Optimisation Non linèaire

Par

Professeur Abdellatif El Afia

Optimisation Non linèaire

Partie1: Optimisation sans contraintes

1. Motivation
    1. Connaissances de base
    2. Notations et définitions
    3. Résultat préliminaire
    4. Théorie de l'Algorithme

2. Algorithmes
    1. Fonction à une seule variable
    2. Fonction à plusieurs variable

Partie 2: Optimisation avec contraintes

1. Analyse Convexe
    1. Ensembles convexes
    2. fonctions convexes
    3. Hyperplan
    4. Théories de séparation
2. Conditions d'optimalité
    1. Multiplicateur de Lagrange
    2. Conditions de Karush-Kuhn-Tucker
3. Algorithmes primales
4. Algorithmes Duales
5. Algorithmes de Pénalités

Optimisation non Linéaire
Partie1: Optimisation sans contraintes

1. **Motivation**
2. **Connaissances de base**
    1. **Notations et définitions**
    2. **Résultat préliminaire**
    3. **Algorithme**

Motivation

- Au fur et à mesure que la puissance de l'ordinateur augmente et que les technologies de collecte de données avancent**, un surcharge de données est toujours présent** dans presque tous les domaines où les ordinateurs sont utilisés.

- **Machine Learning** est l'un des domaines de recherche informatiques les plus avancée qui permet de découvrir (déduire) les connaissances à partir de données importantes (**Big data**). En se basant sur des statistiques et sur les analyses prédictives.

- Parmi ces algorithmes, ils existent ceux qui réalisent la classification et d'autres font la régression

Motivation

- Ils se décomposent généralement en 2 étapes:
  - **Phase d'entraînement:** phase d'apprentissage sur une partie des données en **minimisant l'erreur d'entrainement**
  - **Phase de test:** phase d'approximation de la généralisation sur la deuxième partie de données en minimisant l'erreur de classification ou de régression.

- Malheureusement, **Il est impossible de minimiser l'erreur de généralisation directement** vue que les algorithmes de Machine Learning n'ont l'accès qu'aux données d'entrainement.

Motivation

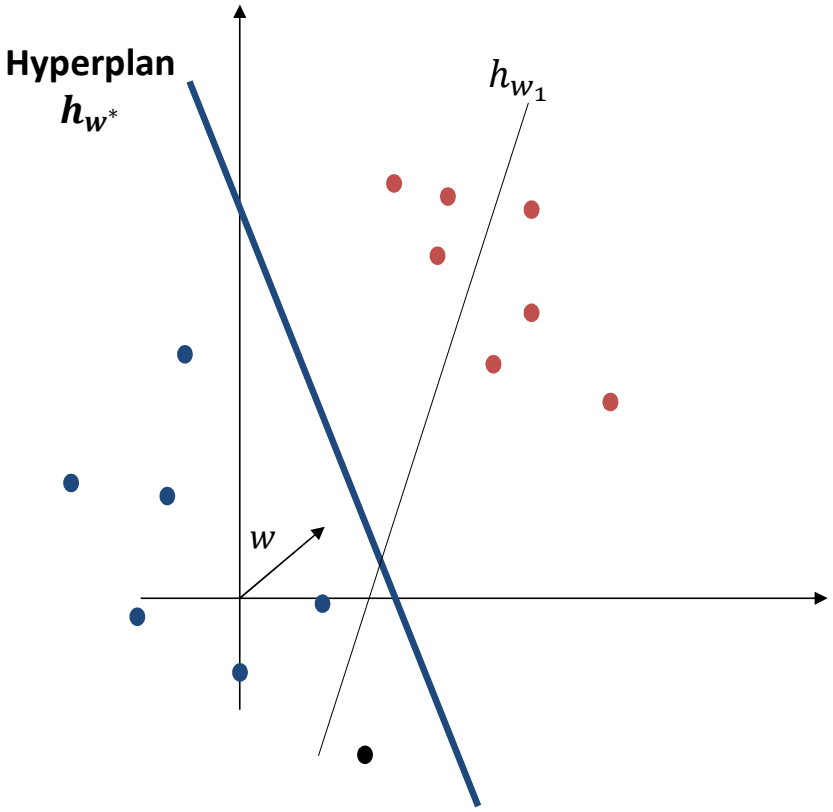
Phase d'entraînement

- $\{x_i, y_i\}_{i=1}^{n}$
  - $\rightarrow x_i \in \mathbb{R}^d$
  - $\rightarrow y_i \in \{-1,1\}$ classification binaire
  - $\rightarrow y_i \in \{0,1,2,3\}$ classification multiclass
  - $\rightarrow y_i \in \mathbb{R}$ regression
  - $\rightarrow y_i \in [0,1]$ distribution, loi de probabilite
- Obj: trouver une relation entre $\forall i \;\; h(x_i) = y_i$
- En Minimisant(optimisation) La fonction de perte
  - $\rightarrow$ l'objet de ce cours

Phase de test:

- Data testing
- Future data

Motivation: Algorithme de Perceptron à marge dure



- **Phase d'entraînement:**

$$S = \{(x_1, y_1) \dots, (x_m, y_m)\}$$

- $y_1 \in Y = \{Rouge, Bleu\ Foncé\}$
- $x_i \in \mathbb{R}^d$
- **Perceptron à marge dure**, cherche un séparateur linéaire, $\boldsymbol{h_w^* \in H}$
- $H = \{h: \mathbb{R}^d \to \mathbb{R} \,|\, h_w(x) = x^T w\}$
- $H = \{w \in \mathbb{R}^d \,|\, h_w(x) = x^T w\}$

$$h(x_i) = y_i$$

$$h(x_i) = \boldsymbol{signe}(\boldsymbol{h_{w^*}}(x_i)) = \boldsymbol{signe}(w^{*T} x_i) = y_i$$

$y_i\ h(x_i) < 0 \ \to y_i w^{*T} x_i < 0$ x_i est mal classifier

$y_i\ h(x_i) > 0 \ \to \ y_i w^{*T} x_i > 0$ x_i est bien classifie

Motivation: Algorithme de Perceptron à marge dure

- $\rightarrow x_i \in \mathbb{R}^d$
- $\rightarrow y_i \in \{-1,1\}$ classification binaire

- $h(x_i) = y_i$
- $h(x_i) = \boldsymbol{signe}(\boldsymbol{h_{w^*}}(x_i)) = \boldsymbol{signe}({w^*}^T x_i) = y_i$
- $y_i\, h(x_i) < 0 \rightarrow y_i {w^*}^T x_i < 0$ x_i est mal classifier
- $y_i\, h(x_i) > 0 \rightarrow y_i {w^*}^T x_i > 0$ x_i est bien classifier
- $L_S(h) = L_S(w)$
- $\max(0, -y_i w^T x_i)$
- $L_S(w) = \frac{1}{m}\sum_{i=1}^{m} \max(0, -y_i w^T x_i)$
- $L_S(w) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{w^T x_i \neq y_i\}}$
- $\min_{w\in\mathbb{R}^d} L_S(w) \Leftrightarrow w^* = \underset{w\in\mathbb{R}^d}{arg\min}\, L_S(w)$

Motivation: Algorithme de Perceptron à marge dure

- Considérons un ensemble d'entrainement $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ la partie de donnée pour l'apprentissage.
- **Avant de parler de cet algorithme il** faut mieux de mentionner qu'il existe plusieurs mesures d'erreurs appelées fonctions de perte ou Erreur Empirique Noté par L_S .
- L'un des fonctions de perte dit « **hinge loss** » est définie comme suivant :

