Math 286
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Fall Semester 2021

Outline

Today's Lecture: Preparations for the Existence and Uniqueness Theorem

Problem Restatement

Consider an explicit first-order ODE $y' = f(t, y)$ with a continuous function $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^2$ open, and the corresponding initial value problems $y' = f(t, y) \wedge y(t_0) = y_0$ for $(t_0, y_0) \in D$.

Observation

$\phi\colon I \to \mathbb{R}$ is a solution of $y' = f(t, y) \wedge y(t_0) = y_0$ if and only if the graph $\mathrm{G}_\phi = \{(t, \phi(t)); t \in I\}$ of ϕ is contained in D and

$$\phi(t) = y_0 + \int_{t_0}^{t} \phi'(s)\mathrm{d}s = y_0 + \int_{t_0}^{t} f\big(s, \phi(s)\big)\,\mathrm{d}s$$

for all $t \in I$. Here $I \subseteq \mathbb{R}$ is an interval containing t_0 in its interior.

Equivalently, $\phi(t)$ is a fixed point ("fixed function") of the operator $\phi \mapsto T\phi$ defined by

$$(T\phi)(t) = y_0 + \int_{t_0}^{t} f\big(s, \phi(s)\big)\,\mathrm{d}s, \quad t \in I.$$

As domain of T we can take the set of continuous functions $\phi\colon I \to \mathbb{R}$ with $\mathrm{G}_\phi = \{(t, \phi(t)); t \in I\} \subseteq D$.

Thus the (local) existence of solutions of the IVP $y' = f(t, y) \wedge y(t_0) = y_0$ reduces to the following

Problem

Given $(t_0, y_0) \in D$, show that there exists an interval $I = (t_0 - \delta, t_0 + \delta)$, $\delta > 0$, such that the corresponding operator T (which depends on I) has a fixed point.

The Existence Theorem for solutions of 1st-order ODE's (and ODE systems) will be proved in this way, but the proof can be given only after several further preparations.

The Uniqueness Theorem is easier to prove and essentially requires only to find the correct condition on the function $f(t, y)$ that implies the uniqueness of solutions. But the proof is also far from being trivial, as you will see.

Order Reduction

Now consider an explicit n-th order ODE $y^{(n)} = f(t, y, y', \ldots, y^{(n-1)})$ with a continuous function $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^{n+1}$ open, and the corresponding initial value problems obtained by prescribing $y^{(i)}(t_0) = y_i$ for $0 \le i \le n-1$ for some $(t_0, y_0, \ldots, y_{n-1}) \in D$.

Observation

Writing the ODE in the vectorial form

$$\begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{pmatrix}' = \begin{pmatrix} y' \\ y'' \\ \vdots \\ y^{(n)} \end{pmatrix} = \begin{pmatrix} y' \\ y'' \\ \vdots \\ f(t, y, y', \ldots, y^{(n-1)}) \end{pmatrix},$$

we see that it is equivalent to the first-order ODE system $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$ with $\mathbf{f}\colon D \to \mathbb{R}^n$ defined by

$$\mathbf{f}(t, y_0, \ldots, y_{n-1}) = \begin{pmatrix} y_1 \\ \vdots \\ y_{n-1} \\ f(t, y_0, y_1, \ldots, y_{n-1}) \end{pmatrix}.$$

Order Reduction Cont'd

This is so because $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$ written out in full means

$$\begin{pmatrix} y_0' \\ y_1' \\ \vdots \\ y_{n-2}' \\ y_{n-1}' \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ f(t, y_0, y_1, \ldots, y_{n-1}) \end{pmatrix},$$

and a solution n-tuple $(y_0, y_1, \ldots, y_{n-1})$ must satisfy

$$\begin{aligned} y_1 &= y_0', \\ y_2 &= y_1' = y_0'', \\ &\vdots \\ y_{n-1} &= y_0^{(n-1)}, \quad \text{and hence} \\ y_0^{(n)} &= y_{n-1}' = f(t, y_0, y_1, \ldots, y_{n-1}) = f(t, y_0, y_0', \ldots, y_0^{(n-1)}). \end{aligned}$$

In other words, the first coordinate function is a solution of the n-th order ODE and the remaining coordinate functions are its derivatives up to order $n-1$.

Order Reduction Cont'd

For the corresponding IVP's the same reduction applies:
A solution of the vectorial IVP

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}^0 = (y_0^0, y_1^0, \ldots, y_{n-1}^0)$$

has as its first component function $y_0(t)$ a solution of the n-th order IVP

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}), \quad y^{(i)}(t_0) = y_i^0 \text{ for } 0 \le i \le n-1.$$

Conclusion

Extending the scope to systems of ODE's allows us to restrict attention to first-order systems only.

The operator view applies also to this case and shows that a solution of $\mathbf{y}' = \mathbf{f}(t, \mathbf{y}) \wedge \mathbf{y}(t_0) = \mathbf{y}^0$ satisfies

$$\mathbf{y}(t) = \mathbf{y}^0 + \int_{t_0}^{t} \mathbf{f}\big(s, \mathbf{y}(s)\big)\,\mathrm{d}s\,.$$

and hence is a fixed point of the operator T defined by $(T\phi)(t) = \mathbf{y}^0 + \int_{t_0}^{t} \mathbf{f}\big(s, \phi(s)\big)\,\mathrm{d}s$.

For the subsequent development it is instructive to recall a similar setting from Calculus I, where solving a fixed-point equation for a certain "operator" (map) was also required:

Newton's Method for finding roots (cf. [Ste16], Ch. 4.8)

Suppose we want to compute a solution of an univariate equation like $\sin(x) = 1/3$. This equation can be rewritten as $f(x) = 0$ with $f(x) = \sin x - 1/3$ and solved as follows.

Suppose we know already a good approximation x_n to the unknown root x^* of f. It is then reasonable to replace $f(x)$ by its linear approximation $\ell(x)$ in x_n and take the root of ℓ as new (hopefully better) approximation to x^*.

$$\ell(x) = f(x_n) + f'(x_n)(x - x_n) = 0 \iff x = x_n - \frac{f(x_n)}{f'(x_n)} =: x_{n+1}$$

Repeating this step gives a sequence $x_0, x_1, x_2, \ldots$, which is determined by the recurrence relation $x_{n+1} = x_n - f(x_n)/f'(x_n)$ and the initial value x_0.

After introducing the operator $T(x) = x - f(x)/f'(x)$, the recurrence relation becomes $x_{n+1} = T(x_n)$.

Newton's Method cont'd

We are interested in the case where the sequence (x_n) converges in $\mathbb{R}$, say to x^*. Passing to the limit in $x_{n+1} = T(x_n)$ and using continuity of T (which requires that f is C^1 and f' has no zero "nearby"), we obtain

$$x^* = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} T(x_n) = T\left(\lim_{n\to\infty} x_n\right) = T(x^*).$$

$\Longrightarrow$ x^* is a fixed point of T.
$\Longrightarrow$ x^* is a root of f, since $T(x) = x - f(x)/f'(x) = x$ is equivalent to $f(x) = 0$.

Thus (x_n) can only converge to a root of f. But how can we be sure that the sequence actually converges (or, rather, how to choose the starting value x_0, so that the sequence must converge?).

First answer: Suppose we know already that f has a root x^*. (For example, if $a < b$ are such that $f(a) < 0$, $f(b) > 0$ then the Intermediate Value Theorem implies $f(x^*) = 0$ for some $x^* \in (a, b)$.)

$$\Longrightarrow\ x_{n+1} - x^* = T(x_n) - T(x^*) = T'(\xi_n)(x_n - x^*)$$

for some ξ_n between x^* and x_n.

Newton's Method cont'd

Since

$$T'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

we have $T'(x^*) = 0$, and hence (provided T' is continuous, which requires f to be of class C^2) $|T'(x)|$ is very small near x^*.
$\Longrightarrow$ (x_n) will converge rapidly to x^* if the starting value x_0 is sufficiently close to x^*.
For example, suppose we know that $x^* \in (a, b)$, $|T'(x)| \leq \frac{1}{2}$ for $x \in [a, b]$, and $T(a) \leq b$. Then the iteration with initial value $x_0 = a$ will converge to x^*. (For the proof consider the sign of $T'(\xi_n)$.)

Speed of convergence: 2nd-order Talor approximation of T in x^* gives, using $T'(x^*) = 0$,

$$x_{n+1} - x^* = \frac{T''(\xi_n)}{2}(x_n - x^*)^2,$$

again with ξ_n between x^* and x_n; moreover, if f is of class C^3 then $T''(\xi_n) \to T''(x^*) = f''(x^*)/f'(x^*)$.
This is called *quadratic convergence* and says that the number of correct digits in the decimal expansion of x_n essentially doubles at every iteration.

