Chapter 2

Metric Spaces

A normed space is a vector space endowed with a norm in which the length of a vector makes sense and a metric space is a set endowed with a metric so that the distance between two points is meaningful. There is always a metric associated to a norm. Normed spaces form a sub-class of metric spaces and metric spaces form a sub-class of topological spaces. In Section 1 the definitions of a normed space and a metric space are given and some examples are present. In Section 2 limit of sequences and continuity of functions in a metric space is defined. Next open and closed sets are introduced and used to describe the convergence of sequences and continuity of functions in Section 3. Relevant notions such as the boundary points, interior points, closure and interior of a set are collected in Section 4.

2.1 Definitions and Examples

By far we are familiar with the Euclidean space and the space of continuous functions. You may have already observed there are certain similarities between these two mathematical entities. For instance, the completeness theorem for $\mathbb{R}$ states that every Cauchy sequence converges. This property can be extended to $\mathbb{R}^n$ without much effort. In fact, it suffices to observe that a Cauchy sequence in $\mathbb{R}^n$ is defined by replacing the absolute value by the Euclidean norm

$$|x| = \sqrt{\sum_{j=1}^{n} x_j^2} \ , \quad x = (x_1, \cdots, x_n) \ .$$

In other words, a sequence $\{x^k\}$ in $\mathbb{R}^n$ is called a Cauchy sequence if for each $\varepsilon > 0$, there is some k_0 such that

$$|x^k - x^m| = \sqrt{\sum_{j=1}^{n} (x_j^k - x_j^m)^2} < \varepsilon \ , \quad \forall k, m \geq k_0 \ .$$

On the other hand, in the space of continuous function $C[a,b]$, a corresponding completeness theorem states every Cauchy sequence of functions converges uniformly to a continuous function. Recall that a sequence $\{f_k\}$ is a Cauchy sequence if for every $\varepsilon > 0$, there is some k_0 such that

$$\|f_k - f_m\|_\infty < \varepsilon \ , \quad \forall k, m \geq k_0 \ .$$

You can see things are the same except the Euclidean norm is now replaced by the sup-norm.

In order to unify the Euclidean spaces and the space of continuous functions, we introduce the general definition of a normed space. This is our first step of abstraction. A **norm** $\|\cdot\|$ is a function on a real vector space X to $[0,\infty)$ satisfying the following three conditions, for all $x, y \in X$ and $\alpha \in \mathbb{R}$,

N1. $\|x\| \geq 0$ and "=" 0 if and only if $x = 0$,

N2. $\|\alpha x\| = |\alpha| \ \|x\|$, and

N3. $\|x + y\| \leq \|x\| + \|y\|$. (Triangle inequality)

A normed space is a vector space endowed with a norm. The pair $(X, \|\cdot\|)$ is called a **normed space**. Here are some examples of normed spaces.

Example 2.1. Let $\mathbb{R}$ be the set of all real numbers. For $x \in \mathbb{R}$, set its Euclidean norm $|x|$ to be the absolute value of x. It is easily seen that $|x|$ satisfies N1-N3 above and so it defines a norm. In particular, N3 is the usual triangle inequality. Thus $(\mathbb{R}, |\cdot|)$ is a normed space. From now on whenever we talk about $\mathbb{R}$, it is understood that it is a normed space endowed with the Euclidean norm.

Example 2.2. More generally, let $\mathbb{R}^n$ be the n-dimensional real vector space consisting of all n-tuples $x = (x_1, \ldots, x_n)$, $x_j \in \mathbb{R}$, $j = 1, \ldots, n$. Introduce the **Euclidean norm**

$$|x| = \sqrt{x_1^2 + \cdots + x_n^2} \ .$$

It reduces to the previous example when $n = 1$. Apparently, N1 and N2 are fulfilled. Moreover, taking square of both sides of the triangle inequality, N3 follows from the Cauchy-Schwarz Inequality

$$\Big| \sum_1^n x_j y_j \Big| \leq \big(\sum_1^n x_j^2 \big)^{1/2} \big(\sum_1^n y_j^2 \big)^{1/2} \ .$$

Example 2.3. In general, for $p \in [1,\infty)$, we may define

$$\|x\|_p = \left(\sum_{j=1}^\infty |x_j|^p \right)^{1/p} \ ,$$

and for $p = \infty$, define $\|x\|_\infty = \max\{|x_1|, \cdots, |x_n|\}$. It can be shown that this is a norm on $\mathbb{R}^n$. To verify N3 one needs the Minkowski's Inequality, see my notes on inequalities. As the 2-norm comes up much more often than other p-norm, and it justifies to use a simpler notation $|\cdot|$.

Example 2.4. Let $C[a, b]$ be the real vector space of all continuous, real-valued functions on $[a, b]$. For $f \in C[a, b]$, define the **sup-norm**

$$\|f\|_\infty \equiv \max\{|f(x)| : \quad x \in [a, b]\}.$$

It is easily checked that it defines a norm on $C[a, b]$. Parallel to $\mathbb{R}^n$, for each $p \in [1, \infty)$, we define the p**-norm** by

$$\|f\|_p = \left(\int_a^b |f|^p\right)^{1/p}, \quad p \in [1, \infty) ,$$

on $C[a, b]$. Each p-norm defines on $C[a, b]$. The triangle inequality is the integral form of Minkowski's Inequality, see the end of the chapter.

Example 2.5. Let E be a nonempty subset of $\mathbb{R}^n$, let $C_b(E)$ be the family of all bounded, continuous functions defined in E. It is readily checked that $C_b(E)$ forms a vector space under the usual addition and scalar multiplication of functions. Moreover, the supnorm is well defined and makes $C_b(E)$ into a normed space. It reduces to $C[a, b]$ when $E = [a, b]$ as continuous functions on a closed, bounded interval must be bounded.

In passing we point out same notations such as $\|\cdot\|_p$ have been used to denote norms on different spaces. As they arise in quite different context, hopefully it would not cause much confusion.