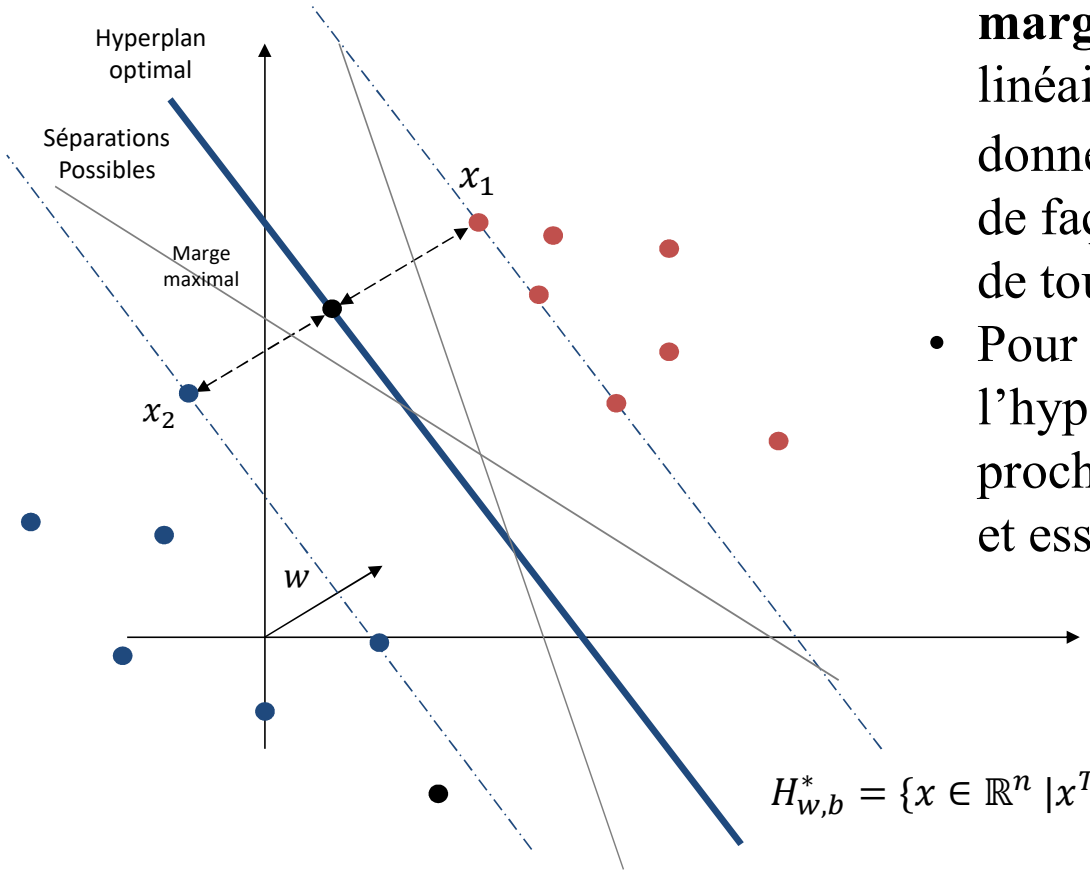
$$\max(0, -y_i w^T x_i) = \begin{cases} -y_i w^T x_i & si \quad -y_i w^T x_i > 0 \\ 0 & sinon \end{cases}$$

- Une variance de perceptron prend cette mesure d'erreur et essaie de la minimiser, ce qui conduit au modèle mathématique suivant:

$$w^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \max(0, -y_i w^T x_i)$$

- Ce dernier représente **un problème d'optimisation sans contrainte**.

Motivation: SVM pour la classification « SVC» à marge dure



- **SVM pour la classification avec marge dure**, cherche un séparateur linéaire,$\boldsymbol{H^*_{w,b}}$, entre les points de données de deux classes différentes de façon qu'il soit loin que possible de tous les points de donnée
- Pour cela, On définit un marge entre l'hyperplan et les points les plus proches de chaque classe (différent) et essaie de la maximiser.

$$H^*_{w,b} = \{x \in \mathbb{R}^n \mid x^T w + b = 0\}$$

Motivation: SVM pour la classification « SVC» à marge dure

- L'objectif la méthode SVM est de Maximisez la Marge Sous contraintes que l'ensemble d'entrainement soient bien classées
- Donc la méthode SVM adopte un modèle **mathématique avec contrainte.**

$$\begin{cases} min\left(\frac{1}{2}w^T w\right) \\ sujet\ \text{à} \quad y_i.(w^T x_i + b) \geq 1 \end{cases}$$

- Problème d'optimisation quadratique sous contraintes linéaires pour lesquels existe une vaste d'algorithmes d'optimisation. On obtient w^* et b^*en définissant la fonction de décision:

$$h(x) = sign\left(w^{*T} x_i + b^*\right)$$

Motivation: **Régression Linéaire**

- L'algorithme de régression linéaire est basé sur la minimisation d'erreur quadratique entre $h(x)$ et y

$$L_G(h) = E[h(x) - y]^2$$

- Pour un ensemble d'entrainement S $= \{(x_1, y_1), \dots, (x_m, y_m)\}$ l'erreur est calculé comme suivant :

$$\boldsymbol{L_S(h) = \frac{1}{n}\sum_{i=1}^{m}(h(x_i) - y_i)^2}$$

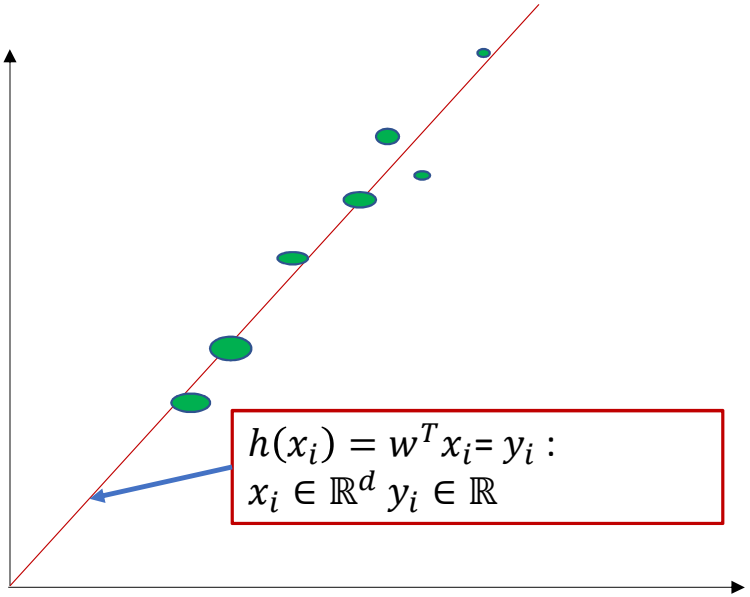
Avec :

$$h(x) = w^T x \Longrightarrow L_S(h) = L_S(w)$$

Une fonction de la régression linéaire est obtenue par minimisation de l'erreur pour toutes les valeur de w d'où le modèle d'optimisation :

