1 Introduction

This is a course on applied functional analysis in the context of optimization. The basic structure here is a vector space, i.e., a set whose elements can be added and mutliplied by scalars, and the two operations of vector addition and multiplication of vectors by scalars obey a couple of additional axioms which we will spell out precisely in due time. The notion of vector spaces is an abstraction of the familiar Euclidean space, for which we have a great deal of intuition already.

We start by listing some examples of optimization problems that will be encountered in the course. The discussion here is rather informal, and we will fill in all the details later on.

Example 1.1 (Allocation problems) Here, we seek a vector $x = (x_1, \ldots, x_n)^\top$ with real coordinates that would minimize the cost function

$$\langle c, x \rangle := \sum_{i=1}^{n} c_i x_i \tag{1}$$

subject to the constraints

$$Ax \leq b, \qquad Cx = d \tag{2}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $C \in \mathbb{R}^{k \times n}$, $d \in \mathbb{R}^k$ are the fixed problem specification. The notation $u \leq v$ for a pair of vectors $u = (u_1, \ldots, u_m)^\top$ and $v = (v_1, \ldots, v_m)^\top$ of the same dimension is shorthand for the m inequalities $u_1 \leq v_1, \ldots, u_m \leq v_m$. This is an instance of a *linear program*; the name comes from the fact that the cost function in (1) is linear in x, while the constraints in (2) are given by a finite collection of linear (in)equalities. The name "allocation problems" comes from operations research, where the ith coordinate x_i is the amount of the ith resource allocated to some process or operation, and we wish to minimize the total cost while satisfying a collection of constraints. A good reference is [BN01].

The object we are minimizing over in (1) is a vector of finite dimension n. We will also consider *infinite-dimensional* linear programs. The next example is the so-called *optimal transportation problem*.

Example 1.2 (Optimal transportation) Let p_0 and p_1 be two probability density functions (pdf's) on the real line, i.e., $p_0(x), p_1(x) \geq 0$ for all $x \in \mathbb{R}$, and

$$\int_{\mathbb{R}} p_0(x)\,\mathrm{d}x = \int_{\mathbb{R}} p_1(x)\,\mathrm{d}x = 1. \tag{3}$$

Let a nonnegative cost function $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be given (typically, one assumes more about c, e.g., continuity). We wish to minimize the cost function

$$\langle c, p \rangle := \int_{\mathbb{R} \times \mathbb{R}} c(x, y) p(x, y)\,\mathrm{d}x\,\mathrm{d}y \tag{4}$$

over all functions $p : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfying the following constraints: for all $x, y \in \mathbb{R}$,

$$p(x,y) \geq 0, \qquad \int_{\mathbb{R}} p(x,y')\,\mathrm{d}y' = p_0(x), \qquad \int_{\mathbb{R}} p(x',y)\,\mathrm{d}x' = p_1(y). \tag{5}$$

It is easy to verify that any p satisfying the constraints (5) is itself a probability density function over the plane $\mathbb{R}^2$; in fact, any such p defines a *joint probability distribution* of two real-valued random variables X and Y, such that X has pdf p_0 and Y has pdf p_1. If any such p achieves the minimum in (4), then we say that it solves the problem of *optimal transportation* between p_0 and p_1, with transportation cost $c(x,y)$. Note that here, despite the fact that we are optimizing over functions of continuous variables, the structure of the optimization problem is similar to the linear program considered in the first example: The objective functions in (1) and in (4) are both linear, and the constraints in (2) and in (5) are given by linear (in)equalities. If instead of pdf's on the real line we consider probability distributions over two finite sets, then the corresponding optimal transportation problem is a linear program; this is, in fact, the origin of modern theory of linear programming, through the work of the Soviet mathematician Leonid Kantorovich on the optimal transportation problem. The book [Vil03] is an excellent reference on optimal transportation problems.

Example 1.3 (Planning/control problems) In control problems, there is a dynamical system whose state evolves in time under the combined influence of its autonomous dynamical law and external inputs (controls), and the goal is to devise a control in order to steer the system towards some desired behavior. As an example, consider the problem of controlling the acceleration of a rocket to reach a desired altitute in a given time while minimizing the fuel expenditure. The dynamical model in question relates $y(t)$, the vertical displacement of the rocket at time t, to the external force $u(t)$, which here plays the role of control:

$$m\ddot{y}(t) = u(t) - mg, \qquad y(0) = 0 \tag{6}$$

where $m > 0$ is the mass of the rocket, g is the constant downward acceleration due to gravity, and $u(t)$ is the external force. The above equation (6) is simply Newton's law that expresses the total force at time t as the product of the mass m and the acceleration $\ddot{y}(t)$. The objective is to choose the control $u(t)$ as a function of time $t \in [0,T]$ to minimize $\int_0^T |u(t)|^2\,\mathrm{d}t$ subject to the dynamics (6) and to the endpoint constraint $y(T) = h$, where h is the desired altitude at time T. This is an infinite-dimensional constrained minimization problem over a suitable class of functions $u : [0,T] \to \mathbb{R}$. See [FR75] for a thorough treatment.

