MATHEMATICS 3103 (Functional Analysis)
YEAR 2012–2013, TERM 2

HANDOUT #1: WHAT IS FUNCTIONAL ANALYSIS?

Elementary analysis mostly studies real-valued (or complex-valued) functions on the real line $\mathbb{R}$ or on n-dimensional space $\mathbb{R}^n$. Functional analysis, by contrast, shifts the point of view: we collect all the functions of a given class (for instance, all bounded continuous functions) into a *space of functions*, and we study that space (and operations on it) as an object in its own right. Since spaces of functions are nearly always infinite-dimensional, we are led to study analysis on infinite-dimensional vector spaces, of which the most important cases are Banach spaces and Hilbert spaces.

Before we get down to the detailed study of functional analysis, here are two examples that show how functional-analysis ideas arise already in elementary analysis:

Ordinary differential equations. Recall that we can solve a first-order linear inhomogeneous ordinary differential equation (ODE)

$$\frac{dy}{dt} + p(t)y = q(t) \tag{1.1}$$

[where $p(t)$ and $q(t)$ are given functions and $y(t)$ is the unknown function] by the method of integrating factors. Now, what about a second-order linear inhomogeneous ODE

$$\frac{d^2y}{dt^2} + p_1(t)\frac{dy}{dt} + p_0(t)y = q(t) \; ? \tag{1.2}$$

In general this is hard, but in case of *constant coefficients*, i.e.

$$\frac{d^2y}{dt^2} + c_1\frac{dy}{dt} + c_0y = q(t) \; , \tag{1.3}$$

we can factor the equation into a pair of first-order ODEs, which can then be solved in succession. You probably recall the method: let α and β be the roots of the quadratic polynomial $\lambda^2 + c_1\lambda + c_0$, so that

$$\lambda^2 + c_1\lambda + c_0 = (\lambda - \alpha)(\lambda - \beta) \; . \tag{1.4}$$

Then we rewrite (1.3) in the form

$$\Big(\frac{d^2}{dt^2} + c_1\frac{d}{dt} + c_0\Big)y = q(t) \; , \tag{1.5}$$

and we factor the differential operator as

$$\frac{d^2}{dt^2} + c_1\frac{d}{dt} + c_0 = \Big(\frac{d}{dt} - \alpha\Big)\Big(\frac{d}{dt} - \beta\Big) \; , \tag{1.6}$$

so that the equation (1.3) becomes

$$\Big(\frac{d}{dt} - \alpha\Big)\Big(\frac{d}{dt} - \beta\Big)y \;=\; q(t)\;. \tag{1.7}$$

If we define

$$z \;=\; \Big(\frac{d}{dt} - \beta\Big)y\;, \tag{1.8}$$

we can rewrite the second-order equation (1.7) as the pair of first-order equations

$$\Big(\frac{d}{dt} - \alpha\Big)z \;=\; q(t) \tag{1.9a}$$

$$\Big(\frac{d}{dt} - \beta\Big)y \;=\; z \tag{1.9b}$$

So we can first solve (1.9a) for the unknown function $z(t)$, and then solve (1.9b) for the unknown function $y(t)$.

Now, what have we done here? In particular, what are the objects

$$\frac{d}{dt}, \quad \frac{d^2}{dt^2}, \quad \frac{d}{dt} - \alpha, \quad \text{etc?} \tag{1.10}$$

The answer is that they are *linear operators* on a *space of functions* (which map it into another, not necessarily the same, space of functions). We have some choices in how we make this precise. For instance, let $C^k(a,b)$ be the vector space of real-valued functions on the interval $(a,b) \subseteq \mathbb{R}$ that are k times continuously differentiable (for $k = 0$ this is just the space of continuous functions); and let $C^\infty(a,b)$ be the vector space of real-valued functions on the interval $(a,b) \subseteq \mathbb{R}$ that are infinitely differentiable. Then d/dt can be considered as a linear operator mapping $C^k(a,b)$ into $C^{k-1}(a,b)$ for any $k \geq 1$, or as a linear operator mapping $C^\infty(a,b)$ into itself. Likewise, d^2/dt^2 can be considered as a linear operator mapping $C^k(a,b)$ into $C^{k-2}(a,b)$ for any $k \geq 2$, or as a linear operator mapping $C^\infty(a,b)$ into itself.

Note that all these spaces of functions are *infinite-dimensional*. (Why? You should supply a proof.) So we are inexorably led to study analysis on infinite-dimensional vector spaces. Furthermore, we see that among the important objects are *linear operators* that map one infinite-dimensional vector space to another.

Remark: Differential operators have the unfortunate property that they *reduce* the "smoothness" of a function. As a result, they do not map any of the spaces C^k into itself; rather, they map C^k into a *larger* space such as C^{k-1} or C^{k-2}. If we insist on having a space that is mapped into itself, we have to work with C^∞, which turns out to be harder to handle than the spaces C^k (it is a "Fréchet space" rather than a "Banach space"). On the other hand, for many purposes we *want* to map a space into itself. One way of solving this dilemma is to rewrite the differential equation (together with its initial conditions) as an *integral equation*, and then apply methods of functional analysis to this integral equation. Since integral operators *increase* the smoothness of a function — for instance, the indefinite integral of a C^k function is C^{k+1} — they *do* map the spaces C^k into themselves (in fact, into proper subsets of themselves), and the functional-analytic treatment of integral equations is straightforward.

Fourier analysis. Let f be a continuous function (either real-valued or complex-valued) defined on the interval $[-\pi, \pi]$. Then you will recall that its Fourier coefficients $\{c_n\}_{n=-\infty}^{\infty}$ are complex numbers defined by

$$c_n \;=\; \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-int} \, f(t) \, dt \tag{1.11}$$

for $n = \ldots, -2, -1, 0, 1, 2, \ldots$. It is easy to see that the sequence $\{c_n\}$ is *bounded*. (Why? You should supply a proof.) Therefore, the Fourier transform for continuous functions on $[-\pi, \pi]$ can be considered as a linear operator $\mathcal{F}$ mapping the space $\mathcal{C}[-\pi, \pi]$ of continuous complex-valued functions on the interval $[-\pi, \pi]$ into the space $\ell^\infty(\mathbb{Z})$ of bounded doubly infinite sequences of complex numbers. (I will explain the funny notation ℓ^∞ next week.)

Once again, both of these spaces are *infinite-dimensional*. (Why? You should supply a proof.) So we are again led to study analysis on infinite-dimensional vector spaces. And we see once again the key role played by *linear operators*.

Similar reasoning applies to the Fourier transform on the whole real line $\mathbb{R}$. For instance, let f be a continuous function (either real-valued or complex-valued) defined on $\mathbb{R}$ that is absolutely integrable, i.e. satisfies

$$\int_{-\infty}^{\infty} |f(t)| \, dt \;<\; \infty \, . \tag{1.12}$$

(Why didn't we need to impose such a condition when we were working on the closed bounded interval $[-\pi, \pi]$? How would things be different if we had chosen instead to work on the open interval $(-\pi, \pi)$?) Then its Fourier transform is the function $\widehat{f}(\omega)$ defined for $\omega \in \mathbb{R}$ by

$$\widehat{f}(\omega) \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\omega t} \, f(t) \, dt \, . \tag{1.13}$$

Then it is easy to see that the function $\widehat{f}$ is well-defined (why?) and bounded (why?); and with some work it can also be shown that $\widehat{f}$ is continuous. Therefore, the Fourier transform for functions on $\mathbb{R}$ can be considered as a linear operator $\mathcal{F}$ mapping the space $\mathcal{C}(\mathbb{R}) \cap L^1(\mathbb{R})$ of complex-valued functions that are continuous and absolutely integrable on $\mathbb{R}$ into the space $\mathcal{C}(\mathbb{R}) \cap L^\infty(\mathbb{R})$ of complex-valued functions that are continuous and bounded on $\mathbb{R}$. (I will explain the funny notations L^1 and L^∞ next week.) Once again, these spaces are infinite-dimensional.

Remark: As always in applying functional analysis, we have some choice concerning the space of functions where we take our operator to act. I chose to make the Fourier transform act on the spaces $\mathcal{C}[-\pi, \pi]$ and $\mathcal{C}(\mathbb{R}) \cap L^1(\mathbb{R})$ solely for illustrative purposes, so that the integral could be understood as the ordinary Riemann integral (together with the standard limit definition when one or both of the limits of integration is infinite). This is not, in fact, the cleanest way to understand the Fourier transform. The cleanest approach uses the spaces $L^1[-\pi, \pi]$ and $L^1(\mathbb{R})$ of functions that are *Lebesgue-measurable* (but *not* necessarily

continuous) and absolutely integrable, or else the spaces $L^2[-\pi, \pi]$ and $L^2(\mathbb{R})$ of functions that are Lebesgue-measurable and whose *squares* are absolutely integrable. But this requires an understanding of Lebesgue integration, which I am not assuming for this course. If you do know a bit about Lebesgue integration, then a few weeks from now (after we have studied the elementary theory of Hilbert spaces) you might want to look at Saxe, Chapter 4, Rynne and Youngson, Section 3.5 or Kreyszig, Sections 3.4 and 3.5.

