Ramp Up Mathematics — Analysis

Dirk Lorenz
with minor additions by Konstantin Merz

April 22, 2024

Contents

Introductory Remarks

These are lecture notes for the four lectures on analysis as part of the ramp up course for mathematics in the data science master's. It is assumed that you are familiar with the basic notions of calculus of a single variable, i.e. that you know the notions of limit, continuity, differentiability and integrals. Moreover, basic knowledge of multivariable calculus will be assumed, i.e. parameterization of curves and surface as well as partial derivatives.

Analysis deals with dynamics and change and thus, is build upon notions that enable us to speak about "small perturbations" for various objects. In data science, we are faced with data from very different sources which can be organized in different mathematical spaces: Data may come from a linear space (i.e where the scaling and sums of data points may be sensible), from a curved manifold (where we can't scale or add, but still want to know which data points are close to each other and which are not), from a discrete structure like a graph or just from a quite general set with no structure whatsoever. Hence, we want to have such notions for nearness in quite general contexts such as general vector spaces or even general sets.

If you need more background in mathematics for data science, I can recommend the book [DFO20]. This back also covers some of the stuff we cover here, although it does not treat infinite dimensional spaces and does not mention the notions Banach and Hilbert space. I do not know of an introduction to infinite dimensional spaces for data science, but a general reference is [Alt16]. See also [Ver18].

Braunschweig, October 24, 2023 Dirk Lorenz
Braunschweig, April 22, 2024 Konstantin Merz

References

[DFO20] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning.* Cambridge University Press, 2020. `https://mml-book.github.io/`

[Alt16] H.W. Alt. *Linear Functional Analysis. An Application-oriented Introduction.* Springer, 2016. `https://link.springer.com/book/10.1007/978-1-4471-7280-2`

[Ver18] Roman Vershynin. *High-Dimensional Probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 2018. An Introduction with Applications in Data Science, With a foreword by Sara van de Geer.

1 Lecture 1: Vector spaces, norms, and convergence

The building blocks of analysis are notions of distances. You may know that the euclidean distance between two points (x_1, x_2) and (y_1, y_2) in $\mathbb{R}^2$ is $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. Here we will deal with *abstract* notions of distances, i.e. they will share properties with the intuitive understanding of distance, but will be useful in greater generality. The notions are the ones of *norms*, *inner products* and *metrics*.

Remark 1.1. We assume that you are familiar with all basic notions of vector spaces, i.e. with linearity, matrices, linear maps, bases, linear independence and such.

Definition 1.2 (Vector space)**.** A set V is called an $\mathbb{R}$ vector space, if it is equipped with an addition $+ : V \times V \to V$ and a scalar multiplication $\cdot : \mathbb{R} \times V \to V$ such that the following properties are satisfied for all $a, b, c \in V$ and $\alpha, \beta \in \mathbb{R}$.

1. $a + b = b + a$ and $a + (b + c) = (a + b) + c$.
2. There is an element $0 \in V$ such that $0 + a = a + 0 = a$.
3. There is an element called $-a$ such that $a + (-a) = (-a) + a = 0$.
4. $\alpha(a + b) = \alpha a + \beta b$ and $(\alpha + \beta)a = \alpha a + \beta a$.
5. $\alpha(\beta a) = (\alpha\beta)a$.

There are many different vector spaces that will pop up in data science:

1. Most importantly: $\mathbb{R}^n$, the *n-dimensional real vector space* that consists of tuples of n real numbers.

   The space $\mathbb{C}^n$ of complex vectors is less prominent but may occur as well.

   These spaces come up all the time, e.g. when measured data comes a tuples of numbers.

2. As important: *Spaces of functions!* The set of all functions $f : A \to \mathbb{R}$ from some set A into the real numbers form a vector space as well! Such a space may be used to model a set of decision functions that we want to build to predict an output y for given data $x \in A$.

   If we want to predict more than just one number for a data point x we consider the space of functions $f : A \to \mathbb{R}^n$, and these functions form a vector space as well.

Definition 1.3 (Norm)**.** Let V be a real vector space. A *norm* on V is a map

$$\|\cdot\| : V \to [0, \infty[$$

with the following properties:

(i) It holds that $\|x\| = 0$ exactly if $x = 0$. [*positive definite*]

(ii) For every $x \in V$ and $\alpha \in \mathbb{R}$ it holds that $\|\alpha x\| = |\alpha| \, \|x\|$. [*absolutely homogeneous*]

(iii) For every $x, y \in V$ it holds that $\|x + y\| \leq \|x\| + \|y\|$. [*triangle inequality*]

Remark 1.4. We could also define norms on complex vector spaces and the only thing we need to change in the definition is that $\alpha \in \mathbb{C}$ has to be allowed.

The notion of a norm is the abstract concept of a *ruler*: Given a vector x, you can tell "the length of the vector x".

Example 1.5. 1. On $\mathbb{R}^n$ we have, for example

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad (\ell^1\text{-norm})$$

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \qquad (\ell^2\text{-norm})$$

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i| \,. \qquad (\ell^\infty\text{-norm})$$

We even have that $\|x\|_p = \left(\sum\limits_{i=1}^{n} |x_i|^p\right)^{1/p}$ is a norm for every $p \in [1, \infty)$, called ℓ^p-norm.

2. For a subset $\Omega \subset \mathbb{R}^d$ we can define norms for the space of functions $u : \Omega \to \mathbb{R}$, e.g.

$$\|u\|_1 = \int\limits_\Omega |u(x)| \,\mathrm{d}x \qquad (L^1\text{-norm})$$

$$\|u\|_2 = \left(\int\limits_\Omega |u(x)|^2 \,\mathrm{d}x\right)^{1/2} \qquad (L^2\text{-norm})$$

$$\|u\|_\infty = \max_{x \in \Omega} |u(x)| \,. \qquad (L^\infty\text{-norm})$$

One can also define L^p-norms for p between 1 and ∞ as well.

Remark 1.6. There is slight subtlety about these norms: As such they are not positive definite since there are functions u which are not constantly zero but for which an L^p norm evaluates to zero (e.g. functions which are nonzero on very small sets such as single points). One can deal with this problem by "factoring out such functions" but we will not go into detail about this here. If we restrict attention to continuous functions, this issue disappears immediately.

3. There are norms on the space of linear maps between vector spaces (called *operator norms*) but we will deal with them in greater detail in the next lecture.

△

Definition 1.7 (Normed space)**.** A vector space V equipped with a norm $\|\cdot\|$ is called a *normed space.*

Norms can be used to define a notion of convergence! We build on the notion of convergence of sequences of real numbers which we shortly recall:

Definition 1.8 (Convergence in $\mathbb{R}$)**.** We say that a sequence a_n of real number converge to a limit a^* if

$$\begin{aligned}&\text{for every } \epsilon > 0 \text{ there exists an } N \in \mathbb{N} \text{ such that}\\ &n \geq N \implies |a_n - a^*| \leq \epsilon.\end{aligned}$$

Remark 1.9. In word: a is the limit of the sequence if we can come as close to the limit as we want if we look far enough back in the sequence.
You may also think of this in terms of approximation: However accurate we would like to approximate a by elements of a_n we can do so if we look far back in the sequence.

Example 1.10. 1. Irrational numbers as π can be approximated by rational numbers: The sequence

$$\begin{aligned}a_0 &= 3\\ a_1 &= 3.1\\ a_2 &= 3.14\\ a_3 &= 3.141\\ &\vdots\end{aligned}$$

where we add more and more correct digits does converge to π.

In terms of the definition: For $\epsilon = 10^{-k}$, we just take $N = k$: Then a_N approximates π with N correct digits and $|a_n - \pi| < 10^{-N}$. All a_n with $n > N$ are even more accurate.

2. The sequence $a_n = 1/n$ converges to 0: For any $\epsilon > 0$ we take $N \geq 1/\epsilon$ and then we see that for $n > N$

$$|a_n - 0| = \tfrac{1}{n} < \tfrac{1}{N} \leq \epsilon.$$

△

We use this to define convergence *with respect to a norm*:

Definition 1.11 (Convergence)**.** Let $(V, \|\cdot\|)$ be a normed space. A sequence (x_n) in V is said to *converge* to $x \in V$ if $\|x_n - x\| \xrightarrow{n\to\infty} 0$. We write "$x_n \to x$" or "$\lim_{n\to\infty} x_n = x$".