Newton's Method cont'd

Second answer: Suppose we don't yet know that f has a root. In this case we consider the difference

$$x_{n+1} - x_n = T(x_n) - T(x_{n-1}) = T'(\xi_n)(x_n - x_{n-1})$$

with ξ_n between x_{n-1} and x_n.
Further we suppose that there exists a constant $C < 1$ such that $|T'(\xi_n)| \le C$ for all n. (For example, this holds if $|T'(x)| \le C < 1$ on $[a, b]$ and $x_n \in [a, b]$ for all n.) Then, using induction, we obtain

$$\begin{aligned}
|x_{n+1} - x_n| &\le C^n |x_1 - x_0|, \\
|x_{n+k} - x_n| &\le \sum_{i=1}^{k} |x_{n+i} - x_{n+i-1}| \\
&\le (C^n + C^{n+1} + \cdots + C^{n+k-1}) |x_1 - x_0| \\
&\le \left(\sum_{i=n}^{\infty} C^i \right) |x_1 - x_0| = \frac{C^n}{1 - C} |x_1 - x_0|.
\end{aligned}$$

Since $\lim_{n\to\infty} C^n = 0$, given $\epsilon > 0$ we can find a response $N \in \mathbb{N}$ such that $|x_m - x_n| < \epsilon$ whenever $m, n > N$. This says that the sequence (x_n) is a Cauchy sequence and hence converges in $\mathbb{R}$.

Definition

A real-valued sequence (a_n) is said to be a *Cauchy sequence* (or to satisfy the *Cauchy criterion* for convergence) if for every $\epsilon > 0$ there exists $N = N_\epsilon \in \mathbb{N}$ such that $|x_m - x_n| < \epsilon$ whenever $m, n > N$.

Theorem

Every Cauchy sequence in $\mathbb{R}$ converges.

Proof.

We have stated and proved this theorem in Calculus III. Here is a different proof: Given a Cauchy sequence (a_n), define two further sequences (ℓ_n), (u_n) by

$$\ell_n = \inf\{a_n, a_{n+1}, a_{n+2}, \dots\},$$
$$u_n = \sup\{a_n, a_{n+1}, a_{n+2}, \dots\}.$$

$$\Longrightarrow\ \ell_1 \le \ell_2 \le \cdots \le \ell_n \le a_n \le u_n \le u_{n-1} \le \dots u_1.$$

Since (ℓ_n) is non-decreasing and bounded from above by u_1, say, the limit $\ell = \lim_{n\to\infty} \ell_n$ exists in $\mathbb{R}$. Similarly, $u = \lim_{n\to\infty} u_n$ exists in $\mathbb{R}$. We claim that $\ell = u$.

Proof cont'd.

Consider $\epsilon > 0$. Then for $n = N_\epsilon + 1$ and $m > n$ we have

$$\begin{aligned} & a_n - \epsilon < a_m < a_n + \epsilon; \\ \Longrightarrow\ & a_n - \epsilon \le \ell_n \le u_n \le a_n + \epsilon. \end{aligned}$$

Hence $u_n - \ell_n \le 2\epsilon$, which (since $\epsilon > 0$ is arbitrary) implies $\ell = u$.

Finally we can apply the squeezing theorem to conclude from $\ell_n \le a_n \le u_n$ that $\lim_{n\to\infty} a_n = \ell = u$ as well. □

Remark

The definition of Cauchy sequences makes also sense for the Euclidean spaces $\mathbb{R}^d$, $d > 1$, and in particular for $\mathbb{C} \triangleq \mathbb{R}^2$. An easy adaption of the previous proof shows that Cauchy sequences in $\mathbb{R}^d$ converge as well; cf. next exercise.

Exercise

Show that every Cauchy sequence $(\mathbf{x}^{(n)})$ in $\mathbb{R}^d$, $d > 1$, converges.

Hint: Show first that for $1 \le i \le d$ the i-th coordinate sequence of $(\mathbf{x}^{(n)})$, which is defined as $x_i^{(1)}, x_i^{(2)}, x_i^{(3)}, \ldots$ where $\mathbf{x}^{(n)} = (x_1^{(1)}, \ldots, x_d^{(n)})$, is a Cauchy sequence in $\mathbb{R}$.

Review of Differentiable Multivariate Functions

Recall that a function $f\colon D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$, is differentiable in a point $\mathbf{x}_0 \in D$ (which must be an inner point of D) if $f(\mathbf{x})$ can be linearly approximated near $\mathbf{x}_0$ with an error $\mathrm{o}(\mathbf{x} - \mathbf{x}_0)$; more precisely, if there exists a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + L(\mathbf{h}) + \mathrm{o}(\mathbf{h}) \quad \text{for } \mathbf{h} \to \mathbf{0}$$

or, equivalently, $\lim_{\mathbf{h}\to\mathbf{0}} |f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - L(\mathbf{h})| \,/\, |\mathbf{h}| = 0$.

If applicable, the linear map L is uniquely determined by this condition. It is called the *differential* of f at the point $\mathbf{x}_0$ and usually denoted by $\mathrm{d}f(\mathbf{x}_0)$. In terms of the differential, the above condition takes the form (rewritten in terms of $\mathbf{x} = \mathbf{x}_0 + \mathbf{h}$)

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \mathrm{d}f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \mathrm{o}(\mathbf{x} - \mathbf{x}_0) \quad \text{for } \mathbf{x} \to \mathbf{x}_0.$$

The matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ representing $L = \mathrm{d}f(\mathbf{x}_0)$ (i.e., $L(\mathbf{h}) = \mathbf{A}\mathbf{h}$ for $\mathbf{h} \in \mathbb{R}^n$) is called *Jacobi matrix* (or *functional matrix*) of f at $\mathbf{x}_0$ and denoted by $\mathbf{J}_f(\mathbf{x}_0)$. The entries a_{ij} of $\mathbf{A}$ turn out to be the partial derivatives of f at $\mathbf{x}_0$: Writing $f = (f_1, \ldots, f_m)$, we have $a_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x}_0)$.

Example (Squaring map in $\mathbb{C}$)

Since $z^2 = (x+y\mathrm{i})^2 = x^2 + 2xy\mathrm{i} + \mathrm{i}^2y^2 = x^2 - y^2 + 2xy\,\mathrm{i}$, it is natural to call the map $s\colon \mathbb{R}^2 \to \mathbb{R}^2$ defined by $s(x,y) = (x^2 - y^2, 2xy)$ the complex squaring map.

$$\begin{aligned} s(x+h_1, y+h_2) &= \begin{pmatrix} (x+h_1)^2 - (y+h_2)^2 \\ 2(x+h_1)(y+h_2) \end{pmatrix} \\ &= \begin{pmatrix} x^2 - y^2 + 2xh_1 - 2yh_2 + h_1^2 - h_2^2 \\ 2xy + yh_1 + xh_2 + h_1h_2 \end{pmatrix} \\ &= \begin{pmatrix} x^2 - y^2 \\ 2xy \end{pmatrix} + \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \begin{pmatrix} h_1^2 - h_2^2 \\ 2h_1h_2 \end{pmatrix} \\ &= s(x,y) + \mathbf{J}_s(x,y)\mathbf{h} + R(\mathbf{h}) \end{aligned}$$

with $R(\mathbf{h}) = \mathrm{o}(\mathbf{h})$.

You can verify that the entries of $\mathbf{J}_s(x,y)$ are the partial derivatives $(s_1)_x$, $(s_1)_y$, $(s_2)_x$, $(s_2)_y$ of $s_1(x,y) = x^2 - y^2$ and $s_2(x,y) = 2xy$.

Newton's Method cont'd

Newton's Method can also be used to solve vectorial equations numerically, e.g.,

$$\begin{array}{rcrcr} 5x & + & e^y & = & -4, \\ x^2 & - & xy & = & 2. \end{array}$$

Setting $f(x,y) = (5x + e^y + 4, x^2 - xy - 2)$ and $\mathbf{x} = (x,y)$, the system becomes $f(\mathbf{x}) = \mathbf{0}$, and we can use the same idea as in the 1-dimensional case (writing $\mathbf{x}^{(n)} = (x_n, y_n)$):

$$\begin{aligned} & \ell(\mathbf{x}) = f(\mathbf{x}^{(n)}) + \mathbf{J}_f(\mathbf{x}^{(n)})(\mathbf{x} - \mathbf{x}^{(n)}) = \mathbf{0} \\ \iff \quad & \mathbf{x} = \mathbf{x}^{(n)} - \mathbf{J}_f(\mathbf{x}^{(n)})^{-1} f(\mathbf{x}^{(n)}) =: \mathbf{x}^{(n+1)}, \end{aligned}$$

provided that $\mathbf{J}_f(\mathbf{x}^{(n)})$ is invertible.

Choosing $\mathbf{x}^{(0)} = (x_0, y_0)$ suitably and assuming that during the execution only invertible matrices $\mathbf{J}_f(\mathbf{x}^{(n)})$, $n = 0, 1, 2, \ldots$, are encountered, the iteration $\mathbf{x}^{(n+1)} = T(\mathbf{x}^{(n)})$, $T(\mathbf{x}) = \mathbf{x} - \mathbf{J}_f(\mathbf{x})^{-1} f(\mathbf{x})$, defines a sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$, of points in $\mathbb{R}^2$, which converges to the unique solution of $f(\mathbf{x}) = \mathbf{0}$; see the subsequent example. (You can verify that the system has a unique solution, e.g., by eliminating y and applying standard Calculus techniques to the resulting equation for x.)

Newton's Method cont'd

The method just outlined generalizes to systems of n equations in n unknowns.