To summarize this discussion, we can say roughly that

> **Elementary analysis** is the study of finite-dimensional vector spaces (i.e. $\mathbb{R}^n$) and maps between them (e.g. continuous functions from $\mathbb{R}^n$ to $\mathbb{R}^m$),

while

> **Functional analysis** is the extension of elementary analysis in which we consider also infinite-dimensional vector spaces (of certain kinds) and maps between them (e.g. continuous functions, but especially *linear* operators).

Now, the space $\mathbb{R}^n$ possesses *two* structures that are relevant here, namely:

- a **vector-space structure**, which allows us to add elements of $\mathbb{R}^n$ and to multiply them by scalars; and
- a **convergence structure**, which allows us to say when (for instance) a sequence $x_1, x_2, \ldots$ of points in $\mathbb{R}^n$ converges to a point $x \in \mathbb{R}^n$. (For historical reasons this convergence structure is usually termed a **topological** structure.)

The vector-space structure of $\mathbb{R}^n$ is the subject of **linear algebra**, while the convergence structure of $\mathbb{R}^n$ is the subject of **elementary real analysis**. These two structures fit together nicely in the sense that

(a) if the sequence $x_1, x_2, \ldots$ converges to x and the sequence $y_1, y_2, \ldots$ converges to y, then the sequence $x_1 + y_1, x_2 + y_2, \ldots$ converges to $x + y$; and

(b) if the sequence $x_1, x_2, \ldots$ converges to x, and α is a real number, then the sequence $\alpha x_1, \alpha x_2, \ldots$ converges to αx.

The study of infinite-dimensional vector spaces will fit this same pattern: each space under study will have both a vector-space structure and a convergence structure; we must study both structures, and we must make sure that they fit together nicely in the sense of (a) and (b) above. Now, the *algebraic* part of this study is not very different from what you already learned in your linear-algebra course: indeed, many of the results of linear algebra hold equally well for finite-dimensional or infinite-dimensional vector spaces. (The exceptions are, of course, results explicitly referring to bases, dimension, etc.) On the other hand, the *topological* behavior of infinite-dimensional spaces is quite different from that of finite-dimensional spaces, and much of this course will be devoted to studying precisely those differences.

Now, the most general setting for studying questions of convergence is the structure known as a **topological space**. So, should we start this course by studying topological spaces? Well, we could do so (and some courses in functional analysis do just that); but the concept of a topological space is rather abstract, and the behavior of general topological spaces can be rather pathological. Therefore, it makes sense to be more modest, and to begin by studying a subclass of topological spaces that is

(a) rich enough to include most (though not all) of the function spaces that are relevant for functional analysis,

and at the same time

(b) sufficiently simple so that convergence can be easily visualized and much (though not all) of the intuition from $\mathbb{R}^n$ can be carried over.

The **metric spaces** are a subclass of topological spaces that have these two properties. Most (though not all) of the important function spaces are metrizable (and indeed are Banach spaces); and the theory of metric spaces is quite a bit simpler than the general theory of topological spaces. So this is where we shall start.

You have already studied the elementary theory of metric spaces in Real Analysis (module 7102), so most of the rest of this handout should constitute review for you. You should make sure that you know this material well and are capable of filling in all the missing proofs. Please consult me as soon as possible if you have trouble with any of this material.

Metric spaces (and normed linear spaces)

Roughly speaking, a metric space is a space in which convergence is defined via a real-valued "distance function". Here is the precise definition:

Definition 1.1 *A* **metric space** (X, d) *is defined to be a set* X *together with a function* $d\colon X \times X \to \mathbb{R}$ *satisfying the following four conditions:*

(i) $d(x, y) \geq 0$ *for all* $x, y \in X$ *(***nonnegativity***);*

(ii) $d(x, y) = 0$ *if and only if* $x = y$ *(***nondegeneracy***);*

(iii) $d(x, y) = d(y, x)$ *for all* $x, y \in X$ *(***symmetry***);*

(iv) $d(x, z) \leq d(x, y) + d(y, z)$ *for all* $x, y, z \in X$ *(***triangle inequality***).*

(A function $d\colon X \times X \to \mathbb{R}$ *satisfying conditions (i)–(iv) is called a* **metric** *on* X*.)*

Conditions (i)–(iii) are usually trivial to check in any concrete example, while the triangle inequality (iv) may require more work.

I stress that the metric space is the *pair* (X, d), i.e. the set X *together with* the metric d. However, we shall often refer informally to "the metric space X" whenever it is understood from the context what the metric d is.

Most of the examples of metric spaces that we will consider are also *normed linear spaces* (or subspaces thereof) — and although we will begin the study of normed linear spaces in earnest about 2 weeks from now, it seems sensible to give the definition without delay, because norms are slightly easier to work with than general metrics and it would be silly to deprive ourselves of this convenience. Roughly speaking, a *norm* is a real-valued function that assigns a "length" to each vector. Here is the precise definition:

Definition 1.2 *Let X be a vector space over the field $\mathbb{R}$ of real numbers (or the field $\mathbb{C}$ of complex numbers). Then a* **norm** *on X is a function that assigns to each vector $x \in X$ a real number $\|x\|$, satisfying the following four conditions:*

(i) $\|x\| \geq 0$ *for all* $x \in X$ *(***nonnegativity***);*

(ii) $\|x\| = 0$ *if and only if* $x = 0$ *(***nondegeneracy***);*

(iii) $\|\lambda x\| = |\lambda| \, \|x\|$ *for all* $x \in X$ *and all* $\lambda \in \mathbb{R}$ *(or* $\mathbb{C}$*) (***homogeneity***);*

(iv) $\|x + y\| \leq \|x\| + \|y\|$ *for all* $x, y \in X$ *(***triangle inequality***).*

The pair $(X, \| \cdot \|)$ *consisting of a vector space X together with a norm $\| \cdot \|$ on it is called a* **normed linear space**.

I stress once again that the normed linear space is the *pair* $(X, \| \cdot \|)$. The same vector space X can be equipped with many different norms, and these give rise to *different* normed linear spaces. However, we shall often refer informally to "the normed linear space X" whenever it is understood from the context what the norm is.

The point now is that every normed linear space can be given the structure of a metric space in an obvious way, namely we define the metric by

$$d(x, y) \;=\; \|x - y\| \,. \tag{1.14}$$

Indeed, properties (i), (ii) and (iv) of the metric follow trivially from the corresponding properties of the norm, while property (iii) of the metric follows from the special case $\lambda = -1$ of property (iii) of the norm (why?).

Of course, the homogeneity property of the norm holds for all real (or complex) numbers λ, not just $\lambda = -1$, so normed linear spaces constitute a special (but important) *subclass* of metric vector spaces.

Example 1. Discrete metric spaces. Let X be any set, and define

$$d(x, y) \;=\; \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases} \tag{1.15}$$

It is easy to see that d is a metric on X (why?). It is called the **discrete metric**. As we shall see, it corresponds to a rather uninteresting "space of isolated points", in which a sequence $\{x_n\}$ can converge to x only if it is eventually *equal* to x (i.e. $x_n = x$ for all $n \geq$ some n_0).

Note that, even if X happens to be a vector space, this metric does *not* arise from a norm (why?).

Example 2. The real line with the usual norm. Let X be the set $\mathbb{R}$ of real numbers, considered as a one-dimensional vector space over the field of real numbers, and define

$$\|x\| \;=\; |x| \;. \tag{1.16}$$

It is easy to see that $\|\cdot\|$ is a norm on $\mathbb{R}$ (you should supply a proof), so $(\mathbb{R}, \|\cdot\|)$ is a normed linear space. The corresponding metric is of course

$$d(x,y) \;=\; |x-y| \;. \tag{1.17}$$

We call these the *usual norm* and the *usual metric* on $\mathbb{R}$.

Let us remark that exactly the same formula defines a norm on the complex numbers $\mathbb{C}$, considered as a one-dimensional vector space over the field of complex numbers.

Example 3. $\mathbb{R}^n$ with the ℓ^1 norm. Fix an integer $n \geq 1$, and let X be the space $\mathbb{R}^n$ of ordered n-tuples of real numbers (considered as a vector space over the field of real numbers). For $x = (x_1, x_2, \ldots, x_n)$, define

$$\|x\|_1 \;=\; \sum_{i=1}^{n} |x_i| \;. \tag{1.18}$$

It is easy to see that $\|\cdot\|_1$ is a norm on $\mathbb{R}^n$ (you should supply a proof). It is sometimes called the "Manhattan norm" on $\mathbb{R}^n$ (do you see why?). More commonly it is called the "ℓ^1 norm", for reasons to be discussed later when we study the ℓ^1 sequence space (Example 12 below).

Let us remark once again that exactly the same formula defines a norm on $\mathbb{C}^n$, considered as a vector space over the field of complex numbers.

