Optimization

Màster de Fonaments de Ciència de Dades

PART 1. Preliminars

Chapter 1. Introduction to Optimization.

Chapter 2. Mathematical notation and background.

Chapter 1

Introduction to Optimization

The goal: Objective and constrains

Optimization is an important issue in decision theory and analysis of physical systems. In fact it has been also an important issue in evolution theory. How to find out an optimal proposal to whatever process?

The natural ingredients are the objective (function) and the (potential) constrains. We make use of variables.

The abstract process of identifying everything (objective, constrains, variables, etc) is known as modelling. That is, how do you go from real world to a problem one can solve. And how you conclude that this solution is real (solve the original problem) ...

The goal: Objective and constrains

So the process to solve an optimization problem is as follows

$$\text{Modelling} \Rightarrow \begin{matrix}\text{Objective}\\ \text{Constrains}\end{matrix} \Rightarrow \text{Algorithm(s)} \Rightarrow \begin{matrix}\text{Optimality}\\ \text{Conditions}\end{matrix}$$

Final considerations might include a Sensibility analysis or discussion of the solution, role of parameters..

Objective and constrains: Mathematical formulation

1. The variable(s) or unkown(s), usually denoted by $x \in \mathbb{R}^n$.
2. The objective (function), usually denoted by f. The scalar function f of the variable x is the one we want to maximize/minimize (optimize).
3. The constrains (functions) are also scalar functions of the variable x defined in terms of equalities and/or inequalities:

$$\begin{aligned} g_j(x) &= 0, \ j = 1, \ldots, n \\ g_j(x) &\leq 0, \ j = n+1, \ldots, m. \end{aligned}$$

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{subject to} \qquad \begin{aligned} c_j(x) &= 0 \quad j = 1, \ldots, n \\ c_j(x) &\leq 0 \quad j = n+1, \ldots, m \end{aligned}$$

Objective and constrains: The transport problem

One product. A pharmacy company has 2 factories, say $F_j,\ j = 1, 2$, and 10 retail outlets, say $R_j,\ j = 1, \ldots 10$.

- We denote by a_j the capacity of factory F_j.
- We denote by b_j the demand of retail R_j.
- We denote by c_{ij} the cost of shipping one ton from F_i to R_j.
- We denote by x_{ij} the tons (in $\mathbb{R}$) of the product shipped from F_i to R_j.

Objective and constrains: The transport problem

The problem we are facing is

$$\min_{x\in\mathbb{R}^{20}} \sum_{i,j} c_{ij}x_{ij} \qquad \text{subject to}$$

$$\sum_{j=1}^{10} x_{ij} \leq a_i, \quad i = 1, 2$$

$$\sum_{i=1}^{2} x_{ij} \geq b_j, \quad j = 1, \ldots, 10$$

$$x_{ij} \geq 0 \quad i = 1, 2, \; j = 1, \ldots 10$$

Remark. This problem is known as Linear Programming and the (universal) solution algorithm is known as Simplex Method.

Objective and constrains: Academic example I

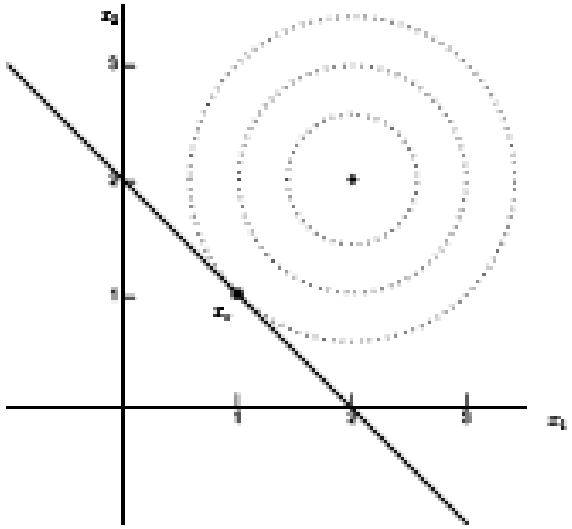
Problem. Find the solution of

$$\min_{(x,y)\in\mathbb{R}^2} \quad f(x,y) = (e^x - 1)^2 + (y-1)^2.$$

1. The set where we must look for the solution, the feasible set, is the entire two-dimensional space $\mathbb{R}^2$.
2. The solution is $(x,y)^T = (0,1)^T$, since the function value is zero only at this point and is positive elsewhere.
3. In this problem, the objective function is f, there are no constrains and the decision variables are x, y.

Objective and constrains: Academic example II

Problem. Find the point on the line $x + y = 2$ that is closest to the point $(2, 2)^T$.



$$\min_{(x,y)\in\mathbb{R}^2} \quad (x-2)^2 + (y-2)^2$$

$$\text{subject to} \quad x + y = 2.$$

Solution. The solution (optimal point) is $(x, y)^T = (1, 1)^T$

Objective and constrains: Current applications

1. Designing a car with minimal air resistance.
2. Designing a bridge of minimal weight that still meets essential specifications.
3. Defining a stock portfolio where the risk is minimal and the expected return is optimal.