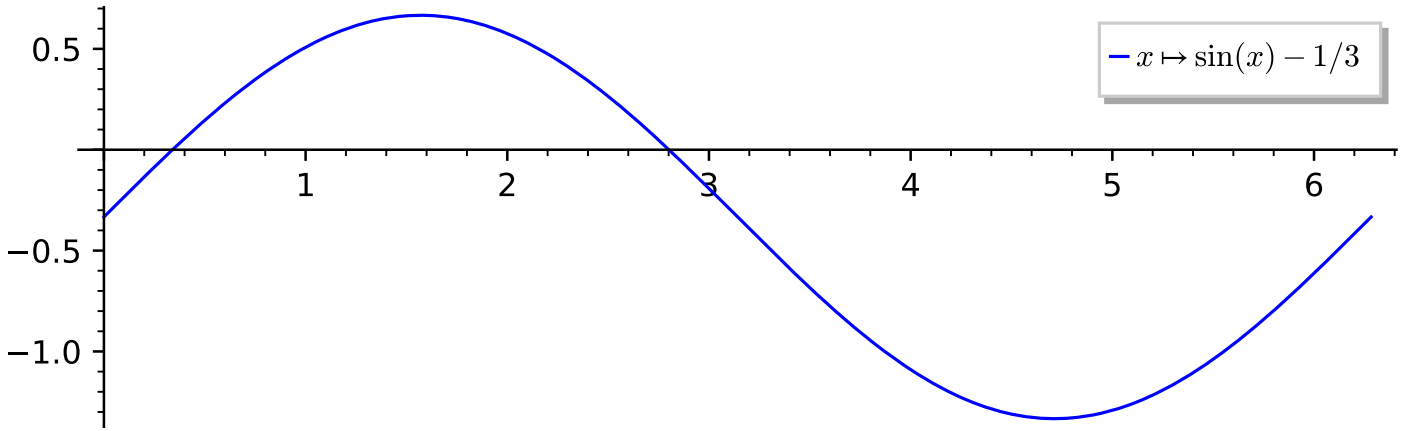
In general, however, the convergence analysis of this higher-dimensional Newton iteration is much more involved than that of the 1-dimensional iteration.
Since there is no analog of the Intermediate Value Theorem for $\mathbb{R}^d$, $d > 1$, we can only use the second method ("second answer") to prove convergence. The "contraction" property $\left|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\right| \leq C \left|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\right|$ for all n, where $C < 1$ is a fixed constant, turns out to work for $d > 1$ as well. A few more details on the method will be provided when we discuss matrix norms and in the exercises.

Example ($f(x) = \sin(x) - 1/3$)

For $f(x) = \sin(x) - 1/3$ we have $T(x) = x - \frac{\sin(x)-1/3}{\cos(x)}$ and the recurrence $x_{n+1} = x_n - \frac{\sin(x_n)-1/3}{\cos(x_n)}$. The following lists the Newton iterates for the starting values $x_0 = 1$ and $x_0 = 2$.

n	x_n
0	1.000000000000000
1	0.059530847905406 3
2	0.333854436356604 0
3	0.339830667137674 8
4	0.339836909447233 6
5	0.339836909454121 9
6	0.339836909454121 9

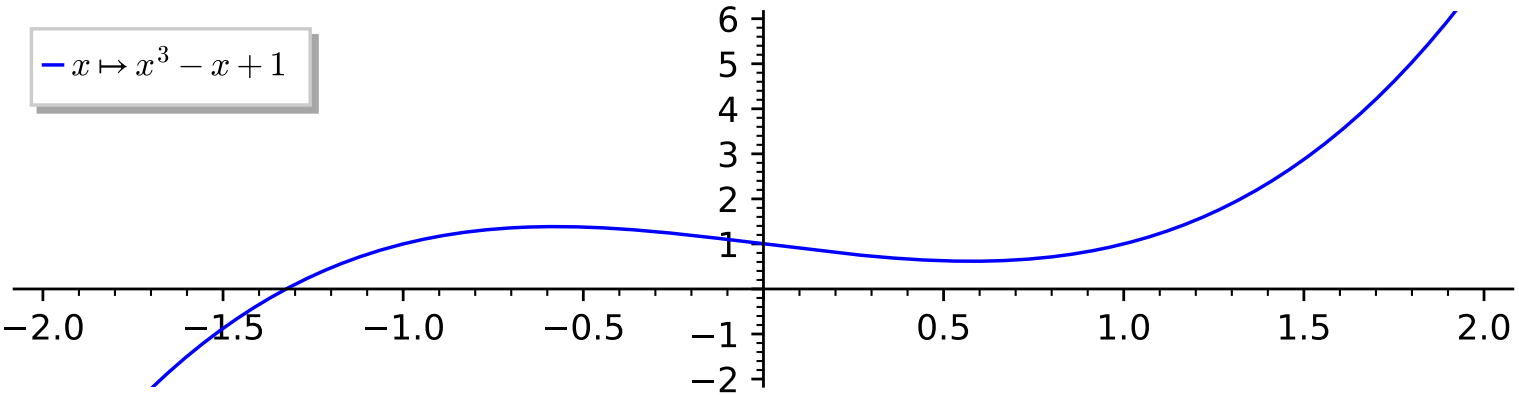
n	x_n
0	2.000000000000000
1	3.384040542687392 0
2	2.793351812048839 0
3	2.801768465049102 4
4	2.801755744164277 0
5	2.801755744135671 3
6	2.801755744135671 3



Example ($f(x) = x^3 - x + 1$)

Here it takes quite a while until quadratic convergence sets in.

n	x_n	n	x_n
0	1.000000000000000	13	−0.742494298720700**9**
1	0.500000000000000	14	−2.781295940677608**3**
2	3.000000000000000	15	−1.982725247043830**6**
3	2.038461538461538**3**	16	−1.536927379758256**3**
4	1.390282147216736**2**	17	−1.357262483187732**5**
5	0.911611897717927**0**	18	−1.325663094428867**9**
6	0.345028496748169**2**	19	−1.324718788615257**2**
7	1.427750704027270**3**	20	−1.324717957245390**2**
8	0.942417912509482**9**	21	−1.324717957244746**0**
9	0.404949357199379**6**	22	−1.324717957244746**0**
10	1.706904645182851**6**	23	−1.324717957244746**0**
11	1.155756361074813**4**	24	−1.324717957244746**0**
12	0.694191813329546**9**	25	−1.324717957244746**0**

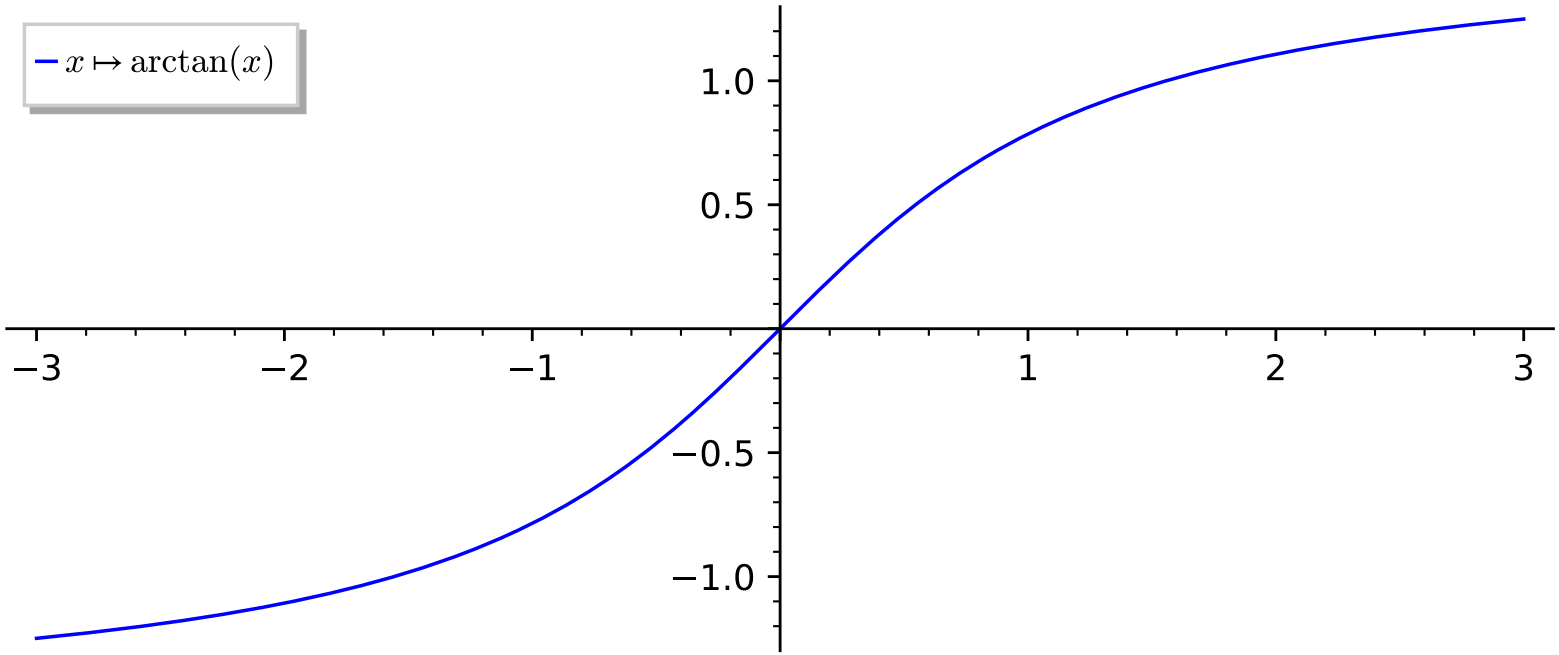


Example ($f(x) = \arctan x$)