Example 4. $\mathbb{R}^n$ with the ℓ^∞ norm. Consider again $\mathbb{R}^n$, but now with the norm

$$\|x\|_\infty \;=\; \max_{1 \leq i \leq n} |x_i| \;. \tag{1.19}$$

Once again you should verify that $\|\cdot\|_\infty$ is a norm on $\mathbb{R}^n$. It is called the "max norm" or "sup norm" or "uniform norm" or "ℓ^∞ norm"; the latter terminology will be clarified later when we study the ℓ^∞ sequence space (Example 11 below). Once again, exactly the same formula defines a norm on $\mathbb{C}^n$.

Example 5. $\mathbb{R}^n$ with the Euclidean norm. Consider again $\mathbb{R}^n$, but now with the usual Euclidean norm

$$\|x\|_2 \;=\; \sqrt{\sum_{i=1}^{n} |x_i|^2} \;. \tag{1.20}$$

The proof that $\|\cdot\|_2$ is a norm (in particular, that it satisfies the triangle inequality) is slightly less trivial than in the preceding examples. It needs the *Cauchy–Schwarz inequality*

$$\left| \sum_{i=1}^{n} x_i y_i \right| \;\leq\; \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} \left(\sum_{i=1}^{n} y_i^2 \right)^{1/2} \tag{1.21}$$

for real numbers $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, which you will prove in Problem 1(a) of Problem Set #1. Assuming this, we then have

$$\begin{aligned}
\sum_{i=1}^n (x_i + y_i)^2 &= \sum_{i=1}^n x_i^2 + 2 \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2 && \text{(1.22a)} \\
&\le \sum_{i=1}^n x_i^2 + 2 \left(\sum_{i=1}^n x_i^2 \right)^{1/2} \left(\sum_{i=1}^n y_i^2 \right)^{1/2} + \sum_{i=1}^n y_i^2 && \text{(1.22b)} \\
&= \left[\left(\sum_{i=1}^n x_i^2 \right)^{1/2} + \left(\sum_{i=1}^n y_i^2 \right)^{1/2} \right]^2 && \text{(1.22c)}
\end{aligned}$$

where the middle step used the Cauchy–Schwarz inequality. Taking square roots, we have

$$\sqrt{\sum_{i=1}^n (x_i + y_i)^2} \le \sqrt{\sum_{i=1}^n x_i^2} + \sqrt{\sum_{i=1}^n y_i^2} \ , \tag{1.23}$$

which is precisely the triangle inequality $\|x + y\|_2 \le \|x\|_2 + \|y\|_2$.

Exactly the same formula (1.20) defines a norm on $\mathbb{C}^n$, considered as a vector space over the field of complex numbers; but the proof requires the Cauchy–Schwarz inequality for *complex* numbers $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, in which (1.21) has to be modified by replacing x_i^2 and y_i^2 by $|x_i|^2$ and $|y_i|^2$. The proof of the complex Cauchy–Schwarz inequality is not really more difficult than that of the real Cauchy–Schwarz inequality, but let us leave it for a few weeks from now when we really need it.

Remark. Examples 3–5 show that the same space X can be equipped with many different norms (and thus many different metrics). We shall see later that *these three* metrics on $\mathbb{R}^n$ are all "equivalent" in a sense to be defined later — in particular, they give rise to the same convergence structure — so it doesn't matter which one we use. But *not all* metrics on a given space X are equivalent! For instance, the discrete metric (Example 1) clearly gives rise to a *different* convergence structure on $X = \mathbb{R}$ than is given by the usual metric (Example 2). Here is another "funny" metric on $\mathbb{R}$, which we will study again later:

Example 6. The real line with the tanh metric. Let X be the set $\mathbb{R}$ of real numbers, with the distance

$$d_\star(x, y) = |\tanh x - \tanh y| \tag{1.24}$$

where tanh is the hyperbolic tangent function

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}} \ . \tag{1.25}$$

It is easy to see that $d_\star$ is a metric on $\mathbb{R}$ (you should supply a proof), which does *not* arise from a norm (you should prove this too). We will see later that this metric is *not* equivalent to the usual metric.

Example 7. Spaces of bounded functions, with the sup norm. Let A be any set, and let $X = \mathcal{B}(A)$ be the set of *bounded* real-valued functions on A (please remind yourself exactly what it means for a real-valued function on A to be bounded!). The set $\mathcal{B}(A)$ is a vector space under the usual operations of pointwise addition and pointwise multiplication by scalars.[1] Then, for any $f \in \mathcal{B}(A)$ we define

$$\|f\|_\infty \;=\; \sup_{t \in A} |f(t)| \;. \tag{1.26}$$

You should prove that $\| \cdot \|_\infty$ is a norm on $\mathcal{B}(A)$. It is called, not surprisingly, the "sup norm" on $\mathcal{B}(A)$. Why did we need to restrict attention here to *bounded* functions?

Note that $\mathbb{R}^n$ with the sup norm (Example 4) is the special case in which $A = \{1, 2, \ldots, n\}$.

Note also that exactly the same formula would define a norm on the space $\mathcal{B}_{\mathbb{C}}(A)$ of bounded *complex*-valued functions on A, considered as a vector space over the field of complex numbers.

Example 8. Spaces of bounded continuous functions, with the sup norm. Let A be any subset of the real line, and let $X = \mathcal{C}(A)$ be the vector space of *bounded continuous* real-valued functions on A. Use once again the sup norm (1.26). Since $\| \cdot \|_\infty$ is a norm on $\mathcal{B}(A)$, it is also a norm on the linear subspace $\mathcal{C}(A) \subseteq \mathcal{B}(A)$ [why?]. Why did we need to restrict attention here to *bounded* continuous functions?

The most important case is when A is a *closed bounded* subset of the real line, such as a closed interval $A = [a, b]$. Then all continuous real-valued functions on A are *automatically bounded* — you proved this nontrivial fact in your first analysis course, and we will generalize it a week or two from now in connection with the concept of *compactness*. (Exercise: Show by example that if A is either not closed or not bounded, then a continuous real-valued function on A need not be bounded.)

We will see soon that, instead of requiring A to be a subset of the real line, we could take A to be *any metric space*. Once we have defined what it means for a real-valued function on a metric space to be continuous, we will see that the set $\mathcal{C}(A)$ of bounded continuous real-valued functions on A, equipped with the sup norm, is always a normed linear space space. So this shows a way of building new (and more complicated) metric spaces (or normed linear spaces) from old ones. For instance, starting from the real line $\mathbb{R}$, we can build the space $\mathcal{C}(\mathbb{R})$ of bounded continuous real-valued functions on $\mathbb{R}$, then the space $\mathcal{C}(\mathcal{C}(\mathbb{R}))$ of bounded continuous real-valued functions on $\mathcal{C}(\mathbb{R})$, and so forth.

Example 9. Spaces of (bounded) continuous functions, with the L^1 norm. Let $[a, b]$ be a closed bounded interval of the real line, and let $\mathcal{C}[a, b]$ be the vector space of

[1] That is, for any $f, g \in \mathcal{B}(A)$ we define

$$(f + g)(t) \;=\; f(t) + g(t) \qquad \text{for all } t \in A \;,$$

and for any $f \in \mathcal{B}(A)$ and $\lambda \in \mathbb{R}$ we define

$$(\lambda f)(t) \;=\; \lambda f(t) \qquad \text{for all } t \in A \;.$$

continuous real-valued functions on $[a, b]$ (which are automatically bounded, as we have just observed). Now define

$$\|f\|_1 \;=\; \int_a^b |f(t)|\, dt \;. \tag{1.27}$$

You should prove that $\| \cdot \|_1$ is a norm on $\mathcal{C}[a, b]$.

Exactly the same formula would define a norm on the space $\mathcal{C}_{\mathbb{C}}(A)$ of bounded continuous *complex*-valued functions on A.

Example 10. Spaces of (bounded) continuous functions, with the L^2 norm. With $[a, b]$ and $\mathcal{C}[a, b]$ as in the preceding example, define

$$\|f\|_2 \;=\; \left(\int_a^b |f(t)|^2\, dt \right)^{1/2} \;. \tag{1.28}$$

The proof that $\| \cdot \|_2$ is a norm will again require the *Cauchy–Schwarz inequality*, this time for continuous functions on $[a, b]$; see Problem 1(b) of Problem Set #1.

Exactly the same formula would define a norm on the space $\mathcal{C}_{\mathbb{C}}(A)$ of bounded continuous *complex*-valued functions on A; to prove this one has to use the complex Cauchy–Schwarz inequality.

Example 11. The sequence spaces ℓ^∞ and c_0. We denote by ℓ^∞ the vector space of *bounded* infinite sequences $x = (x_1, x_2, \ldots)$ of real numbers, and on this set we define

$$\|x\|_\infty \;=\; \sup_i |x_i| \;. \tag{1.29}$$

You should prove that $\| \cdot \|_\infty$ is a norm on ℓ^∞. Can you see that this is actually a special case of Example 7, and also of Example 8?

