OPTIMIZATION FOR DATA SCIENCE

September 29, 2025

Today (Session 1/8)

- Logistics
- Basics of optimization

Key resources: Course webpage

https://www.lamsade.dauphine.fr/~croyer/teachODS.html

LOGISTICS

2022-2024 : Common M2 MiAGE/M2 MiAGE Apprentissage

2024-2025 : Split

2025-2026 : Merge again

Grading

60% Exam (Early Feb. 2026) => ≈ like the exercises

40% Homework (groups of ≤ 2 students, extended lab session)

Schedule

$8^H30 - 11^H45$

$8^H30 - 10^H$: Lecture

$10^H15 - 11^H45$: Exercises

or $8^H30 - 11^H45$: lab session

29/09] Basics of optimization

03/11] Gradient-based algorithms
10/11

Lab sessions

! 08/12] Stochastic gradient techniques
15/12

19/01] Regularization + Recap
21/01
26/01

BASICS OF OPTIMIZATION

↳ Optimization : "Taking the best decision out of a set of alternatives"

Ex) Vaccine dispatch in Wisconsin
Power generation (e.g. EDF)
} "classical optimization models"

Classify pictures of animals
Extract topis from documents
} modern optimization problems, data-driven

→ Mathematical formulation of an optimization problem

Goal (minimize/maximize)

$$\text{minimize}_{w \in \mathbb{R}^d} \quad f(w) \quad \text{s.t.} \quad w \in \mathcal{F}$$

$f(w)$: objective function
s.t.: "subject to"
$w \in \mathcal{F}$: constraint/feasible set, $\mathcal{F} \subseteq \mathbb{R}^d$

f: measure of how good our decision is
⇒ Want the lowest possible value for f

$(f : \mathbb{R}^d \to \mathbb{R})$ — $\mathbb{R}$: float, real number

w: decision variables (vector with d coordinates)

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

$\mathcal{F}$: conditions on the decision variables
⇒ For most of the course, $\mathcal{F} = \mathbb{R}^d$

↳ 3 key aspects in optimization

- Theory: Analyze a problem, show existence of solutions, ...
- Computational: Build a solver to solve a certain class of problems (Gurobi, CVXPY, ...)

- Algorithms : → Build numerical procedures to solve a problem
  → Show theoretical guarantees
  → Validate practical performance

Examples of data science optimization problems

(1) Linear regression / linear least squares

Data $\{(x_i, y_i)\}_{i=1..n}$ n data points

$x_i \in \mathbb{R}^d$: vector of features (input)

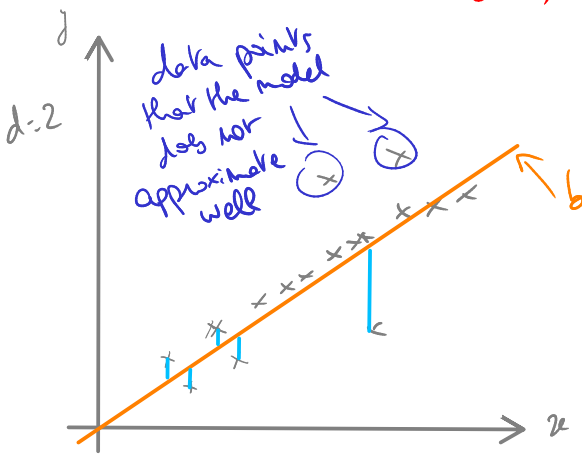
$y_i \in \mathbb{R}$: label (output)

Goal: Build the best linear model of your data

$\Leftrightarrow$ Seek $w \in \mathbb{R}^d$ such that $x_i^T w \approx y_i$ "as close as possible" (exact model may not exist)

$\forall (a,b) \in (\mathbb{R}^d)^2$, $a^T b = \sum_{j=1}^{d} a_j b_j$

Linear model of x_i: combines the coordinates in a linear way, parameterized by w



$x_i \in \mathbb{R}^d$, $x_i^T \in \mathbb{R}^{1 \times d}$

$\begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} = X \in \mathbb{R}^{n \times d}$ (matrix with n rows and d columns)

$y \in \mathbb{R}^n$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

Optimization formulation

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{2n} \|Xw - y\|^2 = \frac{1}{2n} \sum_{i=1}^{n} \left(x_i^T w - y_i\right)^2$$