Objective and constrains: Classical problems

1. Dido's (or isoperimetric) problem. Among all closed plain curves of a given length, find the one that encloses the largest area.
2. Find the maximum of the product of two numbers whose sum is given.
3. In a given circle find a rectangle of maximal area.
4. The Brachistochrone. Let two points A and B be given in a vertical plane. Find the curve that a point M, moving on a path AMB must follow such that, starting from A with zero velocity, it reaches B in the shortest time under its own gravity.
5. The traveling salesman problem Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city?

Continuous versus discrete optimization

Integer optimization. In some optimization problems the variables make sense only if they take on integer values or binary values. Think on the transport problem if instead of tons we have just units. This is what we called Integer Optimization problems. They require particular tools and techniques (extra difficulties).

Discrete optimization. The integer optimization is a particular case of discrete optimization where not only variables are force to be integer (or binary) but some restrictions are discrete, like permutation of an ordered set. So, discrete optimization is about finding optimal objects in finite (large) set of possible solutions.

Remark. We will not deal with this problem in this course.

Unconstrained versus constrained optimization

Unrestricted optimization. This is the classical optimization formulae

$$\max f(x) \quad \text{or} \quad \min f(x)$$

where f is a (smooth) function in $x \in \mathbb{R}^n$.

Restricted optimization. This case shows up when we force the potential solutions to satisfy particular characteristics (like being vectors with positive components). There are two types of restricted problems: the ones determined by equalility constrains and the ones determined by inequality constrains.

Linear programming. This is a particular case of restricted optimization where the objective function and the restrictions are all linear.

The demand function (aggregate individuals demands)

Consumer's problem. Assume a market with two goods A and B. Let q_A and q_B the quantities and p_A and p_B the prices, respectively. If $U(q_A, q_B)$ is the utility function of a individual consumer with total budget m then the problem she is facing is

$$\max_{(q_A,q_B)} U(q_A, q_B) \qquad \text{subject to} \qquad \begin{array}{l} p_A q_A + p_B q_B \leq m \\ \\ q_A \geq 0, \ q_B \geq 0. \end{array}$$

Local versus global optimization

The natural goal when we deal with an optimization problem is to find the solution. That is, which is the max / min of the objective function?

But most of our tools will give only partial solution(s). More precisely we will find local solution(s). That is local extrema of the objective function.

Only in very particular cases we will be able to solve the original problem of finding the global optimal vector for f.

For instance one case will be the convex programming problems.

Non-smooth optimization

Most of our arguments in order to find extrema of objective functions are based on certain regularity (differentiability, say) of the objects we work (objective function, constrains, etc). In certain cases this is not the case.

Example. Compute (complementary goods)

$$\max_{(q_A,q_B)} \min\{q_A, q_B\} \qquad \text{subject to} \qquad \begin{array}{l} p_A q_A + p_B q_B \leq m \\ \\ q_A \geq 0,\ q_B \geq 0. \end{array}$$

Stochastic versus deterministic optimization

In some optimization problems, the model cannot be fully specified because it depends on quantities that are unknown at the time of formulation. We need to deal with this uncertainty.

The way we might attach this problem is by considering the variables as random variables and the stochastic optimization produce solutions that optimize the expected performance of the model.

Example. A first price auction with two players: 1 and 2. We suppose private values v_j, $j = 1, 2$. We assume that $v_j = U[0, 1]$, that is (uniforme) random variables. A strategy for player $j = 1, 2$ is any function $\beta_j(v_j)$ assigning to each value the corresponding bid of player $j = 1, 2$. Find the optimal strategy (Nash equilibrium).

Example. Same with a second price auction.

Infinite-dimensional (versus finite) optimization problem

There is a more general nonlinear optimization problem where the dimension of the space of the solutions is infinite-dimensional.

Problem. Find the state function $x(t)$ and the control function $u(t)$ such that

$$\min \ \Phi(t_0, x_0, t_f, x_f) + \int_{t_0}^{t_f} F(t, x(t), u(t))dt,$$

$$\begin{array}{lll} & t \in [t_0, t_f], & \\ & \dot{x}(t) = f(t, x(t), u(t)), & \text{(dynamic constraints)}, \\ \text{subject to} & b_L \leq b(t_0, x_0, t_f, x_f) \leq b_U, & \text{(boundary conditions)}, \\ & g_L \leq g(t, x(t), u(t)) \leq g_U, & \text{(path constraints)}, \\ & u_L \leq u(t) \leq u_U, & \text{(control constraints)}. \end{array}$$

The infinite-dimensional optimization problem

There are multiple solution approaches for the infinite-dimensional optimization problem.