Remark 1.12. Note how elegant this definition is: We only need to understand the convergence of sequences of real numbers to define convergence with respect to *any norm*!

Example 1.13. 1. One can see that convergence of (x^n) in $\mathbb{R}^d$ with respect to any of the ℓ^p-norms is nothing else as "component-wise convergence", i.e.

$$x^n \to x \iff \text{for all } i \in \{1, \dots, d\} : x_i^n \to x_i.$$

2. Convergence of sequences of functions: For simplicity consider $f_n : [a, b] \to \mathbb{R}$. Then convergence of f_n to f for the L^1-norm from Example 1.5 means that

$$\|f_n - f\|_1 = \int_a^b |f_n(x) - f(x)| \,\mathrm{d}x \to 0.$$

Graphically this means that the area between the two graphs of f_n and f has to go to zero.

Convergence of f_n to f in the L^∞-norm (also from Example 1.5) means something different! It means that

$$\|f_n - f\|_\infty = \max_{a\leq x\leq b} |f_n(x) - f(x)| \to 0.$$

This means that the maximal distance between $f_n(x)$ and $f(x)$ has to go to zero.

One can see that convergence of f_n to f in the L^∞-norm implies convergence in the L^1-norm, but the converse does not hold.

△

Remark 1.14. Convergence with respect to the L^∞-norm is equivalent to the notion of *uniform convergence.*

In the exercises, you will see, at the hand of the example of the ℓ^p-norms over real sequences with n entries, that two different norms on finite-dimensional vector spaces are equivalent. In that regard, it is useful to have the following more general tool. In the following, we write $p' = (1 - 1/p)$ for $p \in (1, \infty)$, $p' = 1$ if $p = \infty$, and $p' = \infty$ if $p = 1$.

Theorem 1.15 (Hölder's inequality)**.** *Let $\Omega \subseteq \mathbb{Z}^d$, $1 \leq p \leq \infty$, $a \in \ell^p(\Omega)$, and $b \in \ell^{p'}(\Omega)$. Then*

$$\|ab\|_1 \leq \|a\|_p \|b\|_{p'}.$$

If $p \in (1, \infty)$, we have equality if and only if the vectors $(|a_j|^p)_{j\in\Omega}$ and $(|b_j|^{p'})_{j\in\Omega}$ are linearly dependent, i.e., if there is $\lambda > 0$ such that $(|a_j|^p)_{j\in\Omega} = \lambda(|b_j|^{p'})_{j\in\Omega}$.

For $p = p'$, Hölder's inequality reduces to the Cauchy–Schwarz inequality. The proof uses the following inequality, which is sometimes called Young inequality or Peter–Paul inequality.

Lemma 1.16. *Let $a, b \geq 0$ and $1 < p < \infty$. Then $ab \leq a^p/p + b^{p'}/p'$ and equality holds if and only if $a^p = b^{p'}$.*

Proof. Without loss of generality, we assume $a, b > 0$. Let $t = 1/p$ and hence $1-t = 1/p'$. Then, by concavity of $\log : (0,\infty) \to \mathbb{R}$, we get

$$\log(ta^p + (1-t)b^{p'}) \geq t\log(a^p) + (1-t)\log(b^{p'}) = \log a + \log b = \log(ab).$$

Applying the monotone function $\exp : \mathbb{R} \to (0,\infty)$ to this inequality, we get the desired estimate. The previous bound also shows that $ap = a^p/p + b^{p'}/p'$ holds if and only if $a^p = b^{p'}$. □

We are now ready to give the

Proof of Theorem 1.15. Without loss of generality, we assume $a, b \neq \vec{0}$. The claim for $p \in \{1, \infty\}$ is obvious. Now let

$$\alpha := \frac{a}{\|a\|_p}, \qquad \beta := \frac{b}{\|b\|_{p'}}.$$

Then $\|\alpha\|_p = \|\beta\|_{p'} = 1$. By Young's inequality (Lemma 1.16),

$$\|\alpha \cdot \beta\|_1 = \sum_{j\in\Omega} |\alpha_j||\beta_j| \leq \sum_{j\in\Omega} \left(\frac{|\alpha_j|^p}{p} + \frac{|\beta_j|^{p'}}{p'} \right) = \frac{\|\alpha\|_p^p}{p} + \frac{\|\beta\|_{p'}^{p'}}{p'} = 1\,.$$

By definition of α, β, this is equivalent to the estimate

$$\left\| \frac{a}{\|a\|_p} \cdot \frac{b}{\|b\|_{p'}} \right\|_1 \leq 1,$$

which shows Hölder's inequality.

By the discussion of the optimality of Young's inequality, we see that Hölder's inequality is saturated if and only if $|\alpha_j|^p = |\beta_j|^{p'}$ holds for all $j \in \Omega$. By definition of α and β, this is equivalent to the existence of $\lambda > 0$ such that $(|a_j|^p)_{j\in\Omega} = \lambda(|b_j|^{p'})_{j\in\Omega}$. □

Related to the concept of limit are the notions *infimum* and *supremum*. These should express maximality and minimality in situations where it is not clear if these are actually attained. Consider the function $f(x) = e^x$. What's the smallest value it has? That's an ill-posed question, since for every value $y_0 = e^{x_0}$ you can find an even smaller value by choosing $x < x_0$ and getting $y = e^x < e^{x_0} = y_0$. But you will never get smaller than 0. So we would like to say something like "the smallest value of e^x is zero", but this is not exactly right. This is where the notion of infimum comes into play. We formulate it for sets of real numbers (there we have the same problem: What's the smallest element in the open interval $(0,1)$?)

Definition 1.17. Let $A \subset \mathbb{R}$. We say that x is a *lower bound* for A if for every $a \in A$ it holds that $x \leq a$. The *infimum* of A is the largest lower bound of A, i.e.

(i) it is a lower bound x^* of A and

(ii) every $x > x^*$ is *not* a lower bound of A anymore.

Similarly, x is an *upper bound* for A if for every $a \in A$ it holds that $a \leq x$. The *supremum* of A is the smallest upper bound of A, i.e.

(i) it is an upper bound x^* of A and

(ii) every $x < x^*$ is *not* an upper bound of A anymore.

We write $\inf A$ or $\inf_{x\in A} x$ for the infimum of A and $\sup A$ or $\sup_{x\in A} x$ for the supremum of A. If $\inf A \in A$, we have $\min A := \inf A$, called minimum of A. Likewise, if $\sup A \in A$, we have $\max A := \sup A$, called maximum of A.

If a set $A \subset \mathbb{R}$ is unbounded from above (i.e. it does not have an upper bound), we write $\sup A = \infty$ and if it is unbounded from below we write $\inf A = -\infty$.

Example 1.18.

1. Coming back to our motivation: it holds that $\inf_{x\in\mathbb{R}} e^x = 0$ since 0 is a lower bound for all values e^x and no number larger than zero is a lower bound.

2. For open and closed intervals we have

$$\begin{aligned} 0 &= \inf\,(0,1) = \inf\,[0,1], \\ 1 &= \sup\,(0,1) = \sup\,[0,1]. \end{aligned}$$

△

In some cases infima and suprema are attained, i.e. there is an element in the set which equals it. We've seen this in the example for the closed interval $A = [0,1]$: $1 = \sup A$ and also $1 \in A$. In this case we call the supremum the *maximum* of A. Similarly, if an infimum is attained we call in the minimum.