Example 2.6. Let $R[a, b]$ be the vector space of all Riemann integrable functions on $[a, b]$ and consider $\|f\|_\infty$ and $\|f\|_1$ as defined above. It is routine to verify that while $\|f\|_\infty$ defines a metric, $\|f\|_1$ is not as N2 and N3 are satisfied but not N1. In fact, we know from the previous chapter that $\|f\|_1 = 0$ if and only if f vanishes almost everywhere. Hence $R[a, b]$ does not form a normed space under $\|\cdot\|_1$. The space of integrable functions, where Riemann integrability is replaced by Lebesgue integrability, will be studied in detailed in Real Analysis.

Observe that whenever $(X, \|\cdot\|)$ is a normed space and Y is a vector subspace of X, we can make Y into a normed space by restricting the norm to Y. However, apparently such hereditary property does not hold when Y is merely a subset but not a subspace of X. We have come to the second step of abstraction. We will detach the vector space

structure from the norm structure by introducing the notion of a metric space as follows.

Let X be a non-empty set. We would like to define a concept of distance which assigns a positive number to every two points in X, that is, the distance between them. In analysis the name metric is used instead of distance. (But "d" not "m" is used in notation. I have no idea why it is so.) A **metric** on X is a function from $X \times X$ to $[0, \infty)$ which satisfies the following three conditions: $\forall x, y, z \in X$,

M1. $d(x, y) \geq 0$ and equality holds if and only if $x = y$,

M2. $d(x, y) = d(y, x)$, and

M3. $d(x, y) \leq d(x, z) + d(z, y)$. (Triangle inequality)

M3 is a key property of a metric. M2 and M3 together imply another form of the triangle inequality, namely,

$$|d(x, y) - d(x, z)| \leq d(y, z) .$$

There is always a metric associated to a norm in a normed space $(X, \|\cdot\|)$. Indeed, letting

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

it is readily checked that d defines a metric on X. This metric is called the **metric induced** by the norm. For instance, in $\mathbb{R}^n$ we have

$$d_p(x, y) = \|x - y\|_p , \quad p \in [1, \infty] .$$

On $C[a, b]$, we have

$$d_p(f, g) = \|f - g\|_p , \quad p \in [1, \infty] .$$

Given a metric space (X, d), the **metric ball** centered at x and with radius r is the set

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

As there could be more than one metrics on a set, it is interesting to compare the metric balls with respect to different metrics. For instance, on $\mathbb{R}^n$ there are infinitely many metrics given by d_p. Denote the metric balls in d_2, d_1 and d_∞ metrics by $B_r(x)$, $B_r^1(x)$ and $B_r^\infty(x)$ respectively. Then $B_r(x)$ is the usual ball of radius r centered at x and $B_r^\infty(x)$ is the cube of length r centered at x. I let you draw $B_r^1(x)$ as an exercise.

In the following we give two examples of metrics defined on a set without the structure of a vector space. They are not metric spaces induced by normed spaces.

Example 2.7. Let X be a non-empty set. For $x, y \in X$, define

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

The metric d is called the **discrete metric** on X. The metric ball $B_r(x)$ consists of x itself for all $r \in (0,1]$ and is equal to X when $r > 1$.

Example 2.8. Let H be the collection of all strings of words in n digits. For two strings of words in H, $a = a_1 \cdots a_n$, $b = b_1 \cdots b_n$, $a_j, b_j \in \{0, 1, 2, \ldots, 9\}$. Define

$$d_H(a,b) = \text{the number of digits at which } a_j \text{ is not equal to } b_j.$$

By using a simple induction argument one can show that (H, d_H) forms a metric space. Indeed, the case $n = 1$ is straightforward. Let us assume it holds for n-strings and show it for $(n+1)$-strings. Let $a = a_1 \cdots a_n a_{n+1}, b = b_1 \cdots b_n b_{n+1}, c = c_1 \cdots c_n c_{n+1}, a' = a_1 \cdots a_n, b' = b_1 \cdots b_n$, and $c' = c_1 \cdots c_n$. As $d_H(a',b') \leq d_H(a,b)$ always holds for any a, b, it suffices to consider two cases, namely, (a) $a_{n+1} = b_{n+1}$ and (b) $a_{n+1} \neq b_{n+1}$. In (a), we have $d_H(a,b) = d_H(a',b')$. By induction hypothesis, $d_H(a,b) = d_H(a',b') \leq d_H(a',c') + d_H(c',b') \leq d_H(a,c) + d_H(c,b)$, done. In (b), $d_H(a,b) = d_H(a',b') + 1$. Since a_{n+1} is not equal to b_{n+1}, either c_{n+1} is not equal to a_{n+1} or b_{n+1}. Assume it is the former. Then $d_H(a,c) = d_H(a',c') + 1$, so $d(a,b) = d(a',b') + 1 \leq d_H(a',c') + d_H(c',b') + 1 \leq d_H(a,c) + d_H(c',b') \leq d_H(a,c) + d_H(c,b)$, done. (Thanks to a student who suggested this simplified proof.)

There is a common way to construct metrics based on the following observation. Let Φ be a bijection from a metric space (X,d) to a set Z. For $z_1, z_2 \in Z$, define $\rho(z_1,z_2) = d(\Psi(z_1), \Psi(z_2))$ where Ψ is the inverse of Φ. Then ρ becomes a metric on Z.

Example 2.9. The Euclidean metric on $\mathbb{R}$ is induced from the Euclidean norm. We now construct another metric on $\mathbb{R}$ as follows. Imagine that $\mathbb{R}$ is the x-axis on the plane and consider the unit circle $S = \{p = (x,y) : \ x^2 + y^2 = 1\}$. The unit circle is endowed with a natural metric, that is, the arclength between two points. (I let you define it rigorously.) For every point p on the circle not equal to the north pole $(0,1)$, the ray connecting the north pole and p would hit the x-axis at a unique point x. It is clear that $p \mapsto x$ sets up a one-to-one correspondence between $S \setminus \{(0,1)\}$ and $\mathbb{R}$, the x-axis. For $x_1, x_2 \in \mathbb{R}$, we define $\rho(x_1,x_2)$ to be the arclength of the arc formed by their corresponding points on the circle. The metric is bounded in the sense that $\rho(x_1,x_2) \leq 2\pi$ for all x_1, x_2.