1. Indirect methods. The initial problem is transformed into a Hamiltonian boundary-value problem that must be solved. These methods equire the derivation of the necessary conditions of optimality using calculus of variations.
2. Direct methods. The original problem is first discretized and then re-written as a finite-dimensional nonlinear optimization problem.

Optimality conditions and optimization algorithms

Most (if not all) of the optimization algorithms are iterative. They begin with an initial seed or guess (initial condition) of the variable x and generate a sequence of improved estimates that eventually get arbitrarily close to **a** solution of the problem. Such solution satisfies the optimality condition.

Most strategies (algorithms) make use of the values of the objective function f, the constrain functions g_j (if any) and the smoothness of those objects (gradients, hessians, etc).

1. Robustness (dependence of the solution on initial conditions)
2. Efficiency (time and storage)
3. Accuracy (precision)

Optimality conditions and optimization algorithms

There are two basic strategies for optimization algorithms.

- Line search algorithms. The iterate strategy is based on assuming there is a best direction to jump from one point to the next of the sequence. Always the same direction. Of course the size of the step would be crucial.
- Trust region. At every step we choose the best direction. This is related with the quadratic expression of the objective function.

Chapter 2

Mathematical notation and background

Scalar product

Let $\boldsymbol{x} = (x_1, \ldots, x_n)^T, \boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$.

- Scalar euclidean (dot) product.

$$< \boldsymbol{x}, \boldsymbol{y} >:= \boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{y} = x_1 y_1 + \cdots + x_n y_n \in \mathbb{R}.$$

- Euclidean norm.

$$\|\boldsymbol{x}\| = \sqrt{< \boldsymbol{x}, \boldsymbol{y} >} = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}} = \sqrt{\boldsymbol{x}^T \boldsymbol{x}} = \sqrt{x_1^2 + \cdots + x_n^2}.$$

- Euclidean distance.
  $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{x}\| = \sqrt{(y_1 - x_1)^2 + \cdots + (y_n - x_n)^2}$.
- Cosinus of the angle.

$$\cos(\widehat{\boldsymbol{x}, \boldsymbol{y}}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}.$$

- Perpendicularity (orthogonality): $\boldsymbol{x} \perp \boldsymbol{y} \quad \Leftrightarrow \quad \boldsymbol{x} \cdot \boldsymbol{y} = 0.$

Cross product

Let $\boldsymbol{x} = (x_1, x_2, x_3), \boldsymbol{y} = (y_1, y_2, y_3) \in \mathbb{R}^3$, we define:

- Cross product.

$$\boldsymbol{x} \times \boldsymbol{y} = \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} = \begin{pmatrix} x_2 y_3 - x_3 y_2 \\ x_3 y_1 - x_1 y_3 \\ x_1 y_2 - x_2 y_1 \end{pmatrix}.$$

Lemma.

$$\boldsymbol{x} \times \boldsymbol{y} \perp \boldsymbol{x} \quad \text{and} \quad \boldsymbol{x} \times \boldsymbol{y} \perp \boldsymbol{y}.$$

Lines in $\mathbb{R}^2$

The line determined by the point $\boldsymbol{a} = (a_1, a_2)^T$ and the vector $\boldsymbol{v} = (v_1, v_2)^T$ writes as $(t \in \mathbb{R})$

$$\boldsymbol{x} = \boldsymbol{a} + t\boldsymbol{v} \quad \Leftrightarrow \quad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + t \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}.$$

Equivalently (euclidean coordinates)

$$\frac{x - a_1}{v_1} = \frac{y - a_2}{v_2} \quad \Leftrightarrow \quad Ax + By + C = 0,$$

with $A = v_2,\ B = -v_1,\ C = -a_1 v_2 + a_2 v_1$.

Lines in $\mathbb{R}^3$

- The line determined by the point $\boldsymbol{a} = (a_1, a_2, a_3)^T$ and the vector $\boldsymbol{v} = (v_1, v_2, v_3)^T$ writes as $(t \in \mathbb{R})$

$$\boldsymbol{x} = \boldsymbol{a} + t\boldsymbol{v} \quad \Leftrightarrow \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + t \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

Equivalently (euclidean coordinates)

$$\frac{x - a_1}{v_1} = \frac{y - a_2}{v_2} = \frac{z - a_3}{v_3}.$$

Lines in $\mathbb{R}^3$

The plane determined by the point $\boldsymbol{a} = (a_1, a_2, a_3)^T$ and the vectors $\boldsymbol{u} = (u_1, u_2, u_3)^T$ and $\boldsymbol{v} = (v_1, v_2, v_3)^T$ writes as $(t,\, s \in \mathbb{R})$

$$\boldsymbol{x} = \boldsymbol{a} + t\boldsymbol{u} + s\boldsymbol{v} \quad \Leftrightarrow \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + t \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} + s \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

The above equation of the plane can also be written as

$$\begin{vmatrix} x - a_1 & y - a_2 & z - a_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = 0$$

or as

$$Ax + By + Cz + D = 0,$$

with $(A, B, C)^T = \boldsymbol{u} \times \boldsymbol{v}$.