$\forall v \in \mathbb{R}^n$, $\|v\|^2 = \sum_{i=1}^{n} v_i^2$ (ℓ_2 norm)

→ Minimize the MSE (mean-squared error), i.e. the average of the squares of the model errors

→ Simple, interpretable model

→ Smooth, convex optimization problem ⇒ "Easy" to solve

→ Statistical perspective: Maximum Likelihood estimation (assuming Gaussian noise)

(2) Neural network classification

Data: $\{(x_i, y_i)\}_{i=1..m}$

$x_i \in \mathbb{R}^{d_x}$

$y_i \in \{0,1\}^{d_y}$

d_y: number of classes

$[y_i]_j = \begin{cases} 1 & \text{if data point } i \text{ is in class } j \\ 0 & \text{otherwise} \end{cases}$

Model: Feed-forward neural architecture

$NN: x^0 \in \mathbb{R}^{d_0} \mapsto x^1 \in \mathbb{R}^{d_1} \mapsto \dots \mapsto x^L \in \mathbb{R}^{d_L}$

$d_0 = d_x$ $\quad d_L = d_y$

L: number of layers

$\forall l = 1, \dots, L, \quad x^l = \sigma(W^l x^{l-1} + b^l)$

Nonlinear activation function $\sigma(x) \sigma(v) = \begin{bmatrix} \tanh(v_1) \\ \tanh(v_p) \end{bmatrix}$

Linear transformation

$W^l \in \mathbb{R}^{d_l \times d_{l-1}}$

$b^l \in \mathbb{R}^{d_l}$

w (model parameters): concatenation of (W^ℓ, b^ℓ)

$w \in \mathbb{R}^d$ $\quad d = d_1 d_0 + d_1 + d_2 d_1 + d_2 + \cdots + d_L d_{L-1} + d_L$

Goal: Given my model $NN(\ ; w)$, find w such that

$\forall i = 1 \dots n, \quad NN(x_i; w) \approx y_i$

Key ≠ with linear regression:

- Model is nonlinear
- $y_i \in \{0,1\}^{d_y} \Rightarrow$ closeness to y_i measured differently than in a regression context

Optimization problem

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} \left\{ \log\left(\sum_{j=1}^{d_y} \exp\left([NN(x_i; w)]_j \right) \right) - \sum_{j=1}^{d_y} [y_i]_j \, [NN(x_i; w)]_j \right\}$$

Finite-sum structure (average over data points)

Negative Log-likelihood function (accounts for the fact that $NN(x_i; w)$ is not necessarily a 0/1 vector)

Model $NN(x_i; w) = x^L$ when given $x^0 = x_i$ as input and w for parameters

→ High-dimensional, highly nonconvex problem

→ Yet it can be solved using stochastic gradient algorithms

1) Solutions of optimization problems

From now on, we consider

$$(P) \quad \min_{w \in \mathbb{R}^d} f(w) \qquad f : \mathbb{R}^d \to \mathbb{R}$$

Def.: $w^* \in \mathbb{R}^d$ is a solution / a global minimum of (P) if

$$\forall w \in \mathbb{R}^d, \quad f(w^*) \leq f(w)$$

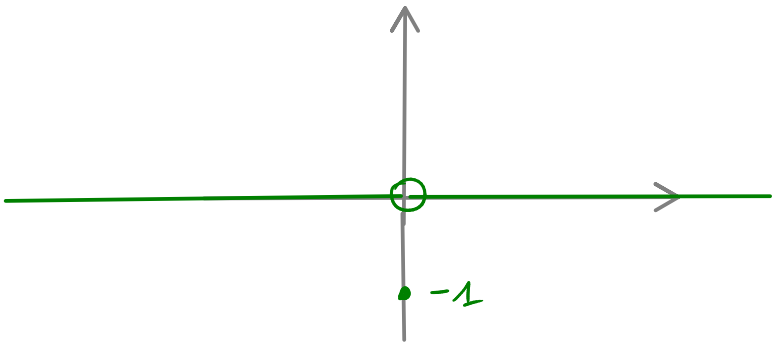
→ Finding global minima is (NP-) hard!