Example 1.19. It holds that

$$0 = \inf_{x\in\mathbb{R}} e^x = \inf\,\{e^x \mid x \in \mathbb{R}\}$$

but the set does not have a minimum. On the other hand we have

$$0 = \inf\left\{x^2 \mid x \in \mathbb{R}\right\} = \min\left\{x^2 \mid x \in \mathbb{R}\right\}$$

since 0 is an element of the set $\left\{x^2 \mid x \in \mathbb{R}\right\}$. △

2 Lecture 2: Inner products and metrics

We can put more structure on a vector space with so-called inner products:

Definition 2.1 (Inner product)**.** Let V be a real vector space. An *inner product* on V is a map

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

with the following properties:

(i) For all $x, y, z \in V$ and $\alpha \in \mathbb{R}$ we have *linearity in the first argument*, i.e.,

$$\langle x + \alpha y, z \rangle = \langle x, z \rangle + \alpha \langle y, z \rangle.$$

(ii) For all $x, y \in V$, we have *symmetry*, i.e.,

$$\langle x, y \rangle = \langle y, x \rangle.$$

(iii) For all $x \in V$, we have *positive definiteness*, i.e.,

$$\langle x, x \rangle \geq 0 \text{ and } \langle x, x \rangle = 0 \text{ only if } x = 0.$$

Remark 2.2. 1. Properties (i) and (ii) together imply that $\langle \cdot, \cdot \rangle$ is also linear in the second argument. Hence, an inner product is a so-called *bilinear map.*

2. For vector spaces V over $\mathbb{C}$, we require *sesquilinearity*, i.e., $\langle x, y \rangle = \overline{\langle y, x \rangle}$ and $\langle \lambda_1 a + b, \lambda_2 c + d \rangle = \overline{\lambda_1} \langle a, d \rangle + \overline{\lambda_1} \lambda_2 \langle a, c \rangle + \lambda_2 \langle b, c \rangle + \langle b, d \rangle$ for $\lambda_1, \lambda_2 \in \mathbb{C}$ and $a, b, c, d \in V$.

Example 2.3. 1. On $\mathbb{R}^n$ there is the *standard inner product*, also called *dot product*

$$\langle x, y \rangle_{\mathbb{R}^n} := \sum_{i=1}^{n} x_i y_i.$$

This inner product is also denoted by $x \cdot y$ (hence the name) and can also be written using matrix-vector multiplication as $x^T y$.

2. On the space of functions $u : \Omega \to \mathbb{R}$ we can define the so-called L^2-inner product

$$\langle u, v \rangle_{L^2} := \int_{\Omega} u(x) v(x) \mathrm{d}x.$$

3. In the spirit of the first point, we can define for any matrix $A \in \mathbb{R}^{n \times n}$ a bilinear map

$$\langle x, y \rangle_A := y^T A x.$$

This map is symmetric when A is so and is an inner product when A is *positive definite*, i.e. when $x^T A x > 0$ for $x \neq 0$.

△

Often one just uses $\langle x, y \rangle$ without any specifying index when the inner product is clear from the context or does not play a role.

It is of fundamental importance that *inner products induce norms*, i.e. for any inner product we can define

$$\|x\| := \sqrt{\langle x, x \rangle}$$

and one can show that this definition does indeed define a norm. This fact relies on positive definiteness of the inner product and the *Cauchy-Schwarz inequality* which states (expressed purely with inner products) that

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle.$$

Remark 2.4. Using the induced norm, the Cauchy-Schwarz inequality reads as

$$|\langle x, y\rangle| \leq \|x\| \, \|y\| \, .$$

Important special cases are

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)$$

and

$$\left(\int_\Omega f(x)g(x)\mathrm{d}x\right)^2 \leq \left(\int_\Omega f(x)^2\mathrm{d}x\right)\left(\int_\Omega g(x)^2\mathrm{d}x\right).$$

A norm $\|x\|$ quantifies the size of an element x in a vector space V. The norm can also be used to measure the distance between two elements $x, y \in V$ via $\|x - y\|$. We can measure the distance between x and y also using other maps, which we call metrics. The notion of a metric actually does not require any vector space structure at all.

Definition 2.5. Let X be a set. A *metric* on X is a map

$$d : X \times X \to (0, \infty)$$

with the properties

(i) For all $x, y \in X$ it holds that $d(x, y) = d(y, x)$. [*symmetry*]

(ii) It holds that $d(x, y) = 0$ exactly if $x = y$. [*positive definite*]

(iii) For all $x, y, z \in X$ it holds that $d(x, z) \leq d(x, y) + d(y, z)$. [*triangle inequality*]

Similarly to the case of a norm, we can define convergence of sequences in metric spaces: We say $x_n \to x$ with respect to the metric d if $d(x_n, x) \to 0$ (cf. Def. 1.11).

Similarly to how inner products induce norms, it is also true that *norms induce metrics*: If the set X is also a vector space and $\|\cdot\|$ is a norm on this space, then

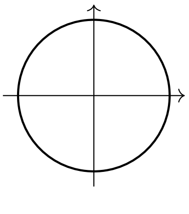
$$d(x, y) := \|x - y\|$$

is in fact a metric. (This is a nice exercise and you should try to prove this.)

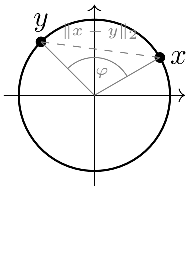
Metrics are often used when the underlying space is not a vector space.

Example 2.6. Let X be the unit circle in $\mathbb{R}^2$, i.e. the set

$$X = \left\{x \in \mathbb{R}^2 \mid \|x\|_2 = 1\right\}.$$



Since X is a subset of the vector space $\mathbb{R}^2$ we could measure the distance of $x, y \in X$ by $\|x - y\|_2$ which would be the distance we measure with a ruler. Alternatively, we could also measure the distance by the *angle* between x, y.



More generally, the angle between two vectors in $\mathbb{R}^d$ is given by

$$\measuredangle(x,y) := \arccos\left(\tfrac{\langle x,y\rangle}{\|x\|\|y\|}\right),$$

and this angle does indeed define a metric on the unit sphere $S^{d-1} = \left\{x \in \mathbb{R}^d \mid \|x\|_2 = 1\right\}$. △

Now we have

$$\text{inner product} \overset{\text{induces}}{\longrightarrow} \text{norm} \overset{\text{induces}}{\longrightarrow} \text{metric},$$

i.e. the metric is the most general notion.

Using the notion of metric, we can construct *topologies*. This concept is essential for a complete understanding of basic notions like “continuity”, “convergence”, “density”, “approximation”, etc. Classical topological notions are “open/closed set”, “neighborhood”, “boundary”, “accumulation point” etc.

Definition 2.7. Let X be a set equipped with a metric d.

1. For $x \in X$ and $\epsilon > 0$ we define

$$B_\epsilon(x) := \{y \in X \mid d(x,y) < \epsilon\}$$

and call $B_\epsilon(x)$ the *open ball* of radius ϵ around x or the *ϵ-neighborhood* of x.

2. A set $U \subset X$ is called *open*, if for all $x \in U$ there exists an $\epsilon > 0$ (which is allowed to depend on x) such hat $B_\epsilon(x) \subset U$.

3. A set $A \subset X$ is called *closed*, if its complement $A^{\complement} = X \setminus A$ is open.

Remark 2.8. Let (X, d) be a metric space.

1. If some given subset $U \subset X$ is open or not depends on the set X which contains U.

2. There are sets which are neither open, nor closed—think of the half-open interval $(a, b]$ as subset of $\mathbb{R}$. On the other hand, there are also sets which are both open and closed, called “clopen”. For instance, both the set X of the original metric space and the empty set $\emptyset$ are both open and closed. (Check that.)

In a metric space X a set A is closed if and only if the limit $x \in X$ of every convergent sequence $x_n \overset{n\to\infty}{\longrightarrow} x$ with $(x_n)_{n\in\mathbb{N}} \subseteq A$, one has $x \in A$. Put differently, convergent sequences in A cannot leave the set.

Example 2.9. We consider the real line $X = \mathbb{R}$ with the metric $d(x,y) = |x - y|$.

- The intervals (a, b) are open. (Given some $x \in (a,b)$, can you find an $\epsilon > 0$ such that $B_\epsilon(x) = (x - \epsilon, x + \epsilon) \subset (a, b)$?).

