Topics in learning theory*

Lecture 3: Recap on linear algebra and differential calculus

Contents

A significant part of learning theory builds upon concepts from algorithmic optimization. To develop a proper intuition about the field, a good command of standard notions of differential calculus is required. This lecture briefly reviews necessary material.

1 Euclidean spaces

Let E be an $\mathbb{R}$-vector space. Recall that a norm on E is a map $\|.\| : E \to \mathbb{R}_+$ satisfying the following properties:

- **Separation.** [1] For all $x \in E$: $\|x\| = 0 \Leftrightarrow x = 0$;
- **Homogeneity.** For all $x \in E$ and all $\lambda \in \mathbb{R}$: $\|\lambda x\| = |\lambda|\|x\|$;
- **Triangular inequality.** For all $x, y \in E$: $\|x + y\| \leq \|x\| + \|y\|$.

A vector space E equipped with a norm is referred to as a normed vector space. In the sequel, we'll be mostly interested in Euclidean spaces, i.e., finite dimensional normed vector spaces whose norm is inherited from a scalar product. Recall that a scalar product on a vector space E is a symmetric, bilinear and positive definite map $\langle ., . \rangle : E \times E \to \mathbb{R}$, i.e., satisfies:

- **Symmetry**. For all $x, y \in E$, $\langle x, y \rangle = \langle y, x \rangle$,
- **Bilinearity**. For all $x, y, z \in E$ and all $\alpha, \beta \in \mathbb{R}$,

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

- **Positive definiteness**. For all $x \in E$, $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Leftrightarrow x = 0$.

Given any scalar product $\langle ., . \rangle : E \times E \to \mathbb{R}$ on a vector space E, a fundamental result, known as the Cauchy-Schwarz inequality, states that

$$\forall x, y \in E, \quad |\langle x, y \rangle| \leq \sqrt{\langle x, x \rangle}\sqrt{\langle y, y \rangle}.$$

This inequality implies in particular that the map $\|.\| : E \to \mathbb{R}_+$, defined by

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

is a norm on E, which we call the norm inherited by (or associated to) the scalar product $\langle ., . \rangle : E \times E \to \mathbb{R}$.

Definition 1.1. *A normed vector space $(E, \|.\|)$ is called Euclidean if E is finite dimensional*[2] *and the norm $\|.\|$ is inherited from a scalar product on E.*

The normed vector space $(\mathbb{R}^d, \|.\|_2)$ is the most standard example of a Euclidean space since the norm

$$\|x\|_2 := \sqrt{\sum_{i=1}^{d} x_i^2}, \quad x = (x_1, \ldots, x_d)^\top,$$

is associated to the scalar product

$$\langle x, y \rangle := x^\top y = \sum_{i=1}^{d} x_i y_i.$$

More generally, given a symmetric and positive definite matrix[3] $A \in \mathbb{R}^{d \times d}$, the expression

$$\langle x, y \rangle_A := x^\top A y$$

defines a scalar product on $\mathbb{R}^d$ and the inherited norm is usually denoted $\|.\|_A$. Actually any d-dimensional Euclidean space $(E, \|.\|)$ can be identified with $(\mathbb{R}^d, \|.\|_A)$ for a certain symmetric and positive definite matrix $A \in \mathbb{R}^{d \times d}$. Indeed, having fixed a particular basis $e_1, \ldots, e_d$ of E, one observes that the map $\phi : E \to \mathbb{R}^d$ defined, for any $x = \sum_{i=1}^{d} x_i e_i \in E$ by

$$\phi(x) = (x_1, \ldots, x_d)^\top,$$

is a linear bijection satisfying

*Teaching material can be found at https://www.qparis-math.com/teaching.

[1] The separation axiom is also called the positivity condition. Note the distinction between the number "0" in $\mathbb{R}$ and the null vector "0" in E in this statement. This distinction is usually clear from the context.

[2] A normed vector space $(E, \|.\|)$, not necessarily finite dimensional and whose norm $\|.\|$ is inherited from a scalar product on E is more generally called a pre-Hilbert space. A Hilbert space is a pre-Hilbert space that is in addition complete for the norm inherited from the scalar product.

