is the whole board. The circles in the knight metric are more complicated. The diameter of the board in the rook or queen metric is two. In the knight metric, the board has diameter six (it takes six moves to get from corner to corner).

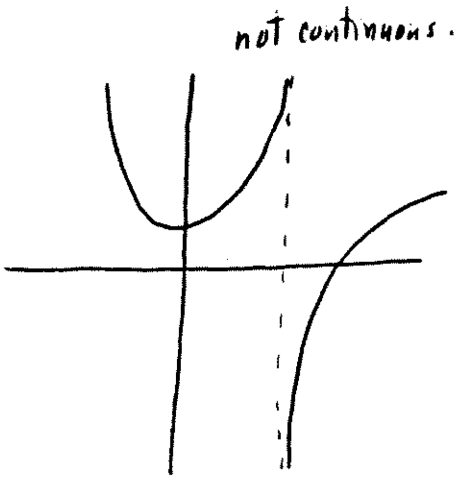
A Metric on Functions

Let S be a collection of functions. If we can find a metric on S. we can compute the distance between any two functions. As a result, we will be able to make sense of the statement "These two functions are a distance X apart".

To keep things somewhat simple and make things definite, let's consider the set of functions that are defined for all $x \in [0,1]$, and are continuous on that interval. A function is continuous on an interval if its graph can be drawn without lifting the pen from the paper. So continuous functions don't have any jumps in their graphs.







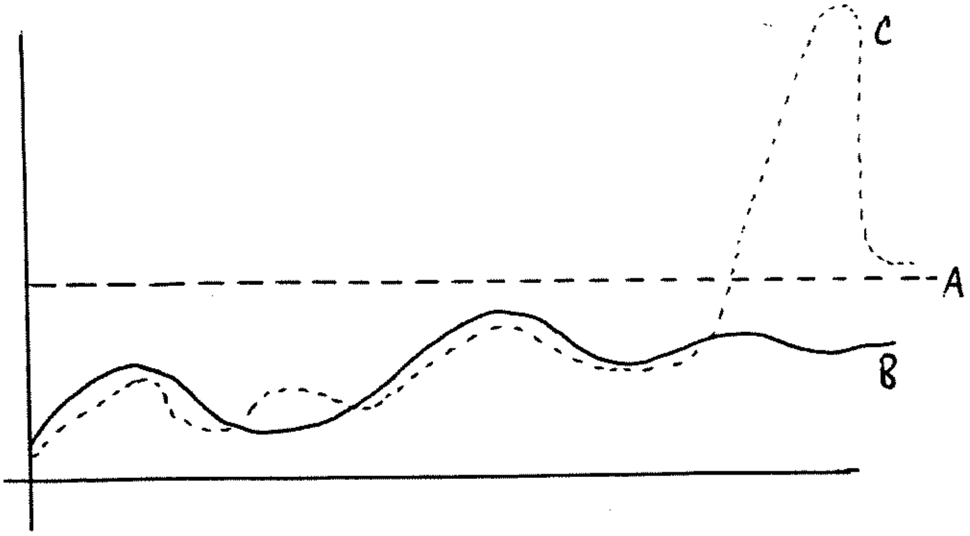
The set of continuous functions on $[0,1]$ is often denoted $C([0,1])$. Continuous functions have some nice properties. The sum or difference of any two continuous functions is continuous. Also, if f is continuous, then so is

$|f|$. Also, a continuous function achieves a maximum and a minimum on a closed and bounded interval. This means that if f is a continuous function on an interval $[a, b]$, then there are numbers x_1 and x_2 in the interval $[a, b]$ such that $f(x_1) \geq f(x)$ for all x in $[a, b]$, and $f(x_2) \leq f(x)$ for all $x \in [a, b]$.

We can define the distance between f and $g \in C([0, 1])$ by

$$d(f, g) = \max_{0 \leq x \leq 1} |f(x) - g(x)| \tag{3}$$

With respect to this distance, function A is closer to function B than function C is.



1. Does (3) define a metric on $C([0, 1])$?

2. What does the circle of radius 1 centered at $f(x) \equiv 0$ look like? What about the circle centered at $g(x) = x^2$?