- The intervals $[a, b]$ are closed (which follows from the fact that if $x_n \leq a$ and $x_n \to x$, then $x \leq a$).

- The intervals $]a, b]$ and $[a, b[$ are neither closed, nor open, while $\mathbb{R}$ and $\emptyset$ are both open and closed.

△

2.1 Optional addenda

The following material is not relevant for the exam but essential for a rigorous study of concepts like continuity, differentiability, etc.

Definition 2.10. Let (X, d) be a metric space and $Y \subseteq X$.

1. A point $x \in X$ is called *boundary point* of Y, if every neighborhood of x contains at least one point in Y and one point in $X \setminus Y$. The set of all boundary points of Y is called "boundary" and denoted by $\partial\Omega$.
2. A point $x \in X$ is called *accumulation/limit point* of Y if every neighborhood of x contains at least one point in Y, which is different from x.
3. A point $x \in X$ is called *adherent point / point of closure* of Y if every neighborhood of x contains at least one point in Y.

Remark 2.11. Adherent points can be isolated. Moreover, we have the following immediate facts.

1. Every limit point is an adherent point, but not vice versa. A limit point is an adherent point, which is not isolated. (The notion of "isolated" is intuitively clear: A point x in a subset $A \subseteq X$ is isolated if there is a neighborhood around x, which does not contain any other points of A.)
2. Every $x \in Y$ is an adherent point of Y.
3. Every adherent point in $X \setminus Y$ is a limit point.

Theorem 2.12. *Let (X, d) be a metric space and $Y \subseteq X$. Then the following statements hold.*

1. *$Y \setminus \partial Y$ is open. Moreover, for every open $U \subseteq Y$, we have $U \subseteq Y \setminus \partial Y$.*
2. *$Y \cup \partial Y$ is closed. Moreover, for every $V \supseteq Y$, we have $V \supseteq (Y \cup \partial Y)$.*
3. *∂Y is closed.*
4. *Y is open if and only if Y does not contain any of its boundary points.*
5. *Y is closed if and only if Y contains all of its boundary points.*
6. *Y is closed if and only if Y contains all of its limit points.*

We have the following reformulation of the equivalence between closedness of a subset $A \subseteq X$ and having the property that it contains all of its limit points.

Theorem 2.13. *Let (X, d) be a metric space and $A \subseteq X$. Then A is closed if and only if the following statement holds: For all convergent sequences $(x_k)_{k\in\mathbb{N}} \subseteq A$ with $\lim_{k\to\infty} x_k =: x \in X$, we have $x \in A$.*

Proof. "$\Rightarrow$": Let A be closed, $(x_k)_{k\in\mathbb{N}} \subseteq A$, and $x = \lim_{k\to\infty} x_k \in X$. Suppose $x \in X \setminus A$, which is, by assumption, open. Then, in particular, $X \setminus A$ is a neighborhood of x. But by the definition of convergence, there is an $N \in \mathbb{N}$ such that $x_k \in X \setminus A$ for all $k \geq N$. But this contradicts the assumption that all elements of the sequence $(x_k)_{k\in\mathbb{N}}$ belong to A.

"$\Leftarrow$": We will show that $X \setminus A$ is open. Let $x \in X \setminus A$; then we show that there is $\epsilon = \epsilon(x)$ such that $B_\epsilon(x) \subseteq X \setminus A$. Suppose this was not true, i.e., for all $\epsilon > 0$ the ball $B_\epsilon(x)$ is not fully contained in $X \setminus A$, i.e., $B_\epsilon(x) \nsubseteq X \setminus A$. Then for each $k \in \mathbb{N}$ we could find $x_k \in A$ with $d(x, x_k) < 1/k$. But by assumption we know that all convergent sequences in A must remain in A, in particular, this sequence $(x_k)_{k\in\mathbb{N}}$ we just constructed must have its limit in A. But this is contradictory to the assumption that the limit x (the point we started our argument with) lies in $X \setminus A$. □

Theorem 2.14. *Let (X, d) be a metric space. Then finite intersections and arbitrary unions of open sets are again open.*

Proof. Finite intersections: Let $U, V \subseteq X$ be open and $x \in U \cap V$. Then there are $\epsilon_1, \epsilon_2 > 0$ such that $B_{\epsilon_1}(x) \subseteq U$ and $B_{\epsilon_2} \subseteq V$. Thus, for $\epsilon = \min\{\epsilon_1, \epsilon_2\}$, we see $B_\epsilon(x) \subseteq U \cap V$. Thus, $U \cap V$ is open.

Arbitrary unions: Let $I \subseteq \mathbb{N}$ be an index set, $\{U_i\}_{i \in I}$ be a family of open subsets $U_i \in X$, and $x \in \bigcup_{i \in I} U_i$. Then there is $j \in I$ such that $x \in U_j$. Since U_j is open, there is an $\epsilon > 0$ such that $B_\epsilon(x) \subseteq U_j \subseteq \bigcup_{i \in I} U_i$. □

Remark 2.15. The intersection of infinitely many open subsets need not be open again. Consider the family of open subsets $(-1/n, 1 + 1/n) \subseteq \mathbb{R}$, indexed by $n \in \mathbb{N}$. For each fixed $n \in \mathbb{N}$, these sets are open, but $\bigcap_{n \in \mathbb{N}} (-1/n, 1 + 1/n) = [0, 1]$ is not open.

3 Lecture 3: Banach spaces, Hilbert spaces and operator norms

By equipping vector spaces with norms, we can use all the notions like convergence, continuity and even differentiability on these spaces (and we will talk about differentiability later). But there is one ingredient missing and this is the notion of *completeness*. Roughly speaking completeness says that "we can't leave the space with a convergent sequence". The correct definition is (in the case of metric spaces):

Definition 3.1 (Cauchy sequence, completeness)**.** Let X be a metric space X with metric d. A sequence x_n in X is a *Cauchy sequence*, if for every $\epsilon > 0$ there exists and N such that for all $m, n \geq N$ it holds that $d(x_n, x_m) < \epsilon$.

The metric space X is *complete* if every Cauchy sequence in X has a limit in X, i.e. if for every Cauchy sequence in X there exists an $x \in X$ such that $x_n \to x$.

A vector space equipped with a norm that is complete with respect to the metric that is induced by that norm is called *Banach space*.

A vector space equipped with an inner product that is complete with respect to the norm induced by the inner product is called *Hilbert space*.

The obvious example of a Banach space is $\mathbb{R}^n$ with any of the ℓ^p-norms. If we use the dot product on $\mathbb{R}^n$ we get a Hilbert space.

We give some examples and non-examples of complete spaces.

Example 3.2.

- $(\mathbb{R}, |\cdot|)$ is complete. For a proof, see any analysis text book. We give an indication, using two ingredients, why this is true. The first one is the fact that Cauchy sequences are bounded—try to prove this. (Roughly speaking, we know that from the N-th element of a Cauchy sequence on, x_n with $n \geq N$ is not much different from x_N; on the other hand, we know the boundedness of finitely many elements, i.e., $|x_1| + ... + |x_N|$.) The second ingredient is a bit deeper: It is the so-called Bolzano–Weierstraß theorem. It says that every bounded sequence has a convergent subsequence. (Consider, e.g., the sequence $(x_n)_{n\in\mathbb{N}}$ given by $(-1)^n$. It has two (trivial) subsequences given by those sequences where we only admit even n and those for which we only admit odd n. In the former case, we consider the sequence $(-1)^{2k} = 1$ and in the latter case we consider $(-1)^{2k+1} = -1$.)
- $(\mathbb{R}^d, \|\cdot\|_p)$ with every $p \in [1, \infty]$ is complete. This is because of the equivalence of $\|\cdot\|_p$ norms, allowing is to infer convergence in any $\|\cdot\|_p$-norm from convergence in the $\|\cdot\|_\infty$-norm, which would be the analog of the norm $|\cdot|$ on $\mathbb{R}$.
- Here is a non-example: Consider $(0, 1) \subseteq \mathbb{R}$ with norm $|\cdot|$. Then the sequence $(x_n)_{n\in\mathbb{N}}$ defined by $x_n = 1/n$ is Cauchy, but its limit, $\lim_{n\to\infty} x_n = 0$ does not belong to $(0, 1)$. Thus, $((0, 1), |\cdot|)$ is not complete.
- On the other hand, one can show that $([0, 1], |\cdot|)$ is indeed complete. We see that completeness of a metric space could be connected to closedness of the space in the sense of topology.
- The rational numbers $(\mathbb{Q}, |\cdot|)$ are not complete. Consider, e.g., the sequence $(x_n)_{n\in\mathbb{N}}$ given by $x_1 = 1$ and $x_{n+1} = x_n/2 + 1/x_n$ with limit $x = x/2 + 1/x$ solving $x = \sqrt{2} \notin \mathbb{Q}$.