Real valued functions: Continuity

Consider the real valued function

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}.$$

- The domain $D := D(f)$ of f is the set of points $\boldsymbol{x} \in \mathbb{R}^n$ where f is defined.
- The graph of f, as the subset of $\mathbb{R}^{n+1}$, is defined as

$$\mathrm{graf}(f) := \{(\boldsymbol{x}, z) \in \mathbb{R}^{n+1} : \begin{array}{l} \boldsymbol{x} = (x_1, ..., x_n)^T \in D \subset \mathbb{R}^n, \\ z = f(\boldsymbol{x}) \in \mathbb{R} \end{array} \}.$$

- The level set of f (of level $c \in \mathbb{R}$) is given by

$$L_c = \{\boldsymbol{x} \in D \, : \, f(\boldsymbol{x}) =\} \subset \mathbb{R}^n.$$

Real valued functions: Continuity

We say that f is continuous at a point $\boldsymbol{a} \in \mathcal{C}$ if and only if

$$\lim_{\boldsymbol{x} \to \boldsymbol{a}} f(\boldsymbol{x}) = f(\boldsymbol{a}).$$

(whenever the limit has meaning)

- The elementary functions of one (or several) variable

  $$x^n, \quad e^x, \quad \log x, \quad \sin x, \quad \cos x$$

  are continuous in their domain.

- Addition, subtraction, product, division (except at the points where the denominator vanishes) and composition of continuous functions are also continuous functions.

Real valued functions: Continuity

Theorem (Bolzano). Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Assume that $f(a)f(b) < 0$. Then there exists $c \in (a, b)$ such that $f(c) = 0$.

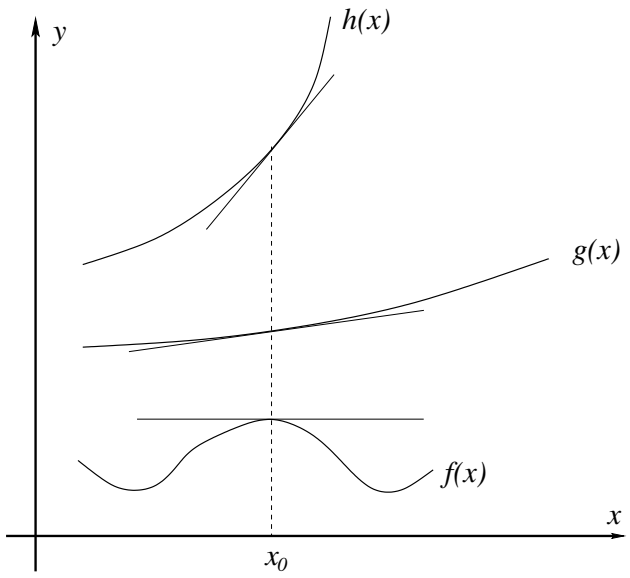
Theorem (Weierstrass). Let

$$f : K \subset \mathbb{R}^n \longrightarrow \mathbb{R},$$

be continuous function such that K is compact (closed and bounded), then f is bounded (there exist M such that $|f(x)| < M$ for all $x \in K$) and f attaints its maximum and minimum values on K.

The notion of derivative $n = 1$

The derivative of a function $y = f(x)$ is a measure of the (infinitesimal) rate at which the value y of the function changes with respect to the change of the variable x.



- Geometrically. The limit of the secant lines is the tangent line and the derivative is the slope of the tangent line.
- Analitically. The derivative of the function f at the point a is

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

The notion of derivative $n = 1$

Easily, one can check that the derivative satisfies the property that

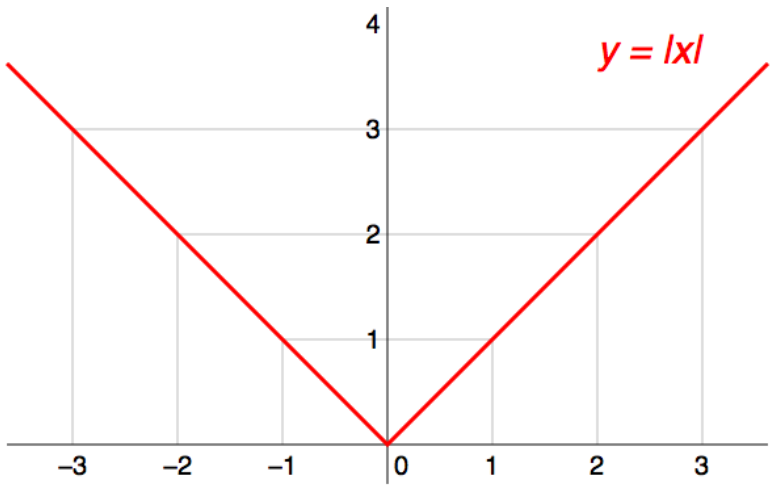
$$\lim_{h \to 0} \frac{f(a+h) - f(a) - f'(a) \cdot h}{h} = 0,$$

which has the intuitive interpretation that the tangent line to f at a gives the best linear approximation to f near a (h small).

