Lecture 1

3rd of October 2018

A. Frangioni

Example 1.0.1 (On derivatives)**.** *Let $f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(x, y) = x^2e^y$. Compute the partial derivatives and the directional derivatives in the two directions $\mathbf{d_1} = (0, 1)^T$ and $\mathbf{d_2} = (1, 0)^T$.*

- $\frac{\partial f}{\partial x} = 2xe^y$
- $\frac{\partial f}{\partial y} = x^2e^y$
- $\frac{\partial f}{\partial \mathbf{d_1}} = \lim\limits_{t\to 0} \frac{f(x+t\cdot 0, y+t\cdot 1)}{t} = \lim\limits_{t\to 0} \frac{f(x,y+t)}{t}$*. An attentive reader may notice that the directional derivative in direction $\mathbf{d_1}$ is equivalent to the derivative on the second component.*
- $\frac{\partial f}{\partial \mathbf{d_2}} = \lim\limits_{t\to 0} \frac{f(x+t\cdot 1, y+t\cdot 0)}{t} = \lim\limits_{t\to 0} \frac{f(x+t,y)}{t}$*. Conversely, with respect to what stated before, the directional derivative of f along the direction $\mathbf{d_2}$ is equivalent to the partial derivative w.r.t the first component.*

Let us compute the scalar product between the gradient and the direction $\mathbf{d_1}$:

$$\frac{\partial f}{\partial \mathbf{d_1}} =< \begin{pmatrix} 2xe^y \\ x^2e^y \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} >= \begin{pmatrix} 2xe^y \cdot 0 \\ x^2e^y \cdot 1 \end{pmatrix} = \begin{pmatrix} 0 \\ x^2e^y \end{pmatrix}$$

The intuition of this example is formalized in the following

Fact 1.0.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$. The directional derivative along a certain direction $\mathbf{d} \in \mathbb{R}^n$ can be computed as the scalar product between the gradient of the function and the direction, formally*

$$\frac{\partial f}{\partial \mathbf{d}} =< \nabla f, \mathbf{d} >$$

Definition 1.0.1 (Vector-valued function)**.** *A function which codomain is multi-dimensional is called **vector-valued function**. Formally, $f : \mathbb{R}^n \to \mathbb{R}^m$, where* $f(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix} \in \mathbb{R}^m$.

For such functions the computation of the derivative requires to specify not only the component with respect to the one the derivation should be performed, but also the index of the function.

Definition 1.0.2 (Partial derivative for vector-valued functions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$, the partial derivative of the j-th function with respect to the i-th component is*

$$\frac{\partial f_j}{\partial x_i}(\mathbf{x}) = \lim_{t \to 0} \frac{f_j(x_1, x_2, \dots, x_{i-1}, \dots, x_i + t, x_{i+1}, \dots, x_n) - f_j(\mathbf{x})}{t}$$

where $t \in \mathbb{R}$.

Definition 1.0.3 (Jacobian)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a vector-valued function. We call **Jacobian** the matrix of all its first-order partial derivatives.*

$$Jf(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla f_1(\mathbf{x}) \\ \nabla f_2(\mathbf{x}) \\ \vdots \\ \nabla f_m(\mathbf{x}) \end{pmatrix}$$

Example 1.0.2. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(x, y) = x^2 e^y$ as before, where the gradient was*

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} 2xe^y \\ x^2 e^y \end{pmatrix}$$

Let us compute the second derivative of this function:

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x \partial x} & \frac{\partial^2 f}{\partial y \partial x} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y \partial y} \end{pmatrix} = \begin{pmatrix} 2e^y & 2xe^y \\ 2xe^y & x^2 e^y \end{pmatrix}$$

The second order derivative of a vector function is a matrix, defined as

Definition 1.0.4 (Hessian)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$, we call **Hessian** of f*

$$\nabla^2 f(\mathbf{x}) := J \nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}) \end{pmatrix}$$

The size of the matrix grows very rapidly with the number of derivatives that we make.

In optimization, handling Hessians is a crucial task, because such matrices are very large hence unfeasible most of the times.

In most cases, the Hessian is symmetric, meaning that the order in which we derive is not relevant and this happens when $\exists \delta > 0$ s.t. $\forall \mathbf{x}' \in \mathcal{B}(x, \delta)$ $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}')$ and $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}')$ exist and $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}')$ and $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}')$ are continuous at $\mathbf{x}$.

Definition 1.0.5 ($\mathcal{C}^2$ functions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$. We say that f* ***belongs to $\mathcal{C}^2$ class*** *iff $\nabla^2 f(x)$ is continuous.*

Property 1.0.2. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ such that $f \in \mathcal{C}^2$, then the Hessian is* symmetric *and the gradient is* continuous*.*

We are now interested in providing some approximations to a generic function f. Such approximations are useful tools to have ah hint on where the function moves along a decreasing direction, but the information is accurate only locally.