△

In many areas of data science, one works with *function spaces*:

Example 3.3. Let Ω be an open subset of $\mathbb{R}^d$.

1. The space

$$C(\Omega) := \{u : \Omega \to \mathbb{R} \mid u \text{ continuous and bounded}\}$$

is a vector space and we can equip it with the *maximum-norm* (also called supremum-norm or ∞-norm)

$$\|u\|_\infty = \max\{|u(x)| \mid x \in \Omega\}.$$

The convergence with respect to this norm is in fact uniform convergence (of sequences of functions) and since the uniform limit of a sequence of continuous functions is again continuous, this space is indeed complete and hence, a Banach space.

2. The space

$$L^2(\Omega) := \left\{ u : \Omega \to \mathbb{R} \mid \int_\Omega |u(x)|^2 \,\mathrm{d}x < \infty \right\}$$

is also a vector space (this is not clear, by the way) and we can equip it with the inner product from Example 2.3

$$\langle u, v \rangle = \int_\Omega u(x)v(x)\mathrm{d}x$$

which induces the L^2 norm

$$\|u\|_2 = \sqrt{\langle u, v \rangle} = \left(\int_\Omega |u(x)|^2 \,\mathrm{d}x \right)^{1/2}.$$

One can show that this space is complete as well, and hence, it is a Hilbert space.

△

Between two vector spaces X and Y we can consider linear maps A (which are called operators in this case). If X and Y are normed spaces, we can speak about continuous linear operators. It turns out that linear operators are continuous exactly if they are bounded in the sense of the following definition:

Definition 3.4. A linear operator $A : X \to Y$ between two normed spaces X and Y with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively, is called *bounded* if there exists a $C > 0$ such that for all $x \in X$ it holds that $\|Ax\|_Y \leq C \|x\|_X$. The infimum over all these constants C is called *operator norm* of A and denoted by $\|A\|_{X \to Y}$, i.e.

$$\|A\|_{X \to Y} := \inf \{ C \geq 0 \mid \|Ax\|_Y \leq C \|x\|_X \}.$$

If the spaces are clear from the context one just writes $\|A\|$.
One can show

- The set

$$L(X, Y) := \{ A : X \to Y | A \text{ linear and bounded} \}$$

is a vector space.

- The map $\|\cdot\|_{X \to Y}$ is a norm on $L(X, Y)$.
- The normed space $(L(X, Y), \|\cdot\|_{X \to Y})$ is complete, i.e., a Banach space.
- The operator norm is submultiplicative, i.e., if $A : X \to Y$ and $B : Y \to Z$ for three normed spaces X, Y, Z, then $\|BA\|_{X \to Z} \leq \|A\|_{X \to Y} \|B\|_{Y \to Z}$.
- Let $X \neq \emptyset$. Then,

$$\|A\| = \sup_{x \neq 0} \tfrac{\|Ax\|_Y}{\|x\|_X} = \sup_{\|x\|_X = 1} \|Ax\|_Y.$$

- (Optional). We call a linear map continuous if for every $\epsilon > 0$ there is a $\delta = \delta(\epsilon) > 0$ such that for all $x, y \in X$ it holds that if $\|x - y\| < \delta$, then $\|Ax - Ay\| < \epsilon$. With this definition, one can show that every bounded linear operator is continuous and every linear continuous operator is bounded.

Linear maps on finite dimensional spaces can be represented by matrices and the notion of operator norm is also interesting in this case.

Example 3.5 (Matrix norms). Let $A \in \mathbb{R}^{m\times n}$ a matrix which we also identify with the linear map $x \mapsto Ax$ from $\mathbb{R}^n$ to $\mathbb{R}^m$. We can equip $\mathbb{R}^n$ and $\mathbb{R}^m$ with different norms and obtain different induced operator norms:

1. We choose in the domain of definition ($\mathbb{R}^n$) and the range ($\mathbb{R}^m$) the same norm, namely the ℓ^1-norm. To calculate the respective operator norm, we examine $\|Ax\|_1$ and try to estimate this expression as tight as possible with $\|x\|_1$:

$$\begin{aligned} \|Ax\|_1 &= \sum_{i=1}^m |(Ax)_i| = \sum_{i=1}^m \left|\sum_{j=1}^n a_{ij}x_j\right| \\ &\leq \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|\,|x_j| = \sum_{j=1}^n \sum_{i=1}^m |a_{ij}|\,|x_j| \\ &\leq \max_{j=1,\dots,n}\left(\sum_{i=1}^m |a_{ij}|\right) \underbrace{\sum_{j=1}^n |x_j|}_{=\|x\|_1}. \end{aligned}$$

   This estimate already shows that $\|A\|_{1\to1} \leq \max_{j=1,\dots,n}\left(\sum_{i=1}^m |a_{ij}|\right)$. To show that this is not just an upper bound, but already the true value of the operator norm, it is enough to find a single vector x such that the above estimate is tight. If j^* is the index where the maximum is attained, we can choose $x = e_{j^*}$ and observe that we get an equality. Hence, we have shown that

$$\|A\|_{1\to1} = \max_{j=1,\dots,n}\left(\sum_{i=1}^m |a_{ij}|\right)$$

   and this norm is called *column-sum norm.*

2. If one chooses the ℓ^∞ norm in the domain and the range, one obtains (with a more or less similar argument) the *row-sum norm*

$$\|A\|_{\infty\to\infty} = \max_{i=1,\dots,m}\left(\sum_{j=1}^n |a_{ij}|\right).$$

3. If we choose the ℓ^2-norm in both the domain and the range we estimate a little differently: We use the fact that the matrix $A^TA \in \mathbb{R}^{n\times n}$ is symmetric positive semi-definite and hence, has an eigenvalue decomposition $A^TA = V\Sigma V^T$ with an orthonormal matrix $V \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ with non-negative entries σ_i (ordered decreasingly). We get that

$$\begin{aligned} \|Ax\|_2^2 &= \langle Ax, Ax\rangle = \langle A^TAx, x\rangle = \langle V\Sigma V^Tx, x\rangle \\ &= \langle \Sigma V^Tx, V^Tx\rangle \leq \sigma_1\langle V^Tx, V^Tx\rangle = \sigma_1 \left\|V^Tx\right\|_2^2 = \sigma_1 \|x\|_2^2. \end{aligned}$$

   By choosing $x = v_1$ (the normalized eigenvector with respect to the first eigenvalue) we get $V^Tx = e_1$ and we get equality above. Hence we have shows that

$$\|A\|_{2\to2} = \sqrt{\sigma_{\max}(A^TA)}$$

   (where $\sigma_{\max}(M)$ denotes the largest eigenvalue of the symmetric positive definite matrix M). This norm is also called *spectral norm.*

△

Remark 3.6 (The Frobenius norm and matrix inner product). Not all matrix norms are operator norms! The *Frobenius norm* defined by

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}$$

is a matrix norm, but not an operator norm.

Remark 3.7. There is a nice connection to the spectral norm: $\|A\|_{2\to 2} = \sqrt{\sigma_{\max}(A^T A)} \leq \sqrt{\sum_{i=1}^{n} \sigma_i(A^T A)} = \|A\|_F$.