Let Y be a non-empty subset of (X,d). Then $(Y, d|_{Y\times Y})$ is again a metric space. It is called a **metric subspace** of (X,d). The notation $d|_{Y\times Y}$ is usually written as d for simplicity. Every non-empty subset of a metric space forms a metric space under the restriction of the metric. In the following we usually call a metric subspace a subspace

for simplicity. Again a metric subspace of a normed space needs not be a normed space. It is so only if the subset is also a vector subspace. For example, consider the subsets $E = \{(x, y) : \ 2x + 3y = 0\}$ and $F = \{(x, y) : \ xy = 1, x, y > 0\}$ in $\mathbb{R}^2$. The restriction of the Euclidean metric to these two sets make them metric spaces. The first one is a one-dimensional vector subspace of $\mathbb{R}^2$, E is a normed space. On the other hand, while the restriction of the Euclidean metric on F makes it a metric space, it is no longer a metric induced from any norm. In fact, F is no longer a vector space. Taking $(1, 1)$ and $(2, 1/2)$ from F, that the point $(3, 2.5) = (1, 1) + (2, 1/2)$ does not satisfy $3 \times 2.5 = 1$ shows F does not inherit the vector space structure of $\mathbb{R}^2$.

2.2 Limits and Continuity

Convergence of sequences of real numbers and uniform convergence of sequences of functions are the main themes in Mathematical Analysis I and II and sequences of vectors were considered in Advanced Calculus I and II. With a metric d on a set X, it makes sense to talk about limits of sequences in a metric space. Indeed, a sequence in (X, d) is a map φ from $\mathbb{N}$ to (X, d) and usually we write it in the form $\{x_n\}$ where $\varphi(n) = x_n$. A sequence $\{x_n\}$ is said to **converge to** x if $\lim_{n\to\infty} d(x_n, x) = 0$, that's, for every $\varepsilon > 0$, there exists n_0 such that $d(x_n, x) < \varepsilon$, for all $n \geq n_0$. When this happens, we write or $\lim_{n\to\infty} x_n = x$ or $x_n \to x$ in X.

Convergence of sequences in $(\mathbb{R}^n, d_2)$ reduces to the old definition encountered before. From now on, we implicitly refer to the Euclidean metric whenever convergence of sequences in $\mathbb{R}^n$ is considered. For sequences of functions in $(C[a, b], d_\infty)$, it is simply the uniform convergence of sequences of functions in $C[a, b]$. Likewise, uniform convergence of sequences in $C_b(E)$ in Example 2.5 is the convergence with respect to d_∞.

As there could be more than one metrics defined on the same set, it is natural to make a comparison among these metrics. Let d and ρ be two metrics defined on X. We call ρ is **stronger** than d, or d is **weaker** than ρ, if there exists a positive constant C such that $d(x, y) \leq C\rho(x, y)$ for all $x, y \in X$. They are **equivalent** if d is stronger and weaker than ρ simultaneously, in other words,

$$d(x, y) \leq C_1\rho(x, y) \leq C_2 d(x, y), \quad \forall x, y \in X,$$

for some positive C_1 and C_2. When ρ is stronger than d, a sequence converging in ρ is also convergent in d. When d and ρ are equivalent, a sequence is convergent in d if and only if it is so in ρ.

Take d_p and d_∞ on $\mathbb{R}^n$ as an example. It is elementary to show that for all $x, y \in \mathbb{R}^n$,

$$d_p(x, y) \leq n^{1/p} d_\infty(x, y) \leq n^{1/p} d_p(x, y),$$

and

$$d_1(x,y) \le n d_\infty(x,y) \le n d_1(x,y),$$

hence all d_p and d_∞ are all equivalent. The convergence of a sequence in one metric implies its convergence in all others.

It is a basic result in functional analysis that every two induced metrics in a finite dimensional normed space are equivalent. Consequently, examples of inequivalent induced metrics can only be found when the underlying space is of infinite dimension.

Let us display two inequivalent metrics on $C[a,b]$. For this purpose it suffices to consider d_1 and d_∞. On one hand, clearly

$$d_1(f,g) \le (b-a) d_\infty(f,g), \quad \forall f,g \in C[a,b],$$

so d_∞ is stronger than d_1. But the reverse is not true. Consider the sequence given by (taking $[a,b]=[0,1]$)

$$f_n(x) = \begin{cases} -n^3 x + n, & x \in [0, 1/n^2], \\ 0, & x \in (1/n^2, 1]. \end{cases}$$

We have $d_1(f_n,0) \to 0$ but $d_\infty(f_n,0) \to \infty$ as $n \to \infty$. Were $d_\infty(f_n,0) \le C d_1(f_n,0)$ true for some positive constant C, $d_1(f_n,0)$ must tend to ∞ as well. Now it tends to 0, so d_1 cannot be stronger than d_2 and these two metrics are not equivalent.

A metric space (X,d) is called **bounded** if $d(x,y) \le M$ for all $x,y \in X$. A set E is bounded if the subspace (E,d) is bounded. It is clear that E is bounded if and only if $E \subset B_R(x)$ for some $x \in X$ and $R > 0$. The **diameter** of a set is defined to be $\operatorname{diam}(E) = \sup\{d(x,y) : \ x,y \in E\} \le \infty$.

It is a little unexpected that two inequivalent metrics may have the same strength of convergence.

Proposition 2.1. *Let (X,d) be a metric space. Define*

$$\rho(x,y) = \frac{d(x,y)}{1+d(x,y)} \ .$$

Then ρ is a metric on X. Moreover, a sequence converges in d if and only it converges in ρ.

I leave the proof of this proposition as an exercise. We note that ρ is always bounded, $\rho(x_1,x_2) < 1$ for all x_1, x_2. For instance, on $\mathbb{R}$ there is the Euclidean metric and the new metric

$$\rho(x,y) = \frac{|x-y|}{1+|x-y|} \ .$$

We have $\rho(x,y) \leq |x-y|, \quad \forall x, y$, but there is no constant C to make $|x-y| \leq C\rho(x,y)$ holds. These two metrics are not equivalent.