Here convergence/divergence of the Newton iteration $x_{n+1} = T(x_n) = x_n - \arctan(x_n)(1 + x_n^2)$ depends on the choice of the initial value x_0.

n	x_n
0	1.0000000000000000
1	−0.5707963267948966
2	0.1168599039989130
3	−0.0010610221170447
4	0.0000000007963096
5	0.0000000000000000

n	x_n
0	2.0000000000000000
1	−3.5357435889704525
2	13.950959086927493
3	−279.34406653361738
4	122016.99891795458
5	−23386004197.933886



Example ($f(x,y) = (5x + e^y + 4, x^2 - xy - 2)$)

Here we have

$$T\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}x\\y\end{pmatrix} - \begin{pmatrix}5 & e^y\\ 2x-y & -x\end{pmatrix}^{-1}\begin{pmatrix}5x+e^y+4\\ x^2-xy-2\end{pmatrix}$$
$$= \begin{pmatrix}x\\y\end{pmatrix} + \frac{1}{5x+(2x-y)e^y}\begin{pmatrix}-x & -e^y\\ y-2x & 5\end{pmatrix}\begin{pmatrix}5x+e^y+4\\ x^2-xy-2\end{pmatrix}.$$

Starting with the "approximate" solution $(x_0, y_0) = (-1, 0)$ (well, rather it solves the first equation exactly), we obtain the sequence

n	x_n	y_n
0	-1.00000000000000	0.00000000000000
1	-1.14285714285714	$0.71428571428571\underline{4}$
2	-1.15343194160013	$0.57938401044252\underline{5}$
3	-1.15552495201267	$0.57528648658840\underline{1}$
4	-1.15552722080764	$0.57528445025\underline{1602}$
5	-1.15552722080795	$0.57528445025\underline{0385}$
6	-1.15552722080795	$0.57528445025\underline{0385}$

You can check that (x_5, y_5) is indeed very close to being a root of f (the entries of $f(x_5, y_5)$ have absolute value $< 10^{-14}$).

Metric Spaces

Definition

A *metric space* (M, d) consists of a set M and a map $d\colon M \times M \to \mathbb{R}$ ("distance function") satisfying the following for all $x, y, z \in M$:

(M1) $d(x,y) \geq 0$; $d(x,y) = 0 \iff x = y$; (*non-negativity*)

(M2) $d(x,y) = d(y,x)$; (*symmetry*)

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

Examples

1. $(\mathbb{R}^n, \mathrm{d_E})$ with $\mathrm{d_E}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (*Euclidean distance*);
   includes $\mathbb{R}$ and $\mathbb{C}$ with $\mathrm{d_E}(x,y) = |x-y|$, resp.,
   $\mathrm{d_E}(z,w) = |z-w| = \sqrt{(\operatorname{Re} z - \operatorname{Re} w)^2 + (\operatorname{Im} z - \operatorname{Im} w)^2}$ as special cases.

Examples (cont'd)

2. $(\mathbb{R}^n, \mathrm{d}_1)$ and $(\mathbb{R}^n, \mathrm{d}_\infty)$ with the metrics

$$\mathrm{d}_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|\,, \quad (\ell^1\text{-distance, "Manhattan distance"})$$

$$\mathrm{d}_\infty(\mathbf{x},\mathbf{y}) = \max\bigl\{|x_i - y_i|\,; 1 \le i \le n\bigr\}. \qquad (\ell^\infty\text{-distance})$$

3. $(\mathbb{R}^n, \mathrm{d}_\mathrm{F})$ with $\mathrm{d}_\mathrm{F}(\mathbf{x},\mathbf{y}) = \begin{cases} \mathrm{d}_\mathrm{E}(\mathbf{x},\mathbf{y}) & \text{if } \mathbb{R}\mathbf{x} = \mathbb{R}\mathbf{y}, \\ \mathrm{d}_\mathrm{E}(\mathbf{x},\mathbf{0}) + \mathrm{d}_\mathrm{E}(\mathbf{0},\mathbf{y}) & \text{if } \mathbb{R}\mathbf{x} \ne \mathbb{R}\mathbf{y}. \end{cases}$

   d_F is sometimes called *"French distance"* or, more accurately, *metric of the French railway network*.

4. The set of all complex-valued, infinite sequences $(a_n)_{n=0}^{\infty}$ satisfying $\sum_{n=0}^{\infty} |a_n|^2 < \infty$ with distance function

$$d\bigl((a_n),(b_n)\bigr) = \sqrt{\sum_{n=0}^{\infty} |a_n - b_n|^2}.$$

   This metric space is known as *Hilbert's Cube* and usually denoted by ℓ^2.

Examples (cont'd)

5. Any set M (for example, $M = \mathbb{R}^n$) with distance

$$d(x,y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases} \qquad (\textit{discrete metric})$$

6. Weighted connected simple graphs (V, E, w) with the shortest path distance. Here $w\colon E \to \mathbb{R}^+$ is a weight function on the edge set E, the weight or length of a path is the sum of the weights of its edges, the underlying set is the vertex set V, and $d(v,w)$ is defined as the length of a shortest path (i.e., path of smallest weight) between v and w. This example includes unweighted graphs with the shortest path distance if we assign weight 1 to all edges.

7. The set of all bit strings of length n with

$$\mathrm{d}_{\mathrm{Ham}}(\mathbf{s},\mathbf{t}) = \left|\{1 \le i \le n; s_i \neq t_i\}\right|. \qquad (\textit{Hamming distance})$$

This is a special case of Example 6, because $\mathrm{d}_{\mathrm{Ham}}(\mathbf{s},\mathbf{t})$ is equal to the length of a shortest path between $\mathbf{s}$ and $\mathbf{t}$ in the hypercube Q_n.

Examples (cont'd)

8. Any subset $M' \subseteq M$ of a metric space (M, d) forms a metric space (M', d') of its own by defining $d'(x, y) = d(x, y)$ for $x, y \in N$ (i.e., the distance on N is the induced distance).

9. The set $\mathrm{C}([a, b])$ of all continuous functions $f : [a, b] \to \mathbb{R}$ on a compact interval $[a, b] \subset \mathbb{R}$ with distance

$$d_\infty(f, g) = \max\{|f(x) - g(x)| ; a \le x \le b\}.$$

d_∞ is also referred to as *metric of uniform convergence*, since $f_n \to g$ in this metric, i.e., $\lim_{n\to\infty} d_\infty(f_n, g) = 0$, is equivalent to $f_n \to g$ uniformly.

This example admits various generalizations, e.g., the domain $[a, b]$ can be replaced by a compact set $K \subset \mathbb{R}^n$, the codomain $\mathbb{R}$ can be replaced by $\mathbb{R}^m$ if we change "absolute value" to "Euclidean length", we could work more generally with bounded functions if we change "maximum" to "supremum", or restrict to C^1-functions and use the maximum of $|f(x) - g(x)|$ and $|f'(x) - g'(x)|$ in the definition of $d_\infty(f, g)$, etc.

Examples

10. A non-connected (weighted) simple graph with the shortest-path distance forms an example of a *generalized metric space*, in which distances are allowed to take the value $\infty = +\infty$. Axioms (M1)–(M3) must still be satisfied—for example, if $d(x,y) = \infty$ then $d(x,z) < \infty \wedge d(y,z) < \infty$ is impossible.

    Further examples of generalized metric spaces are the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ with $d = \mathrm{d_E}$ on $\mathbb{R} \times \mathbb{R}$ and

    $$d(-\infty, +\infty) = d(\pm\infty, x) = \infty \quad \text{for all } x \in \mathbb{R},$$

    and the set of all functions $f\colon X \to \mathbb{R}$ on an arbitrary (but fixed) domain X with distance

    $$d_\infty(f,g) = \sup\bigl\{\,|f(x) - g(x)|\,;\, x \in X\bigr\}.$$

    Similar to the case of $\mathrm{C}\bigl([a,b]\bigr)$, d_∞ captures uniform convergence in the sense that $f_n \to g$ uniformly iff $\lim_{n\to\infty} d_\infty(f_n, g) = 0$ (which requires $d_\infty(f_n, g) < \infty$ for all but finitely many n).

Exercise

A metric space (M, d) (or just the metric d) is said to be *translation-invariant* or *norm-induced* if an addition $(x, y) \to x + y$ is defined on M and $d(x, y) = d(x + z, y + z)$ holds for all $x, y, z \in M$.

1. Which of the preceding examples of metric spaces are translation-invariant?
2. Show that a translation-invariant metric $d\colon M \times M \to \mathbb{R}$ is determined by the corresponding *norm* $n\colon M \to \mathbb{R}$ defined by $n(x) = d(x, 0)$. (The zero element $0 \in M$ is distinguished by $x + 0 = 0 + x = x$ for all $x \in M$.)
3. Which properties should a function $n\colon M \to \mathbb{R}$ satisfy in order to determine a metric on M as in b) ?