→ In practice, we:

- use classes of functions f for which minimization is easier
- use weaker notions of minima

(Ex) $d=1$

$$f(w) = \begin{cases} 0 & \text{if } w \neq 0 \\ -1 & \text{if } w = 0 \end{cases}$$



Def.: $\bar{w} \in \mathbb{R}^d$ is a **local** minimum of (P) if

$$\exists \varepsilon > 0 \text{ such that } \forall w \in \mathbb{R}^d, \ \|w - \bar{w}\| \leq \varepsilon, \ f(\bar{w}) \leq f(w)$$

→ Any global minimum is a local minimum, but the converse is not true

→ Local minima are also difficult to find in general

but it is easier to check that a point is a local minimum, especially for certain classes of functions

C^1 functions

"continuously differentiable"

$f: \mathbb{R}^d \to \mathbb{R}$ is called C^1 if $\forall w \in \mathbb{R}^d$, $\exists \nabla f(w) \in \mathbb{R}^d$ that represents the local variations in f around w

($\nabla f(w)$: Gradient of f at w)

Intuition: If $v \in \mathbb{R}^d$ is close to w, then

$$f(v) \approx f(w) + \nabla f(w)^T (v - w)$$

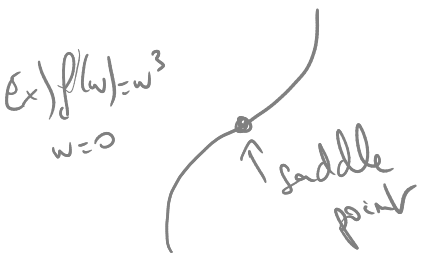
($v \mapsto f(w) + \nabla f(w)^T (v-w)$: Linear function of v)

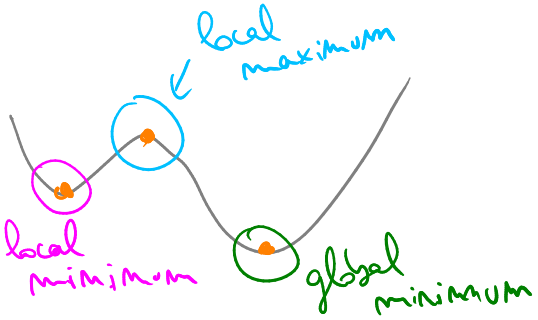
Theorem (for f C^1):

$$\left[\bar{w} \in \mathbb{R}^d \text{ local minimum of } f(P) \right] \Rightarrow \left[\nabla f(\bar{w}) = 0_{\mathbb{R}^d} \right]$$

"First-order optimality conditions"

→ Only an implication, converse not true in general: A point zero gradient can be a minimum, a maximum or a saddle point





• points with $\nabla f(\,) = 0$

→ If we have access to ∇f, we can check whether $\nabla f(w) = 0$

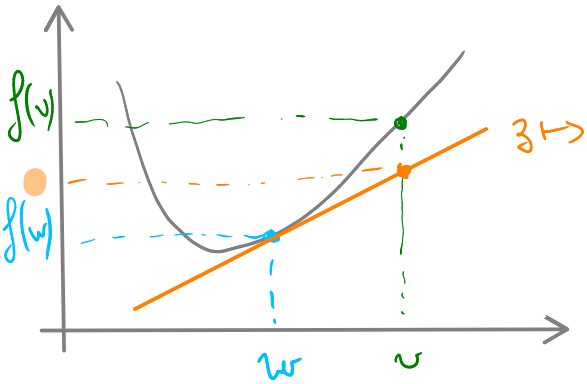
or not => Basic idea behind the gradient descent algorithm !

C^1, convex functions

Def.: $f \in C^1$ is convex if

$$\forall (v, w) \in (\mathbb{R}^d)^2, \quad f(v) \geq f(w) + \nabla f(w)^T (v - w)$$

Linear function of v (Approximation / Tangent of f at w)



$z \mapsto f(w) + \nabla f(w)^T (z - w)$

Ex). Any linear function $w \mapsto a^T w + b$ is convex $a \in \mathbb{R}^d$, $b \in \mathbb{R}$

- $w \mapsto \frac{1}{2} \|w\|^2$ convex
- $w \mapsto \frac{1}{2n} \|Xw - y\|^2$ $\forall X \in \mathbb{R}^{n \times d}$, $\forall y \in \mathbb{R}^n$, is convex