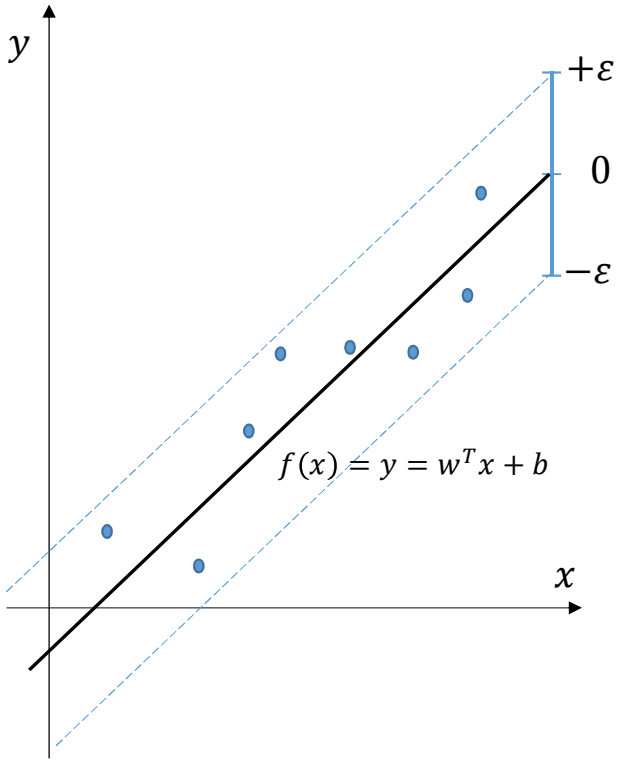
$$w = \arg\min_w L_S(w)$$

Motivation: **Régression Linéaire**



- $d_2(h(x_i), y_i) = (h(x_i) - y_i)^2$
- $d_1(h(x_i), y_i) = |h(x_i) - y_i|$
- $L_S(w) = \frac{1}{n}\sum_{i=1}^{n} d(h(x_i), y_i)$
  - $L_S(w) = \frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$ si $d_2 = d$
  - $L_S(w) = \frac{1}{n}\sum_{i=1}^{n}|h(x_i) - y_i|$ si $d_1 = d$

Motivation: SVM pour la **Régression(SVR) à marge dure**



- La régression SVR a pour objectif de trouver une fonction $f(x)$ qui a au maximum ε de déviation par rapport aux cibles réellement obtenues y_i de tous les donnée d'entrainement.
- le problème est formulé comme suivant:

$$\begin{cases} \min \quad \frac{1}{2} w^T w \\ s.\text{à} \quad y_i - (w^T x_i - b) \leq \varepsilon \\ \qquad\quad w^T x_i + b - y_i \leq \varepsilon \end{cases}$$

Motivation: Modèle mathématique

- Généralement le modèle mathématique correspondant au problèmes de minimisation est décrit comme suivant:

$$(P)\begin{cases} \min f(x) \\ P_1: \quad f_i(x) \leq 0 \quad i \in I_1 \\ P_2: \quad f_i(x) = 0 \quad i \in I_2 \end{cases}$$

$x \in R^n$ qui vérifie P_1 et P_2 est dite réalisable

- Domaine réalisable:

$$D_R = \{x \in \mathbb{R}^n | P_1 \text{ et } P_2 \text{ sont verifiée } en\ x\}$$

- $(f(x), f_i(x)\ i \in I_1 \cup I_2)$ est différentiable entant de fois que c'est nécessaire.

Motivation: Approche de résolution

- Résolution de problèmes d'optimisation à l'aide d'algorithme ou processus itératif générant une suite de points $\{x_k\}$, très souvent dans le domaine réalisable du problème:

$$x^{k+1} = x^k + \alpha_k d^k$$

- x^{k+1} est généré à partir de x^k en choisissant une direction $d^k \in \mathbb{R}^n$ et en prenant un pas $\alpha_k \in \mathbb{R}$ dans cette direction pour s'éloigner de x^k.
- **Les méthodes diffèrent par leurs choix de direction d^k et de pas α_k.**

Motivation : Approche de résolution

- $\min_{x \in D_R} f(x)$
- Générer une suite $\{x^k\}$
- Phase d'exploration:
  - Calcule de d^k telque :
    - d^k est une direction realisable $\Leftrightarrow \exists \bar{\alpha}$ telque $\forall \alpha \in [0, \bar{\alpha}]$
      $$x^{k+1} = x^k + \alpha d^k \in D_R$$
    - d^kest une direction de descente $\Leftrightarrow f(x^{k+1}) < f(x^k)$
- Phase d'exploitation Recherche lineaire(line search):
  - Calcule du pas optimale α_k:
  - $\alpha_k = \underset{\alpha}{\operatorname{argmin}}\, \varphi(\alpha)$ telque $\varphi(\alpha) = f(x^k + \alpha d^k)$
  - Recherche lineaire approché
- Jusqu'à test d'arrêt : les conditions d'optimalités
- $\lim_{k \to \infty} x^k = x^*$
  - Vitesse de convergence
  - Ordre de convergence

Motivation : Classification de modèle

Chaque modèle (P) entre l'une des trois catégories suivantes:

- le problème est sans contraintes ($D_R = \mathbb{R}^n$).

  Ici, il n'y a aucune restriction pour le déplacement d'un point x^0 de départ vers une solution optimale. La difficulté de résolution dépend du degré de non linéarité de f

- Problème avec contraintes linéaires seulement :
  Dans ce cas, il est relativement facile de cerner les déplacements d'un point vers un autre à l'intérieur de D_R.

- Problème avec contraintes non linéaires :

  Généralement difficile d'assurer qu'on demeure dans le domaine réalisable.

Plan de cours

I. Connaissances de base
    1. Notations et définitions
    2. Résultat préliminaire
    3. Algorithme

II. Optimisation sans contraint
    1. Fonction à une seule variable
    2. Fonction à plusieurs variable

III. Ensembles et fonctions convexes

IV. Optimisation avec contraintes
    1. Conditions d'optimalité
    2. Méthodes primales
    3. Méthodes des pénalités et des barrières

Connaissances de base

- Notations et définitions
  - Notations
  - Ensemble ouvert / fermé
  - Minimum Local/Global
- Résultat préliminaire
- Algorithme

Notations et Définitions: ensemble ouvert

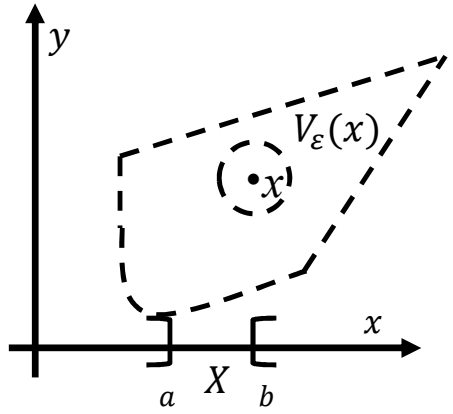
- $X \subset \mathbb{R}^n$ est un ensemble ouvert si $\forall x \in X \; \exists \varepsilon > 0$ tel que tout voisinage de x dénoté:

$$V_\varepsilon(x) = \{z \in \mathbb{R}^n : \|x - z\| < \varepsilon\} \subset X$$

- Dans $\mathbb{R}$ l'intervalle ouvert $]a, b[$, est un ensemble ouvert puisque $\forall\, x \in]a, b[, \exists\, \varepsilon > 0$ tel que

$$V_\varepsilon(x) = \{z \in \mathbb{R} : |x - z| < \varepsilon\} \subset]a, b[$$

- Dans $\mathbb{R}^2$ l'ensemble $\Gamma = \{(x, y) \in \mathbb{R}^2 : x \in]a, b[, y = 0\}$ n'est pas un ensemble ouvert puisque nous ne pouvons pas bouger pour modifier la valeur de y.



- Dans $\mathbb{R}^2$ l'ensemble X est ouvert puisque $\forall x \in X, \exists \varepsilon > 0$ tel que la sphère ouverte de rayon ε,

$$V_\varepsilon(x) = \{z \in \mathbb{R}^2 : \|x - z\| < \varepsilon\} \subset X$$

Notations et Définitions: ensemble fermé/compact

- Considérons une suite de points $x^k \in X, k = 1, \dots, \infty$ telle que $\{x^k\} \to x$.
  - le point x est un point d'accumulation (point limite) de X.
- X est un ensemble fermé si tout point d'accumulation (point limite) de x appartient à X.