3. What happens if we replace the closed interval $[0, 1]$, with the open interval $(0, 1)$?

4. What happens if we replace the bounded interval $[0, 1]$ with the infinite interval $[0, \infty)$?

Answers: This is a metric. Since it is defined by absolute value, $d(f, g) \geq 0$ for any $f, g \in C([0, 1])$. The only way it can be zero is if $f(x) = g(x)$ for all x in the interval, in other words if f and g are the same function. Also, when f

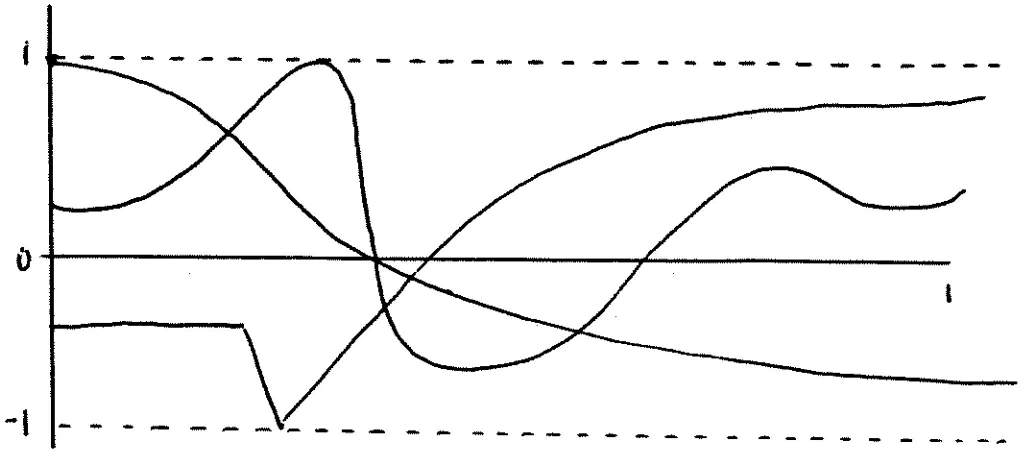
and g are continuous, the function $|f-g|$ is also continuous on $[0,1]$, so it has a maximum value on $[0,1]$, so $d(f,g)$ is defined. For any x, $|f(x)-g(x)| = |g(x)-f(x)|$, so $d(f,g) = d(g,f)$. To see the triangle inequality, let f, g and h be any functions in $C([0,1])$. Then, for any $x \in [0,1]$, by using the triangle inequality for points on the number line,

$$|f(x)-g(x)| \leq |f(x)-h(x)| + |h(x)-g(x)|$$
$$\leq \max_{0\leq x\leq 1} |f(x)-h(x)| + \max_{0\leq x\leq 1} |h(x)-g(x)| = d(f,h)+d(h,g).$$

Now since we started with an arbitrary $x \in [0,1]$, this inequality holds for any point x where $\max|f-g|$ is achieved, and we get

$$d(f,g) = \max_{0\leq x\leq 1} |f(x)-g(x)| \leq d(f,h)+d(h,g)$$

so the triangle inequality holds, and this is a metric. The circle of radius one centered at any function is the set of all functions in $C([0,1])$ whose maximum distance from the center is one (or less than one). For example, the circle with radius one, centered at $f(x) \equiv 0$ is the set of all continuous functions on $[0,1]$ with maximum absolute value 1. Some examples are:



If we consider an open interval, the distance between two functions might not be defined because it could be infinite; take for example the functions $f(x)=0$ and $g(x)=1/x$ on the open interval $(0,1)$. Both of these functions are continuous, but for x close to 0, $g(x)$ is very large and the maximum of $|f(x)-g(x)| = 1/x$ does not exist. We run into similar problems on the infinite interval $[0,\infty)$. The functions $f(x)=0$ and $g(x)=x$ are both continuous on this interval, but $\max|f(x)-g(x)| = \max x$ does not exist.

Metric on Sets

Is there a way to define a metric on the set of regions or sets in the plane? What would such a distance have to do? Suppose S and T are two subsets of the plane. Then if d is a metric on such sets, we must have that $d(S,T) \geq 0$ and is zero if and only if $S = T$. This means that the distance between two sets should be zero only if the sets are exactly the same. We also need to know that $d(S,T) = d(T,S)$, that is that the distance from S to T is the same as the distance from T to S. Finally, we need to know that the triangle inequality holds, in other words, if U is a third set in the plane, then