Key properties of C^1, convex functions

f C^1 convex

- Every local minimum of minimize $f(w)$, $w \in \mathbb{R}^d$ is a global minimum.
- $[w^*$ global minimum of minimize $f(w)$, $w \in \mathbb{R}^d] \iff [\nabla f(w^*) = 0_{\mathbb{R}^d}]$

C^1, strongly convex functions

Def: f C^1 is μ-strongly convex (for $\mu > 0$) if

$$\forall (v,w) \in (\mathbb{R}^d)^2, \quad f(v) \geq f(w) + \nabla f(w)^T (v-w) + \frac{\mu}{2} \|v-w\|^2$$

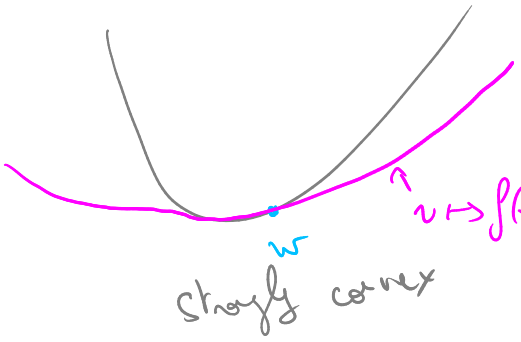
Quadratic term

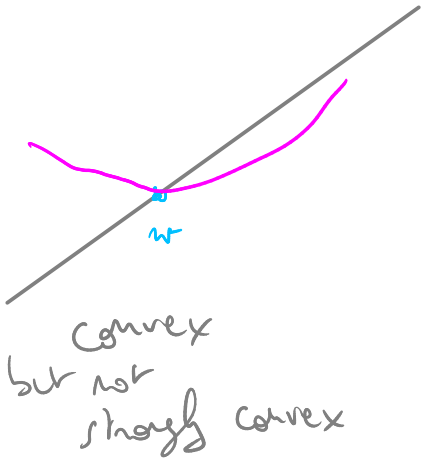
Quadratic function of v

→ [μ-strongly convex] ⇒ convex

→ Examples: $w \mapsto \frac{\mu}{2}\|w\|^2$ μ-strongly convex

$w \mapsto a^T w + b$ not strongly convex





Convex but not strongly convex

Key property of C^1 strongly convex functions.

If f C^1, μ-strongly convex, minimize $f(w)$ $w \in \mathbb{R}^d$ has a unique global minimum, and it is the only point $w^* \in \mathbb{R}^d$ such that $\nabla f(w^*) = 0_{\mathbb{R}^d}$

$$\nabla f(w^*) = 0_{\mathbb{R}^d} \iff \forall j = 1..d, \quad [\nabla f(w^*)]_j = 0 \in \mathbb{R}$$

$$0_{\mathbb{R}^d} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^d$$

$$\iff \|\nabla f(w^*)\| = 0$$

$$\|v\| = \sqrt{\sum_{j=1}^{d} v_j^2}$$

C^2 functions

Hessian matrix of f at w

Square matrix of size $d \times d$

Def: f is C^2 if it is C^1 and $\forall w \in \mathbb{R}^d$, $\exists \nabla^2 f(w) \in \mathbb{R}^{d \times d}$ that reflects the (second-order) changes of f around w

Intuition: $f(v) \approx f(w) + \nabla f(w)^T (v-w) + \frac{1}{2} (v-w)^T \nabla^2 f(w) (v-w)$

$(v-w)^T \in \mathbb{R}^{1 \times d}$, $\nabla^2 f(w) \in \mathbb{R}^{d \times d}$, $(v-w) \in \mathbb{R}^d$

$\nabla^2 f(w) (v-w) \in \mathbb{R}^d$

$(v-w)^T \nabla^2 f(w) (v-w) \in \mathbb{R}$

Theorem: f C^2, (P) $\min_{w \in \mathbb{R}^d} f(w)$

2nd order necessary conditions

i) $[\bar{w}$ local minimum of $(P)] \Rightarrow [\nabla f(\bar{w}) = 0_{\mathbb{R}^d}$ and $\nabla^2 f(\bar{w}) \succeq 0]$

ii) $[\nabla f(\bar{w}) = 0_{\mathbb{R}^d}$ and $\nabla^2 f(\bar{w}) \succ 0] \Rightarrow [\bar{w}$ local minimum of $(P)]$