- Dans $\mathbb{R}$ l'intervalle fermé $[a, b]$ est un ensemble fermé.
- Dans $\mathbb{R}^2$ l'ensemble $\Gamma = \{[x, y] \in \mathbb{R}^2 : x \in [a, b], y = 0\}$ est aussi un ensemble fermé.

- Un ensemble Γ est borné si $\forall x \in \Gamma$, $\mathrm{d}(\vec{0}, \mathrm{x}) = \|x\| < \infty$
- Un ensemble Ψ est compact si Ψ est un ensemble fermé et borné

Notations et Définitions: Norme euclidienne/ Gradient

Soit $Y \subset \mathbb{R}^n$ **un ensemble ouvert** et $f : Y \to \mathbb{R}$ une fonction à valeurs réelles .

Notes :

- La norme euclidienne : $\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_{j=1}^{n} {x_j}^2}$
- $x^T x = \sum_{j=1}^{n} x_j^2$ dénote le produit scalaire de x avec lui-même.
- $\nabla f(x)$ le gradient de f à $x \in R^n$: $\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^T$

Définition: (Hessien d'une matrice)

On appel:

$$\nabla^2 f(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]$$

le Hessien (matrice réelle $n \times n$) au point x d'une fonction f de classe C^2(ayant des dérivées partielles continues d'ordre 2).

Notations et Définitions: Hessien d'une matrice

Définition:

- **Forme quadratique**
  soit D une matrice réelle d'ordre n, **la forme quadratique** associée est la fonction Q: $\mathbb{R}^n \rightarrow \mathbb{R}$ définie par :

$$Q(x) = x^T D x$$

- **Semi-définie positive**
  D est une matrice **semi-définie positive** si : $\boldsymbol{Q(x) \geq 0 \ \forall x \in \mathbb{R}^n}$

- **Définie positive**
  D est une matrice **définie positive** si : $\boldsymbol{Q(x) > 0 \ \forall x \in \mathbb{R}^n \ , x \neq 0}$

Résultat : D est une matrice symétrique et **semi-définie positive** (définie positive) si et seulement si :

- toutes ses valeurs propres sont (≥ 0)non négatives (positives (> 0)) .
- tous les mineurs principaux sont non négatifs (≥ 0)
  (tous les leading mineurs principaux sont positifs (> 0)).

Notations et Définitions: Hessien d'une matrice

Exemple: $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$

Mineurs principaux (≥ 0) **:**

d'ordre 1:

$$a_{11}, a_{22}, a_{33}$$

d'ordre 2:

$$\left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right|, \left|\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}\right|, \left|\begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix}\right|$$

d'ordre 3:

$$\left|\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right|$$

Leading mineurs principaux (> 0) **:**

d'ordre 1 :

$$a_{11}$$

d'ordre 2 :

$$\left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right|$$

d'ordre 3 :

$$\left|\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right|$$

Notations et Définitions: Développement de Taylor

Soit $Y \subset \mathbb{R}^n$ **un ensemble ouvert** et $f : Y \to \mathbb{R}$ une fonction à valeurs réelles.

Si $f \in C^1/Y$ pour $\forall x \in Y$ et $\forall y \in \mathbb{R}^n$

i.e. (f possède des dérivées partielles de premier ordre continues sur Y)

- **Développement de Taylor de premier ordre (version résidu)**

$$\boldsymbol{f\,(y) = f(x) + \nabla f(x)^T (y - x) + \theta(x, (y - x))\|y - x\|}$$

Où le $\theta\big(x, (y - x)\big)$ dit le résidu avec $\lim_{(y-x)\to 0} \theta\big(x, (y - x)\big) = 0$

- **Développement de Taylor sans résidu :**

Il existe un vecteur z sur le segment de droite entre x et y

$$z \in \Im(x, y) = \{t \in \mathbb{R}^n : t = \alpha x + (1 - \alpha)y, \alpha \in [0,1]\}$$

Tel que : $\boldsymbol{f\,(y) = f(x) + \nabla f(z)^T (y - x)}$

Notations et Définitions: Développement de Taylor

Soit $Y \subset \mathbb{R}^n$ **un ensemble ouvert** et $f : Y \to \mathbb{R}$ une fonction à valeurs réelles. Si $f \in C^2/Y$ pour $\forall x \in Y$ et $\forall y \in \mathbb{R}^n$ i.e. (f possède des dérivées partielles de deuxième ordre continues sur Y)

- **Développement de Taylor de $2^{ième}$ ordre (version résidu)**

$$\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\nabla f}(\boldsymbol{x})^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}\ (\boldsymbol{y}-\boldsymbol{x})^T \nabla^2 f(x)(\boldsymbol{y}-\boldsymbol{x}) + \boldsymbol{\theta}(\boldsymbol{x},(\boldsymbol{y}-\boldsymbol{x}))\|\boldsymbol{y}-\boldsymbol{x}\|^{\boldsymbol{2}}$$

Où le $\theta\big(x,(y-x)\big)$ dit le résidu avec $\lim_{(y-x)\to 0}\ \theta\big(x,(y-x)\big) = 0$

- **Développement de Taylor sans résidu :**

Il existe un vecteur z sur le segment de droite entre x et y

$$z \in \Im(x,y) = \{t \in \mathbb{R}^n : t = \alpha x + (1-\alpha)y, \alpha \in [0,1]\}$$

Tel que : $\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\nabla f}(\boldsymbol{x})^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}\ (\boldsymbol{y}-\boldsymbol{x})^T \nabla^2 f(z)(\boldsymbol{y}-\boldsymbol{x})$

Notations et Définitions: Minimum Local/Global

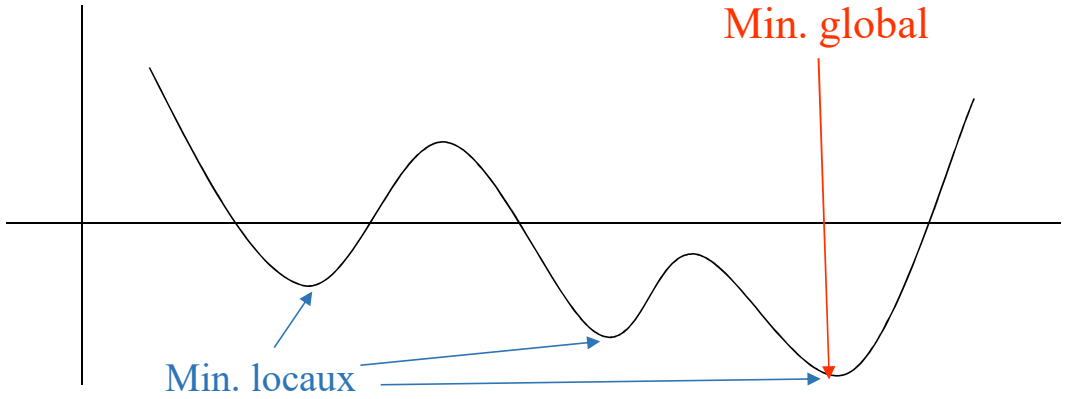
Définition:

- Un point $\bar{x} \in \mathbb{R}^n$ est un minimum local de f s'il existe un voisinage de $\bar{x}$ dénoté

$$V_\varepsilon(\bar{x}) = \left\{ x \in D_R \subset \mathbb{R}^n : \|x - \bar{x}\| = \sqrt{(x-\bar{x})^T (x-\bar{x})} < \varepsilon \right\}$$

  tel que $\boldsymbol{f(\bar{x}) \leq f(x)} \quad \boldsymbol{\forall x \in V_\varepsilon(\bar{x})}$