$$d(S,T) \leq d(S,U) + d(U,T)$$

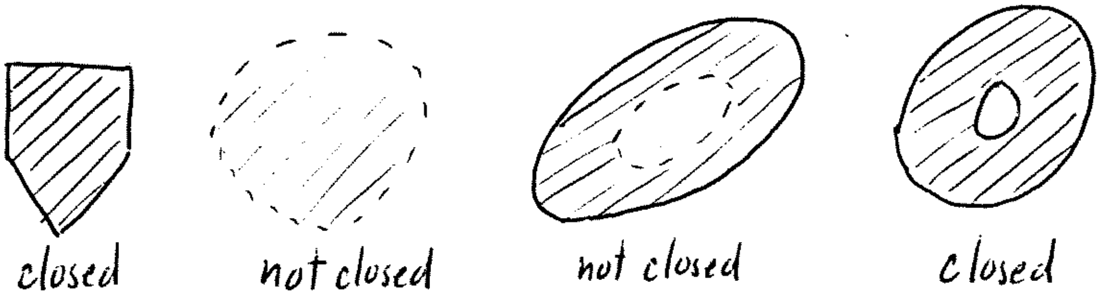
Since we know how to measure distance between points in the plane and sets are made up of points, it might be a good idea to use something we know about distance between points in defining a distance between sets. Let's start with something simple. What if we let S_1 be a square with side length 1 and centered at $(0,0)$, and define S_2 to be the same square translated in the positive x-direction say 10 units? What should the distance between S_1 and S_2 be? Well, if we translated S_2 ten units in the negative x-direction, it would land right on top of S_1, so S_1 and the transported S_2 are the same set, so it makes sense to define the distance between a set and and a translation of the set to be the length of the translation.

The problem of defining a metric on the set of planar sets would not be very difficult if all sets were translations of each other. Unfortunately, this is not the case. Two squares of different areas are not translates of each other, and neither are a square and a triangle. Furthermore, not even all congruent sets are translates of each other: If we take a square and rotate it 45°, we still have a square, but the original square and the rotated one are not the same set, and they are not translates of each other.

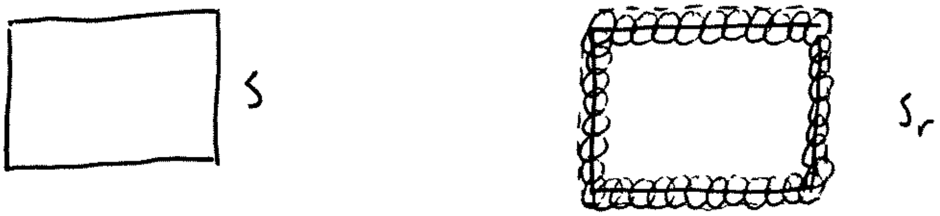
The distance between two sets should be small when the sets are close to being identical. The problem is how to determine when sets are close to being identical. What should the distance be between a set, say a circle with radius one including its interior and another circle (again including its interior) with the same center but with radius three? What about with a circle of radius 100? All of these circles have exactly the same shape, but the largest one is pretty far from being identical to the small one.

For technical reasons, lets confine our attention to what are called *compact* sets. A set of points in the plane is compact if it is both *bounded* and *closed.*

A set S is bounded if a circle with finite radius can be drawn in such a way that S is inside it. A set is closed if it contains all of its boundary points. Some explanation is necessary here. A circle or a polygon. together with all of the points interior to it, is a closed set. The boundary of such a set is the outside of the set (which is the original circle or polygon). However, if my set is just the set of points interior to a circle, this set is not closed because the boundary (the circle itself) is not included. In fact. if even one point on the circle is not included in the set, the set is not closed. These are very simple examples of closed sets – there are much more complicated examples.