Example 1.4 (Estimation) In statistical estimation problems, the goal is to estimate the value of some random variable X on the basis of another random variable Y, which is viewed as a noise-corrupted version of X. For instance, if both X and Y are finite-dimensional random vectors with finite second moments, we may want to find a function g to minimize the expected squared error

$$J(g) := \mathbf{E}\left[|X - g(Y)|^2\right], \tag{7}$$

where $|\cdot|$ denotes the usual Euclidean norm. It is well-known (and we will prove it) that the optimal estimator is the conditional expectation $g^*(Y) = \mathbf{E}[X|Y]$. This optimization problem has a great

deal of geometric structure having to do with the fact that the Euclidean norm $|\cdot|$ is induced by an inner product, so we will be able to take advantage of such notions as orthogonality and orthogonal projection. In fact, we will be able to extend this problem formulation to the infinite-dimensional setting using the machinery of Hilbert spaces. A good exposition of minimum mean-square error estimation can be found in [Haj15].

Example 1.5 (Approximation) In approximation problems, we seek to minimize the error arising from approximating some 'complicated' object by the best element of a class of 'simpler' objects. As an example, let a continuous function $f : [0,1] \to \mathbb{R}$ be given. If we approximate f by some other continuous function $\hat{f} : [0,1] \to \mathbb{R}$, we incur the error $e : [0,1] \to \mathbb{R}$, i.e., $e(t) := f(t) - \hat{f}(t)$. Let $\mathcal{F}$ be some fixed class of continuous real-valued functions on $[0,1]$. Then we may seek $\hat{f} \in \mathcal{F}$ to minimize some criterion $J(e)$, which could have many possible forms, e.g.:

$$J(e) = \begin{cases} \max\limits_{t\in[0,1]} |e(t)| \\ \int_0^1 |e(t)|\,\mathrm{d}t \\ \int_0^1 |e(t)|^2\,\mathrm{d}t \end{cases} . \tag{8}$$

The approximating class $\mathcal{F}$ likewise can take many forms, e.g., it may consist of all polynomials of degree at most k, or of all splines (continuous pieceiwse polynomial functions), where each polynomial has degree at most k and there are at most m pieces, or of all two-layer neural nets with at most k hidden units. See [DL93] for a detailed exposition of function approximation.

Example 1.6 (Games) So far, we have discussed minimization problems, where there is only one decision variable (an element of a finite-dimensional or an infinite-dimensional vector space), and we aim to choose this decision variable so as to minimize some objective function while possibly satisfying some additional constraints. By contrast, game theory considers optimization problems with two or more decision variables, where the goal of optimizing over one of them may conflict with the goal of optimizing over the others. For instance, in a zero-sum two-player game, we have two decision sets X and Y, corresponding to decisions (moves, actions) of Players 1 and 2, respectively. Player 1 is the minimizing player, while Player 2 is the maximizing player. If Player 1 (resp., Player 2) makes the move $x \in \mathsf{X}$ (resp., $y \in \mathsf{Y}$), then Player 1 incurs the cost $c(x,y)$, while Player 2 receives the reward $c(x,y)$, or, equivalently, incurs the cost $-c(x,y)$. If the two players choose their moves simultaneously, then it makes sense for Player 1 to choose her move under the assumption of worst-case behavior by Player 2, i.e., Player 1 would want to choose $x \in \mathsf{X}$ to minimize $\max_{y\in\mathsf{Y}} c(x,y)$. Likewise, Player 2 might want to choose his move under the assumption that Player 1's choice will be as adversarial as possible, i.e., Player 1 will choose $y \in \mathsf{Y}$ to maximize $\min_{x\in\mathsf{X}} c(x,y)$. (We are assuming, of course, that all these minimizations and maximizations are permissible, etc.) Under these considerations, Player 1 (resp., Player 2) hopes to incur the minimax cost (resp., collect the maximin reward)

$$V^+ := \min_{x\in\mathsf{X}} \max_{y\in\mathsf{Y}} c(x,y) \quad \left(\text{resp., } V^- := \max_{y\in\mathsf{Y}} \min_{x\in\mathsf{X}} c(x,y)\right). \tag{9}$$

It is not hard to show that $V^- \leq V^+$; the best-case (or, rather, the least pessimistic) scenario is when these values coincide, i.e., $V^- = V^+$, and there exists a single choice of moves $x^* \in \mathsf{X}, y^* \in \mathsf{Y}$, such that

$$c(x^*, y) \leq c(x^*, y^*) \leq c(x, y^*), \qquad \forall x \in \mathsf{X}, y \in \mathsf{Y}. \tag{10}$$

In this case, we say that the pair $(x^*, y^*) \in \mathsf{X} \times \mathsf{Y}$ is the *saddlepoint* of the game, and the common value of V^- and V^+ is referred to as the *value of the game* and denoted by V.