- Un point $\bar{x} \in \mathbb{R}^n$ est un minimum global de f si $\boldsymbol{f(\bar{x}) \leq f(x)} \quad \boldsymbol{\forall x \in D_R \subset \mathbb{R}^n}$



Résultats préliminaires

- Direction réalisable-descendant
- Condition nécessaire de premier ordre pour les minimums locaux
- Conditions nécessaires de $2^{ième}$ ordre pour les minimums locaux

Approche de résolution

- $\min_{x \in D_R} f(x)$
- $\{x^k\}$
- Phase d'exploration:
- Calcule de d^k telque :
  - **d^k est une direction realisable $\Leftrightarrow \exists \bar{\alpha}$ telque $\forall \alpha \in [0, \bar{\alpha}]$ $x^{k+1} = x^k + \alpha d^k \in D_R$**
  - **d^kest une direction de descente $\Leftrightarrow$ $f(x^{k+1}) < f(x^k)$**
- Phase d'exploitation Recherche lineaire(line search):
  - Calcule du pas optimale α_k:
  - $\alpha_k = \underset{\alpha}{\operatorname{argmin}}\, \varphi(\alpha)$ telque $\varphi(\alpha) = f(x^k + \alpha d^k)$
  - Recherche lineaire approché
  - $x^k \leftarrow x^k + \alpha d^k$
- Jusqu'à test d'arrêt : les conditions d'optimalités $\left\|\nabla f(x^k)\right\| > \delta$
- $\lim_{k \to \infty} x^k = x^*$
  - Vitesse de convergence
  - Ordre de convergence

Résultats préliminaires: Direction

Développement de Taylor de premier ordre (version résidu) :

Si $f \in C^1/Y$ pour tout $x \in Y$ et $d \in \mathbb{R}^n$

$$\begin{cases} f(x+\alpha d) = f(x) + \nabla f(x)^T(x+\alpha d - x) + \alpha(x, x+\alpha d - x)\|x+\alpha d - x\| \\ = f(x) + \alpha \nabla f(x)^T d + \theta(x, \alpha d)\|\alpha d\| \end{cases}$$

$\forall \alpha$, où $\theta(x, \alpha d)$, est une fonction de α prenant des valeurs réelles telles que $\boldsymbol{\lim_{\alpha \to 0} \theta(x, \alpha d) = 0}$

$$\Rightarrow \quad \frac{f(x+\alpha d) - f(x)}{\alpha} = \nabla f(x)^T d + \frac{\theta(x, \alpha d)\|\alpha d\|}{\alpha} = \nabla f(x)^T d \pm \theta(x, \alpha d)\|d\|$$

$$\Rightarrow \quad \lim_{\alpha \to 0} \frac{f(x+\alpha d) - f(x)}{\alpha} = \nabla f(x)^T d$$

Lemme 1 :

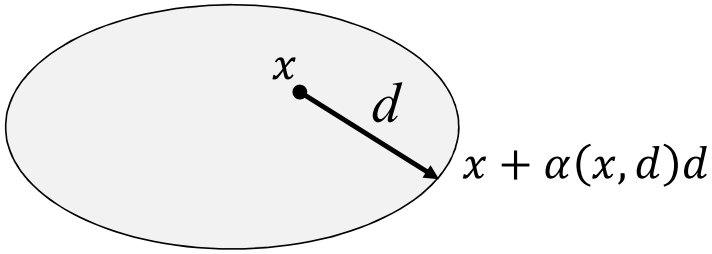
Soient $Y \subset \mathbb{R}^n$ un ensemble ouvert , $f : Y \to \mathbb{R}$, $f \in C^1 / Y$, $x \in Y$ et $d \in \mathbb{R}^n$. Alors :

$$\lim_{\alpha \to 0} \frac{f(x+\alpha d) - f(x)}{\alpha} = \nabla f(x)^T d$$

Résultats préliminaires: Direction

Définition : soient $X \subset \mathbb{R}^n$ et, $f : X \to \mathbb{R}$, étant donné $x \in X$,$d \in \mathbb{R}^n$ est une direction réalisable (admissible) à x s'il existe un scalaire $\alpha(x,d) > 0$ tel que :

$$(x + \alpha d) \in X \quad \forall \alpha \in [0, \alpha(x,d)]$$



Lemme 2 : Soient $X \subset \mathbb{R}^n$,$f \in C^1$ / X , et $x \in X$. Si $d \in \mathbb{R}^n$ est une direction réalisable à x et $\nabla f(x)^T d < 0$,

Puisque $\lim_{\alpha \to 0} \frac{f(x+\alpha d) - f(x)}{\alpha} = \nabla f(x)^T d < 0$ alors $\exists \xi > 0, \forall \alpha \in [-\xi, \xi] \setminus \{0\}$, $\frac{f(x+\alpha d) - f(x)}{\alpha} < 0$

Ainsi $\forall \alpha \in]0, \xi]$

$$f(x + \alpha d) - f(x) < 0 \implies f(x + \alpha d) < f(x)$$

$\implies d$ est une **direction de descente** à x

Résultats préliminaires: Condition nécessaire de premier ordre

Lemme 3 : soient $X \subset \mathbb{R}^n$ un ensemble ouvert, $f \in C^1 / X$.

- si x est minimum local de f sur X , alors $\nabla f(x) = 0$.

Preuve:

Par contradiction, supposons que x est un minimum local et que $\nabla f(x) \neq 0$.
et considérons la direction $d = -\nabla f(x)$.

- Puisque X est un ouvert et $\nabla f(x) \neq 0$, alors d est une direction réalisable à x, et
$$\nabla f(x)^T d = -\nabla f(x)^T \nabla f(x) < 0$$
- Par le lemme 2 , $d = -\nabla f(x)$ est une direction de descente à x et par conséquent il est possible de déterminer un scalaire
$\alpha > 0$ tel que $x + \alpha d \in V_\varepsilon(x)$ et $f(x + \alpha d) < f(x)$, une contradiction.

Lemme 4 : soient $X \subset \mathbb{R}^n$ un ensemble ouvert, $f \in C^1 / X$.

- Si x est maximum local de f sur X ,alors : $\nabla f(x) = 0$

Résultats préliminaires: Conditions nécessaires de $2^{ième}$ ordre

Lemme 5 : soient $X \subset \mathbb{R}^n$ un ensemble ouvert et $f \in C^2 \ / \ X$. Si $x \in X$ est un minimum local de f sur X, alors $\nabla f(x) = 0$ et $\nabla^2 f(x)$ est **une matrice semi-définie positive**.

Preuve:

- La première condition a été démontrée au lemme 3. Puisque $\nabla f(x) = 0$, alors
- La seconde condition est établie à partir du développement de Taylor de $2^{ième}$ ordre :

$$f\ (x + \alpha d) - f(x) = \frac{1}{2}\ \alpha^2 d^T \nabla^2 f(x) d + \bar{\theta}(x, \alpha d)\|\alpha d\|^2$$

$$= \frac{1}{2}\alpha^2 [\ d^T \nabla^2 f(x) d + 2\bar{\theta}(x, \alpha d)\|d\|^2]$$

Où $\lim_{\alpha \to 0}\ \bar{\theta}(x, \alpha d) = 0$

Par contradiction, si $\nabla^2 f(x)$ n'est pas semi-définie positive, alors $\exists d \in \mathbb{R}^n$ tel que $d^T \nabla^2 f(x) d < 0$.