Exercise (Product metric spaces)

Suppose (M_1, d_1) and (M_2, d_2) are metric spaces and $M = M_1 \times M_2$. For $p \in \mathbb{R}^+$ define $d_p \colon M \times M \to \mathbb{R}$ by

$$d_p(\mathbf{x}, \mathbf{y}) = d_p\big((x_1, x_2), (y_1, y_2)\big) = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p}.$$

1. For which $p \in \mathbb{R}^+$ is (M, d_p) a metric space?
2. Which of our 10 examples fall under this product construction?
3. Show that $d_\infty(\mathbf{x}, \mathbf{y}) = \lim_{p \to +\infty} d_p(\mathbf{x}, \mathbf{y})$ also defines a metric on M. To which of our examples does it correspond?

Exercise

Let $d \colon M \times M \to \mathbb{R}$ be a function satisfying $d(a, a) = 0$ for $a \in M$, $d(a, b) \neq 0$ for $a, b \in M$ with $a \neq b$, and $d(a, b) \leq d(b, c)) + d(c, a)$ for $a, b, c \in M$.

1. Show that d is a metric.
2. Does this conclusion also hold if $d(a, b) \leq d(b, c)) + d(c, a)$ is replaced by the ordinary triangle inequality $d(a, b) \leq d(a, c)) + d(c, b)$?

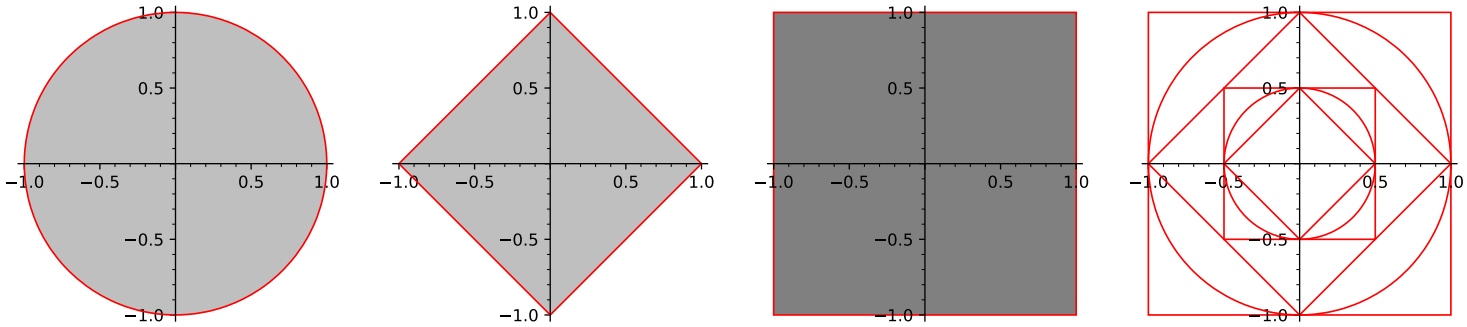
Analysis on Metric Spaces

Observations

- In any metric space we can define balls (and spheres) just as we did in $(\mathbb{R}^n, \mathrm{d_E})$. For example, the *open ball with center* $a \in M$ *and radius* $r \in \mathbb{R}^+$ is defined as $\mathrm{B}_r(a) = \{x \in M; d(x,a) < r\}$.
- Using balls, we can define open sets, closed sets, inner points, boundary points, accumulation points, limits of sequences, and continuity of maps (but not differentiability!) for arbitrary metric spaces. For example, a sequence $(a_n)_{n=0}^{\infty}$ of points $a_n \in M$ *converges to a point* $a \in M$, notation $\lim_{n\to\infty} a_n = a$, if for every $\epsilon > 0$ there exists $N = N_\epsilon \in \mathbb{N}$ such that $a_n \in \mathrm{B}_\epsilon(a)$ for all $n > N$; equivalently, $d(a_n, a) < \epsilon$ for all $n > N$.
- Care must be taken, however, when generalizing some of the less obvious (but important) properties of $(\mathbb{R}^n, \mathrm{d_E})$ to arbitrary metric spaces. An example is the Bolzano-Weierstrass Theorem, which fails to hold in a general metric space; another example is completeness.

Example

The following three figures show the unit balls with respect to the three metrics $\mathrm{d_E}$, $\mathrm{d_1}$, $\mathrm{d_\infty}$ on $\mathbb{R}^2$. (For $\mathrm{d_1}$ the closed unit ball is given by $|x| + |y| \leq 1$, and for $\mathrm{d_\infty}$ by $\max\{|x|, |y|\} \leq 1$.)



(a) $\mathrm{d_E}$ (b) $\mathrm{d_1}$ (c) $\mathrm{d_\infty}$ (d) all

The 4th figure shows the nested structure of the balls of the three metrics: Every ball of one metric contains balls of the other two metrics, possibly of smaller radius. This implies that the three metric spaces have the same convergent sequences, open sets, etc.; they are topologically indistinguishable; cf. also the exercise on strongly equivalent metrics. On the other hand, the French distance $\mathrm{d_F}$ is essentially different from these three. For example, the sequence of points $\big(\cos(1/n), \sin(1/n)\big)$, $n \in \mathbb{N}$, converges to $(1, 0)$ in $\mathrm{d_E}$, $\mathrm{d_1}$, $\mathrm{d_\infty}$ but not in $\mathrm{d_F}$, where all these points have distance 2.

Example (cont'd)

(Unit) Balls of general metric spaces can look quite weird. For the French metric this is discussed in a subsequent exercise. For a discrete metric space (M, d), the closed balls of radius $0 < r < 1$ contain only 1 element (the center) and those of radius $r \geq 1$ are all equal to M. For $\mathrm{C}([a, b])$ equipped with the metric d_∞ of uniform convergence, the unit ball centered at f consists of all continuous functions whose graph is contained in the strip of width 2 symmetrically around the graph of f; see picture.

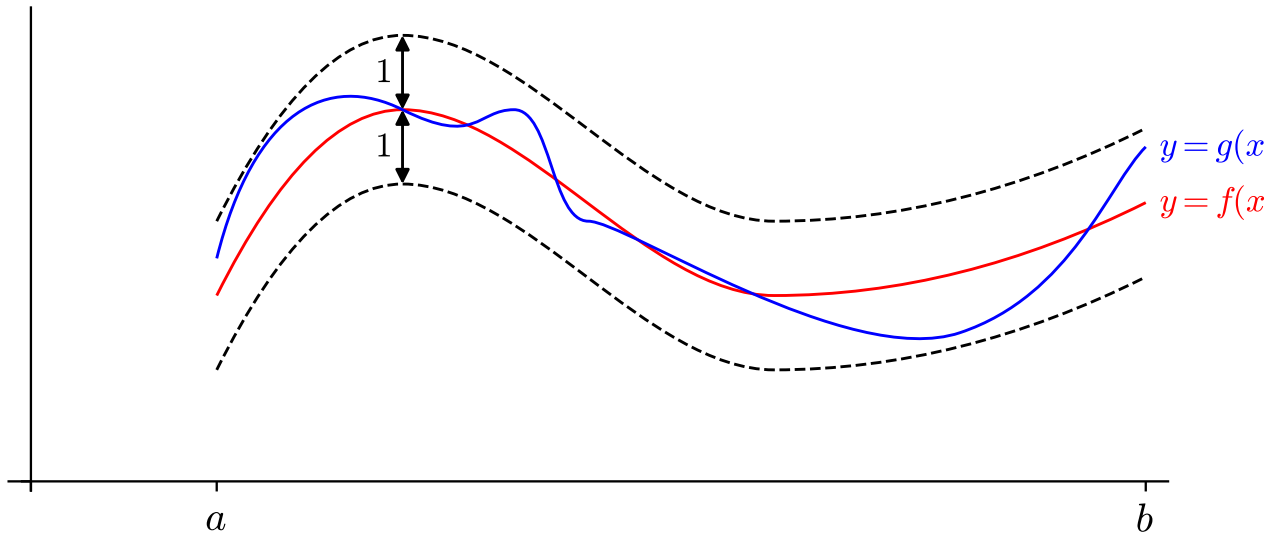


Figure: Illustration of a ball $\mathrm{B}_1(f)$ in $\mathrm{C}([a, b])$ with respect to the uniform metric and a particular function $g \in \mathrm{B}_1(f)$

It is rather obvious that the Bolzano-Weierstrass Theorem fails for discrete metric spaces, but it also fails for the Hilbert cube, which otherwise very much looks like $(\mathbb{R}^n, \mathrm{d_E})$. This is the subject of the following

Exercise

1. Give an example of a set M that when equipped with the discrete metric does not obey the Bolzano-Weierstrass Theorem "Every bounded sequence has a convergent subsequence".
2. Show that the Bolzano-Weierstrass Theorem also fails for the Hilbert cube ℓ^2.

Exercise

Determine the (closed, say) unit balls in the French metric $\mathrm{d_F}$ around each point $(x, y) \in \mathbb{R}^2$.

Exercise (Continuity of a metric)

Let (M, d) be a metric space and $(a, b) \in M \times M$. Show that for every $\epsilon > 0$ there exists a $\delta > 0$ such that $d(x, a) < \delta \wedge d(y, b) < \delta$ implies $|d(x, y) - d(a, b)| < \epsilon$.