[3] Recall that a matrix $A \in \mathbb{R}^{d \times d}$ is called symmetric and positive definite if

- $A^\top = A$,
- For all $x \in \mathbb{R}^d$, $x^\top A x \geq 0$,
- For all $x \in \mathbb{R}^d$, $x^\top A x = 0$ if and only if $x = 0$.

$$\begin{aligned}\langle x, y\rangle &= \langle \sum_{i=1}^{d} x_i e_i, \sum_{i=1}^{d} y_i e_i\rangle \\ &= \sum_{i,j=1}^{d} x_i y_j \langle e_i, e_j\rangle \\ &= \phi(x)^\top A \phi(y),\end{aligned}$$

where

$$A := (\langle e_i, e_j\rangle)_{1\leq i,j\leq d},$$

is symmetric and positive definite.

If $(E, \|.\|)$ is a Euclidean space, there are a number of useful identities connecting the norm with the associated scalar product. For instance, the classical polarization identities, which allow to express the scalar product in terms of the associated norm only, state that for all $x, y \in E$,

- $2\langle x, y\rangle = \|x\|^2 + \|y\|^2 - \|x - y\|^2$,
- $2\langle x, y\rangle = \|x + y\|^2 - \|x\|^2 - \|y\|^2$.

Adding these two identities, we obtain the parallelogram identity:

$$\forall x, y \in E,\ 2\|x\|^2 + 2\|y\|^2 = \|x + y\|^2 + \|x - y\|^2.$$

We end by a few elementary notions of topology which will be usefull in the sequel.

Let $(E, \|.\|)$ be a finite dimensional normed vector space. For $x \in E$ and $r > 0$, we denote

$$B(x, r) := \{y \in E : \|x - y\| < r\},$$

the open ball with center x and radius $r > 0$ in E.

A subset $A \subset E$ is called bounded if there exists $x \in E$ and $r > 0$ such that $A \subset B(x, r)$.

A subset $U \subset E$ is called open if, for all $x \in U$, there exists $r > 0$ (possibly depending on x) such that $B(x, r) \subset U$.

A subset $F \subset E$ is called closed if its complement $E \setminus F$ is open. Equivalently, one shows that a set $F \subset E$ is closed if and only if, for any sequence $(x_n)_{n\geq 0}$ of elements of F that converges to some limit $x_\infty \in E$, the limit x_∞ necessarily belongs to F.

A subset $K \subset E$ is called compact if any sequence $(x_n)_{n\geq 0}$ of elements of K admits a subsequence converging in K, i.e., there exists a strictly increasing function $\varphi : \mathbb{N} \to \mathbb{N}$ and a point $y_\infty \in K$ such that the sequence $(y_n)_{n\geq 0}$ defined by $y_n := x_{\varphi(n)}$ converges to y_∞. In the context of a finite dimensional normed vector space E, a subset $K \subset E$ is compact if and only if it is bounded and closed.

Recall that a sequence $(x_n)_{n\geq 0}$ is called a Cauchy sequence if

$$\lim_{\min\{p,q\}\to+\infty} \|x_p - x_q\| = 0.$$

Recall finally that a finite dimensional normed vector space E is complete: a sequence converges in E if and only if it is a Cauchy sequence.

2 Differentiable maps

At the core of differential calculus lies a basic observation: the simplest maps between two vector spaces are the linear maps (i.e., matrices if one fixes reference bases). In particular, the fundamental idea of differential calculus is to:

- Identify the maps that are locally well approximated by a linear map,
- For those maps, quantify precisely this approximation.

Maps that can be locally well approximated by a linear map are called differentiable and the best local linear approximation of a differentiable map is its differential.

A way to quantify the approximation of a differentiable map by its differential is Taylor's formula.

In this paragraph we address quickly the first point. Consider two finite dimensional normed vector spaces $(E, \|.\|_E)$ and $(F, \|.\|_F)$. Let $U \subset E$ be an open set and $f : U \to F$ be a map.