Ainsi pour $\alpha > 0$ suffisamment petit, $x + \alpha d \in V_\varepsilon(x)$ et $\bar{\theta}(x, \alpha d) \to 0$

Par conséquent : $d^T \nabla^2 f(x) d + 2\bar{\theta}(x, \alpha d)\|d\|^2 < 0$ et alors $f\ (x + \alpha d) - f(x) < 0 \implies f\ (x + \alpha d) < f(x)$, contredisant le fait que x est un minimum local de f.

Conditions nécessaires de $2^{ième}$ ordre

Remarque : les conditions que $\nabla f(x) = 0$ et que $\nabla^2 f(x)$ est une matrice semi-définie positive **ne sont pas suffisantes pour assurer que x est un minimum local**.

Contre-exemple:

Soit $f(x,y) = x^3 + y^3 \implies \nabla f(x,y) = [3x^2, 3y^2]^T \implies \nabla^2 f(x,y) = \begin{pmatrix} 6x & 0 \\ 0 & 6x \end{pmatrix}$

Au point $(x,y) = (0,0)$ on a:

- $f(x,y) = 0$, $\nabla f(0,0) = [0,0]^T$ et $\nabla^2 f(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ semi –def positive.
- Donc les conditions sont satisfaites au point $(x,y) = (0,0)$, mais pour un $\varepsilon > 0$ suffisamment petit ,$\left[\frac{-\varepsilon}{2}, \frac{-\varepsilon}{2}\right] \in V_\varepsilon(0,0)$ et $f\left(\frac{-\varepsilon}{2}, \frac{-\varepsilon}{2}\right) = \left(\frac{-\varepsilon}{2}\right)^3 + \left(\frac{-\varepsilon}{2}\right)^3 = -\frac{2\varepsilon^3}{8} < 0 = f(0,0)$

  Et $[0,0]$ ne peut être un minimum local même si les conditions sont satisfaites.

Conditions nécessaires de $2^{ième}$ ordre

Lemme 6 : soient $X \subset \mathbb{R}^n$ un ensemble ouvert et $f \in C^2 \ / \ X$.

Si $\nabla f(x^*) = 0$ et $\nabla^2 f(x^*)$ est une matrice définie positive, alors il existe un $\varepsilon > 0$ suffisamment petit tel que $f\ (x^*) < f(x)$ pour tout $x \in V_\varepsilon(x^*)$.

Preuve :

Se référant au lemme 5.

$$f\ (x^* + \alpha d) = f(x^*) + \alpha \nabla f(x^*)^T d + \frac{1}{2}\ \alpha^{\ 2} d^T \nabla^2 f(x^*) d + \bar{\theta}(x^*, \alpha d)\|\alpha d\|^2$$

Où: $\lim_{\theta \to 0}\ \bar{\theta}(x^*, \alpha d) = 0$

Puisque $\nabla f(x^*) = 0$, alors :

$$f\ (x^* + \alpha d) - f(x^*) = \frac{1}{2}\alpha^2 [\ d^T \nabla^2 f(x^*) d + 2\bar{\theta}(x^*, \alpha d)\|d\|^2]$$

Conditions nécessaires de $2^{ième}$ ordre

Il est facile de vérifier que $d^T \nabla^2 f(x^*) d \geq \beta \|d\|^2$ où $\beta > 0$ est la plus petite valeur propre de $\nabla^2 f(x^*)$. En effet, sous l'hypothèse que $f \in C^2 \ / \ X$, alors $\nabla^2 f(x^*)$ est symétrique.

Par conséquent , il existe une matrice orthogonale P (i.e., $P^T = P^{-1}$) telle que :

$$\nabla^2 f(x^*) = P \Delta P^T$$

où Δ est une matrice diagonale comportant les valeurs propres de $\nabla^2 f(x^*)$ sur la diagonale et où les colonnes de P sont les vecteurs propres normalisés de $\nabla^2 f(x^*)$.

Donc :

$$d^T \nabla^2 f(x^*) d = d^T P \Delta P^T d = y^T \Delta y = \sum_{i=1}^{n} \lambda_i {y_i}^2$$

Où les λ_i sont les valeurs propres de $\nabla^2 f(x^*)$.

Ainsi si $\beta = \min_i \{\lambda_i\}$ alors :

$$d^T P \Delta P^T d = \sum_{i=1}^{n} \lambda_i {y_i}^2 \geq \beta \sum_{i=1}^{n} {y_i}^2 = \beta d^T P P^T d = \beta \|d\|^2$$

Conditions nécessaires de $2^{ième}$ ordre

Nous venons de vérifier que $d^T \nabla^2 f(x^*) d \geq \beta \|d\|^2$ où $\beta > 0$ est la plus petite valeur propre de $\nabla^2 f(x^*)$.

Donc

$$f(x^* + \alpha d) - f(x^*) \geq \frac{1}{2}\alpha^2 [\beta \|d\|^2 + 2\bar{\theta}(x^*, \alpha d)\|d\|^2]$$

$$\geq \frac{1}{2}\alpha^2 \|d\|^2 [\beta + 2\bar{\theta}(x^*, \alpha d)]$$

Puisque $\beta > 0$, il s'ensuit que pour $\alpha > 0$ suffisamment petit $(x^*, \alpha d) \in V_\varepsilon(x^*)$ et $\beta + 2\bar{\theta}(x^*, \alpha d) > 0$. Ainsi

$$f(x^* + \alpha d) - f(x^*) > 0 \; ou \; f(x^* + \alpha d) > f(x^*)$$

Corollaire :

soient $X \subset R^n$ un ensemble ouvert et $f \in C^2 / X$. Si $\nabla f(x^*) = 0$ et $\nabla^2 f(x^*)$ est une matrice définie positive, alors x^* est un minimum local de f sur X.

Approche de résolution: Algorithme

- $\min_{x \in D_R} f(x)$
- $\{x^k\}$
- Phase d'exploration:
- Calcule de d^k telque :
  - **d^k est une direction realisable $\Leftrightarrow \exists \bar{\alpha}$ telque $\forall \alpha \in [0, \bar{\alpha}]$**
    $x^{k+1} = x^k + \alpha d^k \in D_R$
  - **d^kest une direction de descente $\Leftrightarrow$ $f(x^{k+1}) < f(x^k)$**
- Phase d'exploitation Recherche lineaire(line search):
- Calcule du pas optimale α_k:
  - $\alpha_k = \underset{\alpha}{\operatorname{argmin}}\, \varphi(\alpha)$ telque $\varphi(\alpha) = f(x^k + \alpha d^k)$
  - Recherche lineaire approché
- $x^k \leftarrow x^k + \alpha d^k$
- Jusqu'à test d'arrêt : les conditions d'optimalités $\|\nabla f(x^k)\| > \delta$
- $\lim_{k \to \infty} x^k = x^*$
- **Vitesse de convergence**
- **Ordre de convergence**

Algorithme

- **Algorithmes itératives**
- **Convergence**

Algorithme itérative

Les algorithmes itératifs génèrent des séquences $\{x_k\}$

Définition: (Algorithme)

Un algorithme défini sur un sous-ensemble $X \subseteq \mathbb{R}^n$est un processus itératif qui, partant d'une solution initiale $x^0 \in X$, génère une suite de points $\{x^k\}$ dans X. selon l'itération $x^{k+1} = G_k(x^k)$ où $G_k : X \to$ X est une fonction qui peut dépendre de k.

Exemple:

l'itération d'algorithme de gradient $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$ vise à satisfaire les condition d'optimalité de fonction $f : \mathbb{R}^n \to \mathbb{R}$, ici α_k est un paramètre de pas positif qui est utilisé pour assurez-vous que l'itération avance vers l'ensemble de solutions du problème étudié.