The above set-up, where the two players make their moves deterministically, corresponds to *pure strategies* for the players. Unfortunately, most games do not admit saddlepoints in pure strategies. However, under fairly mild regularity conditions, a saddlepoint will always exist if we allow the players to *randomize* their choices, i.e., to use *mixed strategies*. For simplicity, let us assume that both X and Y are finite sets. A mixed strategy for Player 1 (resp., Player 2) is a probability distribution μ on X (resp., a probability distribution ν on Y). If the players adopt a mixed strategy pair (μ, ν), then we can express the expected cost incurred by Player 1 (or, equivalently, the reward collected by Player 2) as

$$J(\mu, \nu) := \sum_{x \in \mathsf{X}} \sum_{y \in \mathsf{Y}} c(x, y) \mu(x) \nu(y). \tag{11}$$

Then a deep result known as the *von Neumann minimax theorem* states that there always exists a saddlepoint in mixed strategies, i.e., there exists a pair of probability distributions (μ^*, ν^*), such that

$$J(\mu^*, \nu^*) = \min_{\mu} \max_{\nu} J(\mu, \nu) = \max_{\mu} \min_{\nu} J(\mu, \nu). \tag{12}$$

Eq. (12) is a manifestation of *duality*, where a minimization problem can be expressed as a maximization problem by changing the objective and/or the decision variables in a suitable way. We will see duality in action many times over this course. The book [BG54] is a classic reference on the theory of games in the context of statistical decision theory.

The above examples give only a partial view of a wide variety of optimization problems over vector spaces. One common thread that runs through all these problems is that the decision variable(s) are elements of some subset of a vector space, i.e., a set whose elements can be added and multiplied by scalars. The presence of constraints may complicate matters (i.e., it could very well be the case that the constraints in a given problem are not invariant under addition or scalar multiplication), but usually there is some additional property in place, such as convexity, compactness, etc. *Functional analysis*, which is a study of vector spaces from a unified perspective of algebra, geometry, and analysis, is the natural setting for all of these problems.

2 Vector spaces

We start by laying down some definitions. The first basic definition is that of a *vector space* (or *linear space*) over a field $\mathbb{K}$ (which will typically be the real number field $\mathbb{R}$ or the complex number field $\mathbb{C}$). The elements of $\mathbb{K}$ are called *scalars*.

Definition 2.1 *A* vector space *over* $\mathbb{K}$ *is a set* X *together with two binary operations* $\mathsf{X} \times \mathsf{X} \xrightarrow{+} \mathsf{X}$ *(vector addition) and* $\mathbb{K} \times \mathsf{X} \xrightarrow{\cdot} \mathsf{X}$ *(multiplication of vectors by scalars) that obey the following axioms:*

1. *For all* $x, y, u \in \mathsf{X}$, $x + y = y + x$ *(commutativity) and* $(x + y) + u = x + (y + u)$ *(associativity);*
2. *For all* $\alpha, \beta \in \mathbb{K}$ *and all* $x, y \in \mathsf{X}$, $\alpha(\beta x) = (\alpha\beta)x$ *(associativity),* $(\alpha + \beta)x = \alpha x + \beta x$*, and* $\alpha(x + y) = \alpha x + \alpha y$ *(distributivity);*
3. *There exists a unique element of* X *(the* zero vector*), which we denote by* 0*, such that the equation* $x + 0 = x$ *holds for all* $x \in \mathsf{X}$*;*
4. *For any* $x \in \mathsf{X}$*, the equation* $x + y = 0$ *has a unique solution.*

Remark 2.1 Strictly speaking, we should distinguish the zero vector (an element of X) from the zero scalar (an element of $\mathbb{K}$). However, it easily follows from the axioms that multiplying any nonzero vector x by the zero scalar 0 gives the zero vector. Indeed, using the distributive property, we have $x + 0 \cdot x = (1 + 0)x = x$, so $0 \cdot x = 0$ since the equation $x + 0 = x$ has a unique solution.

Remark 2.2 By the same token, we can denote the unique solution of the equation $x + y = 0$, for a fixed $x \in \mathsf{X}$, by $-x \equiv (-1)x$. Indeed, using distributivity again, we have $x + (-x) = x + (-1)x = (1 - 1)x = 0$, so $y = -x$.

Let us now illustrate the above abstract definition via several examples.

Example 2.1 ($\mathsf{X} = \mathbb{K}$) It is not hard to verify that the field of scalars $\mathbb{K}$, with its usual operations of addition and multiplication, is a vector space over itself.