Definition 2.1. *The map $f : U \to F$ is called differentiable at $x \in U$ if there exists a linear function, denoted $d_x f : E \to F$ such that, for all $h \in E$ such that $x + h \in U$,*

$$f(x + h) = f(x) + d_x f(h) + \|h\|_E \varepsilon(h),$$

where $\varepsilon(h) \in F$ satisfies

$$\lim_{h\to 0} \|\varepsilon(h)\|_F = 0.$$

If it exists, the linear function $d_x f : E \to F$ is unique and called the differential of f at x. The map f is called differentiable on U if is differentiable at every point $x \in U$.

Remark 2.2. The reason we ask U to be open is that it is the only condition under which the differential $d_x f : E \to F$ can be well defined on the whole vector space E. To see this, fix $x \in U$ where f is differentiable and note that, since U is open, then for all $h \in E$ there exists $t > 0$ small enough such that $x + th \in U$. Then, we have by definition of $d_x f$ that

$$\begin{aligned}f(x + th) &= f(x) + d_x f(th) + t\|h\|_E \varepsilon(th) \\ &= f(x) + t d_x f(h) + t\|h\|_E \varepsilon(th),\end{aligned}$$

where the last line follows from linearity of $d_x f$. In particular, we deduce that for all $h \in E$, and small enough $t > 0$,

$$d_x f(h) = \frac{f(x + th) - f(x)}{t} - \|h\|_E \varepsilon(th),$$

from which it follows that, for all $x \in U$ and all $h \in E$,

$$d_x f(h) = \lim_{t\to 0, t\neq 0} \frac{f(x + th) - f(x)}{t}. \tag{2.1}$$

Whenever we consider a function $f : U \to F$ defined on a non open set $U \subset E$, it isn't clear a priori what it means for f to be differentiable. The convention in this case is very simple: we say that $f : U \to F$ is differentiable iff there exists an open set $\bar{U} \supset U$ and a differentiable map $\bar{f} : \bar{U} \to F$ such that, for all $x \in U$, we have $f(x) = \bar{f}(x)$ and we define in this case $d_x f(h) := d_x \bar{f}(h)$ for any $x \in U$ and any $h \in E$.

Remark 2.3. Formula (2.1) allows to interpret $d_x f(h)$ as the directional derivative of f at x in the direction of vector h. Indeed, introducing the curve $\gamma(t) := x + th$, defined for t small enough, formula (2.1) reads precisely

$$d_x f(h) = (f \circ \gamma)'(0).$$

Classical computations:

1. Whenever $E = \mathbb{R}$, then for all $x \in U$ and any $h \in \mathbb{R}$, we recover the more familiar formula

   $$d_x f(h) = hf'(x),$$

   where

   $$f'(x) = \lim_{t \to 0, t \neq 0} \frac{f(x+t) - f(x)}{t} \in F.$$

2. Suppose that, there exists a linear map $L : E \to F$ and some $b \in F$ such that $f(x) = L(x) + b$. Then f is differentiable on E and, for all $x, h \in E$,

   $$d_x f(h) = L(h).$$

3. Suppose that E is a Euclidean space, with scalar product $\langle ., . \rangle$. Let $L : E \to E$ be a symmetric linear map, i.e., satisfying for all $x, y \in E$.

   $$\langle L(x), y \rangle = \langle x, L(y) \rangle.$$

   Let $v \in E$ and $c \in \mathbb{R}$ be fixed and consider the map $f : E \to \mathbb{R}$ defined by $f(x) = \langle x, L(x) \rangle + \langle x, v \rangle + c$. Then f is differentiable on E and, for all $x, h \in E$,

   $$d_x f(h) = 2\langle h, L(x) + v \rangle.$$

The chain rule is often used in computations.

Theorem 2.4 (Chain rule)**.** *Let E, F, G be finite dimensional normed vector spaces. Let $U \subset E$ and $V \subset F$ be open sets, let $f : U \to F$ and $g : V \to G$ be differentiable maps such that $f(U) \subset V$. Then $g \circ f : U \to G$ is differentiable and, for any $x \in U$ and any $h \in E$,*

$$d_x(g \circ f)(h) = d_{f(x)} g(d_x f(h)).$$

3 Gradients

From now on, we fix a Euclidean space E with scalar product $\langle ., . \rangle$ and associated norm $\|.\|$.