Now we define continuity in a metric space. First of all, recall that there are two ways to describe it, namely, the behavior of sequences and the ε-δ formulation. Specifically, the function f is continuous at $x \in E$ if for every sequence $\{x_n\} \subset E$ satisfying $\lim_{n\to\infty} x_n = x$, $\lim_{n\to\infty} f(x_n) = f(x)$. Equivalently, for every $\varepsilon > 0$, there exists some $\delta > 0$ such that $|f(y) - f(x)| < \varepsilon$ whenever $y \in E, |y - x| < \delta$. Both definition can be formulated on a metric space. We will adapt the sequence approach. Let (X, d) and (Y, ρ) be two metric spaces and $f : (X, d) \to (Y, \rho)$. Let $x \in X$. We call f is **continuous at** x if $f(x_n) \to f(x)$ in (Y, ρ) whenever $x_n \to x$ in (X, d). It is **continuous** on a set $E \subset X$ if it is continuous at every point of E.

First we show the sequence formulation is equivalent to the ε-δ-formulation.

Proposition 2.2. *Let f be a mapping from (X, d) to (Y, ρ) and $x_0 \in X$. Then f is continuous at x_0 if and only if for every $\varepsilon > 0$, there exists some $\delta > 0$ such that $\rho(f(x), f(x_0)) < \varepsilon$ for all $x, d(x, x_0) < \delta$.*

Proof. $\Leftarrow$) Let ε be given and δ is chosen accordingly. For any $\{x_n\} \to x_0$, given $\delta > 0$, there exists some n_0 such that $d(x_n, x_0) < \delta$, $\forall n \geq n_0$. It follows that $\rho(f(x_n), f(x_0)) < \varepsilon$ for all $n \geq n_0$, so f is continuous at x_0.

$\Rightarrow$) Suppose that the implication is not valid. There exist some $\varepsilon_0 > 0$ and $\{x_k\} \in X$ satisfying $d(f(x_k), f(x_0)) \geq \varepsilon_0$ and $d(x_k, x_0) < 1/k$. However, the second condition tells us that $\{x_k\} \to x_0$, so by the continuity at x_0 one should have $d(f(x_k), f(x_0)) \to 0$, contradiction holds. □

Shortly we will face a third formulation, that is, by open/closed sets to describe continuity in a metric space.

As usual, continuity of functions is closed under compositions of functions.

Proposition 2.3. *Let $f : (X, d) \to (Y, \rho)$ and $g : (Y, \rho) \to (Z, m)$ be given.*

(a) If f is continuous at x and g is continuous at $f(x)$, then $g \circ f : (X, d) \to (Z, m)$ is continuous at x.

(b) If f is continuous in X and g is continuous in Y, then $g \circ f$ is continuous in X.

Proof. It suffices to prove (a). Let $x_n \to x$. Then $f(x_n) \to f(x)$ as f is continuous at x. Then $(g \circ f)(x_n) = g(f(x_n)) \to g(f(x)) = (g \circ f)(x)$ as g is continuous at $f(x)$. □

To end this section, we consider the following question: Are there any continuous functions in a metric space? Of course, constant functions are obviously continuous. But we want some non-trivial ones. It turns out that many continuous functions can be constructed from the distance function. (What else?) Indeed, let A be a non-empty set in (X, d). We define the distance from a point x to A by

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

We claim:

$$|\rho_A(x) - \rho_A(y)| \leq d(x, y), \quad \forall x, y \in X.$$

It shows that ρ_A is not only continuous but also "Lipschitz continuous". To prove the claim, let $\varepsilon > 0$, we pick $z \in A$ such that $\rho_A(y) + \varepsilon > d(y, z)$. The

$$\begin{aligned} \rho_A(x) &\leq d(x, z) \\ &\leq d(x, y) + d(y, z) \\ &\leq d(x, y) + \rho_A(y) + \varepsilon \ . \end{aligned}$$

Since ε could be any positive number, we conclude that $\rho_A(x) - \rho_A(y) \leq d(x, y)$, and the claim follows after noting that the roles of x and y are exchangeable.

It is convenient to introduce the notations:

$$d(x, F) = \inf\{d(x, y) : \ y \in F\} \ ,$$

and

$$d(E, F) = \inf\{d(x, y) : x \in E, y \in F\} \ .$$

The collection of continuous functions is large in the sense that they separate points. Indeed, let x_1 and x_2 be two distinct points in X. The continuous function $\rho_{\{x_1\}}$ satisfies $\rho_{\{x_1\}}(x_1) = 0$ and $\rho_{\{x_1\}}(x_2) > 0$. Further related results can be found in the exercise.

2.3 Open and Closed Sets

The existence of a metric on a set enables us to talk about the convergence of a sequence and continuity of a map. It turns out that, in order to define continuity, a structure less stringent then a metric structure is needed. It suffices the set be endowed with a topological structure. In a word, a metric induces a topological structure on the set but not every topological structure comes from a metric. In a topological space, continuity can no longer be defined via the convergence of sequences. Instead one uses the notion of open and closed sets in the space. As warm up for topology we discuss how to use the language of open/closed sets to describe the convergence of sequences and the continuity of functions in this section.

First the definition. Let (X, d) be a metric space. A set $G \subset X$ is called an **open set** if for each $x \in G$, there exists some ρ such that $B_\rho(x) \subset G$. The number ρ may vary depending on x. We also define the empty set ϕ to be an open set. Roughly speaking, an open set is a subset in which every point is surrounded by points in the set.

Proposition 2.4. *Let (X, d) be a metric space. We have*

(a) X and ϕ are open sets.

(b) If $\bigcup_{\alpha\in\mathcal{A}} G_\alpha$ is an open set provided that all G_α, $\alpha \in \mathcal{A}$, are open where $\mathcal{A}$ is an arbitrary index set.

(c) If $G_1, \ldots, G_N$ are open sets, then $\bigcap_{j=1}^N G_j$ is an open set.

Note the union in (b) of this proposition is over an arbitrary collection of sets while the intersection in (c) is a finite one.

Proof. (a) Obvious.

(b) Let $x \in \bigcup_{\alpha\in\mathcal{A}} G_\alpha$. There exists some α_1 such that $x \in G_{\alpha_1}$. As G_{α_1} is open, there is some $B_\rho(x) \subset G_{\alpha_1}$. But then $B_\rho(x) \subset \bigcup_{\alpha\in\mathcal{A}} G_\alpha$, so $\bigcup_{\alpha\in\mathcal{A}} G_\alpha$ is open.