Definition 1.0.6 (First order Taylor model)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \in C^1$, we define the* ***first-order Taylor's formula*** *as a linear approximation of the function f in a neighbourhood of $\mathbf{x}$.*

Formally, for a given ball $\mathcal{B}(\mathbf{x}, \delta)$, $\forall \mathbf{x}' \in \mathcal{B}(\mathbf{x}, \delta)$ $\exists$ $\alpha \in (0, 1)$ s.t.

$$f(\mathbf{x}') =< \nabla f(\alpha \mathbf{x} + (1 - \alpha)\mathbf{x}'), \mathbf{x}' - \mathbf{x} > + f(\mathbf{x})$$

Equivalently, we can write the so-called ***remainder version of first-order Taylor formula*** *as*

$$f(\mathbf{x}') =< \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} > + f(\mathbf{x}) + R(\mathbf{x}' - \mathbf{x})$$

where $\lim\limits_{h \to 0} \frac{R(h)}{\|h\|} = 0$, in other words the error that we make is at most quadratic.

Notice that the Taylor's approximation works *only locally*: the furthest we get from $\mathbf{x}$ the more distant the function and the model are.

Definition 1.0.7 (Second order Taylor model)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \in C^1$, we define the* ***second-order Taylor's formula*** *as a quadratic approximation of the function f in a neighbourhood of $\mathbf{x}$. Formally, given a ball $\mathcal{B}(\mathbf{x}, \delta)$, $\forall \mathbf{x}' \in \mathcal{B}(\mathbf{x}, \delta)$*

$$f(\mathbf{x}') = L_x(\mathbf{x}') + \frac{1}{2}(\mathbf{x}' - \mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{x}' - \mathbf{x}) + R(\mathbf{x}' - \mathbf{x})$$

with $\lim\limits_{h \to 0} R(h) \|h\|^2 = 0$ or, equivalently, the remainder vanishes at least cubically, or the error is $O(\|\mathbf{x}' - \mathbf{x}\|^3)$.

Example 1.0.3. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(x, y) = x^2 e^y$, as above and $\nabla f(\mathbf{x}) = \begin{pmatrix} 2xe^y \\ x^2 e^y \end{pmatrix}$ and $\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 2e^y & 2xe^y \\ 2xe^y & x^2 e^y \end{pmatrix}$. In $\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in \mathbb{R}^2$ we have $f(\mathbf{x}) = 1$ and $\nabla f(\mathbf{x}) = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$.*

The linear model has the following shape

$$L_{(1,0)}(x,y) = 1 < \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} x-1 \\ y-0 \end{pmatrix} > +1 = 2x - 2 + y + 1 = 2x + y - 1$$

And the quadratic model is

$$\begin{aligned} Q_{[1,0]}(x,y) &= 2x + y - 1 + \frac{1}{2} \cdot \begin{pmatrix} x-1, & y-0 \end{pmatrix} \cdot \begin{pmatrix} 2 & 2 \\ 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} x-1 \\ y-0 \end{pmatrix} \\ &= 2x + y - 1 + \frac{1}{2} \cdot \begin{pmatrix} 2x-2+2y, & 2x-2+y \end{pmatrix} \cdot \begin{pmatrix} x-1 \\ y-0 \end{pmatrix} \\ &= 2x + y - 1 + \frac{1}{2} \cdot \Big((2x-2+2y) \cdot (x-1) + (2x-2+y) \cdot y \Big) \\ &= 2x + y - 1 + \frac{1}{2} \cdot (2x^2 - 2x + 2xy - 2x + 2 - 2y + 2xy - 2y + y^2) \\ &= 2x + y - 1 + \frac{1}{2} \cdot (2x^2 - 4x + 4xy + 2 - 4y + y^2) \\ &= 2x + y - 1 + x^2 - 2x + 2xy + 1 - 2y + \frac{1}{2}y^2) \\ &= x^2 + 2xy - y + \frac{1}{2}y^2 \end{aligned} \tag{1.0.1}$$

The Taylor model can be extended to the more general k-th order. Such an expansion requires the computation of the k-th order derivatives, but $\nabla^k f(\mathbf{x})$ is a tensor of order $k \equiv n^k$ numbers. For $k > 2$ this approach is practically unfeasible.

As it happens often in computer science, it is important to find a good trade-off between the complexity of the model and the accuracy of the approximation: if the model is too simple (but efficient) the approximation is not very good; conversely, complex and accurate approximations are computationally heavy.

Fact 1.0.3. *Let $f : \mathbb{R}^n \to \mathbb{R} \in \mathcal{C}^1$. If f is Lipschitz continuous on $S \in \mathbb{R}^n$, then in such set the norm of the gradient is bounded by the Lipshitz constant.*

Formally, $\sup\{\|\nabla f(\mathbf{x})\| \ : \ \mathbf{x} \in S\} \leq L$*. If S convex* $\sup\{\|\nabla f(\mathbf{x})\| \ : \ \mathbf{x} \in S\} = L$*.*

Moreover, we can prove

Fact 1.0.4. *Let $f : \mathbb{R}^n \to \mathbb{R} \in \mathcal{C}^1$. f is Lipschitz continuous on $S \in \mathbb{R}^n$ iff its Hessian in bounded on S. Formally,* $\sup\{\left\|\nabla^2 f(\mathbf{x})\right\| \ : \ \mathbf{x} \in S\} \leq L$*.*