Hint: First derive the so-called *quadrangle inequality* $|d(x, y) - d(a, b)| \leq d(x, a) + d(y, b)$.

Complete Metric Spaces

We have defined completeness of $\mathbb{R}$ using the natural ordering $\leq$. This does not generalize to arbitrary metric spaces, but there is a reformulation of the completeness property which does:

Definition

Let (M, d) be a metric space.

1. A sequence $(a_n)_{n=0}^{\infty}$ of points $a_n \in M$ is said to be a *Cauchy sequence* (or satisfy the *Cauchy criterion*) if for every $\epsilon > 0$ there exists $N = N_\epsilon \in \mathbb{N}$ such that $d(a_m, a_n) < \epsilon$ for all $m, n > N$.
2. (M, d) is said to be *complete* if every Cauchy sequence in M converges (i.e., has a limit $a \in M$).

Note

When dealing with series $\sum_{n=1}^{\infty} a_n$ rather than sequences, we must check the Cauchy criterion for the sequence of partial sums $s_n = \sum_{k=1}^{n} a_k$. This requires bounding

$$s_n - s_m = \sum_{k=m+1}^{n} a_k \quad \text{for } n > m > N.$$

Examples/Counterexamples

- We have proved that $\mathbb{R}$ is complete according to the new definition and mentioned that, more generally, the Euclidean spaces $(\mathbb{R}^d, \mathrm{d_E})$, $d = 1, 2, 3, \ldots$, are complete. In particular the field $\mathbb{C}$ of complex numbers is complete (the case $d = 2$). Here we are tacitly assuming that the underlying metric is the Euclidean metric $\mathrm{d_E}$. (Otherwise the assertion could be false.)
- Any subset M of $\mathbb{R}^n$ forms a metric space of its own with the metric induced by $\mathrm{d_E}$ (i.e., distances between points in M are the same as in $\mathbb{R}^n$). Such a metric subspace is complete iff M is a closed subset of $\mathbb{R}^n$; cf. subsequent exercise. (Recall that M is closed if the boundary ∂M is contained in M or, equivalently, M contains with any convergent sequence also its limit).

Examples/Counterexamples Cont'd

- The "punctured" real line $\mathbb{R} \setminus \{0\}$ forms an incomplete metric space (since it is not closed in $\mathbb{R}$). We can prove this directly as follows: Consider the sequence $x_n = 1/n \in \mathbb{R} \setminus \{0\}$. This sequence is a Cauchy sequence, since it has a limit in $\mathbb{R}$, viz. $lim_{n\to\infty} 1/n = 0$, and the definition of "Cauchy sequence" makes no reference to the ambient metric space M (we could even take $M = \{1/n; n \in \mathbb{N}\}$). But it has no limit in $\mathbb{R} \setminus \{0\}$, and hence $\mathbb{R} \setminus \{0\}$ is incomplete.

  On the other hand, $\mathbb{R} \setminus (0,1) = (-\infty, 0] \cup [1, +\infty)$ is complete since it is closed in $\mathbb{R}$. The analogous incompleteness "proof" using the sequence $x_n = 1/2 + 1/n$ is invalid (can you see where the argument breaks down?).

- The Hilbert cube H is complete. The proof of this is a bit technical, since the elements of H are itself sequences and hence a Cauchy sequence in H is sort of an infinite matrix of real numbers with a particular property.

Examples/Counterexamples Cont'd

- The metric spaces $\mathrm{C}\big([a,b]\big)$ of Example 9 are complete. This can be seen as follows:
  (f_n) is a Cauchy sequence w.r.t. d_∞ if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$|f_n(x) - f_m(x)| < \epsilon \quad \text{for all } m, n > N \text{ and } x \in [a,b]. \qquad \text{(C)}$$

  $\Longrightarrow$ All sequences $\big(f_n(x)\big)$, $x \in [a,b]$, are Cauchy sequences in $\mathbb{R}$ and hence convergent, showing that (f_n) has a point-wise limit function $f\colon [a,b] \to \mathbb{R}$.
  Letting $n \to \infty$ in (C) gives $|f(x) - f_m(x)| \le \epsilon$ for all $m > N$ and $x \in [a,b]$, showing that $f_n \to f$ uniformly.
  Now the Continuity Theorem can be applied to conclude that f is continuous, i.e., $f \in \mathrm{C}\big([a,b]\big)$. Thus every Cauchy sequence in $\mathrm{C}\big([a,b]\big)$ converges.

  For continuous functions on unbounded intervals (and on other domains such as $\mathbb{R}^n$) similar assertions hold: A sequence (f_n) of continuous functions that forms a Cauchy sequence w.r.t. to the generalized metric d_∞, converges uniformly and hence has a continuous limit function.

Examples/Counterexamples Cont'd

The subsequent exercises contain still further examples and counterexamples. Discrete metric spaces (see Example 5 and a subsequent exercise) are complete, and so are the metric spaces arising from graphs (Example 6). It is possible to change the Euclidean metric $\mathrm{d_E}$ on $\mathbb{R}$ in such a way that convergence of sequences is not affected but the new metric space $(\mathbb{R}, d)$ is bounded and incomplete; see subsequent exercises.

Exercise

Two metrics d_1, d_2 on a set M are said to be *strongly equivalent* if there exist constants $\alpha, \beta > 0$ such that

$$\alpha d_1(x,y) \le d_2(x,y) \le \beta d_1(x,y) \quad \text{for all } x, y \in M.$$

a) Show that the metric spaces (M, d_1), (M, d_2) have the same open (closed) sets, the same set of convergent sequences (Cauchy sequences), and are either both complete or both incomplete.

b) Show that the Euclidean metric $\mathrm{d_E}$ and the metrics d_1, d_∞ in Example 2 are strongly equivalent.

Exercise

For $x, y \in \mathbb{R}$ set

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

Show that d defines a metric on $\mathbb{R}$, which is not strongly equivalent to $\mathrm{d_E}$, but that nevertheless the conclusions in Part (1) of the previous Exercise hold for $d_1 = d$ and $d_2 = \mathrm{d_E}$.

Exercise

For $x, y \in \mathbb{R}$ set

$$d(x,y) = \mathrm{d_E}\left(\frac{x}{1+|x|}, \frac{y}{1+|y|}\right) = \left|\frac{x}{1+|x|} - \frac{y}{1+|y|}\right|.$$

a) Show that d defines a metric on $\mathbb{R}$.

b) Show that $(\mathbb{R}, d)$ has the same open sets and the same convergent sequences as $(\mathbb{R}, \mathrm{d_E})$.

c) Show that $(\mathbb{R}, d)$ is incomplete.

Hint: Consider the sequence $a_n = n$.

Exercise

Let (M, d) be a discrete metric space; cf. Example 5. Describe convergent sequences and Cauchy sequences in (M, d) in an alternative way (without using ϵ), and conclude that (M, d) is complete.

Exercise

a) Show that a closed subset N of a complete metric space (M, d) is complete in the induced metric $N \times N \to \mathbb{R}$, $(x, y) \mapsto d(x, y)$.

b) Conversely, show that a subset of a metric space that is complete in the induced metric must be closed.

Exercise

A metric space (M, d) is said to be an *ultrametric* space if it satisfies the following sharper variant of the triangle inequality:

$$d(a,b) \le \max\{d(a,c), d(c,b)\} \quad \text{for } a,b,c \in M.$$

a) Which of our ten introductory examples are ultrametric spaces?

b) For a prime number p the *p-adic absolute value* on $\mathbb{Q}$ is defined by $|0|_p = 0$ and

$$|x|_p = p^{-m} \quad \text{if } x = p^m \frac{a}{b} \text{ with } m \in \mathbb{Z} \text{ and } p \nmid ab.$$

Show that $d_p(x,y) = |x-y|_p$ turns $\mathbb{Q}$ into an ultrametric space.

c) Show that an infinite series $\sum_{n=0}^{\infty} x_n$ in $(\mathbb{Q}, d_p)$ satisfies the Cauchy criterion for convergence iff $x_n \to 0$ for $n \to \infty$.

d) Show that the metric spaces $(\mathbb{Q}, d_p)$, p prime, are not complete.

BANACH's Fixed-Point Theorem

Also called "Contraction Mapping Theorem"

Definition

Let (M, d) be a metric space. A map ("transformation") $T\colon M \to M$ is said to be a *contraction* if there exists a constant $0 \le C < 1$ such that

$$d\big(T(x), T(y)\big) \le C \cdot d(x, y) \quad \text{for all } x, y \in M.$$

Note

The condition in the definition is stronger than $d\big(T(x), T(y)\big) < d(x, y)$ for all $x, y \in M$ with $x \neq y$. For example, the transformation $T(x) = x + 1/x$ of $[1, +\infty)$ (equipped with the Euclidean metric) has this property, since

$$\left|x + \frac{1}{x} - y - \frac{1}{y}\right| = \left|x - y + \frac{y - x}{xy}\right| = \left(1 - \frac{1}{xy}\right)|x - y|$$

and $0 \le 1 - \frac{1}{xy} < 1$ for all $x, y \ge 1$. But T is not a contraction since, given $0 \le C < 1$, the numbers x, y can be chosen to satisfy $1 - \frac{1}{xy} > C$ (take, e.g., $x = 1$ and $y > (1 - C)^{-1}$).