We denote by c_0 the set of infinite sequences $x = (x_1, x_2, \ldots)$ of real numbers that are *convergent to zero* (i.e. $\lim_{n\to\infty} x_n = 0$). We have $c_0 \subsetneq \ell^\infty$ (why?). Since $\| \cdot \|_\infty$ is a norm on ℓ^∞, it follows that it is also a norm on the linear subspace $c_0 \subset \ell^\infty$.

Once again, we can here replace real numbers by complex numbers if we wish.

Example 12. The sequence space ℓ^1. We denote by ℓ^1 the set of infinite sequences $x = (x_1, x_2, \ldots)$ of real numbers that are *absolutely summable*, i.e. satisfy $\sum_{i=1}^{\infty} |x_i| < \infty$. On this set we define

$$\|x\|_1 \;=\; \sum_{i=1}^{\infty} |x_i| \;. \tag{1.30}$$

(Why do we need to restrict attention to absolutely summable sequences in order that $\| \cdot \|_1$ be well-defined?) You should prove that $\| \cdot \|_1$ is a norm on ℓ^1. Once again, we can here replace real numbers by complex numbers.

Example 13. The sequence space ℓ^2. We denote by ℓ^2 the set of infinite sequences $x = (x_1, x_2, \ldots)$ of real numbers that are *absolutely square-summable*, i.e. satisfy $\sum_{i=1}^{\infty} |x_i|^2 < \infty$. On this set we define

$$\|x\|_2 \;=\; \left(\sum_{i=1}^{\infty} |x_i|^2 \right)^{1/2} . \tag{1.31}$$

(Why do we need to restrict attention to absolutely square-summable sequences in order that $\| \cdot \|_2$ be well-defined?) To prove that $\| \cdot \|_2$ is a norm on ℓ^2, we first recall from Example 5 the triangle inequality for *finite* sequences of real numbers:

$$\left(\sum_{i=1}^{n} (x_i + y_i)^2 \right)^{1/2} \;\le\; \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} \;+\; \left(\sum_{i=1}^{n} y_i^2 \right)^{1/2} , \tag{1.32}$$

which holds for all n. Next we observe that the finite sums on the right-hand side are bounded above by the corresponding infinite sums, so that

$$\left(\sum_{i=1}^{n} (x_i + y_i)^2 \right)^{1/2} \;\le\; \left(\sum_{i=1}^{\infty} x_i^2 \right)^{1/2} \;+\; \left(\sum_{i=1}^{\infty} y_i^2 \right)^{1/2} \tag{1.33}$$

for all n. Now we can take $n \to \infty$: the left-hand side converges to $\left(\sum_{i=1}^{\infty} (x_i + y_i)^2 \right)^{1/2}$ [why?], which proves what is needed.

This is a typical type of argument that is routinely used to deduce inequalities for infinite sequences from those for finite sequences, and you should make sure that you understand well all the steps in it.

Example 14. The sequence spaces ℓ^p for $1 \le p < \infty$. We will define these sequence spaces later; they include ℓ^1 and ℓ^2 as the most important special cases.

Example 15. The sequence space $\mathbb{R}^{\mathbb{N}}$. We denote by $\mathbb{R}^{\mathbb{N}}$ the set of *all* infinite sequences $x = (x_1, x_2, \ldots)$ of real numbers, bounded or unbounded. We define

$$d(x, y) \;=\; \sum_{j=1}^{\infty} 2^{-j} \, \frac{|x_j - y_j|}{1 + |x_j - y_j|} \, . \tag{1.34}$$

First of all, do you see why $d(x, y)$ is well-defined and *finite* for all $x, y \in \mathbb{R}^{\mathbb{N}}$? In Problem 2 of Problem Set #1 you will prove that d is a metric on $\mathbb{R}^{\mathbb{N}}$.

Example 15 is an interesting example of a metrizable topological vector space whose topology does *not* arise from a norm.

Open and closed sets, interior and closure

We now begin the study of analysis on metric spaces. Most of the elementary definitions will be completely analogous to those employed in analysis on $\mathbb{R}$ or $\mathbb{R}^n$, with which you are

assumed to be familiar; and *most* (but not all) of the facts that hold in $\mathbb{R}^n$ will carry over to general metric spaces. So you can use your knowledge of $\mathbb{R}^n$, as well as rough sketches on a piece of paper ($\mathbb{R}^2$), to gain intuition about general metric spaces that is *often* (but not always) correct. Later we will examine in detail some results in $\mathbb{R}^n$ that do *not* carry over to general metric spaces (notably those involving compactness).

Most of the proofs of results in this section are fairly easy, and they will be left as exercises for you. You should make sure you can do them! (If you need help, consult one of the textbooks such as Kolmogorov–Fomin, Kreyszig, Giles, or Dieudonné; and if that doesn't help, come see me.) These elementary proofs illustrate techniques that you will need to use later, in more complicated contexts.

We fix, once and for all, a metric space (X, d). Given a point $x \in X$ and a real number $r > 0$, we define the **open ball of center x and radius r**

$$B(x, r) \;=\; \{y \in X\colon\, d(x, y) < r\} \tag{1.35}$$

and the **closed ball[2] of center x and radius r**

$$\overline{B}(x, r) \;=\; \{y \in X\colon\, d(x, y) \le r\}\;. \tag{1.36}$$

A subset $A \subseteq X$ is said to be **open** if, for every $x \in A$, there exists $r > 0$ (depending in general on x) such that $B(x, r) \subseteq A$.

Proposition 1.3 *Every open ball is an open set.*

Proof. Consider an open ball $B(x, r)$. We need to check that for each $y \in B(x, r)$, there exists $r' > 0$ such that $B(y, r') \subseteq B(x, r)$. Before reading further, you should draw a picture of this situation and figure out what r' should be!

Did you guess that we should take $r' = r - d(x, y)$? To see that this works, consider any $z \in B(y, r')$, i.e. any z for which $d(y, z) < r - d(x, y)$. Then by the triangle inequality we have

$$d(x, z) \;\le\; d(x, y) + d(y, z) \;<\; r\;, \tag{1.37}$$

which implies that $z \in B(x, r)$. Since z was an arbitrary point of $B(y, r')$, we have proven that $B(y, r') \subseteq B(x, r)$. $\square$

Proposition 1.4

(a) The empty set $\varnothing$ and the whole space X are open sets.

(b) The union of an arbitrary collection of open sets is open.

[2] **Warning:** Kolmogorov–Fomin use the term "closed sphere" for what I (and nearly everyone else) call a "closed ball". This is potentially misleading, as most authors use the term "sphere" to denote the set

$$S(x, r) \;=\; \{y \in X\colon\, d(x, y) = r\}\;.$$

I haven't even bothered to introduce this latter concept here, as it is of little relevance to most of our work.

(c) The intersection of a finite *collection of open sets is open.*

OK, this time *you* should supply the proofs!

Please note that an *infinite* intersection of open sets need not be open — you should construct a simple example of this in the real line $\mathbb{R}$.

Note also that in a discrete metric space (Example 1), *every* set is open — you should supply a proof of this fact too. This shows that discrete metric spaces are rather uninteresting.

Remark. A **topological space** is defined to be a set X equipped with a collection $\mathcal{U}$ of subsets that satisfy properties (a), (b) and (c) above. So Proposition 1.4 shows that the open sets of a metric space give rise to a topological space. But not every topological space comes from a metric space! (Those that do are called **metrizable**.) The deeper study of functional analysis requires the study of nonmetrizable topological spaces; but we will leave that for a future course.

Given a point $x \in X$, an **open neighborhood** of x is any open set containing x. A **neighborhood** of x is any set containing an open neighborhood of x.

More generally, if A is any nonempty subset of X, an open neighborhood of A is any open set containing A; and a neighborhood of A is any set containing an open neighborhood of A.

For any subset $A \subseteq X$, the **interior** of A, denoted A°, consists of all points $x \in A$ for which A is a neighborhood of x.[3] Equivalently, A° consists of all points $x \in A$ for which there exists $r > 0$ such that $B(x, r) \subseteq A$ (why is this equivalent?).

Proposition 1.5 *For any set A, A° is the largest open set contained in A.*

You should supply the proof. Note that you have to prove *two* things: that A° is indeed open; and that if B is any open set contained in A, then $B \subseteq A^\circ$.

Proposition 1.6

(a) A set A is open if and only if $A = A^\circ$.

(b) For any set A, we have $(A^\circ)^\circ = A^\circ$.

(c) If $A \subseteq B$, then $A^\circ \subseteq B^\circ$.

(d) For any pair of sets A, B, we have $(A \cap B)^\circ = A^\circ \cap B^\circ$.

Question: Is it necessarily true that $(A \cup B)^\circ = A^\circ \cup B^\circ$? Give either a proof or a counterexample.

A subset $A \subseteq X$ is said to be **closed** if its complement $X \setminus A$ is open.

Proposition 1.7 *Every closed ball is a closed set.*

[3] Dieudonné uses the notation $\overset{\circ}{A}$ in place of A°. Some authors write $\mathrm{int}(A)$ or $\mathrm{Int}(A)$.