- Si $x^{k+1} = x^k - \alpha_k \nabla f(x^k) \to G(x^k) = x^{k+1}$
- Si $x^{k+1} = x^k - \hat{\alpha} \nabla f(x^k)$ telque $\hat{\alpha} \in [a, b]$
- $\to A(x^k) = \{x^k - \hat{\alpha} \nabla f(x^k), \hat{\alpha} \in [a, b]\}$

Algorithme itérative

un algorithme correspond une multi application (multi fonction ou « mapping »)

$$A: X \to X$$

Associant à un point $x^k \in X$ un sous ensemble $A(x^k) \subset X$. Dénotons par $X^* \subset X$ l'ensemble des solutions recherchées.

- Si $x^{k+1} = x^k - \alpha_k \nabla f(x^k) \to A(x^k) = x^{k+1}$
- Si $x^{k+1} = x^k - \hat{\alpha} \nabla f(x^k)$ telque $\hat{\alpha} \in [a, b]$
- $\to A(x^k) = \{x^k - \hat{\alpha} \nabla f(x^k), \hat{\alpha} \in [a, b]\}$

Exemple : $X = \{ x \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0\}$ $\quad x^0 = (100,100).$

$$A(x) = \left\{ \left[0, \frac{x^1}{2}\right], \left[0, \frac{x^2}{2}\right] \right\} \subset X$$

$$X^* = (0,0)$$

Algorithme itérative

un algorithme correspond une multi application (multi fonction ou « mapping »)

$$A: X \to X$$

Associant à un point $x^k \in X$ un sous ensemble $A(x^k) \subset X$. Dénotons par $X^* \subset X$ l'ensemble des solutions recherchées.

- Si $x^{k+1} = x^k - \alpha_k \nabla f(x^k) \to A(x^k) = x^{k+1}$
- Si $x^{k+1} = x^k - \hat{\alpha} \nabla f(x^k)$ telque $\hat{\alpha} \in [a, b]$
- $\to A(x^k) = \{x^k - \hat{\alpha} \nabla f(x^k), \hat{\alpha} \in [a, b]\}$

Exemple : $X = \{ x \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0\}$ $\quad x^0 = (100,100).$

$$A(x) = \left\{ \left[0, \frac{x^1}{2}\right], \left[0, \frac{x^2}{2}\right] \right\} \subset X$$

$$X^* = (0,0)$$

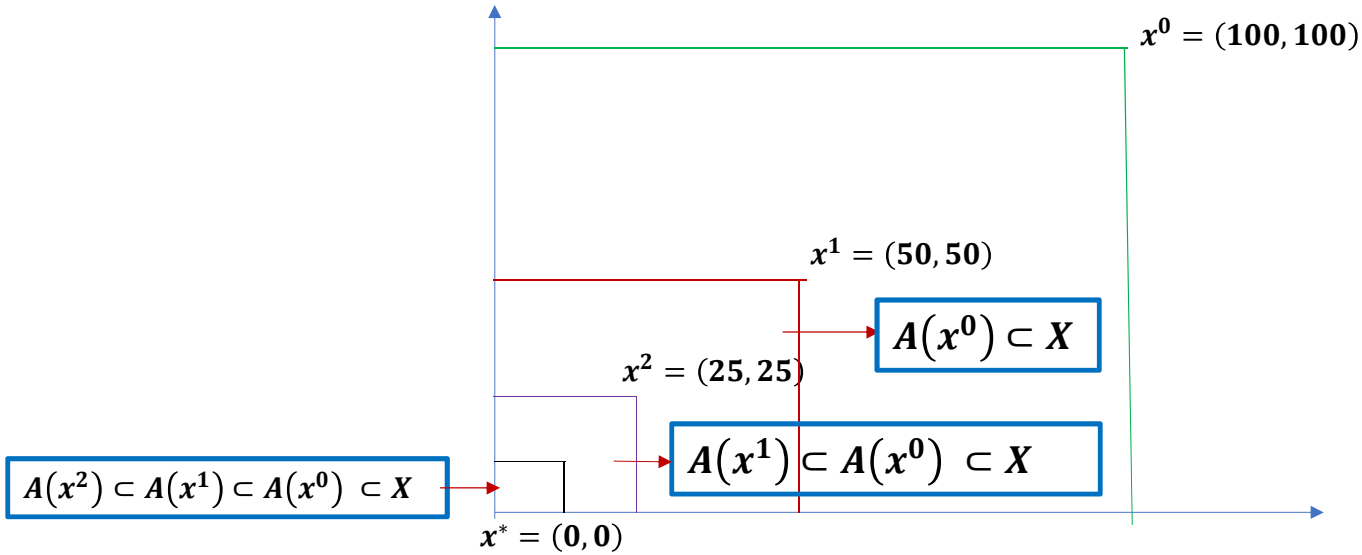
Algorithme itérative

$X = \{ x \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0\}$, $x^0 = (100,100)$

$\longrightarrow A(x^0) = [0;50] \times [0;50] \rightarrow x^1 = (50,50) \in A(x^0) \longrightarrow A(x^1) = [0;25] \times [0;25] \subset A(x^0)$

$\longrightarrow x^2 = (25,25) \in A(x^1) \subset A(x^0)$

$$\longrightarrow \cdots \longrightarrow A\left(x^{k+1}\right) = \left\{\left[0; \frac{x_1^{k+1}}{2}\right], \left[0; \frac{x_2^{k+1}}{2}\right]\right\} \subset A\left(x^k\right) = \left\{\left[0; \frac{x_1^k}{2}\right], \left[0; \frac{x_2^k}{2}\right]\right\}$$



Algorithme itérative

Définition 1 : (Algorithme de descente)
un algorithme A est un algorithme de descente par rapport à une fonction $z: X \to \mathbb{R}$ continue si :

$$x \notin X^* \text{et } y \in A(x) \Longrightarrow z(y) < z(x)$$

$$x \in X^* \text{et } y \in A(x) \Longrightarrow z(y) \leq z(x)$$

Note : souvent la fonction z que nous utilisons est la fonction économique du problème que nous voulons résoudre à l'aide de l'algorithme.

Exemple :

$$A(x) = \left\{\left[0, \frac{x^1}{2}\right], \left[0, \frac{x^2}{2}\right]\right\}, \qquad z(x) = x^1 + x^2$$

$$\min_{x \in X} z(x) = x^1 + x^2$$

Algorithme itérative: Algorithme de descente: $z: X \subseteq \mathbb{R}^n \to \mathbb{R}$