Example 2.2 ($\mathsf{X} = \mathbb{K}^N$) For any $N \in \mathbb{N}$, let $\mathbb{K}^N$ be the collection of all N-tuples $x = (\xi_1, \ldots, \xi_N)$ with $\xi_k \in \mathbb{K}$ for all k, and with operations defined coordinatewise: for $x = (\xi_1, \ldots, \xi_N), y = (\eta_1, \ldots, \eta_N)$, $\alpha \in \mathbb{K}$, let

$$x + y := (\xi_1 + \eta_1, \ldots, \xi_N + \eta_N), \quad \alpha x := (\alpha\xi_1, \ldots, \alpha\xi_N). \tag{13}$$

Example 2.3 (infinite sequences) Consider infinite sequences $x = (\xi_1, \xi_2, \ldots)$ over $\mathbb{K}$. We can construct a wide variety of vector spaces consisting of subsets of such sequences. Some examples:

1. The space of all bounded sequences, i.e., all x such that $\max_{n \geq 1} |\xi_n| < \infty$;
2. The space of all sequences converging to zero, i.e., all x such that $\lim_{n \to \infty} \xi_n = 0$;
3. The space of all sequences x, such that all the entries ξ_n are zero except for finitely many values of n;
4. The space of all square-summable sequences, i.e., all x such that the infinite series $\sum_{n=1}^{\infty} |\xi_n|^2$ converges.

It is a useful exercise to verify that all of the above are vector spaces over $\mathbb{K}$ with operations defined coordinatewise.

Example 2.4 ($C[a,b]$) Given two real numbers $-\infty < a < b < \infty$, let $C[a,b]$ denote the set of all continuous functions $x : [a,b] \to \mathbb{R}$. Then it is easy to verify that this is a vector space over the reals, with operations defined pointwise: for all $\alpha \in \mathbb{R}$, all $x, y \in C[a,b]$, and all $t \in [a,b]$,

$$(x+y)(t) := x(t) + y(t), \qquad (\alpha x)(t) := \alpha x(t). \tag{14}$$

Example 2.5 (random variables) Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A *random variable* is a function $X : \Omega \to \mathbb{R}$, such that the set $\{\omega \in \Omega : X(\omega) \leq r\}$ belongs to the σ-algebra $\mathcal{F}$ for every $r \in \mathbb{R}$. The collection of all such random variables is a vector space over the reals, with operations defined pointwise:

$$(X+Y)(\omega) := X(\omega) + Y(\omega), \qquad (\alpha X)(\omega) := \alpha X(\omega). \tag{15}$$

3 Normed vector spaces and convergence

The next order of business is to introduce the notion of *length* of a vector, again with the goal of preserving as much Euclidean intuition as possible. In what follows, we will stick to real vector spaces, but everything applies to complex vector spaces verbatim.

Definition 3.1 *Let X be a real vector space. A* norm *on X is a function $\|\cdot\| : \mathsf{X} \to \mathbb{R}$ that obeys the following axioms:*

1. *Positive definiteness — $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$.*
2. *Homogeneity — for all $\alpha \in \mathbb{R}$ and $x \in \mathsf{X}$, $\|\alpha x\| = |\alpha| \cdot \|x\|$.*
3. *Triangle inequality — for all $x, y \in \mathsf{X}$, $\|x+y\| \leq \|x\| + \|y\|$.*

We often use the notation $(\mathsf{X}, \|\cdot\|)$ whenever we need to emphasize the norm $\|\cdot\|$.

Remark 3.1 Evidently, $\|-x\| = \|(-1)x\| = \|x\|$ by homogeneity. For any two vectors $x, y \in \mathsf{X}$, the norm $\|x-y\|$ can be naturally thought as the *distance* between x and y.

Remark 3.2 A simple induction argument shows that, for any finite collection of vectors $x_1, \ldots, x_n$,

$$\left\| \sum_{i=1}^n x_i \right\| \leq \sum_{i=1}^n \|x_i\|. \tag{16}$$

The following consequence of the triangle inequality is useful:

Proposition 3.1 (Generalized triangle inequality) *For all $x, y \in \mathsf{X}$,*

$$\big|\|x\| - \|y\|\big| \leq \|x \pm y\| \leq \|x\| + \|y\|. \tag{17}$$

Proof: The upper bound on $\|x \pm y\|$ is a simple consequence of the triangle inequality and the fact that $\|-x\| = \|x\|$. For the lower bound, observe that

$$\|x\| = \|(x-y)+y\| \leq \|x-y\| + \|y\|, \tag{18}$$

which implies that $\|x-y\| \geq \|x\| - \|y\|$. Similarly,

$$\|y\| = \|y-x+x\| \leq \|x-y\| + \|x\|, \tag{19}$$

which implies that $\|x-y\| \geq \|y\| - \|x\|$. This proves the inequality $\|x-y\| \geq \big|\|x\| - \|y\|\big|$. The corresponding lower bound on $\|x+y\|$ is proved analogously. ■

Let us go through a few examples of normed spaces:

Example 3.1 Let $\mathsf{X} = \mathbb{R}^N$, and for a vector $x = (\xi_1, \ldots, \xi_N)$ let

$$|x|_\infty := \max_{1 \leq k \leq N} |\xi_k|. \tag{20}$$

This is known as the ℓ^∞ *norm*. It is an easy exercise to verify that this is, indeed, a norm.