PROOF. Consider a closed ball $\overline{B}(x,r)$. We need to show that its complement $X \setminus \overline{B}(x,r)$ is open, i.e. that for each $y \in X \setminus \overline{B}(x,r)$ there exists $r' > 0$ such that $B(y,r') \subseteq X \setminus \overline{B}(x,r)$. Before reading further, you should draw a picture of this situation and figure out what r' should be!

Did you guess that we should take $r' = d(x,y) - r$? Note that $d(x,y) > r$ because $y \notin \overline{B}(x,r)$, hence $r' > 0$. To see that this works, consider any $z \in B(y,r')$. Then $d(x,z) \geq d(x,y) - d(y,z)$ (why is this a consequence of the triangle inequality?); and since $d(y,z) < r' = d(x,y) - r$, we have $d(x,z) > r$, hence $z \notin \overline{B}(x,r)$, hence $z \in X \setminus \overline{B}(x,r)$. Since z was an arbitrary point of $B(y,r')$, we have proven that $B(y,r') \subseteq X \setminus \overline{B}(x,r)$. $\square$

We have the following "dual" of Proposition 1.4, which is obtained by taking complements everywhere and using the standard rules of Boolean algebra (hence the roles of union and intersection get reversed):

Proposition 1.8

(a) The empty set $\varnothing$ and the whole space X are closed sets.

(b) The intersection of an arbitrary collection of closed sets is closed.

(c) The union of a finite *collection of closed sets is closed.*

A **cluster point** of a set $A \subseteq X$ is a point $x \in X$ such that every neighborhood of x has a nonempty intersection with A.[4] The set of all cluster points of A is called the **closure** of A and written $\overline{A}$.[5] To say that x is *not* a cluster point of A means that it is interior to $X \setminus A$ (why?), or in other words:

Proposition 1.9 *The closure of A is the complement of the interior of the complement of A.*

Thus, just as open and closed sets are "dual" under complementation, so interior and closure are "dual" under complementation.

Warning: The closure of an open ball $B(x,r)$ is always *contained in* the closed ball $\overline{B}(x,r)$ (you should verify this!); but in a general metric space it is not necessarily *equal* to $\overline{B}(x,r)$. For instance, in a discrete metric space we have $B(x,1) = \{x\}$ for each point x (why?), and the closure of $\{x\}$ is itself (why?); but $\overline{B}(x,1) = X$ (why?).

We have the following "duals" of Propositions 1.5 and 1.6:

Proposition 1.10 *For any set A, $\overline{A}$ is the smallest closed set containing A.*

[4]Kolmogorov–Fomin use the term "contact point", and Rynne–Youngson use the term "closure point". **Warning:** Do not confuse this with the related but different notion of a **limit point** (or **accumulation point**) of A, which is a point $x \in X$ such that every neighborhood of x contains *infinitely many* points of A.

[5]Kolmogorov–Fomin use the notation $[A]$ in place of $\overline{A}$. Rynne–Youngson use both of the notations $\overline{A}$ and A^-.

Proposition 1.11

(a) A set A is closed if and only if $A = \overline{A}$.

(b) For any set A, we have $\overline{\overline{A}} = \overline{A}$.

(c) If $A \subseteq B$, then $\overline{A} \subseteq \overline{B}$.

(d) For any pair of sets A, B, we have $\overline{A \cup B} = \overline{A} \cup \overline{B}$.

Question: Is it necessarily true that $\overline{A \cap B} = \overline{A} \cap \overline{B}$? Give either a proof or a counterexample.

Distance and neighborhoods

If A, B are any two nonempty subsets of X, we define the **distance from A to B** as

$$d(A, B) \;=\; \inf_{x \in A,\, y \in B} d(x, y) \tag{1.38}$$

When A is reduced to a single point x, we write $d(x, B)$ as a synonym for $d(\{x\}, B)$. Note that the infimum in $d(A, B)$ need not be attained, i.e. there need not exist any pair $x \in A$ and $y \in B$ such that $d(x, y) = d(A, B)$; all we know, *a priori*, is that there exist pairs x, y that make $d(x, y)$ arbitrarily close to $d(A, B)$. See Problem 3 on Problem Set #1.

Lemma 1.12 *If A is a nonempty subset of X, then for all $x, y \in X$ we have*

$$|d(x, A) - d(y, A)| \;\le\; d(x, y)\;. \tag{1.39}$$

We will see later, after defining continuous functions, that this says that the function $x \mapsto d(x, A)$ is a continuous (in fact, Lipschitz-continuous) real-valued function on X.

Proof of Lemma 1.12. For every $z \in A$ we have $d(x, z) \le d(x, y) + d(y, z)$, hence

$$\begin{aligned} d(x, A) &= \inf_{z \in A} d(x, z) && \text{(1.40a)} \\ &\le \inf_{z \in A} [d(x, y) + d(y, z)] && \text{(1.40b)} \\ &= d(x, y) + \inf_{z \in A} d(y, z) && \text{(1.40c)} \\ &= d(x, y) + d(y, A)\;, && \text{(1.40d)} \end{aligned}$$

so $d(x, A) - d(y, A) \le d(x, y)$. Doing the same thing with the roles of x and y reversed, one concludes that $d(x, A) - d(y, A) \ge -d(x, y)$. Putting these two inequalities together proves the Lemma. □

Proposition 1.13 *For any nonempty set $A \subseteq X$ and any $r > 0$, the set $V_r(A) = \{x \in X\colon d(x, A) < r\}$ is an open neighborhood of A.*

You should prove Proposition 1.13. Obviously $V_r(A)$ contains A (why?), so you need only prove that $V_r(A)$ is open; that is, you need to prove that for any $x \in V_r(A)$, there exists $r' > 0$ such that $B(x, r') \subseteq V_r(A)$. You should draw a picture to figure out what r' should be taken to be, and then use Lemma 1.12 to complete the proof.

Proposition 1.14 *Let A be a nonempty subset of X; then a point x belongs to $\overline{A}$ if and only if $d(x, A) = 0$.*

PROOF. If $d(x, A) = 0$, then for every $\epsilon > 0$ there exists $y \in A$ with $d(x, y) < \epsilon$; or in other words, for every $\epsilon > 0$ we have $B(x, \epsilon) \cap A \neq \varnothing$. This proves that x is a cluster point of A (why?), i.e. $x \in \overline{A}$.

Conversely, if $x \in \overline{A}$, then for every $\epsilon > 0$ we have $B(x, \epsilon) \cap A \neq \varnothing$, i.e. there exists $y \in A$ with $d(x, y) < \epsilon$. But this shows that $d(x, A) = 0$. $\square$

Proposition 1.14 can be rephrased as saying that

$$\overline{A} \;=\; \bigcap_{r>0} V_r(A) \;=\; \bigcap_{n\geq 1} V_{1/n}(A) \tag{1.41}$$

(why is this the case?). It follows that:

Corollary 1.15 *In a metric space X,*

(a) Every closed set is the intersection of a decreasing sequence *of open sets.*

(b) Every open set is the union of an increasing sequence *of closed sets.*

Indeed, (a) follows from (1.41), while (b) follows from (a) by taking complements.

The key word here is "sequence": that is, every closed set is the intersection of a *countably infinite* collection of open sets. **Warning:** This does *not* hold in arbitrary topological spaces.

Subspaces of a metric space

If (X, d) is a metric space and Y is a nonempty subset of X, then the restriction of d to $Y \times Y$ is obviously a metric on Y; the metric space defined by this induced metric is called the **subspace** Y of the metric space X. (We have already used this idea to discuss the subspace c_0 of ℓ^∞.)

The following lemma will be useful:

Lemma 1.16 *A set $B \subseteq Y$ is open in the subspace Y if and only if there exists an open set A in X such that $B = A \cap Y$.*

PROOF. If $y \in Y$ and $r > 0$, then $B(y, r) \cap Y$ is the open ball of center y and radius r *in the subspace Y* (why?).

Now, if A is open in X and $y \in A \cap Y$, then there exists $r > 0$ such that $B(y, r) \subseteq A$, hence $y \in B(y, r) \cap Y \subseteq A \cap Y$, which shows that $A \cap Y$ is open in Y.

Conversely, if $B \subseteq Y$ is open in the subspace Y, then for each $y \in B$ there exists a number $r(y) > 0$ such that $B(y, r(y)) \cap Y \subseteq B$. This shows that

$$B = \bigcup_{y \in B} (B(y, r(y)) \cap Y) = \left(\bigcup_{y \in B} B(y, r(y)) \right) \cap Y \tag{1.42}$$

(you should explain both of these equalities!), or in other words $B = A \cap Y$ where $A = \bigcup_{y \in B} B(y, r(y))$ is open in X (why?). $\square$

Dense subsets; separable metric spaces

If A, B are subsets of a metric space X, we say that A is **dense with respect to** B if B is contained in the closure of A (i.e. $B \subseteq \overline{A}$).

You should prove the following, using the elementary properties of the closure operator (Proposition 1.11):

Proposition 1.17 *If A is dense with respect to B, and B is dense with respect to C, then A is dense with respect to C.*

A set A that is dense with respect to the whole space X is called **everywhere dense**, or simply **dense** in X. Such sets are characterized by the fact that $\overline{A} = X$, or equivalently that every nonempty open set contains a point of A.