Banach's Fixed-Point Theorem applies to contractions of complete metric spaces.

Theorem (BANACH, 1922)

Suppose (M, d) is a complete metric space and $T\colon M \to M$ a contraction.

1. *T has a unique fixed point, i.e., there exists precisely one element $x^* \in M$ satisfying $T(x^*) = x^*$.*
2. *For every point $x_0 \in M$ the sequence $x_0, x_1, x_2, \ldots$ defined recursively by $x_{n+1} = T(x_n)$ converges to x^*, and we have the error estimates*

$$d(x_n, x^*) \leq \begin{cases} \dfrac{C^n}{1-C}\, d(x_1, x_0), \\ \dfrac{C}{1-C}\, d(x_n, x_{n-1}). \end{cases}$$

Proof.

(1) Choose $x_0 \in M$ and define the sequence (x_n) in M recursively by $x_n = T(x_{n-1}) = T^2(x_{n-2}) = \cdots = T^n(x_0)$. (Here $T^2 = T \circ T$, $T^3 = T \circ T \circ T$, etc.) For $m < n$ the triangle inequality (used successively) and the contraction property of T give

$$\begin{aligned} d(x_m, x_n) &= d\big(T(x_{m-1}), T(x_{n-1})\big) \le C\, d(x_{m-1}, x_{n-1}) \\ &\le C^2\, d(x_{m-2}, x_{n-2}) \le \cdots \le C^m\, d(x_0, x_{n-m}) \\ &\le C^m \big[d(x_0, x_1) + d(x_1, x_2) + \cdots + d(x_{n-m-1}, x_{n-m})\big] \\ &\le C^m \big[1 + C + C^2 + \cdots + C^{n-m-1}\big]\, d(x_0, x_1) \\ &= \frac{C^m - C^n}{1 - C}\, d(x_0, x_1) \\ &\le \frac{C^m}{1 - C}\, d(x_0, x_1). \end{aligned}$$

Since $0 \le C < 1$ we have $\lim_{m\to\infty} C^m = 0$.
$\Longrightarrow$ For given $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $d(x_m, x_n) < \epsilon$ for all $n > m > N$. This means that (x_n) is a Cauchy sequence in the complete metric space (M, d) and hence converges.

Proof cont'd.

Let $x^* = \lim_{n\to\infty} x_n$.
From $d\big(T(x), T(y)\big) \le C\, d(x,y) \le d(x,y)$ it is clear that T is continuous ($\delta = \epsilon$ works).

$$\Longrightarrow T(x^*) = T\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} T(x_n) = \lim_{n\to\infty} x_{n+1} = x^*.$$

Suppose that also $T(x') = x'$.

$$d(x^*, x') = d\big(T(x^*), T(x')\big) \le C\, d(x^*, x')$$

$\Longrightarrow d(x^*, x') = 0 \Longrightarrow x^* = x'$.

(2) The first assertion is clear from the proof of Part (1).
Since metrics are continuous, we get from
$d(x_m, x_n) \le \frac{C^m}{1-C}\, d(x_0, x_1)$ by passing to the limit:

$$d(x_m, x^*) = d\left(x_m, \lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} d(x_m, x_n) \le \frac{C^m}{1-C}\, d(x_0, x_1).$$

The second inequality follows by applying the first inequality to the shifted sequence $x_{n-1}, x_n, x_{n+1}, \ldots$, which also has the limit x^*. □

Notes

- The "weak contraction" property $d\big(T(x),T(y)\big) < d(x,y)$ is not sufficient for the existence of a fixed point of T. A counterexample is the previously considered map $T(x) = x + 1/x$ on $[1,+\infty)$. (For this note that closed intervals in $\mathbb{R}$ form complete metric spaces of their own.)
- The 2nd error estimate $d(x_n,x^*) \le \frac{C}{1-C}\,d(x_n,x_{n-1})$ is particularly useful, since $d(x_n,x_{n-1})$ can be read off by looking at the last two iterates. (In fact, this is what in the examples allowed us to conclude from equality of the floating-point representations of x_n and x_{n-1} that x^* has the same floating point representation.)
- In many applications, e.g. Newton's method, the map T becomes a contraction only when restricted to a suitable complete subspace M of its domain. In this case the condition $T(M) \subseteq M$, which is often difficult to verify, can be relaxed to "$x_n = T^n(x_0) \in M$ for all $n = 0,1,2,\ldots$"; that is, we are now looking at particular sequences.
  Specifically, if $M = \overline{\mathrm{B}_r(a)}$ is a ball and $x_0 = a$, it suffices to check the single condition $d\big(a,T(a)\big) = d(x_0,x_1) \le (1-C)r$. This is proved on the next slide.

Notes cont'd

- (cont'd)
  As in the proof of Banach's Theorem, this condition gives $d(x_n, a) \leq (C^{n-1} + C^{n-2} + \cdots + 1)d(x_1, x_0) \leq (1 - C^n)r \leq r$, i.e., $x_n \in \overline{\mathrm{B}_r(a)}$, and the proof of Part (1) of Banach's Theorem goes through.
  $\Longrightarrow$ (x_n) converges to $x^* \in \overline{\mathrm{B}_r(a)}$, and x^* is the unique fixed point of T in $\overline{\mathrm{B}_r(a)}$.
  Part (2), however, is not necessarily true in this setting, since for a different sequence (y_n), $y_0 \in \overline{\mathrm{B}_r(a)} \setminus \{a\}$, the contraction property of T on $\overline{\mathrm{B}_r(a)}$ doesn't exclude the possibility that some iterate y_n falls outside $\overline{\mathrm{B}_r(a)}$.
  In fact, if T doesn't map $\overline{\mathrm{B}_r(a)}$ into itself, there exists $y_0 \in \overline{\mathrm{B}_r(a)}$ such that $y_1 = T(y_0) \notin \overline{\mathrm{B}_r(a)}$.
- For the analysis of iterations on subsets of $\mathbb{R}^n$ we can use any metric on $\mathbb{R}^n$ that is strongly equivalent to the Euclidean metric d_E (cf. previous exercise), e.g., also d_1 or d_∞. Convergence proofs may become easier by choosing a metric different from d_E.

Example

Consider the squaring map $T\colon \mathbb{C} \to \mathbb{C}$, $z \mapsto z^2$. (The metric on $\mathbb{C}$ is taken as the usual Euclidean one.)

$\mathbb{C} \triangleq \mathbb{R}^2$ is complete, but T is not a contraction since

$$|T(z) - T(w)| = \left|z^2 - w^2\right| = |z + w|\,|z - w|$$

and $z + w$ can have arbitarily large absolute value.
Hence we cannot use Banach's Theorem to find the fixed points of T, which are 0 and 1.

However, we can can restrict the domain of T suitably and then apply Banach's Theorem:

Suppose $0 < r < 1/2$ and let $M = \overline{\mathrm{B}_r(0)} = \{z \in \mathbb{C}; |z| \le r\}$.

- M is complete, since it is a closed subset of $\mathbb{C}$.
- For $z \in M$ we have $\left|z^2\right| = |z|^2 \le r^2 \le r$ and hence $T(M) \subseteq M$.
- For $z, w \in M$ we have $|z + w| \le |z| + |w| \le 2r < 1$.
  $\Longrightarrow T\colon M \to M$ is a contraction (take $C = 2r$).

Example (cont'd)

Hence Banach's Theorem gives that any sequence $z_n = z_{n-1}^2 = z_{n-2}^4 = \cdots = z_0^{2^n}$, $z_0 \in M$, converges to the unique fixed point of T in M, which is 0.

This is of course rather trivial and true for all $z_0 \in \mathbb{C}$ with $|z_0| < 1$.

Definition

Suppose (M, d) is a metric space, $T : M \to M$ a map and x^* a fixed point of T.

1. x^* is said to be *attracting* if there exists a neighborhood U of x^* such that any sequence $x_n = T^n(x_0)$ $(n \in \mathbb{N})$ with initial point $x_0 \in U$ converges to x^*;
2. x^* is said to be *repelling* if there exists a neighborhood U of x^* such that any sequence $x_n = T^n(x_0)$ $(n \in \mathbb{N})$ with initial point $x_0 \in U$ eventually leaves U (i.e., $x_n \notin U$ for some n).

Exercise

a) Decide whether the fixed points 0 and 1 of $T : \mathbb{C} \to \mathbb{C}$, $z \to z^2$ are attracting or repelling, and prove your assertions.

b) For which $z_0 \in \mathbb{C}$ does $z_n = T^n(z_0) = z_0^{2^n}$ converge to $z^* = 1$?

Exercise

The system of equations

$$\begin{aligned} x &= 0,01\,x^2 + \sin(y) \\ y &= \cos(x) + 0,01\,y^2 \end{aligned}$$

has a unique solution (x^*, y^*) with $0,5 \leq x^* \leq 1$, $\pi/6 \leq y^* \leq 1$. Prove this statement and compute (x^*, y^*)

a) with simple fixed-point iteration;

b) with Newton Iteration.

Matrix Norms—Motivation

The quest for the norm ("absolute value") of a square matrix arises naturally during the convergence analysis of higher-dimensional Newton iteration.