(c) When $\bigcap_{j=1}^N G_j$ is empty, it is open by definition. On the other hand, let $x \in \bigcap_{j=1}^N G_j$. For each j, there exists $B_{\rho_j}(x) \subset G_j$. Let $\rho = \min\{\rho_1, \ldots, \rho_N\}$. Then $B_\rho(x) \subset \bigcap_{j=1}^N G_j$, so $\bigcap_{j=1}^N G_j$ is open. □

The complement of an open set is called a **closed set**. Taking the complement of Proposition 2.4, we have

Proposition 2.5. *Let (X, d) be a metric space. We have*

(a) X and ϕ are closed sets.

(b) If F_α, $\alpha \in \mathcal{A}$, are closed sets, then $\bigcap_{\alpha\in\mathcal{A}} F_\alpha$ is a closed set.

(c) If $F_1, \ldots, F_N$ are closed sets, then $\bigcup_{j=1}^N F_j$ is a closed set.

Note that X and ϕ are both open and closed. The terminology of a closed set will become evident soon.

Example 2.10. Every ball in a metric space is an open set. Let $B_r(x)$ be a ball and $y \in B_r(x)$. We claim that $B_\rho(y) \subset B_r(x)$ where $\rho = r - d(y,x) > 0$. For, if $z \in B_\rho(y)$,

$$\begin{aligned} d(z,x) &\leq d(z,y) + d(y,x) \\ &< \rho + d(y,x) \\ &= r \ , \end{aligned}$$

by the triangle inequality, so $z \in B_r(x)$ and $B_\rho(y) \subset B_r(x)$ holds. Next, the set $E = \{y \in X : \ d(y,x) > r\}$ for fixed x and $r \geq 0$ is an open set. For, let $y \in E$, $d(y,x) > r$. We claim $B_\rho(y) \subset E, \ \rho = d(y,x) - r > 0,$. For, letting $z \in B_\rho(y)$,

$$\begin{aligned} d(z,x) &\geq d(y,x) - d(y,z) \\ &> d(y,x) - \rho \\ &= r, \end{aligned}$$

shows that $B_\rho(y) \subset E$, hence E is open. Finally, consider $F = \{x \in X : \ d(x,z) = r > 0\}$ where z and r are fixed. Observing that F is the complement of the two open sets $B_r(z)$ and $\{x \in X : \ d(x,z) > r\}$, we conclude that F is a closed set.

Example 2.11. In the real line every open interval (a,b), $-\infty \leq a \leq b \leq \infty$, is an open set. Other intervals such as $[a,b)$, $[a,b]$, $(a,b]$, $a, b \in \mathbb{R}$, are not open. It can be shown that every open set G in $\mathbb{R}$ can be written as a disjoint union of open intervals. Letting $(a_n, b_n) = (-1/n, 1/n)$,

$$\bigcap_{n=1}^{\infty} \left(-\frac{1}{n}, \frac{1}{n} \right) = \{0\}$$

is not open. It shows that Proposition 2.3(c) does not hold when the intersection is over infinite many sets. On the other hand, $[a,b]$ is a closed set since it is the complement of the open set $\mathbb{R} \setminus (-\infty, a) \cup (a, \infty)$. Setting $a = b$, $\{a\}$ is closed, that is, a single point is always a closed set. Finally, some sets we encounter often are neither open nor closed. Take the set of all rational numbers as example, as every open interval containing a rational number also contains an irrational number, we see that $\mathbb{Q}$ is not open. The same reasoning shows that the set of all irrational numbers is not open, hence $\mathbb{Q}$ is also not a closed set.

Example 2.12. When we studied multiple integrals in Advanced Calculus II, we encountered many domains or regions as the domain of integration. These domains are bounded by nice curves in the plane or by nice surfaces in the space. Without counting its boundary points, the interior of these domains form open sets. The exterior of these domains, again excluding the boundary points, are also open sets. Therefore, the set consisting of all boundary points is a closed set as it is the complement of the union of two open sets, namely, the interior and exterior of the domain.

Example 2.13. Consider the set $E = \{f \in C[a,b] : \ f(x) > 0, \ \forall x \in [a,b]\}$ in $C[a,b]$. We claim that it is open. For $f \in E$, it is positive everywhere on the closed, bounded interval $[a,b]$, hence according to extremal value theorem it attains its minimum at some x_0. It follows that $f(x) \geq m \equiv f(x_0) > 0$. Letting $r = m/2$, for $g \in B_r(f), \ d_\infty(g,f) < r = m/2$ implies

$$\begin{aligned} g(x) &\geq f(x) - |g(x) - f(x)| \\ &> m - \frac{m}{2} \\ &= \frac{m}{2} > 0, \end{aligned}$$

for all $x \in [a,b]$, hence $g \in E$ which implies $B_r(f) \subset E$, E is open. Likewise, sets like $\{f : \ f(x) > \alpha, \ \forall x\}, \{f : \ f(x) < \alpha, \ \forall x\}$ where α is a fixed number. On the other hand, by taking complements of these open sets, we see that the sets $\{f : \ f(x) \geq \alpha, \ \forall x\}, \{f : f(x) \leq \alpha, \ \forall x\}$ are closed.

Example 2.14. Consider the extreme case where the space X is endowed with the discrete metric. We claim that every set is open and closed. Clearly, it suffices to show that every singleton set $\{x\}$ is open. But, this is obvious because the ball $B_{1/2}(x) = \{x\} \subset \{x\}$.

We now use open sets to describe the convergence of sequences.

Proposition 2.6. *Let (X,d) be a metric space. A sequence $\{x_n\}$ converges to x if and only if for each open G containing x, there exists n_0 such that $x_n \in G$ for all $n \geq n_0$.*

Proof. Let G be an open set containing x. According to the definition of an open set, we can find $B_\varepsilon(x) \subset G$. It follows that there exists n_0 such that $d(x_n, x) < \varepsilon$ for all $n \geq n_0$, i.e., $x_n \in B_\varepsilon(x) \subset G$ for all $n \geq n_0$. Conversely, taking $G = B_\varepsilon(x)$, we see that $x_n \to x$. □

From this proposition we deduce the following result which explains better the terminology of a closed set.

Proposition 2.7. *The set A is a closed set in (X,d) if and only if whenever $\{x_n\} \subset A$ and $x_n \to x$ as $n \to \infty$ implies that x belongs to A.*

Proof. $\Rightarrow$). Assume on the contrary that x does not belong to A. As $X \setminus A$ is an open set, we can find a ball $B_\varepsilon(x) \subset X \setminus A$. However, as $x_n \to x$, there exists some n_0 such that $x_n \in B_\varepsilon(x)$ for all $n \geq n_0$, contradicting the fact that $x_n \in A$.

