Mathématiques pour l'Intelligence Artificielle
Notes de cours – CM5

Rédigé par : BELLILI Idir, DIALLO Ismaila, SAAL Racim

25 septembre 2025

Table des matières

1 Un mot sur l'optimisation non lisse

1.1 Qu'est-ce qu'un bon algorithme d'optimisation ?

Un algorithme d'optimisation est généralement jugé comme **bon** selon plusieurs critères, parmi lesquels on cite :

1. Convergence : converger vers une solution selon les hypothèses du problème.
2. Vitesse de convergence : Un bon algorithme est un algorithme qui réussit à converger rapidement (un nombre minimal d'itérations).

1.2 Dualité : les cas d'application

La dualité permet l'optimisation dans différents scénarios :

— **SVM (Support Vector Machine)**

La formulation duale du problème permet de transformer un problème de dimensions élevées en un problème où seules les paires de points apparaissent, ce qui est utilise pour l'implémentation, l'analyse, le noyau.

— **Méthodes a noyaux**

La dualité permet de remplacer les produits scalaires dans l'espace original par des noyaux, sans repasser explicitement dans des dimensions élevées (`kernel trick`). Le duale permet d'exprimer les solutions comme combinaisons linéaires des vecteurs de support.

— **Optimisation convexe**

formulation d'un problème sous contraintes en problème duale et utilisation de la dualité lagrangienne pour résoudre et analyser les solutions plus facilement, avec le passage a la fin vers la solution primal a partir de la solution duale.

— **Relaxations de problèmes difficiles**

pour un problème non convexe, on peut le relâcher en une version convexe pour obtenir une approximation.

Donc, en résumé, la dualité sert a :

— déduire les conditions d'optimalité (KKT, duales).

— exploitation du noyau dans les méthodes d'apprentissage automatique.

— obtenir des relaxations pour des problèmes difficiles.

1.3 Optimisation non lisse

L'optimisation non lisse concerne les fonctions non différentiables partout, telles que la fonction de régularisation `L1` et la fonction d'activation `ReLU`.

1.3.1 Cas convexe non lisse

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction convexe, possiblement non différentiable.

Sous-gradient

Un vecteur $g \in \mathbb{R}^d$ est un *sous-gradient* de f en $x \in \mathbb{R}^d$ si, pour tout $y \in \mathbb{R}^d$,

$$f(y) \ \geq \ f(x) + g^\top (y - x).$$

L'ensemble de tous les sous-gradients en x est appelé *sous-différentiel* de f en x, noté

$$\partial f(x) \ = \ \{g \in \mathbb{R}^d \mid f(y) \geq f(x) + g^\top (y - x), \ \forall y \in \mathbb{R}^d\}.$$

Condition d'optimalité pour les fonctions convexes

Soit $f : \mathbb{R}^d \to \mathbb{R}$ convexe. Si x^* est un minimum local de f, alors

$$0 \ \in \ \partial f(x^*).$$

En particulier, tout minimum local est aussi un minimum global.

1.3.2 Cas non convexe non lisse : somme lisse + régularisation convexe

Considérons le problème

$$\min_{x \in \mathbb{R}^d} \ F(x) := f(x) + g(x),$$

où :

— $f : \mathbb{R}^d \to \mathbb{R}$ est différentiable (mais possiblement non convexe),

— $g : \mathbb{R}^d \to \mathbb{R}$ est convexe mais possiblement non différentiable (exemple : norme ℓ_1).

Condition d'optimalité

Un point x^* est solution optimale locale si et seulement si

$$0 \ \in \ \nabla f(x^*) + \partial g(x^*).$$

C'est-à-dire qu'il existe $v \in \partial g(x^*)$ tel que

$$\nabla f(x^*) + v = 0.$$

Exemple : Le *Lasso* correspond à

$$\min_x \ \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1,$$

2 Rappels de probabilités

2.1 Marginales et conditionnelles

Marginale (continu) : $p(X_i) = \displaystyle\int p(X_1, \ldots, X_n)\, dX_{j \neq i}$.