The Frobenius norm is widely used not only because it is simple to understand and straightforward to compute but also since it is actually induced by an inner product: The standard inner product on the space of $m \times n$ matrices is

$$\langle A, B\rangle = \operatorname{trace}(A^T B).$$

Here, $\operatorname{trace}(M)$ denotes the trace of a square matrix M, i.e. the sum of entries on the diagonal. Moreover, we have

$$\langle A, A\rangle = \operatorname{trace}(A^T A) = \sum_{j=1}^{n}\sum_{i=1}^{m} |a_{ij}|^2 = \|A\|_F^2 .$$

4 Lecture 4: Derivatives of maps

The well known definition of the derivative of a function $f : \mathbb{R} \to \mathbb{R}$ is the one as limit of the difference quotient:

$$f'(x) = \lim_{h\to 0} \tfrac{f(x+h)-f(x)}{h}.$$

The intuition behind this definition is that the derivative is the limit of slopes approximating the slope of the tangent.

A different but equivalent definition is the following: A function f is differentiable at x_0 if there exists some real c such that

$$f(x+h) = f(x) + ch + \varphi(h) \quad \text{with} \quad \lim_{h\to 0} \tfrac{|\varphi(h)|}{|h|} = 0.$$

The intuition behind this is: f is differentiable, if there exists a line with slope c which is a good approximation to f in the sense that the difference $f(x+h) - (f(x) + ch)$ vanishes quicker than h for $h \to 0$. In this case it holds that $f'(x) = c$.

This equivalent definition can easily be generalized to maps between Banach spaces:

Definition 4.1 (Derivative)**.** Let $f : X \to Y$ be a function between two Banach spaces X and Y. We say that f is differentiable in $x \in X$ if there exists $A \in L(X, Y)$ such that

$$f(x+h) = f(x) + Ah + \varphi(h) \quad \text{with} \quad \lim_{h\to 0} \tfrac{\|\varphi(h)\|_Y}{\|h\|_X} = 0.$$

(As always, $h \to 0$ means $\|h\|_X \to 0$.) The map A is called *derivative* of f at x and we write $A = Df(x)$. In this context we will often write $Ah = Df(x)[h]$, i.e. we use $[h]$ to plug in an h.

Remark 4.2. An alternative definition of differentiability is as follows. We call a map $f : X \to Y$ differentiable at $x \in X$ if there is $A \in L(X, Y)$ such that

$$\lim_{\|h\|_X\to 0} \frac{\|f(x+h) - f(x) - Ah\|_Y}{\|h\|_X} = 0.$$

Note that $Df(x)$ is a map (which maps an element of X to an element of Y) but this map is different in every point x. Df is also a map, but this map maps an element of X to an element of the space of all linear and bounded maps from X to Y. In formulae: $Df(x) \in L(X, Y)$ and $x \mapsto Df(x)$ is a map between X and $L(X, Y)$.

A simpler (but not as useful) notion of derivative is the *directional derivative.*

Definition 4.3. Let $f : X \to Y$ be differentiable. The directional derivative at $x \in X$ in direction $h \in X$ is defined as

$$D_h f(x) := \lim_{t\to 0} \tfrac{f(x+th)-f(x)}{t}.$$

Example 4.4 (The Jacobian I). Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, where we equip $\mathbb{R}^n$ with any norm $\|\cdot\|$ and $\mathbb{R}$ with the norm given by the absolute value $|\cdot|$. Recall that all norms on finite dimensional vector spaces, like $\mathbb{R}^n$ are equivalent. (For instance, we proved $\|x\|_p \leq c_{d,q}\|x\|_q$ for any $p, q \in [1, \infty]$.) Hence, we do and will not have to restrict ourselves to some special norm, like the euclidean norm, on $\mathbb{R}^n$.

We wish to compute the derivative $Df(x)$ at $x \in \mathbb{R}^n$. By definition, there is $A \in L(\mathbb{R}^n \to \mathbb{R})$ such that

$$f(x+h) = f(x) + Ah + \varphi(h) \quad \text{with} \lim_{\|h\|\to 0} \frac{|\varphi(h)|}{\|h\|} = 0 \quad \text{for all } h \in \mathbb{R}^n.$$

We know that every linear map between finite-dimensional spaces can be represented as a matrix. So, let us represent $A \in L(\mathbb{R}^n \to \mathbb{R})$ as a $\mathbb{R}^{1\times n}$ matrix in the canonical basis $\{e_j\}_{j=1,\dots,n}$ with $e_j = (0\,0 \ldots 0\,1\,0 \ldots 0)^T$, which is the vector having 1 in its j-th component and zero elsewhere.

Thus, we write $A = (a_1 \dots a_n)$ with $a_1, ..., a_n \in \mathbb{R}$. Let us also decompose $h = \sum_{j=1}^n h_j e_j$ with $h_j = \langle e_j, h \rangle \in \mathbb{R}$. Thus,

$$f(x+h) = f(x) + \sum_{j=1}^n a_j h_j + \varphi(h).$$

Thus, to find the derivative $Df(x)$, it suffices to find the a_ℓ for all $\ell \in \{1, ..., n\}$. To that end, we take $h = h_\ell e_\ell$. Then,

$$f(x + h_\ell e_\ell) = f(x) + a_\ell h_\ell + \varphi(h),$$

or, equivalently,

$$a_\ell = \frac{f(x + h_\ell e_\ell) - f(x) - \varphi(h)}{h_\ell}.$$

Since the left-hand side does not depend on h_ℓ, we can take the limit $h_\ell \to 0$ on both sides and obtain

$$a_\ell = \lim_{h_\ell \to 0} \frac{f(x + h_\ell e_\ell) - f(x) - \varphi(h)}{h_\ell} = \frac{\partial f}{\partial x_\ell}(x),$$

where we used that $\lim_{h\to 0} |\varphi(h)| \|h\|^{-1} = 0$. Thus,

$$Df(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \dots & \frac{\partial f}{\partial x_n} \end{pmatrix} = (\nabla f(x))^T,$$

where the right-hand side is just the transpose of the gradient of f, defined by

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}.$$

△

Example 4.5 (The Jacobian II). We can generalize the previous example and consider differentiable vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$. In this case, the derivative $Df(x) \in L(\mathbb{R}^n \to \mathbb{R}^m)$ can be represented as a $\mathbb{R}^{m\times n}$ matrix. By writing $f(x) = (f_1(x)\, f_2(x) \dots f_m(x))$, we get

$$Df(x) = \begin{pmatrix} (\nabla f_1(x))^T \\ \vdots \\ (\nabla f_m(x))^T \end{pmatrix}.$$

△

Remark 4.6. If $f : X \to Y$ is differentiable, then the directional derivatives exist for all directions h and one has $D_h f(x) = Df(x)[h]$. The converse is not true, i.e., there are functions for which the directional derivatives exist in all directions, but which are not differentiable. Here is an example. Consider

$$\mathbb{R}^2 \ni (x,y) \mapsto f(x,y) := \begin{cases} 0 & \text{for } (x,y) = (0,0), \\ \frac{x^3}{x^2+y^2} & \text{for } (x,y) \neq= (0,0). \end{cases}$$

The directional derivatives of f in the x- and y-direction are given by the gradient

$$\nabla f(x) = (x^2+y^2)^{-2} \begin{pmatrix} 3x^2(x^2+y^2) - 2x^4 \\ -2x^3 y \end{pmatrix}.$$

In particular, the gradient can be used to contruct all other directional derivatives of f.

However, f is not differentiable at zero. Indeed, if f was differentiable at zero, then, since $\nabla f(0) = 0$, the right-hand side of

$$\frac{|f(h) - f(0) - (\nabla f(0))^T h|}{\|h\|} = \frac{|f(h)|}{\|h\|}$$

would have to vanish as $h \to 0$. But consider $h = (h_1, 0)$. Then,

$$\frac{|f((h_1\, 0)^T)|}{\|(h_1\, 0)^T\|} = 1,$$

which does not vanish as $h_1 \to 0$. Thus, f is not differentiable at zero although all of its directional derivatives exist.