$\Leftarrow$). If $X \setminus A$ is not open, say, we could find a point $x \in X \setminus A$ such that $B_{1/n}(x) \bigcap A \neq \phi$ for all n. Pick $x_n \in B_{1/n}(x) \bigcap A$ to form a sequence $\{x_n\}$. Clearly $\{x_n\}$ converges to x. By assumption, $x \in A$, contradiction holds. Hence $X \setminus A$ must be open.

□

Now we use open sets to describe continuity.

Proposition 2.8. *Let $f : (X, d) \to (Y, \rho)$.*

(a) f is continuous at x if and only if for every open set G containing $f(x)$, $f^{-1}(G)$ contains $B_\rho(x)$ for some $\rho > 0$.

(b) f is continuous in X if and only if for every open G in Y, $f^{-1}(G)$ is an open set in X.

These statements are still valid when "open" is replaced by "closed".

Proof. We consider (a) and leave (b) as an exercise.

$\Rightarrow$). Suppose there exists some open G such that $f^{-1}(G)$ does not contain $B_{1/n}(x)$ for all $n \geq 1$. Pick $x_n \in B_{1/n}(x)$, $x_n \notin f^{-1}(G)$. Then $x_n \to x$ but $f(x_n)$ does not converge to $f(x)$, contradicting the continuity of f.

$\Leftarrow$). Let $\{x_n\} \to x$ in X. Given any open set G containing $f(x)$, we can find $B_r(x) \subset f^{-1}(G)$. Thus, there exists n_0 such that $x_n \in B_r(x)$ for all $n \geq n_0$. It follows that $f(x_n) \in G$ for all $n \geq n_0$. By Proposition 2.7, f is continuous at x.

□

This proposition shows in particular that for a continuous function $F : (X, d) \to \mathbb{R}$, the sets $F^{-1}((a, b))$ are open and $F^{-1}([a, b]), -\infty \leq a < b \leq \infty$, are closed. This gives an effective way to determine whether a set is open or closed. Let us look at the following examples.

Example 2.15. Consider Example 2.11 again. In $\mathbb{R}^2$, the domains are obtained as the intersection of several curves. A curve may be described as the zero set of some function. A typical description of a domain would be like

$$D = \left\{(x, y) \in \mathbb{R}^2 : \ f(x) < y < g(x), \ x \in (a, b) \ \right\}$$

where f, g are continuous function from $\mathbb{R}^n$ to $\mathbb{R}$ satisfying $f(x) < g(y)$. If we let $A_1 = \{(x, y) : \ y > f(x)\}$, $A_2 = \{(x, y) : \ y < g(x)\}$ and $B = \{(x, y) : \ x \in (a, b), y \in (-\infty, \infty)\}$. It is not hard to see that A_1 and A_2 are open. In fact, let $F(x, y) = y - f(x)$ which is clearly continuous, so $A_1 = F^{-1}(0, \infty)$ is open. Similarly A_2 is open. (You may also verify it using the definition of an open set.) Since it is clear that B is open, $D = A_1 \cap A_2 \cap B$ is also open.

Some open sets are the regions bounded by closed curves. In this case, a single function is sufficient to define them. For instance, the region bounded by an ellipse is described as $\Omega = \{(x, y) \in \mathbb{R}^2 : \ x^2/a^2 + y^2/b^2 < 1\}$. As $F(x, y) = x^2/a^2 + y^2/b^2$ is obviously a continuous function, and by $\Omega = F^{-1}(-\infty, 1)$, it is an open set.

Example 2.16. Let us consider an example in $C[a,b]$. In Example 2.12 we showed that the set $P = \{f : \ f(x) > 0, \ \forall x \in [a,b]\}$ is open in $C[a,b]$. We claim that the set $D = \{f : \ f^2(x) - \sin f(x) > 0\}$ is also an open set. For, the map $\Phi(f) = f^2 - \sin f$ defines a map from $C[a,b]$ to itself (check it). Now we realize that $D = \{f : \ (\Phi \circ f)(x)) > 0 \ , \ \forall x \in [a,b]\}$. From $D = \Phi^{-1}P$ we conclude that D is open too.

2.4 Points in a Metric Space

We describe some further useful notions associated to sets in a metric space.

Let E be a set in (X,d). A point x is called a **boundary point** of E if $G \cap E$ and $G \setminus E$ are non-empty for every open set G containing x. Of course, it suffices to take G of the form $B_\varepsilon(x)$ for all small ε or $\varepsilon = 1/n, \ n \geq 1$. We denote the boundary of E by ∂E. The **closure** of E, denoted by $\overline{E}$, is defined to be $E \cup \partial E$. Clearly $\partial E = \partial(X \setminus E)$. The boundary of the ball $B_r(x)$ in $\mathbb{R}^n$ is the sphere $S_r(x) = \{y \in \mathbb{R}^n : \ d_2(y,x) = r\}$. Hence, the closed ball $\overline{B_r(x)}$ is given by $B_r(x) \bigcup S_r(x)$, which is precisely the closure of $B_r(x)$.

Example 2.17. Let $E = [0,1) \times [0,1) \subset \mathbb{R}^2$. It is easy to see that $\partial E = [0,1] \times \{0,1\} \cup \{0,1\} \times [0,1]$. Thus some points in ∂E belong to E and some do not. The closure of E, $\overline{E}$, is equal to $[0,1] \times [0,1]$.

It can be seen from definition that the boundary of the empty set is the empty set. Also, the boundary of a set is always a closed set. For, let $\{x_n\}$ be a sequence in ∂E converging to some x. For any ball $B_r(x)$, we can find some x_n in it, so the ball $B_\rho(x_n), \ \rho = r - d(x_n,x) > 0$, is contained in $B_r(x)$. As $x_n \in \partial E$, $B_\rho(x_n)$ has non-empty intersection with E and $X \setminus E$, so does $B_r(x)$ and $x \in \partial E$ too. The following proposition characterizes the closure of a set as the smallest closed set containing this set.