Marginale (discret) : $p(X_i) = \displaystyle\sum_{x_{j \neq i}} p(X_1, \ldots, X_n)$.

Conditionnelle : $p(A \mid B) = \dfrac{p(A, B)}{p(B)}$.

2.2 Formule de Bayes

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}.$$

2.3 Inclusion–exclusion et indépendance

$$p(A \vee B) = p(A) + p(B) - p(A \wedge B).$$

Deux événements sont indépendants ssi $p(A, B) = p(A)p(B)$.

2.4 Interprétation d'une densité continue

1. **La densité n'est pas une probabilité, c'est une « hauteur ».** La probabilité sur un intervalle $[a, b]$ est l'*aire* sous la courbe :

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx.$$

2. **Probabilité d'un point = 0 (en continu).** $\mathbb{P}(X = c) = 0$. On ne « voit » la probabilité qu'en intégrant sur un *intervalle*.
3. **Une densité peut dépasser 1 (ce n'est pas un problème).** Ce qui doit valoir 1, c'est l'*aire totale* $\int_{-\infty}^{+\infty} f_X(x)\, dx = 1$.
   Ex. Si $X \sim \mathcal{U}([0,0{,}1])$, alors $f(x) = 10$ sur $[0,0{,}1]$. Ainsi, $\mathbb{P}(X \in [0,0{,}01]) = \int_0^{0.01} 10\, dx = 0.1$ (10%).

Exemples

Exemple A — Uniforme $[0, 1]$

- $f(x) = 1$ sur $[0, 1]$ (et 0 sinon).
- Probabilité d'une fenêtre de *même largeur* 0,01 n'importe où dans $[0, 1]$:

$$\mathbb{P}(X \in [0.30, 0.31]) = \int_{0.30}^{0.31} 1\, dx = 0.01 \quad (1\%).$$

- **Conclusion :** sous l'uniforme, *toutes* les fenêtres de même largeur ont la *même probabilité*.

Exemple B — Normale standard $X \sim \mathcal{N}(0, 1)$
Densité $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, avec $\varphi(0) = \frac{1}{\sqrt{2\pi}} \approx 0.399$ et $\varphi(2) = \frac{1}{\sqrt{2\pi}} e^{-2} \approx 0.054$. Pour une *même largeur* $h = 0{,}1$:

$$\mathbb{P}(X \in [-0.05, 0.05]) \approx \varphi(0)\, h \approx 0.0399, \qquad \mathbb{P}(X \in [1.95, 2.05]) \approx \varphi(2)\, h \approx 0.0054.$$

Conclusion : la fenêtre autour de 0 est $\approx 7{,}4$ fois plus probable (car $\varphi(0)/\varphi(2) \approx 7.4$) : à *largeur égale*, la *hauteur* de la densité fixe l'ordre de grandeur des probabilités locales.

3 Modèles graphiques dirigés

Si chaque nœud X_i a pour parents X_{π_i}, alors la loi jointe se **factorise** :

$$p(X_1, \ldots, X_n) = \prod_i p(X_i \mid X_{\pi_i}).$$

Cette factorisation encode les indépendances conditionnelles et facilite les calculs de marginales/conditionnelles.