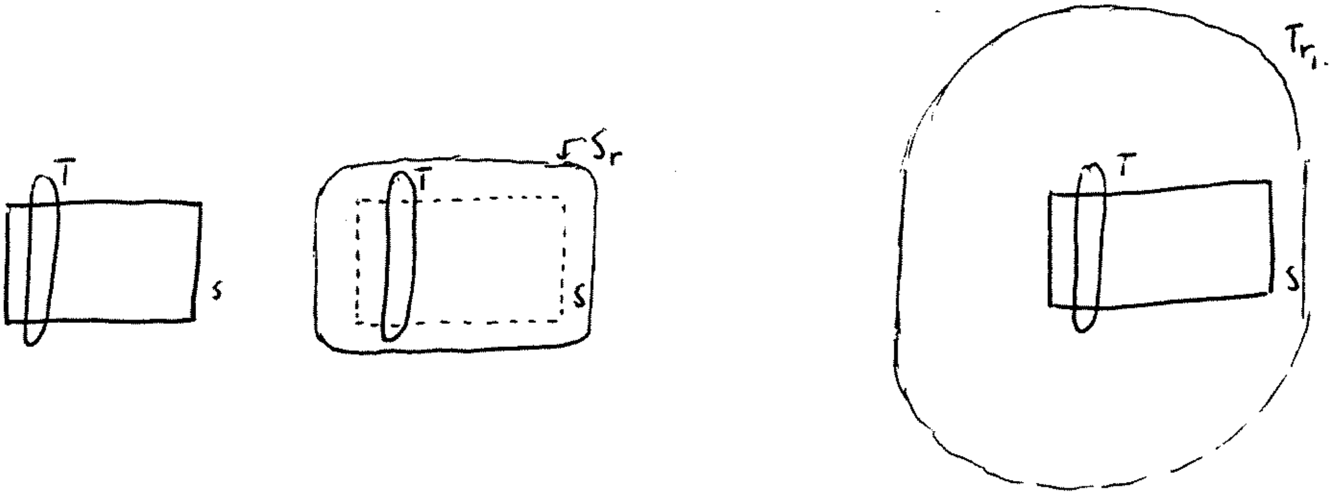
Finding a metric on the set of compact sets is not easy, and one was discovered fairly recently (say in the last 100 years or so). This metric is called the **Hausdorff metric**. To define it we need to define the r-expansion of a set. Let S be a set of points in the plane. The *r-expansion* of S (let's call it S_r) is the set of all points (x, y) in the plane that are a distance less than or equal to r away from some point in S. A way to visualize this is by drawing a circle of radius r (including its interior) at each point of S. Then S_r is the union of all points contained in these circles.



We can now define the Hausdorff metric on compact sets.

$$d(S,T) = \min\{r : S \subset T_r \text{ and } T \subset S_r\}$$

So the Hausdorff distance between S and T is the smallest number r for which the r-expansion of S contains T AND the r-expansion of T contains S. A special case is when $T \subset S$. Then we don't have to expand S at all in order for it to swallow up T, so the distance between T and S is just the amount we have to expand T in order for T_r to contain S. In the picture below, a small expansion of S will contain T, but a much larger expansion of T is needed in order to swallow S.



1. Check that this is a metric.

2. What is the distance between a set and its translation by x units? Is this what we thought earlier?

3. What do circles in this metric look like?

4. Do we still have a metric if sets that aren't bounded are allowed?

5. What happens if sets that aren't closed are allowed? Hint: What is the distance between a circle plus its interior and its interior?

Answers: This distance can't be negative, since only $r \geq 0$ makes sense for the r-expansions. Also, because the sets are bounded, the distance between any two of them is finite. Since the definition is symmetric. we have that $d(S,T) = d(T,S)$ for any S and T. If $S = T$. then $S \subset T$ and $T \subset S$ and $d(S,T) = 0$ (like it should be). Also. if $d(S.T) = 0$. then $S \subset T$ and $T \subset S$. and therefore $S = T$. The triangle inequality can be proved by contradiction. Suppose that $d(S,T) = r_0$ for some number r_0, and suppose that there is a third set U such that the triangle inequality does not hold for S, T and U. This means that

$$d(S,T) \not\leq d(S,U) + d(U,T)$$

Let $r_1 = d(S,U)$ and $r_2 = d(U,T)$. Then $r_1 + r_2 < r_0$. Therefore, $S \subset U_{r_1}$, and since $U \subset T_{r_2}$, $U_{r_1} \subset T_{r_1+r_2}$. Therefore, $S \subset T_{r_1+r_2}$. Similarly, $T \subset S_{r_1+r_2}$. Therefore, since the distance is defined as a minimum, $d(S,T) \leq r_1 + r_2$, but we already know that $d(S,T) = r_0 > r_1 + r_2$, but this is a contradiction. If unbounded sets are allowed, the distance between them could be infinite; an example is a point and a half-plane. If sets that aren't closed are allowed, the minimum may not be defined: For example, let S be a circle and its interior and let T be just the interior of the circle. Then $T \subset S$, and for any $r > 0$, $S \subset T_r$, so there is no smallest r, and the distance is not defined. If this minimum is taken to be zero, then we have two sets that are not identical (but are very close) that are a distance zero apart and this distance is no longer a metric.