Proposition 2.9. *Let E be a set in (X,d). We have*

(a) $x \in \overline{E}$ if and only if $B_r(x) \cap E \neq \phi$ for all $r > 0$.

(b) $A \subset B$ implies $\overline{A} \subset \overline{B}$.

(c) $\overline{E}$ is a closed set.

(d) $\overline{E} = \bigcap\{C : \ C$ is a closed set containing $E\}$.

Proof. (a) If $x \in E$, then x is always contained in $B_r(x) \cap E$ for all r. On the other hand, if $x \in \partial E$, by definition $B_r(x) \cap E \neq \phi$ for all r too. Conversely, it is trivial when $x \in E$. When x does not belong to E but $B_r(x) \cap E \neq \phi$, then $x \in \partial E \subset \overline{E}$, done.

(b) Let $x \in \overline{A}$. By (a) $B_r(x) \cap A \neq \phi$ for all r. But as $A \subset B$, $B_r(x) \cap B \neq \phi$ for all r. By (a), $x \in \overline{B}$.

(c) Let $x_n \in \overline{E}$ and $x_n \to x$. We need to show $x \in \overline{E}$. If not true, x neither belong to E nor ∂E. So there is some $B_r(x)$ disjoint from E, contradicting $x_n \to x$.

(d) Denote the right hand side by F. It is a closed set. By (c) $\overline{E}$ is a closed set containing E, so $F \subset \overline{E}$ already. On the other hand, for any closed C satisfying $E \subset C$, (b) implies $\overline{E} \subset \overline{C} = C$, so $\overline{E} \subset \bigcap C = F$.

□

A point x is called an **interior point** of E if there exists an open set G such that $x \in G \subset E$. It can be shown that all interior points of E form an open set call the **interior** of E, denoted by E^o. It is not hard to see that $E^0 = E \setminus \partial E$. The interior of a set is related to its closure by the following relation: $E^o = X \setminus \big(\overline{X \setminus E}\big)$. Using this relation, one can show that the interior of a set is the largest open set sitting inside E. More precisely, $G \subset E^0$ whenever G is an open set in E.

Example 2.18. Consider the set of all rational numbers E in $[0, 1]$. It has no interior point since there are irrational numbers in every open interval containing a rational number, so I^o is the empty set. On the other hand, since every open interval contains some rational numbers, the closure of E, $\overline{E}$, is $[0, 1]$. It shows the interior and closure of a set could be very different.

Example 2.19. In Example 2.11 we consider domains in $\mathbb{R}^2$ bounded by several continuous curves. Let D be such a domain and the curves bounding it be S. It is routine to verify that $\partial D = S$, that is, the set of all boundary points of D is precisely the S and the closure of D, $\overline{D}$, is $D \cup S$. The interior of $\overline{D}$ is D.

Example 2.20. For any two sets E and F in the same space, it is not hard to show $\overline{E \cup F} = \overline{E} \cup \overline{F}$. But $(E \cup F)^o$ may not always equal to $E^o \cup F^o$. As an extreme case, take $E = \mathbb{Q}$ and $\mathbb{I}$ in $\mathbb{R}$. We have $(\mathbb{Q} \cup \mathbb{I})^0 = \mathbb{R}^0 = \mathbb{R}$, but $\mathbb{Q}^o \cup \mathbb{I}^o = \phi \cup \phi = \phi$. In general, we only have $E^o \cup F^0 \subset (E \cup F)^o$.

Example 2.21. Let

$$S = \{f \in C[0, 1]) : \ 1 < f(x) \leq 5, \quad x \in [0, 1]\}.$$

We have

$$\overline{S} = \{f \in C[0, 1] : \ 1 \leq f(x) \leq 5, \quad x \in [0, 1]\} \ .$$

For, denote this set by A. As we have $A = f^{-1}([1, 5]$ where f is continuous, A is a closed set containing S. On the other hand, let $f \in A$, the functions $f_n(x) = \max\{f(x), 1 + 1/n\} \in S, n \geq 1$, and $f_n \to f$ in sup-norm. (Recall the fact that $\max\{f, g\}$ is continuous when

f and g are continuous.) Hence every function in A is the limit of some sequence in S, so A is contained in any closed set containing S. We conclude that A is the smallest closed set containing S, that is, $A = \overline{S}$. On the other hand, the interior of S is given by $S^o = \{f \in C[0,1] : \ 1 < f(x) < 5, x \in [0,1]\}$. Denoting this set by B, from $B = f^{-1}((1,2))$ we see that B is open set in S. On the other hand, if g is an interior point of S, there is some $\delta > 0$ such that $B_\delta(g) \subset S$. In other words, $1 < h(x) \leq 5$ for all $h, \|h - g\|_\infty < \delta$. In particular, $g(x) < h(x) - \delta \leq 5$ on $[0,1]$. So g belongs to B, that is, B is the largest open set in S.

2.5 Elementary Inequalities for Functions.

We start with the Young's Inequality covered in MATH2060.

Young's Inequality. *For $a, b > 0$ and $p > 1$,*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \ , \quad \frac{1}{p} + \frac{1}{q} = 1,$$

and equality sign holds if and only if $a^p = b^q$.

The number q is called the conjugate of p. Note that $q > 1$. The proof of this inequality is left to you. Basically, we use calculus to show the function

$$\varphi(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab \ ,$$

where b is fixed, has a unique minimum over $(0, \infty)$ at the point $a = b^{1/(1-p)}$, that is, $a^p = b^q$.

Theorem 2.10. ***(Hölder's Inequality.)*** *Let $f, g \in R[a,b]$ and $p > 1$. Then*

$$\int_a^b |f(x)g(x)|\, dx \leq \left(\int_a^b |f(x)|^p \, dx \right)^{1/p} \left(\int_a^b |g(x)|^q \, dx \right)^{1/q} \ , \quad q \text{ is conjugate to } p \ .$$

Equality sign in this inequality holds if and only if either (a) f or g vanish almost everywhere, or (b) there is some positive λ such that $|g|^q = \lambda |f|^p$ almost everywhere.

Proof. Assume $\|f\|_p$ and $\|g\|_q$ are positive, otherwise the inequality holds trivially.