4 Loi Gaussienne multivariée

4.1 Vecteurs aléatoires (Random Vectors)

Soient $X_1, \ldots, X_n$ des variables aléatoires. On définit le vecteur aléatoire :

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

L'espérance d'une fonction $g : \mathbb{R}^n \to \mathbb{R}^m$ appliquée à X est définie composante par composante :

$$\mathbb{E}[g(X)] = \begin{bmatrix} \mathbb{E}[g_1(X)] \\ \mathbb{E}[g_2(X)] \\ \vdots \\ \mathbb{E}[g_m(X)] \end{bmatrix}$$

4.2 Matrices de covariance (Covariance Matrices)

Pour un vecteur aléatoire $X \in \mathbb{R}^n$, la matrice de covariance Σ est la matrice $n \times n$ dont l'élément (i, j) est $\mathrm{Cov}[X_i, X_j]$:

$$\Sigma = \begin{bmatrix} \mathrm{Cov}[X_1, X_1] & \cdots & \mathrm{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_n, X_1] & \cdots & \mathrm{Cov}[X_n, X_n] \end{bmatrix}$$

En utilisant la linéarité de l'espérance et le fait que :

$$\mathrm{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$$

on obtient l'expression matricielle :

$$\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$

4.2.1 Propriétés de la matrice de covariance

— Σ est symétrique et semi-définie positive
— Si $X_i \perp X_j$ pour tous i, j, alors $\Sigma = \mathrm{diag}(\mathrm{Var}[X_1], \ldots, \mathrm{Var}[X_n])$
— La variabilité totale est donnée par la trace de Σ

4.3 Définition de la loi Gaussienne multivariée

4.3.1 Définition par densité

Un vecteur aléatoire $X \in \mathbb{R}^n$ suit une loi Gaussienne multivariée $X \sim \mathcal{N}(\mu, \Sigma)$ si sa densité est donnée par :

$$p(x; \mu, \Sigma) = \frac{1}{\det(\Sigma)^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

4.3.2 Cas univarié

Le cas univarié $X \sim \mathcal{N}(\mu, \sigma^2)$ est un cas particulier lorsque $n = 1$:

$$p(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

On vérifie que si $\Sigma \in \mathbb{R}^{1\times 1}$, alors $\Sigma = \mathrm{Var}[X_1] = \sigma^2$, et donc :
— $\Sigma^{-1} = \frac{1}{\sigma^2}$
— $\det(\Sigma)^{\frac{1}{2}} = \sigma$

4.4 Propriétés importantes des Gaussiennes multivariées

— Les marginales et conditionnelles d'une Gaussienne multivariée sont également Gaussiennes

— Une Gaussienne d-dimensionnelle $X \sim \mathcal{N}(\mu, \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2))$ est équivalente à une collection de d Gaussiennes indépendantes $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$

— Les isocontours d'une Gaussienne multivariée sont des ellipsoïdes n-dimensionnels avec des axes principaux dans les directions des vecteurs propres de Σ

— Les longueurs relatives des axes dépendent des valeurs propres de Σ

4.5 Distributions conditionnelles et marginales

Soit $X \sim \mathcal{N}(\mu, \Sigma)$ partitionné comme suit :

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

avec $X_1 \in \mathbb{R}^q$ et $X_2 \in \mathbb{R}^{n-q}$.

4.5.1 Distribution conditionnelle

La distribution conditionnelle de X_1 sachant $X_2 = a$ est :

$$p(X_1 \mid X_2 = a) = \mathcal{N}(\bar{\mu}, \bar{\Sigma})$$

avec

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{+}(a - \mu_2)$$
$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{+}\Sigma_{21}$$

où Σ_{22}^{+} désigne la pseudo-inverse de Σ_{22}.

4.5.2 Distributions marginales

Les distributions marginales de X_1 et X_2 sont simplement :

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}), \quad X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$$

4.6 Visualisations et interprétations

Les visualisations des Gaussiennes multivariées montrent comment la forme des isocontours change selon Σ :

— Si $\Sigma = \sigma^2 I$, les isocontours sont des sphères

— Si Σ est diagonale avec des variances différentes, les isocontours sont des ellipsoïdes alignés avec les axes

— Si Σ a des covariances non nulles, les ellipsoïdes sont inclinés selon les directions des vecteurs propres

— Les corrélations positives inclinent l'ellipsoïde le long de la première bissectrice, les négatives le long de la seconde bissectrice