$$f(a+h) \approx f(a) + f'(a)h,$$

Definition. We say that f is differentiable at the point $a \in D(f)$ if there exists the derivative of f at the point a (the limit exists).

Continuity and diferentiability $n = 1$



- Theorem. If f is differentiable at the point $a \in D(f)$ then f is continuous at the point a.
- The converse is not necessarily true.

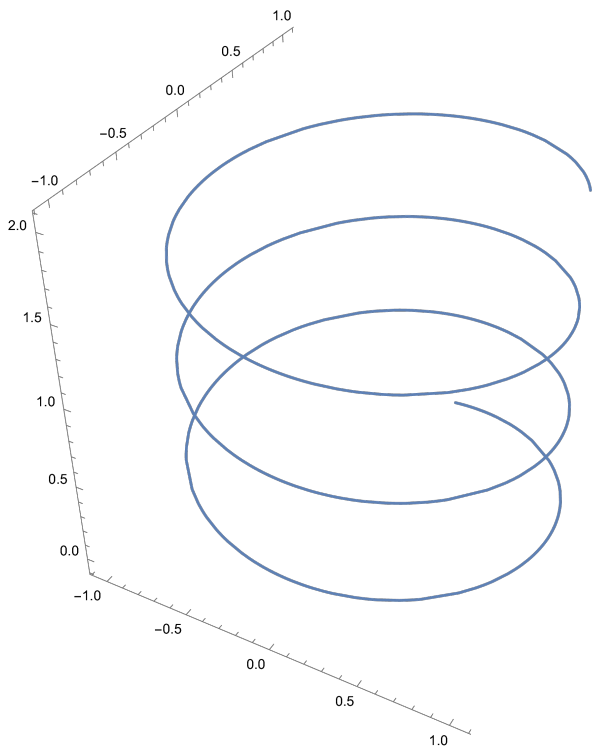
We say that $f : (a, b) \to \mathbb{R}$ is differentiable (in (a, b)) if it is differentiable at every point.

Diferentiability in $\mathbb{R}^n$

Let $\gamma : (a, b) \to \mathbb{R}^n$ be a curve in $\mathbb{R}^n$ with all components being differentiable. We write

$$\gamma(t) = (\gamma_1(t), \ldots, \gamma_n)^T \quad \text{and} \quad \gamma'(t) = (\gamma_1'(t), \ldots, \gamma_n')^T$$

and the vector $\gamma'(t)$ is tangent at every point of the curve $\gamma(t)$.



Differentiability in $\mathbb{R}^n$: Partial derivatives

Let f be a real valued function that depends on n variables

$$\begin{array}{rccl} f: & \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & \boldsymbol{x} & \longrightarrow & f(\boldsymbol{x}) = f(x_1, \ldots, x_n), \end{array}$$

Let $\boldsymbol{a} = (a_1, \ldots, a_n)^T$ be an interior point of $D(f)$. We want to extend the notion of differentiability introduced above for $n = 1$.

Partial derivative(s). The partial derivative of $f(\boldsymbol{x})$ in the direction x_j at the point $\boldsymbol{a}$ is defined to be:

$$f_{x_j}(\boldsymbol{a}) := \frac{\partial f}{\partial x_j}(\boldsymbol{a}) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_j + h, \ldots, a_n) - f(a_1, \ldots, a_j, \ldots, a_n)}{h}.$$

Differentiability in $\mathbb{R}^n$: The gradient vector

The gradient vector. Let f and $\boldsymbol{a}$ as before. The vector formed by the partial derivatives of f at the point a (assuming it exists) is known as the gradient vector

$$\nabla f(a) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{a}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{a}) \right)^T$$

Differentiability in $\mathbb{R}^n$. Let f and $\boldsymbol{a}$ as before. We say that f is differentiable at $\boldsymbol{a}$ if

$$\lim_{||\boldsymbol{h}||\to 0} \frac{\left| f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a}) - (\nabla f(a))^T \boldsymbol{h} \right|}{||h||} = 0, \quad h \in \mathbb{R}^n.$$

One can show that if partial derivatives of f are continuous functions at $\boldsymbol{a}$, then f is (continuous) differentiable.