I assume that you are familiar, from your previous courses in analysis (or set theory), with the classification of sets as **finite**, **countably infinite**, or **uncountably infinite**.[6] We say that a set is **countable** if it is either finite or countably infinite.[7] We then make the following important definition: A metric space X is said to be **separable** if there exists in X a countable dense set. As we shall see, separable metric spaces are easier to work with than general metric spaces, because they are in a certain sense "not too large", i.e. they can be "well approximated" by a countable subset.

For example, the real line $\mathbb{R}$ with the usual metric (Example 2) is separable, because the set $\mathbb{Q}$ of rational numbers is dense in $\mathbb{R}$ (why?), and $\mathbb{Q}$ is countably infinite (why?). More generally, $\mathbb{R}^n$ with any of the usual metrics (Examples 3–5) is separable (why?). At the other extreme, a discrete metric space (Example 1) is separable if and only if the underlying set X is countable (why?).

We now proceed to study the separability of the sequence spaces ℓ^∞, c_0, ℓ^1 and ℓ^2. Let us first show that the space ℓ^∞ of bounded sequences (Example 11) is *not* separable:

[6] If not, then you should carefully study Kolmogorov–Fomin, Chapter 1, Sections 1 and 2 (or the equivalent discussion in another book) without delay. In this course I will make use of basic facts about finite and infinite sets without further comment. See also Handout #0.

[7] The terms **denumerable** and **nondenumerable** are also used as synonyms for "countably infinite" and "uncountably infinite", respectively; and the term **at most denumerable** is also used as a synonym for "finite or countably infinite".

Proposition 1.18 *ℓ^∞ is not separable.*

PROOF. For each subset I of the positive integers $\mathbb{N}$, define $e_I \in \ell^\infty$ by

$$(e_I)_i = \begin{cases} 1 & \text{if } i \in I \\ 0 & \text{if } i \notin I \end{cases} \tag{1.43}$$

Then $d_\infty(e_I, e_J) = 1$ whenever $I \neq J$ (why?). So $\mathcal{B} = \{B(e_I, \frac{1}{2})\colon I \subseteq \mathbb{N}\}$ is an uncountably infinite (why?) collection of disjoint (why?) open balls in ℓ^∞. Now let S be any dense subset of ℓ^∞: then each ball in the family $\mathcal{B}$ must contain at least one element of S (why?), and these elements must all be distinct (why?); so S must be uncountably infinite. This shows that ℓ^∞ is not separable. □

By a similar technique you will prove, in Problem 5 of Problem Set #1, that the space $\mathcal{C}(\mathbb{R})$ of bounded continuous real-valued functions on the whole real line $\mathbb{R}$ (Example 8) is not separable. And in Problem 6 you will abstract this construction to prove a more general result about separability of metric spaces.

On the other hand, the subspace $c_0 \subsetneq \ell^\infty$, which consists of sequences that are convergent to zero, *is* separable:

Proposition 1.19 *c_0 is separable.*

PROOF. Let S be the subset of c_0 consisting of sequences with *rational* entries, of which *at most finitely many are nonzero*. This is a countably infinite set (why?).[8] We will show that S is dense in c_0. To do this, we must show that for any $x \in c_0$ and any $\epsilon > 0$, there exists $y \in S$ such that $d_\infty(x, y) \leq \epsilon$. We construct the needed y in two steps:

Firstly, $x = (x_1, x_2, \ldots) \in c_0$ means that $\lim_{n\to\infty} x_n = 0$, i.e. for any $\epsilon > 0$ there exists an integer N such that $|x_n| \leq \epsilon$ whenever $n > N$. Secondly, we choose rational numbers $y_1, y_2, \ldots, y_N$ such that $|x_i - y_i| \leq \epsilon$ for $1 \leq i \leq N$. So if we define $y = (y_1, y_2, \ldots, y_N, 0, 0, \ldots)$, we have $y \in S$ and $d_\infty(x, y) \leq \epsilon$ (why?). This completes the proof. □

This proof is typical of many proofs in analysis: First we "cut off" an infinite sequence by approximating it (to within a small error ϵ) by a finite sequence; and then we further approximate that finite sequence (again to within a small error ϵ) by a finite sequence of some desired special type. The total error committed in this approximation is at worst the

[8] Since this is a very important argument that we will use over and over again throughout this course, let me make the reasoning explicit in case you were unable to answer the "why?" for yourself:

Let S_N be the subset of c_0 consisting of sequences with rational entries of which *at most the first N entries are nonzero*, i.e. sequences of the form $(x_1, x_2, \ldots, x_N, 0, 0, 0, \ldots)$ with $x_1, \ldots, x_N \in \mathbb{Q}$. Then S_N is in obvious bijection with $\mathbb{Q}^N$ (why?), and $\mathbb{Q}^N$ is a countably infinite set because it is a *finite* Cartesian product of countably infinite sets [cf. Theorem 0.3(c) from Handout #0]. But then $S = \bigcup_{N=1}^{\infty} S_N$ is a countably infinite union of countably infinite sets, hence also countably infinite [cf. Theorem 0.3(b) from Handout #0].

This two-step process — countable union of *finite* Cartesian products — is crucial, since a countably infinite Cartesian product of countably infinite sets is in general *not* countably infinite.

sum of the errors committed in the two steps, by the triangle inequality.[9] You will use this technique, in Problem 7 of Problem Set #1, to prove the separability of the sequence spaces ℓ^1 and ℓ^2.

Later in this course we will see that the spaces $\mathcal{C}[a, b]$ of (bounded) continuous real-valued functions on a *closed bounded* interval of the real line are also separable, but the proof is more difficult.

Continuous mappings

One of the most fundamental concepts in analysis on $\mathbb{R}$ (or $\mathbb{R}^n$) is that of a continuous real-valued function. Intuitively, the idea is that a function f is continuous at a point x_0 in case $f(x)$ can be made arbitrarily close to $f(x_0)$ by taking x sufficiently close to x_0. This idea can be made more precise in either of two equivalent ways: the familiar ϵ–δ way, or by using neighborhoods:

> **ϵ–δ definition.** A function f is continuous at a point x_0 if, for each $\epsilon > 0$, there exists $\delta > 0$ such that $|x - x_0| < \delta$ implies $|f(x) - f(x_0)| < \epsilon$.
>
> **Neighborhood definition.** A function f is continuous at a point x_0 if, for each neighborhood V of $f(x_0)$, there exists a neighborhood U of x_0 such that $f[U] \subseteq V$.

You should make sure that you understand why these two definitions are equivalent. That is, you should sketch for yourself the proof that each property implies the other.

Here we will generalize the concept of continuity so that both the domain and the range are allowed to be arbitrary metric spaces. The definition will be an almost exact copy of the definition from elementary analysis; all we do is replace the absolute value by the metric. Here is the precise definition: Let (X, d_X) and (Y, d_Y) be two metric spaces.

> **ϵ–δ definition.** A mapping $f\colon X \to Y$ is said to be **continuous at the point** $x_0 \in X$ if, for each $\epsilon > 0$, there exists $\delta > 0$ such that $d_X(x, x_0) < \delta$ implies $d_Y(f(x), f(x_0)) < \epsilon$.
>
> **Neighborhood definition.** A mapping $f\colon X \to Y$ is said to be **continuous at the point** $x_0 \in X$ if, for each neighborhood V of $f(x_0)$ in Y, there exists a neighborhood U of x_0 in X such that $f[U] \subseteq V$.

You should make sure, once again, that you understand why the two definitions are equivalent. (The reasoning will be the same as it was in $\mathbb{R}$.) Here is another rephrasing:

> **Neighborhood definition (rephrased).** A mapping $f\colon X \to Y$ is **continuous at the point** $x_0 \in X$ if, for each neighborhood V of $f(x_0)$ in Y, $f^{-1}[V]$ is a neighborhood of x_0 in X.

[9] In the specific case above, we didn't need to use the triangle inequality, because we were dealing with the sup metric. But it would have been no harm if we had used it: the final bound would have been $d_\infty(x, y) \leq 2\epsilon$ instead of ϵ, but that would have been no problem, because we could have used $\epsilon/2$ instead of ϵ in the two approximation steps.

You should make sure that you understand why this is equivalent to the previous phrasing of the neighborhood definition.

Remark: In general topological spaces, the ϵ–δ definition makes no sense because there is in general no metric available, but the neighborhood definition still makes sense and is taken as the definition of a continuous mapping.

A mapping $f\colon X \to Y$ is said to be **continuous on** X (or simply **continuous** without further qualification) if it is continuous at *every* point of X. Continuity is characterized by any one of the following equivalent conditions:

Proposition 1.20 *Let X and Y be metric spaces, and let f be a mapping of X into Y. Then the following properties are equivalent:*

(a) f is continuous;

(b) for every open set V in Y, $f^{-1}[V]$ is an open set in X;

(c) for every closed set C in Y, $f^{-1}[C]$ is a closed set in X;

(d) for every set A in X, we have $f[\overline{A}] \subseteq \overline{f[A]}$.