Theorem 3.1 (Riesz-Fréchet representation Theorem[4])**.** *Let $\ell : E \to \mathbb{R}$ be linear. Then, there exists a unique vector $v \in E$ such that, for all $x \in E$,*

$$\ell(x) = \langle x, v \rangle.$$

Definition 3.2 (Gradients)**.** *Let $f : U \subset E \to \mathbb{R}$ be differentiable and fix $x \in U$. Then, the gradient of f at x is the unique vector $\nabla f(x) \in E$ such that, for all $h \in E$,*

$$d_x f(h) = \langle h, \nabla f(x) \rangle,$$

whose existence is guaranteed by the linearity of $d_x f$ and Theorem 3.1.

[4]This is actually a simple version of a more general result stating that a linear map from a Hilbert space H to $\mathbb{R}$ is continuous iff it is of the form $x \mapsto \langle x, v \rangle$ for some vector $v \in H$.

The gradient of f at x has a very clear physical interpretation:

- If $\nabla f(x) \neq 0$, then denoting $\gamma_v(t) := x + tv$, we have

  $$\frac{\nabla f(x)}{\|\nabla f(x)\|} \in \operatorname*{arg\,max}_{v \in E : \|v\| = 1} (f \circ \gamma_v)'(0).$$

- Furthermore

  $$\|\nabla f(x)\| = \max_{v \in E : \|v\| = 1} (f \circ \gamma_v)'(0).$$

In other words, the gradient of f at x points in the direction of the largest increase of f at x and the norm of the gradient is equal to the value of this maximal increase.

We end by connecting this maybe abstract definition of gradients with the classical interpretation of the gradient as the vector whose coordinates are partial derivatives. Beware, however, that this interpretation is valid only if $(E, \|.\|) = (\mathbb{R}^d, \|.\|_2)$.

So suppose that $(E, \|.\|) = (\mathbb{R}^d, \|.\|_2)$ and denote

$$e_1 = (1, 0, \ldots, 0)^\top, \ldots, e_d = (0, \ldots, 0, 1)^\top,$$

the standard basis of $\mathbb{R}^d$. Then, given a differentiable map $f : U \subset E \to \mathbb{R}$, we may consider its partial derivatives

$$\partial_i f(x) = \lim_{t \to 0, t \neq 0} \frac{f(x + te_i) - f(x)}{t}.$$

Note that these partial derivative may be represented as

$$\partial_i f(x) = d_x f(e_i).$$

Now, note that for any $h = (h_1, \ldots, h_d)^\top \in \mathbb{R}^d$, we have of the one hand that, by linearity of $d_x f$,

$$\begin{aligned} d_x f(h) &= d_x f(\sum_{i=1}^d h_i e_i) \\ &= \sum_{i=1}^d h_i d_x f(e_i) \\ &= \sum_{i=1}^d h_i \partial_i f(x). \end{aligned}$$

On the other hand, by definition of the gradient $\nabla f(x) = ((\nabla f(x))_1, \ldots, (\nabla f(x))_d)$, we have that

$$d_x f(h) = h^\top \nabla f(x) = \sum_{i=1}^d h_i (\nabla f(x))_i.$$

Hence, by uniqueness of the gradient, we indeed recover the classical fact that, in $\mathbb{R}^d$ equipped with the classical scalar product $\langle x, y \rangle = x^\top y$, we have

$$\nabla f(x) = (\partial_1 f(x), \ldots, \partial_d f(x))^\top.$$

4 Higher order differentiability

Let E, F be two finite dimensional normed vector spaces and denote $\mathcal{L}(E, F)$ the vector space of all linear maps $L : E \to F$.