Differentiability in $\mathbb{R}^n$: Directional derivatives

The partial derivatives of f measure the variation in f in the axis directions. But in many occasions we need to measure the variation of f in any direction, represented by a vector $v \in \mathbb{R}^n$. One can show that w.l.o.g. we may assume $||v|| = 1$ (unitary vector).

Choose a unitary vector $\boldsymbol{v} = (v_1, \ldots, v_n)^T$. The directional derivative of f in the direction of $\boldsymbol{v}$ at the point $\boldsymbol{a}$ is defined by

$$D_{\boldsymbol{v}} f(\boldsymbol{a}) = \lim_{h \to 0} \frac{f(\boldsymbol{a} + h\boldsymbol{v}) - f(\boldsymbol{a})}{h}.$$

Differentiability in $\mathbb{R}^n$: The chain rule

Theorem. Let $\alpha : (a, b) \longrightarrow \mathbb{R}^n$ be a differentiable curve, $\alpha(t) = (\alpha_1(t), \ldots, \alpha_n(t))^T$. Let f be a real valued differentiable function as above. Then

$$f(\alpha(t)) = f(\alpha_1(t), \ldots, \alpha_n(t)),$$

and

$$\frac{d}{dt} f(\alpha(t)) = \frac{\partial f}{\partial x_1}(\alpha(t))\alpha_1'(t) + \cdots + \frac{\partial f}{\partial x_n}(\alpha(t))\alpha_n'(t).$$

Corollary. In the above notation (assume f is differentiable) we have

$$D_{\boldsymbol{v}} f(\boldsymbol{x}) = \left.\frac{d}{dt}\right|_{t=0} f(\boldsymbol{x} + t\boldsymbol{v}) = \sum_{j=1}^{n} \frac{\partial f(\boldsymbol{x})}{\partial x_j} v_j = < (\nabla f(\boldsymbol{a}))^T, v > .$$

The gradient vector

Theorem. Let $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ be a differentiable function at $\boldsymbol{a} \in D$, and $\boldsymbol{u} \in \mathbb{R}^n$ is an unitary vector. The following statements hold.

(a)
$$D_{\boldsymbol{u}} f(\boldsymbol{a}) = (\nabla f(\boldsymbol{a})) \cdot \boldsymbol{u} = \|\nabla f(\boldsymbol{a})\| \cos\theta,$$
where θ is the angle between $\boldsymbol{u}$ and $\nabla f(\boldsymbol{a})$.

(b) The gradient vector $\nabla f(\boldsymbol{a})$ gives the maximum direction variation of f at the point $\boldsymbol{a}$.

(c) The gradient vector at the point $\boldsymbol{a} \in D$ is orthogonal to the level curve passing through $\boldsymbol{a}$.

Proof. Statements (a) and (b) are direct. Let $\boldsymbol{r}(t) = (x_1(t), x_2(t), \ldots, x_n(t))^T$ be the level curve passing through $\boldsymbol{a} \in D$. Then
$$0 = \frac{d}{dt} f(\boldsymbol{r}(t)) = \cdots = < \nabla f(\boldsymbol{r}(t))^T, \boldsymbol{r}'(t) > .$$

The tangent plane to the $\mathrm{graf}(f)$

Lemma. Let f be a differentiable function and let $\boldsymbol{a} \in D \subset R^n$. Denote $x_{n+1} = f(\boldsymbol{x}) = f(x_1, \ldots, x_n)$. The equation of the tangent plane of $\mathrm{graf}(f)$ at the point $a \in D$ is given by