Example 3.2 Again taking $\mathsf{X} = \mathbb{R}^N$, consider

$$|x| := \left(\sum_{k=1}^N |\xi_k|^2 \right)^{1/2} \tag{21}$$

– this is the ℓ^2*, or Euclidean, norm*, often denoted also as $|x|_2$ or $\|x\|_2$. The first two properties are easy to verify, so we will establish only the triangle inequality. To that end, consider two vectors $x = (\xi_1, \ldots, \xi_N)$ and $y = (\eta_1, \ldots, \eta_N)$. Then

$$|x+y|^2 = \sum_{k=1}^N (\xi_k + \eta_k)^2 \tag{22}$$

$$= \sum_{k=1}^N \left(\xi_k^2 + 2\xi_k\eta_k + \eta_k^2 \right) \tag{23}$$

$$= |x|^2 + 2\sum_{k=1}^N \xi_k\eta_k + |y|^2. \tag{24}$$

By the Cauchy–Schwarz inequality, we can upper-bound the sum in the middle term by

$$\sum_{k=1}^N \xi_k\eta_k \leq \left(\sum_{k=1}^N \xi_k^2 \right)^{1/2} \left(\sum_{k=1}^N \eta_k^2 \right)^{1/2} = |x| \cdot |y|. \tag{25}$$

Using this in (24), we obtain the inequality

$$|x+y|^2 \leq |x|^2 + 2|x||y| + |y|^2 = (|x| + |y|)^2. \tag{26}$$

Since both sides are nonnegative, we get the triangle inequality upon taking the square root.

The two norms on $\mathbb{R}^N$ introduced above can be bounded in terms of each other:

Proposition 3.2 *For any vector $x \in \mathbb{R}^N$,*

$$|x|_\infty \leq |x| \leq \sqrt{N}|x|_\infty. \tag{27}$$

Proof: For the first inequality:

$$|x|_\infty^2 = \max_{1\leq k\leq N} |\xi_k|^2 \tag{28}$$

$$\leq \sum_{k=1}^N |\xi_k|^2 \tag{29}$$

$$= |x|^2. \tag{30}$$

For the second inequality:

$$|x|^2 = \sum_{k=1}^N |\xi_k|^2 \tag{31}$$

$$\leq N \max_{1\leq k\leq N} |\xi_k|^2 \tag{32}$$

$$= N|x|_\infty^2. \tag{33}$$

■

The inequality (27) expresses the important fact that the ℓ^∞ and the ℓ^2 norms on $\mathbb{R}^N$ are *equivalent*. Later on, we will prove a more general result about norms on finite-dimensional vector spaces that contains this as a special case.

Example 3.3 The space $\mathsf{X} = C[a, b]$ of continuous real-valued functions on a bounded interval $[a, b]$ can be normed in a variety of ways. For example, we can consider the sup (or uniform) norm

$$\|x\|_\infty := \max_{t\in[a,b]} |x(t)|, \tag{34}$$

the 1-norm

$$\|x\|_1 := \int_a^b |x(t)|\,\mathrm{d}t, \tag{35}$$

or the 2-norm

$$\|x\| \equiv \|x\|_2 := \left(\int_a^b |x(t)|^2\,\mathrm{d}t\right)^{1/2}. \tag{36}$$

Here, we no longer have an analog of Proposition 3.2. Indeed, while it is easy to upper-bound the 2 norm by the sup norm,

$$\|x\|^2 = \int_a^b |x(t)|^2\,\mathrm{d}t \leq (b-a) \max_{t\in[a,b]} |x(t)|^2 = (b-a)\|x\|_\infty, \tag{37}$$

there is no way of upper-bounding $\|x\|_\infty$ by $\|x\|$ without imposing extra assumptions on x.

The introduction of a norm allows us to speak not only about the 'length' of a vector or the 'distance' between two vectors, but also about the notion of convergence of a sequence of elements of a vector space.

Definition 3.2 *Let $(x_n)_{n\geq 1}$ be a sequence of vectors in a normed space $(\mathsf{X}, \|\cdot\|)$. We say that (x_n)* converges *to $x \in \mathsf{X}$ as $n \to \infty$ if*

$$\lim_{n\to\infty} \|x_n - x\| = 0. \tag{38}$$

We also denote this by $x_n \to x$ (sometimes adding 'as $n \to \infty$' to avoid confusion).