Note that $\mathcal{L}(E,F)$ is finite dimensional and that it can be naturally equipped with the so called operator norm $\|.\|_{\text{op}} : \mathcal{L}(E,F) \to \mathbb{R}_+$ defined by

$$\|L\|_{\text{op}} := \sup_{x\in E, x\neq 0} \frac{\|L(x)\|_F}{\|x\|_E}.$$

Let $U \subset E$ be an open set and $f : U \to F$. If f is differentiable on U, then we can construct a new map $df : U \to \mathcal{L}(E,F)$ defined, for all $x \in U$, by

$$df(x) := d_x f.$$

Definition 4.1. *Let $f : U \to F$ be differentiable on U.*

- *We say that f is continuously differentiable on U, and we denote $f \in C^1(U,F)$, if the map $df : U \to \mathcal{L}(E,F)$ is continuous.*
- *We say that f is twice differentiable on U if the map $df : U \to \mathcal{L}(E,F)$ is differentiable on U.*
- *We say that f is twice continuously differentiable on U, and we denote $f \in C^2(U,F)$, if $df \in C^1(U, \mathcal{L}(E,F))$.*

We can define higher orders of differentiability by iterating the above process but this will be enough for our purposes.

Now suppose $f : U \to F$ is twice differentiable on U. Then, for $x \in U$ we denote

$$d_x^2 f := d_x(df).$$

By construction, note that $d_x^2 f : E \to \mathcal{L}(E,F)$ is a linear map, i.e., $d_x^2 f \in \mathcal{L}(E, \mathcal{L}(E,F))$. This implies that, for any $u \in E$, $d_x^2 f(u)$ is a linear map from E to F, mapping $v \in E$ to

$$d_x^2 f(u)(v) \in F.$$

It is often much more convenient (and equivalent) to interpret $d_x^2 f$ as a bilinear map $E \times E \to F$ by setting

$$d_x^2 f[u,v] := d_x^2 f(u)(v).$$

Note however that, if f is only twice differentiable, this bilinear map isn't necessarily symmetric. However, if $f \in C^2(U,F)$, this nice property is guaranteed.

Theorem 4.2 (Schwarz)**.** *If $f \in C^2(U,F)$ then for all $x \in U$ the second order differential $d_x^2 f$ of f at x is a symmetric bilinear map. In other words, for all $x \in U$ and all $u, v \in E$,*

$$d_x^2 f[u,v] = d_x^2 f[v,u].$$

We end by discussing the special case where $(E, \|.\|)$ is the standard Euclidean space $(\mathbb{R}^d, \|.\|_2)$ and $F = \mathbb{R}$.

Remark 4.3. Let $(E, \|.\|) = (\mathbb{R}^d, \|.\|_2)$, let $U \subset \mathbb{R}^d$ be open and $f : U \to \mathbb{R}$. Suppose $f \in C^2(U, \mathbb{R})$. Then, introducing

$$e_1 = (1, 0, \dots, 0)^\top, \dots, e_d = (0, \dots, 0, 1)^\top,$$

the standard basis of $\mathbb{R}^d$, we define the second order partial derivatives of f by

$$\partial_{i,j}^2 f(x) := d_x^2 f[e_i, e_j].$$

One may show that, for all $x \in U$,

$$\partial_{i,j}^2 f(x) = \partial_i(\partial_j f)(x) = \lim_{t\to 0, t\neq 0} \frac{\partial_j f(x + te_i) - \partial_j f(x)}{t}.$$

By Theorem 4.2, we have that, for all $1 \le i, j \le d$ and all $x \in U$,

$$\partial_{i,j}^2 f(x) = \partial_{j,i}^2 f(x).$$

Now, for all $u = \sum_{i=1}^d u_i e_i$ and $v = \sum_{i=1}^d v_i e_i$, we get by bilinearity of $d_x^2 f$ that

$$\begin{aligned} d_x^2 f[u,v] &= \sum_{i,j=1}^n u_i v_j d_x^2 f[e_i, e_j] \\ &= \sum_{i,j=1}^n u_i v_j \partial_{i,j}^2 f(x) \\ &= u^\top \nabla^2 f(x) v, \end{aligned}$$

where $\nabla^2 f(x)$ denotes the Hessian matrix of f at x defined by

$$\nabla^2 f(x) := (\partial_{i,j}^2 f(x))_{1\le i,j\le d}.$$

It follows from Theorem 4.2 that $\nabla^2 f(x)$ is a symmetric matrix for all $x \in U$.