Proof. (a) $\implies$ (d): Let x_0 be any point in $\overline{A}$, and let V be any neighborhood of $f(x_0)$ in Y. Then, by hypothesis, $f^{-1}[V]$ is a neighborhood of x_0 in X; so there exists a point $x \in A \cap f^{-1}[V]$ (why?), which means that $f(x) \in f[A] \cap V$. So every neighborhood of $f(x_0)$ contains a point of $f[A]$, which means that $f(x_0) \in \overline{f[A]}$. [**Remark:** This argument uses only the continuity of f at x_0.]

(d) $\implies$ (c): Let C be a closed set in Y, and define $A = f^{-1}[C]$. Then, by (d), $f[\overline{A}] \subseteq \overline{f[A]}$. Now, $f[A] = f[f^{-1}[C]] = C \cap f[X] \subseteq C$ (why?), so $\overline{f[A]} \subseteq \overline{C} = C$ (why?). So we have shown that $f[\overline{A}] \subseteq C$, which implies that $\overline{A} \subseteq f^{-1}[C] = A$. Since the reverse inclusion $A \subseteq \overline{A}$ is obvious, we have $\overline{A} = A$ and hence A is closed.

(c) $\implies$ (b): In fact, each of (b) and (c) immediately implies the other, because $f^{-1}[Y \setminus S] = X \setminus f^{-1}[S]$ for any subset $S \subseteq Y$. (You should think carefully about why this is true, and why the corresponding equality for *direct* images is *not* in general true.)

(b) $\implies$ (a): Consider any $x_0 \in X$, and let V be a neighborhood of $f(x_0)$. Then there exists an open neighborhood W of $f(x_0)$ that is contained in V (why?); and $f^{-1}[W]$ is an open set containing x_0 that is contained in $f^{-1}[V]$, so $f^{-1}[V]$ is a neighborhood of x_0 in X. This proves that f is continuous at x_0; and since x_0 is arbitrary, this proves that f is continuous everywhere. $\square$

It should be stressed that parts (b) and (c) of Proposition 1.20 refer to the *inverse* images of sets under the function f. By contrast, the *direct* image of an open (resp. closed) set by a continuous mapping need not be open (resp. closed). For instance, the map $f(x) = x^2$ is continuous from $\mathbb{R}$ to $\mathbb{R}$, but the image of the open set $(-1, 1)$ is the non-open set $[0, 1)$. Likewise, the map $g(x) = \tanh x$ is continuous from $\mathbb{R}$ to $\mathbb{R}$, but the image of the closed set $\mathbb{R}$ is the non-closed set $(-1, 1)$. [See, however, next week concerning the direct images of the special class of closed sets called "compact sets".]

Convergence of sequences

Another important concept in analysis is that of convergence of sequences. I assume that you recall the definition of convergence of sequences in $\mathbb{R}$ or $\mathbb{R}^n$, and I pass immediately to give the definition for arbitrary metric spaces.

Let (X, d) be a metric space, let $(x_n)_{n=1}^{\infty}$ be a sequence of points of X, and let a be a point of X. Then:

> **ϵ–n_0 definition.** The sequence $(x_n)_{n=1}^{\infty}$ **converges to** a (or **has** a **as a limit**) in case, for every $\epsilon > 0$, there exists an integer n_0 such that $d(x_n, a) < \epsilon$ whenever $n \geq n_0$.
>
> **Neighborhood definition.** The sequence $(x_n)_{n=1}^{\infty}$ **converges to** a (or **has** a **as a limit**) in case, for every neighborhood V of a, there exists an integer n_0 such that $x_n \in V$ whenever $n \geq n_0$.

Once again you should make sure you understand why the two definitions are equivalent.

We can also use the standard definition of convergence of sequences in $\mathbb{R}$ (which is a special case of the metric-space definition) to rephrase the ϵ–n_0 definition as follows:

> **ϵ–n_0 definition (rephrased).** The sequence $(x_n)_{n=1}^{\infty}$ converges to a if and only if $\lim_{n\to\infty} d(x_n, a) = 0$.

Proposition 1.21 *A sequence can have at most one limit.*

Proof. Suppose that the sequence $(x_n)_{n=1}^{\infty}$ converges to both a and b. Then, for each $\epsilon > 0$, there exists an integer n_0 such that $d(x_n, a) < \epsilon$ whenever $n \geq n_0$, and also an integer n_1 such that $d(x_n, b) < \epsilon$ whenever $n \geq n_1$. Now take any $n \geq \max(n_0, n_1)$: we have $d(x_n, a) < \epsilon$ and $d(x_n, b) < \epsilon$, hence by the triangle inequality $d(a, b) < 2\epsilon$. But since ϵ was arbitrary, we can conclude that $d(a, b) = 0$ (why?), which means that $a = b$ (why?). □

Since a sequence can have at most one limit, it makes sense to write $\lim_{n\to\infty} x_n = a$ as a shorthand for the statement that the sequence $(x_n)_{n=1}^{\infty}$ converges to a. We also use the notation $x_n \overset{n\to\infty}{\longrightarrow} a$ or simply $x_n \to a$.

The concept of the closure of a set can be rephrased in terms of sequences:

Proposition 1.22 *Let A be a subset of a metric space X. Then*

$$\overline{A} = \{x \in X\text{: } \textit{there exists a sequence } (x_n)_{n=1}^{\infty} \textit{ of points of } A \textit{ that converges to } x\}\ .$$

Proof. Suppose first that $(x_n)_{n=1}^{\infty}$ is a sequence of points of A that converges to a point $x \in X$. This means that for every neighborhood V of x, there exists an integer n_0 such that $x_n \in V$ whenever $n \geq n_0$. But this means, in particular, that $V \cap A$ is nonempty. Hence $x \in \overline{A}$.

Conversely, suppose that $x \in \overline{A}$. Then, for every positive integer n, the set $A \cap B(x, 1/n)$ is nonempty, so choose (arbitrarily) a point $x_n \in A \cap B(x, 1/n)$. Then it is easy to see (e.g. using the ϵ–n_0 definition) that the sequence $(x_n)_{n=1}^{\infty}$ converges to x. □

Warning: This result does *not* hold in arbitrary topological spaces. Or rather, only half of it does: every limit of a sequence of points in A indeed belongs to $\overline{A}$, but points in $\overline{A}$ need not arise as limits of *sequences* of points in A. Rather, the concept of "sequence" has to be replaced by the more general concept of **net**, and then an analogue of Proposition 1.22 holds.

Likewise, the continuity of functions can be rephrased in terms of sequences:

Proposition 1.23 *Let X and Y be metric spaces, let f be a map from X to Y, and let $x_\star$ be a point in X. Then f is continuous at $x_\star$ if and only if, for each sequence $(x_n)_{n=1}^\infty$ of points of X that converges to $x_\star$, we have $\lim_{n\to\infty} f(x_n) = f(x_\star)$.*

PROOF. Suppose first that f is continuous at $x_\star$ and that the sequence $(x_n)_{n=1}^\infty$ converges to $x_\star$. Then, for every $\epsilon > 0$ we can choose $\delta > 0$ such that $d(x, x_\star) < \delta$ implies $d(f(x), f(x_\star)) < \epsilon$ (why?). And given δ we can choose n_0 such that $d(x_n, x_\star) < \delta$ whenever $n \geq n_0$. It follows that $d(f(x_n), f(x_\star)) < \epsilon$ whenever $n \geq n_0$. But this shows that $\lim_{n\to\infty} f(x_n) = f(x_\star)$.

Conversely, suppose that f is *not* continuous at x_0. Then there exists $\epsilon > 0$ such that for all $\delta > 0$ there exists a point x with $d(x, x_\star) < \delta$ such that $d(f(x), f(x_\star)) \geq \epsilon$. [You should go carefully through the quantifiers here to understand why this is true!] Taking $\delta = 1/n$ for each positive integer n, we can choose points x_n satisfying $d(x_n, x_\star) < 1/n$ and $d(f(x_n), f(x_\star)) \geq \epsilon$. We then have $x_n \to x_\star$ but $f(x_n) \not\to f(x_\star)$ (why?). $\square$

Warning: Once again, this result does *not* hold in arbitrary topological spaces; rather, sequences must once again be replaced by nets.

Cauchy sequences and completeness

One of the inconvenient aspects of the definition of convergence of sequences is that, to verify it, we need to know *in advance* what the limit a is going to be. This can be a bit frustrating, because in many applications we want to prove that a given sequence $(x_n)_{n=1}^\infty$ converges to *some* limit, without necessarily knowing what this limit is. How should we proceed? Can we find a condition that applies directly to the sequence $(x_n)_{n=1}^\infty$, and which implies — or better yet, is equivalent to — the assertion that the sequence $(x_n)_{n=1}^\infty$ converges to *some* (possibly unknown) limit?

Here is one good idea: rather than looking at the distance between x_n and the putative limit a — which we can't do because we don't know a — let's look instead at the distances between the various x_n. Maybe if *these* distances go to zero, we can be certain that the sequence (x_n) converges to *some* limit.