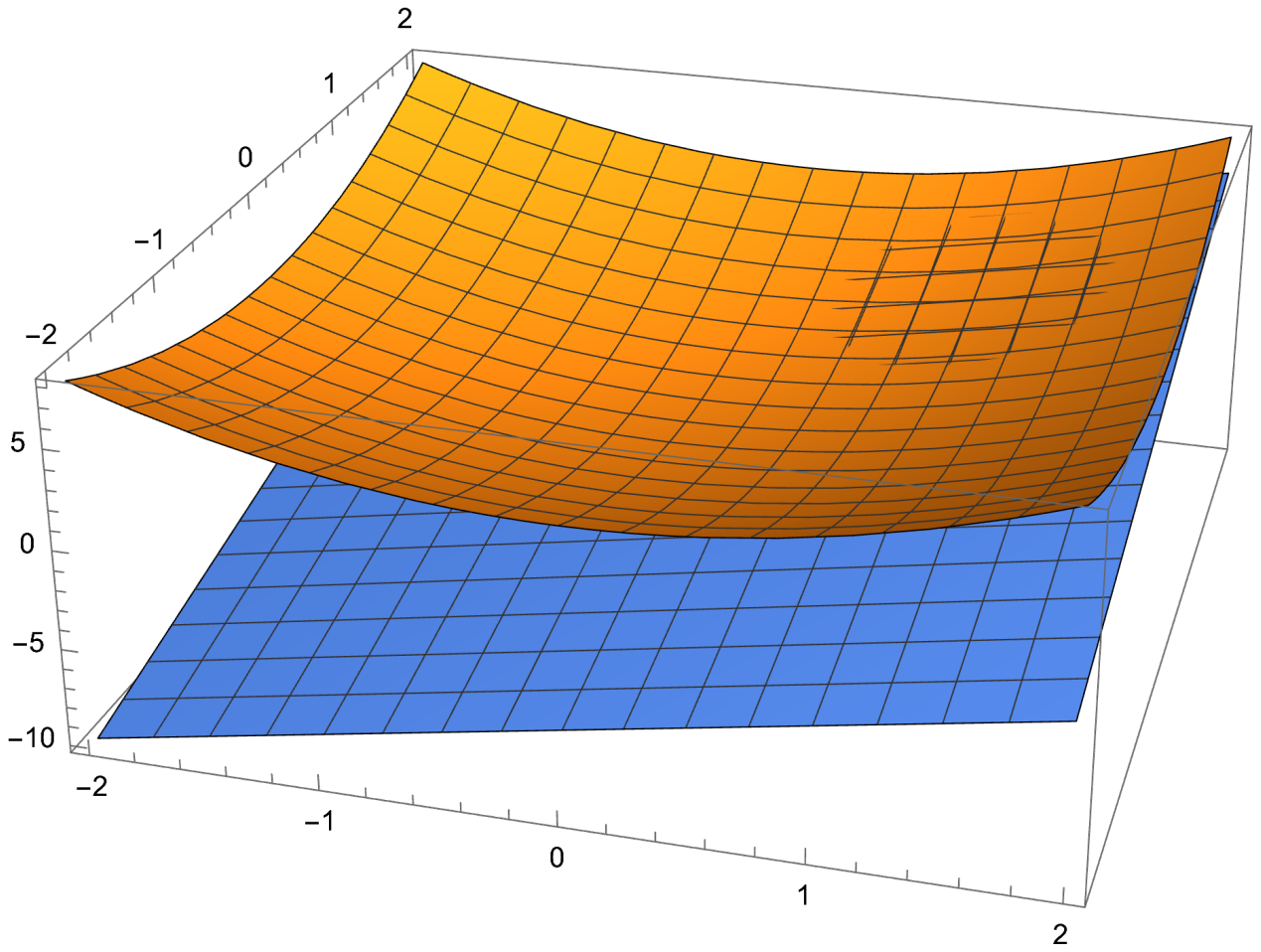
$$< \nabla(F(\boldsymbol{a})), (\boldsymbol{x} - \boldsymbol{a}) >= 0$$

where $F(\boldsymbol{x}, x_{n+1}) = f(\boldsymbol{x}) - x_{n+1}$. In other words

$$x_{n+1} = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\boldsymbol{a})(x_j - a_j).$$

The tangent plane to the $\mathrm{graf}(f)$

Example. Let $f(x, y) = x^2 + y^2$. The tangent plane to the $\mathrm{graf}(f)$ at the point $(1, 1)$ is given by $2x + 2y - z = 2$.



Gradient vector: Linear approximation

- If $n = 1$ the linear approximation of the function f at a point $\boldsymbol{a} \in D \subset R$ is defined by the linear function

$$L(x) = f(x_0) + f'(x_0)(x - x_0).$$

- In dimension n we have

$$L(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{x}_0))^T(\boldsymbol{x} - \boldsymbol{x}_0).$$

Gradient vector: Linear approximation

Example: Let $f(\boldsymbol{x}) = f(x, y, z) = ze^x\sqrt{y}$. Estimate the value of $f(0.01, 24.8, 1.02)$.

We take $\boldsymbol{a} = (0, 25, 1)^T$ and we use the linear approximation of f given by

$$L(\boldsymbol{x}) = f(\boldsymbol{a}) + (\nabla f(\boldsymbol{a}))^T(\boldsymbol{x} - \boldsymbol{a}) = 5 + (5,\ 1/10,\ 5)\begin{pmatrix} x - 0 \\ y - 25 \\ z - 1 \end{pmatrix}$$

$$= 5 + 5x + \frac{1}{10}(y - 25) + 5(z - 1)$$

Finally

$$L(0.01, 24.8, 1.02) = 5.13 \approx f(0.01, 24.8, 1.02) = 5.1306.$$

The differential matrix

Let

$$\begin{array}{rccc} f: & D \subset \mathbb{R}^n & \longrightarrow & \mathbb{R}^m \\ & \boldsymbol{x} & \longrightarrow & f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})), \end{array}$$

where D i open.