For $\varepsilon > 0$ to be chosen, by Young's Inequality,

$$|f(x)g(x)| = |\varepsilon f(x) \varepsilon^{-1} g(x)| \leq \frac{\varepsilon^p |f(x)|^p}{p} + \frac{\varepsilon^{-q} |g(x)|^q}{q} \ .$$

Integrate this inequality to get

$$\int_a^b |f(x)g(x)|\,dx \le \frac{\varepsilon^p}{p}\int_a^b |f(x)|^p\,dx + \frac{\varepsilon^{-q}}{q}\int_a^b |g(x)|^q\,dx\ . \tag{2.1}$$

We now choose ε so that

$$\varepsilon^p \int_a^b |f(x)|^p\,dx = \varepsilon^{-q}\int_a^b |g(x)|^q\,dx\ ,$$

that is,

$$\varepsilon^{p+q} = \left(\int_a^b |g(x)|^q\,dx\right)\left(\int_a^b |f(x)|^p\,dx\right)^{-1}\ .$$

Using this epsilon to plug in (1), the right hand side becomes

$$\frac{\varepsilon^p}{p}\int_a^b |f(x)|^p\,dx + \frac{\varepsilon^{-q}}{q}\int_a^b |g(x)|^q\,dx\ = \left(\int_a^b |f(x)|^p\,dx\right)^{1/p}\left(\int_a^b |g(x)|^q\,dx\right)^{1/q}\ . \tag{2.2}$$

The Hölder's Inequality follows. □

To characterize the inequality sign in this inequality, observe case (a) is obvious so let us assume $\|f\|_p, \|g\|_q$ are both positive, so $|f(x)|$ and $|g(x)|$ are positive almost everywhere. From (2.1) and (2.2) we see that the inequality sign in (1) becomes equality, that is,

$$\int_a^b \left(\frac{\varepsilon^p|f(x)|^p}{p} + \frac{\varepsilon^{-q}|g(x)|^q}{q} - |f(x)g(x)|\right)\,dx = 0\ .$$

The integrand is a non-negative function by Young's Inequality. The vanishing of this integral implies that the integrand must vanish almost everywhere, that is,

$$\frac{\varepsilon^p|f(x)|^p}{p} + \frac{\varepsilon^{-q}g(x)^q}{q} - |f(x)g(x)| = 0\ a.e.\ .$$

By the equality sign condition in Young's Inequality, we conclude that

$$\varepsilon^p|f(x)|^p = \varepsilon^{-q}g(x)^q\ a.e.,$$

that is, $|g(x)|^q = \lambda|f(x)|^p$ almost everywhere where $\lambda = \varepsilon^{-p-q}$.

Remarks. (a) We have used the following proposition proved in Chapter 1: For $f \in R[a,b]$,

$$\int_a^b |f|\,dx = 0 \text{ if and only if } f = 0\ a.e.\ .$$

We also point out, when $f \in C[a,b]$,

$$\int_a^b |f|\,dx = 0 \text{ if and only if } f = 0 \text{ everywhere}\ .$$

(b) When f and g in Hölder's Inequality are continuous, almost everywhere in the characterization of equality sign becomes everywhere.

(c) The inequality still holds in the limiting cases. In fact, when $g \in C[a,b]$ and $p = 1$, we have

$$\int_a^b |f(x)g(x)|\,dx \le \int_a^b |f(x)|\,dx \|g\|_\infty \ .$$

When $f \in C[a,b]$ and $p = \infty$,

$$\int_a^b |f(x)g(x)|\,dx \le \|f\|_\infty \int_a^b |g(x)|\,dx \ .$$

But there is no clean characterization of the equality sign.

Theorem 2.11. *(**Minkowski's Inequality.**) For $f, g \in R[a,b]$ and $p > 1$,*

$$\|f+g\|_p \le \|f\|_p + \|g\|_p \ .$$

Equality sign in this inequality holds if and only if either (a) f or g vanishes almost everywhere, or (b) $\|f\|_p, \|g\|_p > 0$ and there is some positive λ such that $g(x) = \lambda f(x)$ almost everywhere.

Proof. Using

$$|f+g|^p = |f+g|^{p-1}|f+g| \le |f+g|^{p-1}|f| + |f+g|^{p-1}|g| \ ,$$

integrate both sides to get

$$\int_a^b |f+g|^p\,dx \le \int_a^b |f+g|^{p-1}|f|\,dx + \int_a^b |f+g|^{p-1}|g|\,dx \ . \tag{2.3}$$

Applying the Hölder's Inequality to the two integrals on the right separately, we have

$$\int_a^b |f+g||f|\,dx \le \left(\int_a^b |f+g|^q\,dx\right)^{1/q} \left(\int_a^b |f|^p\,dx\right)^{1/p} \ ,$$

and

$$\int_a^b |f+g||g|\,dx \le \left(\int_a^b |f+g|^q\,dx\right)^{1/q} \left(\int_a^b |g|^p\,dx\right)^{1/p} \ ,$$

where q is conjugate to p. Putting this back to (2.3), we obtain the desired inequality after some simplifications. □

The equality case, in principle, could be treated as in the Hölder's case. It is easy to get $|g(x)|^q = \lambda|f(x)|^p$ almost everywhere, but rather tedious (or need to use Lebsegue integral) to get $f(x)^p = \lambda g(x)^q$. Luckily, this property has no consequence in our later

development.

Comments on Chapter 2. A topology on a set X is a collection of sets τ consisting the empty set and X itself which is closed under arbitrary union and finite intersection. Each set in τ is called an open set. The pair (X, τ) is called a topological space. From Proposition 2.2 we see that the collection of all open sets in a metric space (X, d) forms a topology on X. This is the topological space induced by the metric. Metric spaces constitute a large class of topological spaces, but not every topological space comes from a metric. However, from the discussions in Section 3 we know that continuity can be defined solely in terms of open sets. It follows that continuity can be defined for topological spaces, and this is crucial for many further developments. In the past, metric spaces were covered in Introductory Topology. Feeling that the notion of a metric space should be learned by every math major, we move it here.

Although the metric space is a standard topic, I found it difficult to fix upon a single reference book. Rudin's Principles covers some metric spaces, but his attention is mainly on the Euclidean space. There are many text books or lecture notes available in the internet with more or less the same content. Here I simply list the old but very readable book by E.T. Copson, Metric Spaces, as the main reference.