2nd order sufficient conditions

$\forall A \in \mathbb{R}^{d \times d}$, $A \succeq 0$ = "A is positive semi-definite"

Löwner order

$= A$ symmetric ($A_{ij} = A_{ji}$ $\forall (i,j)$) and $v^T A v \geq 0$ $\forall v \in \mathbb{R}^d$

⚠ $A \succeq 0$ does not imply that $A_{ij} \geq 0 \ \forall (i,j)$ $\quad A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$

$A \succ 0$ = "A positive definite"

$= A \succeq 0$ and $v^T A v > 0 \ \forall v \in \mathbb{R}^d, v \neq 0_{\mathbb{R}^d}$

C^2, convex functions

$f \ C^2$

i) f convex if $\nabla^2 f(w) \succeq 0 \quad \forall w \in \mathbb{R}^d$

ii) f μ-strongly convex (with $\mu > 0$)

if $\nabla^2 f(w) \succeq \mu I \quad \forall w \in \mathbb{R}^d$

$\Leftrightarrow \nabla^2 f(w) - \mu I \succeq 0$

$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{d \times d}$

(identity matrix "eye")

TAKEAWAYS

For convex functions: . $\nabla^2 f(w) \succeq 0$ true $\forall w$

. 1st order conditions $\Leftrightarrow$ 2nd order necessary conditions

. local $\Leftrightarrow$ global $\Leftrightarrow \nabla f(\ ) = 0$

For μ-strongly convex: . $\exists$ unique global minimum

. Special class of convex functions

C^1 / C^2: . classes of functions for which we have conditions linked to global minima

Exercise (number 1 in Exercise sheet 1)

Linear least squares

$\{(x_i, y_i)\}_{i=1..m}$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

(P) $\quad \underset{w \in \mathbb{R}^d}{\text{minimize}} \; f(w) = \frac{1}{2m}\|Xw - y\|^2 = \frac{1}{2m}\sum_{i=1}^{m}(x_i^T w - y_i)^2$

a) Suppose that there exists w^* such that $\underbrace{X}_{m\times d}\underbrace{w^*}_{d\times 1} = \underbrace{y}_{m\times 1}$.

What is the link between w^* and problem (P) ?

- $f(w^*) = \frac{1}{2m}\|Xw^* - y\|^2 = \frac{1}{2m}\|y - y\|^2 = \frac{1}{2m}\|0\|^2 = 0$
- $\forall w \in \mathbb{R}^d$, $f(w) = \frac{1}{2m}\sum_{i=1}^{m}(x_i^T w - y_i)^2 \geq 0$ (average of squares ≥ 0)

$\Rightarrow \forall w \in \mathbb{R}^d$, $f(w^*) \leq f(w)$

Hence w^* is a global minimum of (P)

b) $\nabla f(w) = \frac{1}{m}\underbrace{X^T}_{d\times m}\underbrace{(Xw - y)}_{m\times 1} \quad \forall w \in \mathbb{R}^d$ $\quad (d\times 1)$

If $\bar{w}$ is a local minimum of (P), what condition involving $\nabla f(\bar{w})$ holds? Does this condition hold for w^*?

$$\left[\bar{w} \text{ local minimum of (P)}\right] \Rightarrow \left[\nabla f(\bar{w}) = 0_{\mathbb{R}^d}\right]$$

If w^* is the point from a),

$$\nabla f(w^*) = \frac{1}{n} X^T (X w^* - y) = \frac{1}{n} X^T (y - y) = \frac{1}{n} X^T 0_{\mathbb{R}^n} = 0_{\mathbb{R}^d}$$

c) $\nabla^2 f(w) = \frac{X^T X}{n}$ for every $w \in \mathbb{R}^d$ (constant because f is a quadratic function)

i) $\frac{1}{n} X^T X \geq 0$. What does this imply on f and (P)?

$\nabla^2 f(w) \geq 0 \ \forall w \Rightarrow f$ is convex

Then, any local minimum of (P) is global

ii) If we had $\frac{1}{n} X^T X \geq \mu I$ with $\mu > 0$, what would that imply on f and (P)?

$\nabla^2 f(w) \geq \mu I \ \forall w \Rightarrow f$ is μ-strongly convex

Then, (P) has a unique global minimum