1. The differential of f at the point $\boldsymbol{a} \in D$ is (if exists)

$$Df(\boldsymbol{a}) = \begin{pmatrix} \nabla f_1(\boldsymbol{a}) \\ \vdots \\ \nabla f_m(\boldsymbol{a}) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1(\boldsymbol{a})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{a})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\boldsymbol{a})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{a})}{\partial x_n} \end{pmatrix}.$$

2. We say that f is differentiable in D if $f_1, \ldots, f_m$ are differentiable functions in D. In particular if all partial derivatives exist and they are continuous functions in D then f is differentiable.

The differential matrix and the chain rule

- Let $f : D(f) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ and $g : D(g) \subset \mathbb{R}^m \longrightarrow \mathbb{R}^p$ differentiable functions on the open domains $D(f)$ and $D(g)$, respectively.

  Theorem (Chain rule). Then the composition $h = g \circ f$ (wherever it is well defined $f(D(f)) \subset D(g)$) is also differentiable and

$$Dh(\boldsymbol{a}) = Dg(f(\boldsymbol{a}))Df(\boldsymbol{a}).$$

Real valued functions (again): High order derivatives

Let f be a real valued function that depends on n variables

$$\begin{array}{rccl} f: & D \subset \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & \boldsymbol{x} & \longrightarrow & f(\boldsymbol{x}) = f(x_1, \ldots, x_n), \end{array}$$

Let $\boldsymbol{a} = (a_1, \ldots, a_n)^T$ be an interior point of $D(f)$. Assume that not only f but all its partial derivatives are also differentiable functions at a point $\boldsymbol{a} \in D$.

We consider the derivatives of order two (second order derivatives) as follows

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{a}) := \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)(\boldsymbol{a}).$$

Real valued functions (again): The Hessian

Let $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ be a real valued function. Assume D is open and that f admits up to second order derivatives. Then we define the Hessian matrix at the point $\boldsymbol{a} \in D$ as follows

$$H(f)(a) := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{a}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{a}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{a}) \end{pmatrix}$$

Theorem (Schwartz's Lemma). If f admits up to second order derivatives in D and those functions are continuous in D, then the Hessian matrix is symmetric. Indeed,

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{a}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{a}), \quad a \in D.$$

Taylor's expansion for f: Linear

Assume all needed regularity of f in the next slides. When we say that

$$L(\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})$$

is the linear approximation of the function f at a point $\boldsymbol{a}$ we mean (roughly speaking) that

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) + \mathcal{O}\left(|\boldsymbol{x} - \boldsymbol{a}|^2\right).$$

This is known as Taylor's expansion of order one.

Taylor's expansion for f: Quadratic

The second order approximation of f at the point $\boldsymbol{a} \in D$ is

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) + \frac{1}{2} H(f)(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})^2 + \mathcal{O}\left(|\boldsymbol{x} - \boldsymbol{a}|^3\right)$$

where the value of $H(f)(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})^2 \in \mathbb{R}$ is given by

$$(x_1 - a_1, \ldots, x_n - a_n) \begin{pmatrix} \frac{\partial^2 f(\boldsymbol{a})}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\boldsymbol{a})}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{a})}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f(\boldsymbol{a})}{\partial x_n^2} \end{pmatrix} \begin{pmatrix} x_1 - a_1 \\ \vdots \\ x_n - a_n \end{pmatrix}.$$

We say that

$$Q(\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) + \frac{1}{2} H(f)(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})^2$$

is the quadratic approximation of the function f at a point $\boldsymbol{a}$.

Quadratic functions

As it will be clear during the course it is crucial for getting the optimality conditions to determine whether the expression $H(f)(\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{a})^2$ is positive or negative for $\boldsymbol{x} \in \mathbb{R}^n$ (or at least for $\boldsymbol{x}$ near $\boldsymbol{a}$). Remember that $H(f)(\boldsymbol{a})$ is a symmetric matrix. The key notion is the following.

Definition. Let Q any symmetric $n \times n$ matrix.

1. Q is positive semidefinite (PSD) if and only if $\boldsymbol{x}^T Q \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.
2. Q is positive definite (PD) if and only if $\boldsymbol{x}^T Q \boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{x} \neq 0$.

Theorem. Let Q any symmetric $n \times n$ matrix. Then all aigenvalues of Q are real. Moreover Q is PD if and only if all eigenvalues of Q are positive.

Quadratic functions

We say that f is a real quadratic function if f writes as

$$f(\boldsymbol{x}) = \boldsymbol{x}^T Q \boldsymbol{x} + c^T \boldsymbol{x} + d$$

where $Q \in \mathbb{R}^{n \times n}$, $\boldsymbol{c} \in \mathbb{R}^n$ and $d \in \mathbb{R}$.

1. f is linear if and only if $Q = 0$ and $d = 0$.
2. f is affine if and only if $Q = 0$, $\boldsymbol{c} \neq 0$, $d \neq 0$.
3. f is convex if and only if Q is PSD and $f(\boldsymbol{x}) = \boldsymbol{x}^T Q \boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{x} + d$.