Thus, the convergence of (x_n) to x in an abstract normed space $(\mathsf{X}, \|\cdot\|)$ reduces to the usual notion of convergence of sequences of reals. Recalling the definition of the latter, we can say that $x_n \to x$ iff, for any $\varepsilon > 0$, there exists some $n_0 = n_0(\varepsilon) \geq 1$, such that

$$\sup_{n\geq n_0} \|x_n - x\| < \varepsilon. \tag{39}$$

Here are some immediate consequences of convergence:

Proposition 3.3 *Let (x_n) be a sequence of elements of a normed space $(\mathsf{X}, \|\cdot\|)$, such that $x_n \to x$.*

1. *Limits are unique: if we also have $x_n \to x'$, then $x' = x$.*
2. *Convergence of norms: $\|x_n\| \to \|x\|$.*
3. *Convergent sequences are bounded in norm: $\sup_{n\geq 1} \|x_n\| < \infty$.*

Proof:

1. For an arbitrary $\varepsilon > 0$, there exists some $n_0 \geq 1$, such that $\|x_n - x\| < \varepsilon$ and $\|x_n - x'\| < \varepsilon$ for all $n \geq n_0$. Then

$$\|x - x'\| \leq \|x_n - x\| + \|x_n - x'\| < 2\varepsilon. \tag{40}$$

   Since $\varepsilon > 0$ was arbitrary, we conclude that $\|x - x'\| = 0$, i.e., $x = x'$.

2. Fix some $\varepsilon > 0$ and let $n_0 \geq 1$ be large enough, so that $\|x_n - x\| < \varepsilon$ for all $n \geq n_0$. By the generalized triangle inequality,

$$\sup_{n\geq n_0} \big|\|x_n\| - \|x\|\big| \leq \sup_{n\geq n_0} \|x_n - x\| < \varepsilon, \tag{41}$$

   so $\|x_n\| \to \|x\|$.

3. Since $x_n \to x$, there exists $n_0 \geq 1$, such that $\sup_{n\geq n_0} \|x_n - x\| < 1$. Let $R := \max_{n\leq n_0 - 1} \|x_n\|$. Then

$$\|x_n\| \leq \|x_n - x\| + \|x\|, \tag{42}$$

   and therefore $\|x_n\| \leq \max\big\{R, 1 + \|x\|\big\} < \infty$.

■

Another key property of convergent sequences is that their elements get closer together as we take larger and larger indices:

Definition 3.3 *A sequence (x_n) of vectors in a normed space $(\mathsf{X}, \|\cdot\|)$ is a* Cauchy sequence *if, for any $\varepsilon > 0$, there exists an $n_0 \geq 1$ such that*

$$\sup_{m,n \geq n_0} \|x_m - x_n\| < \varepsilon. \tag{43}$$

Proposition 3.4 *Any convergent sequence is Cauchy.*

Proof: Suppose $x_n \to x$. Given an $\varepsilon > 0$, choose $n_0 \geq 1$ so that $\|x_n - x\| < \varepsilon$ for all $n \geq n_0$. Then, for all $m, n \geq n_0$,

$$\|x_m - x_n\| \leq \|x_m - x\| + \|x_n - x\| < 2\varepsilon. \tag{44}$$

■

However, the converse is not always true — a Cauchy sequence is not necessarily convergent (we will see examples of this soon). The class of normed spaces where all Cauchy sequences have limits is important:

Definition 3.4 *A normed space $(\mathsf{X}, \|\cdot\|)$ is called* complete *if any Cauchy sequence (x_n) in X has a limit.*

Complete normed spaces are also referred to as *Banach spaces* after the Polish mathematician Stefan Banach who had introduced and studied them. Let us see some examples and counterexamples.

Example 3.4 ($\mathbb{R}$) The real numbers with the usual norm given by the absolute value is a Banach space. This is a consequence of the Cauchy convergence criterion: any Cauchy sequence of real numbers has a limit. (This, in turn, is a consequence of the so-called *completeness axiom* for the real numbers.)

Example 3.5 ($\mathbb{R}^N, |\cdot|_\infty$) The vector space $\mathbb{R}^N$ with the ℓ^∞ norm $|x|_\infty := \max_{1 \leq k \leq N} |\xi_k|$ is a Banach space.