Fact 1.0.5. *Let $f : \mathbb{R}^n \to \mathbb{R} \in \mathcal{C}^1$. If the gradient of f is Lipschitz continuous in all points $\mathbf{x}' \in \mathbb{R}^n$ that are close to $\mathbf{x} \in \mathbb{R}^n$, then $f(\mathbf{x}') \leq L_x(\mathbf{x}') + \frac{L}{2}\left\|\mathbf{x}' - \mathbf{x}\right\|^2$.*

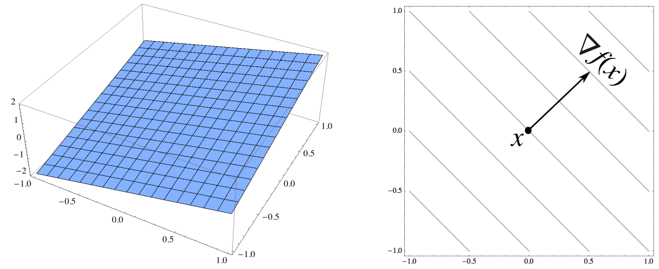
1.1 Simple functions

In the rest of the course we will deal mostly with some linear and quadratic approximations of the objective function. In this section, we introduce a couple of examples and considerations on such functions.

1.1.1 Linear functions

In this scenario, $f : \mathbb{R}^n \to \mathbb{R}$ has the following shape: $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$, for a fixed $\mathbf{c} \in \mathbb{R}^n$.

It holds that $\nabla f(\mathbf{x}) = \mathbf{c}$, $\nabla^2 f(\mathbf{x}) = 0$ and that level sets are parallel hyperplanes orthogonal to $\mathbf{c}$.



(a) Linear function (b) Level sets

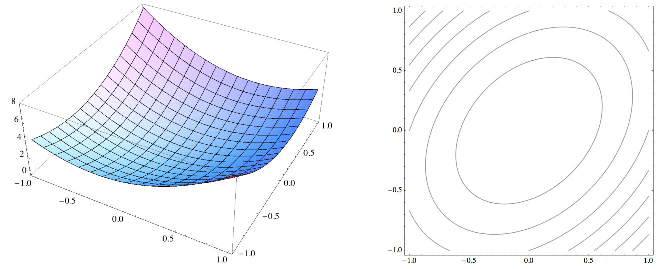
FIGURE 1.1: Graphical example of linear function.

1.1.2 Quadratic functions

A quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ is formalized as $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{q}\mathbf{x}$, where $Q \in \mathcal{M}(n, \mathbb{R})$, $\mathbf{q}^T \in \mathbb{R}^n$.

In Figure 1.2 we can see the plot of the quadratic function where

$$Q = \begin{bmatrix} 6 & -2 \\ -2 & 6 \end{bmatrix} q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$



(a) Quadratic function (b) Level sets

FIGURE 1.2: Graphical example of a quadratic function.

In quadratic functions the gradient is a linear function on $\mathbf{x}$ $\nabla f(\mathbf{x}) = Q\mathbf{x} + \mathbf{q}$, while the Hessian is just Q and the level sets are ellipsoids. Sometimes such ellipses can degenerate to lines, in the case that one of the axis has become $+\infty$.

Fact 1.1.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function with parameters $Q \in \mathcal{M}(n, \mathbb{R})$ and $\mathbf{q}^T \in \mathbb{R}^n$. If Q is symmetric, then it has spectral decomposition.*

Proof. $\mathbf{x}^T Q\mathbf{x} = \frac{[(\mathbf{x}^T Q\mathbf{x}) + (\mathbf{x}^T Q\mathbf{x})^T]}{2} = \mathbf{x}^T [\frac{(Q+Q^T)}{2}]\mathbf{x} = H\Lambda H^T$ □

In the rest of the course we will assume for simplicity that Q is symmetric. Another "lucky case" is when Q is non singular (i.e. all its eigenvalues are not 0): in this case, $\bar{\mathbf{x}} = -Q^{-1}\mathbf{q}$, where $\bar{\mathbf{x}}$ is called the center of the ellipsoid. Let us assume that we moved the origin in such $\bar{\mathbf{x}}$, so $\mathbf{x}' = \mathbf{x} - \bar{\mathbf{x}}$ and $f_{\bar{\mathbf{x}}}(\mathbf{x}') = \frac{1}{2}\mathbf{x}'^T Q\mathbf{x}'$.

Fact 1.1.2. *Along H_i: $f(\alpha) = f_{\bar{\mathbf{x}}}(\alpha H_i) = \alpha^2\lambda_i$*

Moreover, the size of the axes of the level curves is proportional to $\sqrt{1/\lambda_i}$, where λ_i are the eigenvalues. If the eigenvalues are close to 0, then the axes get very long, while when they are negative, we no longer have axes.

This is formalized in

Fact 1.1.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{q}\mathbf{x}$. If Q is positive definite $\bar{\mathbf{x}} = -Q^{-1}\mathbf{q}$ is the minimum of f, while if Q is* indefinite *(i.e. $\exists\lambda_i < 0$) f is unbounded below.*