After this quite abstract introduction, let us come to more down to earth examples where our definition of derivative as a linear map comes in handy:

Example 4.7 (Functions of matrices). We consider functions f defined on sets of $m \times n$ matrices. For a map $f : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times k}$, for example, we could in principle identify $\mathbb{R}^{m\times n} \sim \mathbb{R}^{mn}$, $\mathbb{R}^{p\times k} \sim \mathbb{R}^{pk}$ and express the derivative as a $pk \times mn$ Jacobian matrix. This may get quite complicated and knowing the elements of the Jacobian may not be very helpful. Our definition of derivative says, that the derivative of such a function has to be a linear map from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{p\times k}$.

Here is concrete example. We consider

$$f : \mathbb{R}^{m\times n} \to \mathbb{R}^{n\times n}, f(R) = R^T R.$$

Let us calculate the derivative of f using Definition 4.1:

$$f(R + H) = (R + H)^T (R + H) = \underbrace{R^T R}_{=f(R)} + \underbrace{R^T H + H^T R}_{\text{linear in } H} + \underbrace{H^T H}_{?}.$$

We note that the map $H \mapsto R^T H + H^T R$ is indeed linear in the variable H! (Check this.) Hence, this has to be the derivative and we only need to show that $\varphi(H) = H^T H$ decays fast enough. But this is quite simple with our knowledge on operator norms:

$$\left\| \tfrac{\varphi(H)}{\|H\|} - 0 \right\| = \tfrac{\|H^T H\|}{\|H\|} \le \tfrac{\|H^T\| \|H\|}{\|H\|} = \left\|H^T\right\| \to 0 \quad \text{for} \quad H \to 0.$$

Hence we have

$$Df(R)[H] = R^T H + H^T R.$$

Note that we can't simplify this any further since the multiplication of matrices in not commutative.

Not having the derivative as a concrete object may not always be satisfying, but quite often one only needs to evaluate the derivative in certain directions. △

Example 4.8 (The least squares functional). For a matrix $A \in \mathbb{R}^{m\times n}$, and vectors $b \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ we consider the functional

$$\tfrac{1}{2} \|Ax - b\|_2^2 .$$

We would like to take the derivative of this functional with respect to all variables. Let us write $F(A, x, b) = \frac{1}{2} \|Ax - b\|_2^2$ and take the derivative with respect to b:

$$\begin{aligned} F(A, x, b + h) &= \tfrac{1}{2} \|Ax - (b + h)\|_2^2 = \tfrac{1}{2} \|Ax - b - h\|_2^2 \\ &= \underbrace{\tfrac{1}{2} \|Ax - b\|_2^2}_{F(A,x,b)} \underbrace{- \langle Ax - b, h\rangle}_{D_b F(A,x,b)[h]} + \underbrace{\tfrac{1}{2} \|h\|_2^2}_{=\varphi(h)} . \end{aligned}$$

Since $\varphi(h)/\left\|h\right\|_2 = \frac{1}{2}\left\|h\right\|_2 \to 0$ for $h \to 0$ we get

$$D_bF(A,x,b)[h] = -\langle Ax-b, h\rangle = -(Ax-b)^T h$$

and we can also write

$$D_bF(A,x,b) = -(Ax-b)^T, \quad \text{and} \quad \nabla_b F(A,x,b) = -(Ax-b).$$

For the derivative with respect to x we proceed analogously:

$$\begin{aligned} F(A,x-h,b) &= \tfrac{1}{2}\left\|A(x+h)-b\right\|_2^2 = \tfrac{1}{2}\left\|Ax+Ah-b\right\|_2^2 \\ &= \underbrace{\tfrac{1}{2}\left\|Ax-b\right\|_2^2}_{F(A,x,b)} + \underbrace{\langle Ax-b, Ah\rangle}_{D_xF(A,x,b)[h]} + \underbrace{\tfrac{1}{2}\left\|Ah\right\|_2^2}_{=\varphi(h)}. \end{aligned}$$

Again we have $\varphi(h)/\left\|h\right\|_2 = \frac{1}{2}\left\|Ah\right\|_2^2/\left\|h\right\|_2 \le \frac{1}{2}\left\|A\right\|^2\left\|h\right\|_2 \to 0$ for $h \to 0$ and hence

$$D_xF(A,x,b)[h] = \langle Ax-b, Ah\rangle = (Ax-b)^T Ah$$

and we can also write

$$D_xF(A,x,b) = (Ax-b)^T A, \quad \text{and} \quad \nabla_x F(A,x,b) = A^T(Ax-b).$$

Now for the derivative with respect to A:

$$\begin{aligned} F(A+H,x,b) &= \tfrac{1}{2}\left\|(A+H)x-b\right\|_2^2 = \tfrac{1}{2}\left\|Ax+Hx-b\right\|_2^2 \\ &= \underbrace{\tfrac{1}{2}\left\|Ax-b\right\|_2^2}_{F(A,x,b)} + \underbrace{\langle Ax-b, Hx\rangle}_{D_AF(A,x,b)[H]} + \underbrace{\tfrac{1}{2}\left\|Hx\right\|_2^2}_{=\varphi(H)}. \end{aligned}$$

We see that $2\Phi(H)/\left\|H\right\| = \left\|Hx\right\|_2^2/\left\|H\right\| \le \left\|H\right\|\left\|x\right\|_2^2 \to 0$ for $H \to 0$ an thus the derivative is

$$D_AF(A,x,b)[H] = \langle Ax-b, Hx\rangle.$$

While this is correct (and indeed a linear function in H) we can rewrite this as

$$\begin{aligned} \langle Ax-b, Hx\rangle &= (Ax-b)^T Hx = (x^TA^T - b^T)Hx \\ &= \operatorname{trace}((x^TA^T-b^T)Hx) \\ &= \operatorname{trace}((xx^TA^T - xb^T)H) \\ &= \langle Axx^T - bx^T, H\rangle_{\text{Frobenius}}, \end{aligned}$$

using the inner product form matrices from Remark 3.6. Moreover, we used the two facts:

1. A scalar is its own trace.
2. The trace is cyclic, i.e., $\operatorname{trace}(ABC) = \operatorname{trace}(CAB)$, whenever meaningfully defined.

Hence, we can write

$$D_AF(A,x,b)[H] = \langle Axx^T - bx^T, H\rangle_{\text{Frobenius}}, \quad \text{and} \quad \nabla_A F(A,x,b) = Axx^T - bx^T.$$

Convince yourself that all dimensions in this formula check out. △

5 Optional: Lecture 5: Higher order derivatives and derivatives of functionals

The following material is not relevant for the exam.

We have seen that the derivative of $f : X \to Y$ at $x \in X$ is $Df(x) \in L(X, Y)$. So it makes sense to write

$$Df : X \to L(X, Y).$$

Since $L(X, Y)$ is again a Banach space, we can apply the notion of derivative to the object Df again and get a higher order derivative:

Definition 5.1 (Higher order derivatives)**.** Let $f : X \to Y$ be a map between two Banach spaces X and Y with derivative Df. The *second derivative* if f is

$$D^2 f(x) := D(Df)(x)$$

and higher order derivatives are defined recursively in the same manner.

This definition may be hard to digest. Let us unwrap it a bit:

- Recall that Df is a map that maps some $x \in X$ to a bounded linear map from X to Y, hence we can plug an $h \in X$ into $Df(x)$ and get $Df(x)[h] \in Y$.
- Now $D^2 f$ is a map from X to the space of bounded linear maps from X to $L(X, Y)$. Hence, if we plug some $h \in X$ into $D^2 f(x)$ we still get a bounded linear map from X to Y. Hence, we can plug in another $k \in X$ and get $D^2 f(x)[h][k] \in Y$. Usually one writes $D^2 f(x)[h, k]$ instead of $D^2 f(x)[h][k]$.
- The expression $D^2 f(x)[h, k]$ is linear in both h and k and as such it is an (in fact bounded) *bilinear map* from $X \times X$ to Y.