5 Taylor's identity

Let $(E, \|.\|)$ be a finite dimensional normed vector space, let $U \subset E$ be open and let $f : U \to \mathbb{R}$ be differentiable. We have established that f can be expanded around any $x \in U$ as

$$f(x+h) = f(x) + d_x f(h) + \text{peanuts}(h),$$

where

$$\lim_{h\to 0} \frac{|\text{peanuts}(h)|}{\|h\|_E} = 0.$$

The goal of Taylor's identity is to provide a more explicit expression for peanuts(h) which is very useful as we'll see in the next lectures. Before we embark on the proof, let us remind the reader about the fundamental theorem of calculus which states that, for a differentiable map $g : I \to \mathbb{R}$, defined on an open interval $I \subset \mathbb{R}$, we have, for all $s < t \in I$,

$$g(t) = g(s) + \int_s^t g'(u)\,\mathrm{d}u. \tag{5.1}$$

This basic result is in fact all we need to estalish Taylor's identity. We also need the property that U is a convex subset of E, i.e., that for all $x, y \in U$ and all $t \in [0,1]$,

$$(1-t)x + ty \in U.$$

We start with Taylor's identity at order 1.

Theorem 5.1 (Taylor's identity at order 1)**.** *Suppose that $f : U \to \mathbb{R}$ is differentiable. Suppose U is convex. Let $x \in U$ and $h \in E$ be such that $x + h \in U$. Then,*

$$f(x+h) = f(x) + \int_0^1 d_{x+th} f(h)\,\mathrm{d}t.$$

Proof. Let $g : [0,1] \to \mathbb{R}$ be defined by

$$g(t) = f(x + th).$$

Then,

$$g'(t) = d_{x+th} f(h).$$

The result then follows immediately from formula (5.1). □

Remark 5.2. *In the context where E is a Euclidean space with scalar product $\langle ., . \rangle$, it follows immediately from the definition of gradients that the above formula can be written as,*

$$f(x + h) = f(x) + \int_0^1 \langle h, \nabla f(x + th) \rangle \, \mathrm{d}t,$$

under the same assumptions.

We now arrive to Taylor's formula at order 2.

Theorem 5.3 (Taylor's identity at order 2)**.** *Suppose that $f : U \to \mathbb{R}$ is twice differentiable. Suppose U is convex. Let $x \in U$ and $h \in E$ be such that $x + h \in U$. Then,*

$$f(x + h) = f(x) + d_x f(h) + \int_0^1 (1 - t) d^2_{x+th} f[h, h] \mathrm{d}t.$$

Proof. Let $g : [0,1] \to \mathbb{R}$ be defined by

$$g(t) = f \circ \gamma(t) + (1 - t)(f \circ \gamma)'(t),$$

where $\gamma(t) = x + th$. Then,

$$g'(t) = (1 - t)(f \circ \gamma)''(t).$$

Hence, applying (5.1), we obtain

$$\begin{aligned} f(y) &= g(1) \\ &= g(0) + \int_0^1 g'(t) \mathrm{d}t \\ &= f(x) + (f \circ \gamma)'(0) + \int_0^1 (1 - t)(f \circ \gamma)''(t) \mathrm{d}t. \end{aligned}$$

It remains to observe (prove it), that

$$(f \circ \gamma)'(0) = d_x f(h)$$

and that

$$(f \circ \gamma)''(t) = d^2_{\gamma(t)} f[h, h].$$

□

A classical application of the above result is for instance when U is a convex subset of $\mathbb{R}^d$, equipped with its standard Euclidean structure. In this case, by definition of the gradient and the Hessian matrix introduced above, the previous formula reads

$$f(x + h) = f(x) + \langle h, \nabla f(x) \rangle + \int_0^1 (1 - t) h^\top \nabla^2 f(x + th) h \, \mathrm{d}t.$$