Proof: Let $(x_n)_{n \geq 1}$ be a sequence of elements of $\mathbb{R}^N$, with $x_n = (\xi_{n1}, \ldots, x_{nN})$. We first show that this sequence converges to $x = (\xi_1, \ldots, \xi_N)$ if and only if

$$\xi_{nk} \to x_k, \qquad \text{as } n \to \infty \tag{45}$$

for all $k = [N] := \{1, \ldots, N\}$. Indeed, if (45) holds for all k, then for every $\varepsilon > 0$ and every $k \in [N]$ there exists some $n_k \geq 1$, such that $|\xi_{nk} - x_n| < \varepsilon$ for all $n \geq n_k$. Then evidently

$$|x_n - x|_\infty = \max_{1 \leq k \leq N} |\xi_{nk} - \xi_k| < \varepsilon, \qquad \forall n \geq \max\{n_1, \ldots, n_k\} \tag{46}$$

so $x_n \to x$ in $|\cdot|_\infty$. Conversely, if $x_n \to x$, then, for every $k \in [N]$,

$$|\xi_{nk} - \xi_k| \leq \max_{1 \leq k \leq N} |\xi_{nk} - \xi_k| \xrightarrow{n\to\infty} 0. \tag{47}$$

We now show that any Cauchy sequence (x_n) in $(\mathbb{R}^N, |\cdot|_\infty)$ has a limit $x \in \mathbb{R}^N$. An argument similar to the one above shows that (x_n) is Cauchy if and only if each of the N coordinate sequences $(\xi_{nk})_{n\geq 1}$, $k \in [N]$, is Cauchy. But each such sequence is a Cauchy sequence of real numbers, and therefore for every k there exists $\xi_k \in \mathbb{R}$, such that $\xi_{nk} \to \xi_k$ as $n \to \infty$. Therefore, letting $x := (\xi_1, \ldots, \xi_N)$, we see that $x_n \to x$ as $n \to \infty$. ■

Example 3.6 $(\mathbb{R}^N, |\cdot|)$ Consider now the vector space $\mathbb{R}^N$ with the ℓ^2 (Euclidean) norm $|x| = \left(\sum_{k=1}^N \xi_k^2\right)^{1/2}$. We claim that this is also a real Banah space.

Proof: Recall the equivalence of the ℓ^2 and the ℓ^∞ norms in $\mathbb{R}^N$ (Prop. 3.2): for every $x \in \mathbb{R}^N$,

$$|x|_\infty \leq |x| \leq \sqrt{N}|x|_\infty. \tag{48}$$

Let (x_n) be a Cauchy sequence in $(\mathbb{R}^N, |\cdot|)$. Then it is readily verified using the above inequality that it is also Cauchy in $(\mathbb{R}^N, |\cdot|_\infty)$. Thus, there exists some $x \in \mathbb{R}^N$, such that

$$\lim_{n\to\infty} |x_n - x|_\infty = 0. \tag{49}$$

Consequently, given an $\varepsilon > 0$, there exists some n_0 such that $|x_n - x|_\infty < \varepsilon$ for all $n \geq n_0$ and thus

$$|x_n - x| \leq \sqrt{N}|x_n - x|_\infty < \sqrt{N}\varepsilon, \qquad n \geq n_0. \tag{50}$$

This implies that (x_n) converges to x in ℓ^2 norm. In fact, the equivalence of the two norms gives a much stronger result: a sequence (x_n) in $\mathbb{R}^N$ converges to x in ℓ^2 iff it converges to x in ℓ^∞. Thus,

$$\lim_{n\to\infty} |x_n - x| = 0 \quad \Longleftrightarrow \quad \lim_{n\to\infty} \xi_{nk} = \xi_k, \, \forall k \in [N]. \tag{51}$$

■

Example 3.7 $(C[a,b], \|\cdot\|_\infty)$ Consider the space $C[a,b]$ of continuous functions $x : [a,b] \to \mathbb{R}$ with the sup norm $\|x\|_\infty := \max_{t\in[a,b]} |x(t)|$. We claim that this is a real Banach space.

Proof: Let (x_n) be a sequence of functions in $C[a,b]$ that converges to $x \in C[a,b]$ in the $\|\cdot\|_\infty$ norm:

$$\lim_{n\to\infty} \|x_n - x\|_\infty = \lim_{n\to\infty} \max_{t\in[a,b]} |x_n(t) - x(t)| = 0 \tag{52}$$

— in other words, the convergence is *uniform* in $t \in [a,b]$: for any $\varepsilon > 0$ there exists some $n_0 \geq 1$ *independent of* t such that

$$|x_n(t) - x(t)| < \varepsilon, \qquad \forall t \in [a,b], n \geq n_0. \tag{53}$$

Now let (x_n) be a Cauchy sequence w.r.t. the $\|\cdot\|_\infty$ norm. Then, for every $t \in [a,b]$, the sequence $(x_n(t))$ is a Cauchy sequence of real numbers and therefore has a limit, which we denote by $x(t)$. We will now prove two things: (a) the convergence is uniform, i.e.,

$$\lim_{n\to\infty} \max_{t\in[a,b]} |x_n(t) - x(t)| = 0; \tag{54}$$

and (b) the function $t \mapsto x(t)$ is continuous.