Recall that this iteration has the form $\mathbf{x}^{(k+1)} = T\big(\mathbf{x}^{(k)}\big)$ with

$$T(\mathbf{x}) = \mathbf{x} - \mathrm{d}f(\mathbf{x})^{-1}\big(f(\mathbf{x})\big) = \mathbf{x} - \mathbf{J}_f(\mathbf{x})^{-1} f(\mathbf{x})$$

As in the 1-dimensional case we have $f(\mathbf{x}^*) = \mathbf{0} \iff T(\mathbf{x}^*) = \mathbf{x}^*$ (clear from the definition of T) and $f(\mathbf{x}^*) = \mathbf{0} \Longrightarrow \mathrm{d}T(\mathbf{x}^*) = 0$ (i.e., $\mathbf{J}_T(\mathbf{x}^*) = \mathbf{0} \in \mathbb{R}^{n\times n}$), as one can show with some effort.
Now we would like to show that near a zero $\mathbf{x}^*$ of f the map T defines a contraction, because then Banach's Fixed-Point Theorem implies $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$, provided only that $\mathbf{x}^{(0)}$ (or some other iterate) is sufficiently close to $\mathbf{x}^*$.

The Mean Value Theorem of n-variable calculus gives

$$\begin{aligned} T(\mathbf{x}) - T(\mathbf{y}) &= \left(\int_0^1 \mathbf{J}_T\big(\mathbf{x} + t(\mathbf{y}-\mathbf{x})\big)\,\mathrm{d}t\right)(\mathbf{x}-\mathbf{y}) \\ &= \mathbf{A}(\mathbf{x}-\mathbf{y}) \quad \text{for some matrix } \mathbf{A} = \mathbf{A}(\mathbf{x},\mathbf{y}). \end{aligned}$$

Motivation Cont'd

In the 1-dimensional case one continues with taking absolute values, viz. $|T(x) - T(y)| = |T'(\xi)|\,|x - y|$, and using continuity of T' to conclude $|T'(\xi)| \le C < 1$ provided x, y are near x^*.

Here we postulate the existence of a real number $\|\mathbf{A}\|$ such that "taking Euclidean lengths" yields the inequality

$$|T(\mathbf{x}) - T(\mathbf{y})| = |\mathbf{A}(\mathbf{x} - \mathbf{y})| \le \|\mathbf{A}\|\,|\mathbf{x} - \mathbf{y}|\,.$$

If such a norm ("absolute value") of $\mathbf{A} = \mathbf{A}(\mathbf{x}, \mathbf{y})$ exists and satisfies $\|\mathbf{A}(\mathbf{x}, \mathbf{y})\| \le C < 1$ for $\mathbf{x}, \mathbf{y}$ near $\mathbf{x}^*$ then the analysis in the 1-dimensional case carries over to the n-dimensional case.

Replacing the particular matrices $\mathbf{A}(\mathbf{x}, \mathbf{y})$ by an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and setting $\mathbf{v} = \mathbf{x} - \mathbf{y}$ turns the inequality into

$$|\mathbf{Av}| \le \|\mathbf{A}\|\,|\mathbf{v}| \ \text{ for } \mathbf{v} \in \mathbb{R}^n \quad\Longleftrightarrow\quad \|\mathbf{A}\| \ge \frac{|\mathbf{Av}|}{|\mathbf{v}|} \ \text{ for } \mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}.$$

It turns out that the set $\left\{ |\mathbf{Av}| \,/\, |\mathbf{v}| \,;\, \mathbf{v} \in \mathbb{R}^n, \mathbf{v} \ne \mathbf{0} \right\}$ contains a maximum, which then clearly provides the best definition of $\|\mathbf{A}\|$.

Definition (norm of a matrix or linear map)

The *norm* of $\mathbf{A} \in \mathbb{R}^{n\times n}$ (more precisely, the *matrix norm subordinate to the Euclidean length on* $\mathbb{R}^n$) is defined as

$$\|\mathbf{A}\| = \max\left\{\frac{|\mathbf{Ax}|}{|\mathbf{x}|}; \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}\right\} = \max\{|\mathbf{Ax}|\,; \mathbf{x} \in \mathbb{R}^n, |\mathbf{x}| = 1\}.$$

The norm of a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^n$ is defined as the norm of its representing matrix, i.e., if $L(\mathbf{x}) = \mathbf{Ax}$ then

$$\|L\| = \|\mathbf{A}\| = \max\left\{\frac{|L(\mathbf{x})|}{|\mathbf{x}|}; \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}\right\}.$$

Notes

- The second equality in the definition follows from the linearity of L:

$$\frac{|L(\mathbf{x})|}{|\mathbf{x}|} = \left|\frac{1}{|\mathbf{x}|}L(\mathbf{x})\right| = \left|L\left(\frac{\mathbf{x}}{|\mathbf{x}|}\right)\right|, \quad \text{with } \frac{\mathbf{x}}{|\mathbf{x}|} \text{ of length 1.}$$

  Since linear maps are continuous and the unit sphere in $\mathbb{R}^n$ is compact, the maximum is attained.

Notes cont'd

- The definition of $\|\mathbf{A}\|$ trivially implies $|\mathbf{Ax}| \leq \|\mathbf{A}\|\,|\mathbf{x}|$ for all $\mathbf{x} \in \mathbb{R}^n$, and similarly for the corresponding linear map L, as desired.
- The function $\mathbb{R}^{n\times n} \to \mathbb{R}$, $\mathbf{A} \mapsto \|\mathbf{A}\|$ satisfies the same axioms as the Euclidean length function:

  (N1) $\|\mathbf{A}\| \geq 0$ with equality iff $\mathbf{A} = \mathbf{0}$;
  (N2) $\|c\mathbf{A}\| = |c|\,\|\mathbf{A}\|$ for $c \in \mathbb{R}$;
  (N3) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

  Hence the definition $d(\mathbf{A}, \mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|$ turns $\mathbb{R}^{n\times n}$ into a (translation-invariant) metric space.
- A further important property of $\|\ \|$ is

  (N4) $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|$. (*submultiplicativity*).
- If you wonder how to actually compute matrix norms—the answer is not easy. It uses the so-called *Spectral Theorem for Symmetric Matrices*, which is beyond the scope of our present Linear Algebra knowledge. A few particular examples, which have an adhoc solution, are discussed in the exercises.

Notes cont'd

- Functions on $\mathbb{R}^n$ satisfying the same axioms as the Euclidean length are called *vector norms* and denoted in the same way. Examples are

$$\|\mathbf{x}\|_1 = |\mathbf{x}|_1 = |x_1| + |x_2| + \cdots + |x_n|,$$
$$\|\mathbf{x}\|_\infty = |\mathbf{x}|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}.$$

  For the Euclidean length the notation $|\mathbf{x}|_2$ or $\|\mathbf{x}\|_2$ is frequently used in place of $|\mathbf{x}|$. With any vector norm one may associate a subordinate matrix norm in the same way as for the Euclidean length, for example $\|\mathbf{A}\|_1 = \max\{|\mathbf{Ax}|_1 ; \mathbf{x} \in \mathbb{R}^n, |\mathbf{x}|_1 = 1\}$ for $\mathbf{A} \in \mathbb{R}^{n\times n}$.

- Reading $\mathbf{A} \in \mathbb{R}^{n\times n}$ as an n^2-dimensional vector with entries a_{ij}, it is quite natural to consider the Euclidean length of this vector. This quantity is called *Frobenius norm* of $\mathbf{A}$ and denoted by $\|\mathbf{A}\|_{\mathrm{F}}$, i.e., one defines

$$\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2} \quad \text{for } \mathbf{A} \in \mathbb{R}^{n\times n}.$$

  One can show that $\mathbf{A} \mapsto \|\mathbf{A}\|_{\mathrm{F}}$ satisfies Axioms (N1)–(N4).

Exercise

Prove that $\mathbb{R}^{n\times n} \to \mathbb{R}$, $\mathbf{A} \mapsto \|\mathbf{A}\|$ satisfies (N1)–(N4).

Exercise

Prove that $\mathbb{R}^{n\times n} \to \mathbb{R}$, $\mathbf{A} \mapsto \|\mathbf{A}\|_{\mathrm{F}}$ satisfies (N1)–(N4).

Exercise

Compute the norms $\|\mathbf{A}\|$ of the following matrices $\mathbf{A} \in \mathbb{R}^{2\times 2}$ and compare them with their Frobenius norms $\|\mathbf{A}\|_{\mathrm{F}}$:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 \\ 0 & -3 \end{pmatrix}, \quad \begin{pmatrix} \frac{1}{2} & \pm 1 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Exercise

Show that the norm of a diagonal matrix is the largest absolute value of the entries on the diagonal.

Exercise

Show that $\|\mathbf{A}\| \leq \|\mathbf{A}\|_{\mathrm{F}}$ for all matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ or, equivalently, $|\mathbf{A}\mathbf{x}| \leq \|\mathbf{A}\|_{\mathrm{F}}\,|\mathbf{x}|$ for all $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{x} \in \mathbb{R}^n$.

Hint: Use $\|\mathbf{A}\| = \max\{|\mathbf{A}\mathbf{x}|\,;\mathbf{x} \in \mathbb{R}^n, |\mathbf{x}| = 1\}$ and the Cauchy-Schwarz Inequality for vectors in $\mathbb{R}^n$.