We see that the second derivative is an object of the type $D^2 f(x) \in L(X, L(X, Y))$ and we can identify it with a bilinear map from $X \times X$ to Y. In fact it even holds that

$$L(X, L(X, Y)) \simeq L_2(X, Y) := \{\text{bilinear maps from } X \times X \text{ to } Y\}.$$

Things may become even more simple if we consider the following special case:

Example 5.2 (The Hessian). We consider $X = \mathbb{R}^n$ and $Y = \mathbb{R}$, i.e. $f : \mathbb{R}^n \to \mathbb{R}$. We have $Df(x) \in \mathbb{R}^{1\times n}$. The second derivative is a bilinear map from $\mathbb{R}^n$ to $\mathbb{R}$. You may recall that all such bilinear maps $\Phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ are of the form $(x, y) \mapsto x^T A y$ for some matrix $A \in \mathbb{R}^{n\times n}$. You get this matrix by plugging in standard basis vectors: $a_{ij} = \Phi(e_i, e_j)$. Hence, we can represent the second derivative $D^2 f(x)$ by some $n \times n$ matrix!

[That all bilinear maps from $\mathbb{R}^n$ to $\mathbb{R}$ are of the form $\varphi(h, k) = h^T A k$ can be seen by writing $h = \sum h_i e_j$, $k = \sum k_i e_i$, plugging this into φ and using bilinearity to get $\varphi(h, k) = \sum_{ij} h_i k_j \varphi(e_i, e_j)$. Now just define $A_{ij} = \varphi(e_i, e_j)$.]

We have that

$$D^2 f(x)[h, k] = h^T H_f(x) k$$

and the matrix $H_f(x) = (\frac{\partial^2 f(x)}{\partial x_i \partial x_j})_{ij}$ is called the *Hessian* of f at x. △

Even higher derivatives are

$$D^k f : X \to L_k(X, Y), \quad L_k(X, Y) := \begin{Bmatrix} \text{maps from } X \times \cdots \times X \text{ to } Y \text{ which} \\ \text{are linear in each argument} \end{Bmatrix}.$$

For $X = \mathbb{R}^n$ and $Y = \mathbb{R}$ we have the quite long expression

$$D^k f(x)[h^1, \ldots, h^k] = \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \frac{\partial^k f(x)}{\partial x_{i_1} \cdots \partial x_{i_k}} h^1_{i_1} \cdots h^k_{i_k}.$$

Example 5.3. Let us look at the function $f(x,y) = \exp(-x^2 - y^2)$. The first and second derivatives are

$$\begin{aligned} Df(x) &= \begin{bmatrix} -2xe^{-x^2-y^2} & -2ye^{-x^2-y^2} \end{bmatrix}, \\ D^2f(x) &= \begin{bmatrix} (-2+4x^2)e^{-x^2-y^2} & 4xye^{-x^2-y^2} \\ 4xye^{-x^2-y^2} & (-2+4y^2)e^{-x^2-y^2} \end{bmatrix}. \end{aligned}$$

△

Higher derivatives can be used to define Taylor's formula: The expansion of second order of a function $f : X \to \mathbb{R}$ reads as

$$f(x+h) = f(x) + Df(x)[h] + \tfrac{1}{2}D^2f(x)[h,h] + \varphi(h), \quad \text{with} \quad \tfrac{\varphi(h)}{\|h\|^2} \to 0.$$

The case $\mathbb{R}^n$ can be written as

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \tfrac{1}{2}\langle h, H_f(x)h\rangle + \varphi(h), \quad \text{with} \quad \tfrac{\varphi(h)}{\|h\|^2} \to 0.$$

We can push this further with some effort: The n-th order Taylor expansion of $f : \mathbb{R}^d \to \mathbb{R}$is

$$T_nf(x_0+h) = \sum_{k=0}^{n} \tfrac{1}{k!}D^kf(x_0)\cdot h^k$$

but note that the objects $D^kf(x_0)$ and δ^k are

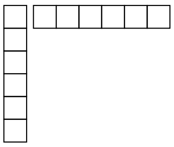
$$\begin{aligned} D^kf(x_0) &= \left(\tfrac{\partial^k f(x)}{\partial x_{i_1}\cdots\partial x_{i_k}}\right)_{\substack{i_1=1,\dots,d \\ \vdots \\ i_k=1\dots,d}} \\ h^k &= (h_{i_1}\cdots h_{i_k})_{\substack{i_1=1,\dots,d \\ \vdots \\ i_k=1\dots,d}}, \end{aligned}$$

respectively. More explicitly, we have

$$D^kf(x_0)\cdot h^k = \sum_{i_1=1}^{d}\cdots\sum_{i_k=1}^{d} \tfrac{\partial^k f(x)}{\partial x_{i_1}\cdots\partial x_{i_k}} h^1_{i_1}\cdots h^k_{i_k}.$$

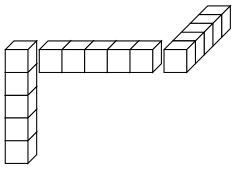
The object h^k is also denoted by $h\otimes\cdots\otimes h$. We can visualize this for $h \in \mathbb{R}^d$ and $k = 2$ as

$$h\otimes h = hh^T = (h_ih_j)_{i,j} = \begin{bmatrix} h_1h_1 & \cdots & h_1h_d \\ \vdots & & \vdots \\ h_1h_d & \cdots & h_dh_d \end{bmatrix} \in \mathbb{R}^{d\times d},$$



and for $k = 3$ as

$$h\otimes h\otimes h = (h_ih_jh_k)_{i,j,k} \in \mathbb{R}^{d\times d\times d},$$



As last topic we consider function spaces, i.e.

$$X = \{\text{some set of function from } [a,b] \text{ to } \mathbb{R}\}.$$

We do we need such things? Here is an example:

Example 5.4. Consider the map, that maps a function $u : [a, b] \to \mathbb{R}$ to the integral $\int_a^a |u'(x)|^2 \, \mathrm{d}x$, i.e. the map

$$F : X \to \mathbb{R}, \quad u \mapsto \int_a^a |u'(x)|^2 \, \mathrm{d}x.$$

If we would have such a function as part in optimization problem (where we would like to find some optimal function) we would like to be able to take the derivative of this functional. So we ask: What is $DF(u)$?

$$\begin{aligned} F(u+h) &= \int\limits_a^b (u'(x) + h'(x))^2 \mathrm{d}x = \int\limits_a^b (u'(x))^2 + 2u'(x)h'(x) + (h'(x))^2 \mathrm{d}x \\ &= \underbrace{\int\limits_a^b (u'(x))^2 \mathrm{d}x}_{=F(u)} + \underbrace{\int\limits_a^b 2u'(x)h'(x) \mathrm{d}x}_{=DF(u)[h]?} + \underbrace{\int\limits_a^b (h'(x))^2 \mathrm{d}x}_{=\varphi(h)} . \end{aligned}$$

Let us have closer look at the middle term: Using integration by parts we get

$$\int\limits_a^b 2u'(x)h'(x) \mathrm{d}x = 2u'(x)h(x)\Big|_{x=a}^b - \int\limits_a^b 2u''(x)h(x) \mathrm{d}x.$$

If we assume that $h(a) = h(b) = 0$ we get (using the inner product $\langle f, g \rangle = \int_a^b f(x)g(x)\mathrm{d}x$)

$$\int\limits_a^b 2u'(x)h'(x) \mathrm{d}x = \langle -2u'', h \rangle.$$

Hence, it makes sense to conclude that $\nabla F(u) = -2u''$. Here we left out some technical details which can be found in books on functional analysis or optimization in function space.]

△

Here is another example:

Example 5.5. Consider that we are given a function $f : [a, b] \to \mathbb{R}$ which is noisy and we would like to find a function $u : [a, b] \to \mathbb{R}$ that is close to f but smooth. We could, for example, solve this problem by solving a minimization problem like this:

$$\min_u \int\limits_a^b |u(x) - f(x)|^2 \, \mathrm{d}x + \lambda \int\limits_a^b |u'(x)|^2 \, \mathrm{d}x, \text{for some } \lambda > 0.$$

If we define the objective function of this problem as F, we can calculate (similar to the previous example)

$$\nabla F(u) = 2(u - f) - 2\lambda u'' = 0.$$

Hence, an optimal u is characterized by the differential equation $u - \lambda u'' = f$. △