To prove that the convergence is uniform, fix some $\varepsilon > 0$. Since (x_n) is a Cauchy sequence, there exists some $n_0 \geq 1$, such that $\|x_m - x_n\|_\infty < \varepsilon$ for all $m, n \geq n_0$. For any such n and any $t \in [a,b]$,

$$\begin{aligned} |x_n(t) - x(t)| &\leq |x_m(t) - x_n(t)| + |x_m(t) - x(t)| &(55)\\ &\leq \|x_m - x_n\|_\infty + |x_m(t) - x(t)|. &(56)\end{aligned}$$

Now, while keeping $n \geq n_0$ *fixed*, we can choose some $m_0 = m_0(\varepsilon, t)$, such that $|x_m(t) - x(t)| < \varepsilon$ for all $m \geq m_0$. This is possible since $x_m(t) \to x(t)$ for every $t \in [a,b]$. Then, for any $n \geq n_0$ and $m \geq \max\{n_0, m_0\}$, we see that $\|x_m - x_n\|_\infty + |x_m(t) - x(t)| < 2\varepsilon$, and therefore

$$\max_{t\in[a,b]} |x_n(t) - x(t)| < 2\varepsilon, \qquad \forall n \geq n_0. \tag{57}$$

This proves that the convergence is uniform, but we still need to show that $x(t)$ is a continuus function of t. To that end, for any $t \in (a,b)$ and $\delta > 0$ such that $t + \delta \in [a,b]$, we have

$$|x(t+\delta) - x(t)| \leq |x(t+\delta) - x_n(t+\delta)| + |x_n(t+\delta) - x_n(t)| + |x_n(t) - x(t)| \tag{58}$$

for all $n \geq 1$. Now, since $x_n(t) \to x(t)$ uniformly in t, for any $\varepsilon > 0$ there exists some $n_0 \geq 1$, such that $\max_{r\in[a,b]} |x_n(r) - x(r)| < \varepsilon$ for all $n \geq n_0$. Now, since all the functions x_n are continuous, we can choose a small enough $\delta > 0$ (depending on ε and on n), such that $|x_n(t+\delta) - x_n(t)| < \varepsilon$. Putting this together, we see that

$$|x(t+\delta) - x_n(t+\delta)| + |x_n(t+\delta) - x_n(t)| + |x_n(t) - x(t)| < 3\varepsilon \tag{59}$$

for $\delta > 0$ small enough. This proves that $x(t)$ is, indeed, continuous. ■

Example 3.8 ($C[a,b], \|\cdot\|_1$) As we saw earlier, we can define many other norms on $C[a,b]$, such as the 1-norm

$$\|x\|_1 := \int_a^b |x(t)|\,\mathrm{d}t. \tag{60}$$

Now, it turns out that the normed space $(C[a,b], \|\cdot\|_1)$ is *not* complete, so this is an example of a normed space which is not a Banach space.

Proof: To prove this fact, it suffices to provide an example of a sequence of functions $x_n \in C[a,b]$,

which is Cauchy in $\|\cdot\|_1$, but which does not have a continuous limit. We define our functions as follows:

$$x_n(t) := \begin{cases} 0, & 0 \le t \le \frac{1}{2} - \frac{1}{2n}, \\ 2nt + 1 - n, & \frac{1}{2} - \frac{1}{2n} \le t \le \frac{1}{2}, \\ 1, & \frac{1}{2} \le t \le 1 \end{cases} \tag{61}$$

It is not hard to verify that these functions are continuous and form a Cauchy sequence. On the other hand, while the sequence (x_n) has a limit as $n \to \infty$, this limit is not continuous (it is, rather, a step function with a discontinuity at $t = \frac{1}{2}$). Thus, $C[a,b]$ with the $\|\cdot\|_1$ norm is not complete. Later on, we will see that, in order to have completeness in the $\|\cdot\|_1$ norm, we will need to enlarge our space to admit functions that are discontinuous yet integrable (in the sense of Lebesgue) on $[a,b]$. This space, denoted by $L^1[0,1]$, is a Banach space with the norm $\|\cdot\|_1$. ■

References

[BG54] David H. Blackwell and Meyer A. Girshick. *Theory of Games and Statistical Decisions*. Wiley, 1954.

[BN01] Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization: Analyais, Algorithms, and Engineering Applications*. SIAM, 2001.

[DL93] Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*. Springer, 1993.

[FR75] Wendell H. Fleming and Raymond W. Rishel. *Deterministic and Stochastic Optimal Control*. Springer, 1975.

[Haj15] Bruce Hajek. *Random Processes for Engineers*. Cambridge University Press, 2015.

[Vil03] Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.