One's first thought might be to look at the distance between x_n and the next element x_{n+1}, but it is easy to see that it is *not* enough for *this* distance to go to zero. For instance, in the real line $\mathbb{R}$, let $x_n = \sqrt{n}$. Then

$$|x_n - x_{n+1}| \;=\; \sqrt{n+1} - \sqrt{n} \;=\; (\sqrt{n+1} - \sqrt{n}) \frac{\sqrt{n+1} + \sqrt{n}}{\sqrt{n+1} + \sqrt{n}} \;=\; \frac{1}{\sqrt{n+1} + \sqrt{n}} \;\leq\; \frac{1}{2\sqrt{n}} \;\to\; 0 \tag{1.44}$$

as $n \to \infty$, but the sequence (x_n) does not converge to any finite limit.

Clearly, what we need to look at is not the distance from x_n to the *next* element x_{n+1}, but rather the distances from x_n to *all* subsequent elements in the sequence. Here is the precise definition:

Let (X, d) be a metric space, and let $(x_n)_{n=1}^{\infty}$ be a sequence of points of X. Then the sequence (x_n) is called a **Cauchy sequence** if, for every $\epsilon > 0$, there exists an integer n_0 such that $d(x_m, x_n) < \epsilon$ whenever $m, n \geq n_0$.

Proposition 1.24 *Any convergent sequence is a Cauchy sequence.*

PROOF. Suppose that the sequence (x_n) converges to a point a. Then, for each $\epsilon > 0$ there exists an integer n_0 such that $d(x_n, a) < \epsilon/2$ for all $n \geq n_0$ (why?). But then, by the triangle inequality, $d(x_m, x_n) < \epsilon$ for all $m, n \geq n_0$. $\square$

Unfortunately, the converse is not in general true: in a general metric space, a Cauchy sequence need not converge. For instance, consider the sequence $(x_n)_{n=1}^{\infty}$ of real numbers defined by

$$x_n \;=\; \sum_{k=1}^{n} \frac{1}{k^2} \;. \tag{1.45}$$

Then (x_n) is an increasing sequence that converges to the limit

$$a \;=\; \sum_{k=1}^{\infty} \frac{1}{k^2} \;=\; \frac{\pi^2}{6} \;. \tag{1.46}$$

And by Proposition 1.24, (x_n) is a Cauchy sequence as well. So far everything looks fine; but now consider (x_n) as a sequence in the metric space $\mathbb{Q}$ of *rational* numbers (equipped with the usual metric $d(x, y) = |x - y|$ that it inherits as a subspace of the metric space $\mathbb{R}$). It is still a Cauchy sequence (why?), but it no longer converges (why?).[10]

The culprit in this example is not the specific sequence (x_n), but rather the metric space of rational numbers; in the real numbers this kind of disaster could not happen, as we shall soon see. Let us therefore make a definition that characterize the "good" metric spaces: A metric space (X, d) is called **complete** if every Cauchy sequence converges.

We have just shown that the metric space $\mathbb{Q}$ of rational numbers is *not* complete. On the other hand, the Cauchy convergence criterion from elementary real analysis states that a sequence of real numbers is convergent *if and only if* it is Cauchy, or in other words:

[10]Maybe you don't like this example because the proof of (1.46) is not completely trivial (the easiest proof uses Fourier series) and the proof of the irrationality of $\pi^2/6$ is decidedly nontrivial. If so, here is a more elementary example: Define the sequence $(x_n)_{n=1}^{\infty}$ by the initial condition $x_1 = 1$ and the recursion

$$x_{n+1} \;=\; \frac{1}{2}\left(x_n \,+\, \frac{2}{x_n}\right) \;.$$

Can you see what limit a this sequence is going to converge to? [*Hint:* Assume that $x_n \to a$, pass to the limit $n \to \infty$ in the recursion equation, and then solve the resulting equation for a. This doesn't prove that (x_n) converges, but it does show what the limit must be *if* the sequence converges at all. With a bit more work you can then prove that the sequence does indeed converge.] You may recognize this recursion as Newton's method $x_{n+1} = x_n - f(x_n)/f'(x_n)$ for the function $f(x) = x^2 - 2$.

Proposition 1.25 *The real line $\mathbb{R}$ (with its usual metric as in Example 2) is complete.*

I won't prove this here, since it would take me too far afield and I assume that you have seen this proof in your previous course of real analysis. Let me just recall that the proof of the Cauchy convergence criterion obviously must depend on some "completeness" property of the real numbers that distinguishes them from the "incomplete" rationals. Any one of the following three properties could be used as the starting point from which one could prove the other two properties as well as the Cauchy convergence criterion:[11]

> **The least upper bound property.** Every nonempty set of real numbers which has an upper bound has a least upper bound.
>
> **The nested-interval property.** Every sequence $I_1 \supseteq I_2 \supseteq I_3 \supseteq \ldots$ of closed bounded intervals $I_n = [a_n, b_n]$ has a nonempty intersection.
>
> **The Bolzano–Weierstrass property.** Every bounded sequence (x_n) of real numbers has a convergent subsequence.

For instance, the completeness of $\mathbb{R}$ can be derived from the Bolzano–Weierstrass property by using the following two facts, which are valid in arbitrary metric spaces (their proofs are easy and I leave them to you):

Proposition 1.26 *Let (X, d) be a metric space and let $(x_n)_{n=1}^{\infty}$ be a Cauchy sequence in X. Then the sequence (x_n) is* **bounded** *(in the sense that the sequence of real numbers $d(x_n, a)$ is bounded for every point $a \in X$).*

Proposition 1.27 *Let (X, d) be a metric space and let $(x_n)_{n=1}^{\infty}$ be a Cauchy sequence in X. If some subsequence of (x_n) converges to a point a, then the whole sequence (x_n) converges to a.*

As a corollary of the completeness of $\mathbb{R}$, one can easily prove:

Proposition 1.28 *The metric space $\mathbb{R}^n$ with any of its usual metrics (Examples 3–5) is complete.*

Let me do it for the sup metric d_∞; I leave the proof to you for the metrics d_1 and d_2.

PROOF. Let $x^{(1)}, x^{(2)}, x^{(3)}, \ldots$ be a Cauchy sequence in $\mathbb{R}^n$, where $x^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$. This means that, for each $\epsilon > 0$ there exists n_0 such that $d_\infty(x^{(i)}, x^{(j)}) < \epsilon$ whenever $i, j \geq n_0$. But by the definition of d_∞ this means that

$$|x_k^{(i)} - x_k^{(j)}| \;<\; \epsilon \tag{1.47}$$

[11] In case you wonder where these properties come from, the best approach is to develop the real numbers "constructively" from the rationals. One way is to use the method of Dedekind cuts: see e.g. Landau, *Foundations of Analysis*. Another way is to define the real numbers as equivalence classes of Cauchy sequences of rationals — this is a special case of the general construction of "completion of a metric space" that we will discuss very soon.

for each index k ($1 \le k \le n$) whenever $i, j \ge n_0$. Therefore the sequence $(x_k^{(i)})_{i=1}^{\infty}$ is a Cauchy sequence in $\mathbb{R}$ for each index k, so it converges to some limit x_k. Moreover, taking the limit $j \to \infty$ in (1.47), we conclude that

$$|x_k^{(i)} - x_k| \;\le\; \epsilon \tag{1.48}$$

for each index k ($1 \le k \le n$) whenever $i \ge n_0$ (why? why did the $<$ turn into $\le$?). But this means that if we define $x = (x_1, \ldots, x_n)$, we have proven that $d_\infty(x^{(i)}, x) \le \epsilon$ whenever $i \ge n_0$. And since ϵ was arbitrary, this shows that $x^{(i)} \to x$ in the metric space $(\mathbb{R}^n, d_\infty)$. $\square$

Next week we shall use a similar technique to prove the completeness of the sequence spaces ℓ^∞, c_0, ℓ^1 and ℓ^2 as well as of the space $\mathcal{B}(A)$ of bounded functions on an arbitrary set A. The space $\mathcal{C}(A)$ of bounded continuous functions, equipped with the sup norm, is also complete, and we shall prove this a few weeks from now (it is not difficult). On the other hand, the spaces $\mathcal{C}[a, b]$ of bounded continuous functions with the L^1 or L^2 metric (Examples 9 and 10) are *not* complete, as we shall also show next week.

Completion of a metric space

Complete metric spaces are *much* more convenient for doing analysis than incomplete spaces — for the same reason that the real numbers $\mathbb{R}$ are much more convenient for doing analysis than the rational numbers $\mathbb{Q}$! It is therefore a very important fact that every metric space can be embedded in an essentially unique way into a complete metric space: this important construction is called the *completion of a metric space*. (It generalizes one of the ways of constructing the reals starting from the rationals.) An excellent treatment of this subject is given by Kreyszig, Sections 1.5 and 1.6, to which I have nothing (at least for now) to add. You should pay attention in particular to the proofs of completeness of the sequence spaces ℓ^∞, c_0, ℓ^1 and ℓ^2, and to the general construction of completing a metric space.