Bute : $X^* \subset X$ l'ensemble des solutions recherchées

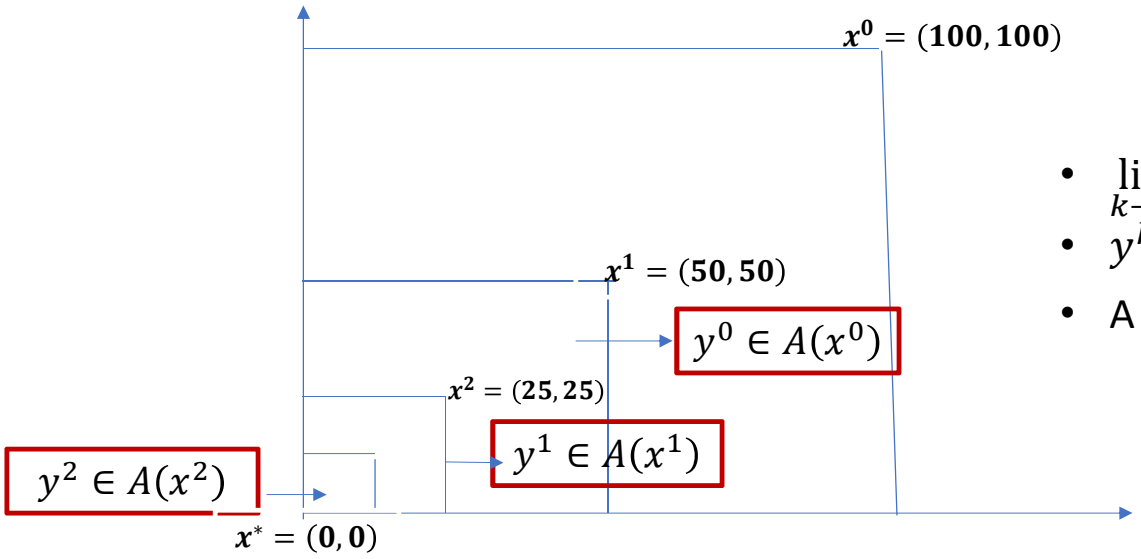
- $x^k \in X \subseteq \mathbb{R}^n, A(x^k)$
- $x^k \notin X^*, y \in A(x^k) \Longrightarrow z(y) < z(x^k)$
- $x^k \in X^*, y \in A(x^k) \Longrightarrow z(y) \leq z(x^k)$

- While $x^k \notin X^* or \left\|\nabla z(x^k)\right\| > \varepsilon$ do
  - $y \in A(x^k)$
  - $x^k = y$
  - $k \leftarrow k + 1$

- $x^k \notin X^* \Leftrightarrow d(x^{k+1}, x^k) > \varepsilon$
- Si $X^* = \{x^*\}$alors
- $x^k = x^* \to y \in A(x^*) = \{x^*\} \to y = x^*$
- $z(y) = z(x^k)$

Algorithme itérative: Algorithme de descente: $z: X \subseteq \mathbb{R}^n \to \mathbb{R}$

- $A(x^k) = \left\{\left[0; \frac{x_1^k}{2}\right], \left[0; \frac{x_2^k}{2}\right]\right\}$
- $X = \{x \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0\}, \quad x^0 = (100,100)$
- $A(x^0) = [0;50] \times [0;50] \to x^1 = (50,50) \in A(x^0), (25,25) \in A(x^0)$
- $A(x^1) = [0;25] \times [0;25] \to x^2 = (25,25) \in A(x^1)$



- $\lim_{k\to\infty} x^k = x = (0,0)$
- $y^k \in A(x^k)$ $\lim_{k\to\infty} y^k = y$
- A est fermé ssi $y \in A(0,0) = (0,0)$

Algorithme itérative

Définition 2 : (**algorithme fermé**)

Un algorithme est fermé au point $x \in X$ si :

- $\{x^k\} \in X$ a la propriété que $x^k \to x$ quand $k \to \infty$
- $\{y^k \in A(x^k)\}$ a la propriété que $y^k \to y$ quand $k \to \infty$

Alors $y \in A(x)$.

Note : la notion de fermeture pour les multi applications correspond à celle de la continuité pour les fonctions.

Exemple : $A(x) = \left\{\left[0, \frac{x_1}{2}\right], \left[0, \frac{x_2}{2}\right]\right\}$ $y^k \in A(x^k)$ y $= (1,1) \in A(2,2)$

$$x^k = \left(2 - \frac{1}{k}, 2 - \frac{1}{k}\right) \text{ avec } k = 1, \ldots \quad ; \ (1,1), \left(\frac{3}{2}, \frac{3}{2}\right), \left(\frac{5}{3}, \frac{5}{3}\right), \ldots, (2,2)$$

$$y^k = \left(1 - \frac{1}{2k}, 1 - \frac{1}{2k}\right) \text{ avec } k = 1, \ldots \quad ; \left(\frac{1}{2}, \frac{1}{2}\right), \left(\frac{3}{4}, \frac{3}{4}\right), \left(\frac{5}{6}, \frac{5}{6}\right), \ldots, (1,1)$$

Convergence d'algorithme (l'ordre de convergence)

Définition: Soit $\{x^k\}$ une suite de vecteurs dans $\mathbb{R}^n$ convergente vers x^*.
L'ordre de convergence de $\{x^k\}$ est le supremum des nombres non négatifs p satisfaisant la relation.

$$0 \leq \lim_{k\to\infty} \frac{\left\|x^{k+1} - x^*\right\|}{\|x^k - x^*\|^p} < \infty$$

Plus la valeur p est grande, plus la convergence est rapide à la limite puisque la distance à r^* décroît plus rapidement. En effet, si

$$\lim_{k\to\infty} \frac{\left\|x^{k+1} - x^*\right\|}{\|x^k - x^*\|^p} = \beta$$

Alors asymptotiquement

$$\left\|x^{k+1} - x^*\right\| = \beta\left\|x^k - x^*\right\|^p$$

Convergence d'algorithme (l'ordre de convergence)

Exemples:

Pour une valeur de a, telque $0 < a < 1$:

- $\{a^k\} \to 0$ avec un ordre de convergence au moins égal à 1 puisque:

$$\frac{\|a^{k+1} - 0\|}{\|a^k - 0\|} = \frac{a^{k+1}}{a^k} = a \text{ or } \frac{\|a^{k+1} - 0\|}{\|a^k - 0\|^2} = \frac{a^{k+1}}{a^{2k}} = \frac{1}{a^{k-1}} \to \infty$$

- $\left\{a^{2^k}\right\} \to 0$ avec un ordre de convergence au moins égal à 2, puisque

$$\frac{\left\|a^{2^{k+1}} - 0\right\|}{\left\|a^{2^k} - 0\right\|^2} = \frac{a^{2^{k+1}}}{\left(a^{2^k}\right)^2} = \frac{a^{2^{k+1}}}{a^{2\times 2^k}} = \frac{a^{2^{k+1}}}{a^{2^{k+1}}} = 1$$

Convergence d'algorithme (l'ordre de convergence)

Définition: Soit $\{x^k\}$ une suite de vecteurs dans $\mathbb{R}^n$ convergente vers x^* telle que:

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^p} = \beta < 1$$

- si $p = 1$ Alors la suite $\{x^k\}$ converge linéairement avec un rapport de convergence de β.

Notes:

1. convergence linéaire aussi dénotée convergence géométrique
2. Le cas $\beta = 0$ est dénoté par convergence super linéaire

- Si p =2 Alors la suite $\{x^k\}$ converge quadratique
- Si p =3 Alors la suite $\{x^k\}$ converge Cubique

Convergence d'algorithme (l'ordre de convergence)

- La convergence d'ordre 1 mais pas linéaire : $\left\{\frac{1}{k}\right\} \to 0$

$$\lim_{k\to\infty} \frac{\frac{1}{k+1}}{\frac{1}{k}} = \lim_{k\to\infty} \frac{k}{k+1} = \lim_{k\to\infty} \left(1 - \frac{1}{k+1}\right) = 1 - \lim_{k\to\infty} \frac{1}{k+1} = 1 - 0 = 1$$

- La convergence d'ordre 1 et super linéaire : $\left\{\left(\frac{1}{k}\right)^k\right\} \to 0$

$$\lim_{k\to\infty} \frac{\left(\frac{1}{k+1}\right)^{k+1}}{\left(\frac{1}{k}\right)^k} = \lim_{k\to\infty} \frac{\left(\frac{1}{k+1}\right)\left(\frac{1}{k+1}\right)^k}{\left(\frac{1}{k}\right)^k} = \lim_{k\to\infty} \left(\frac{\frac{1}{k+1}}{\frac{1}{k}}\right)^k \frac{1}{k+1} = \lim_{k\to\infty} 1\frac{1}{k+1} = 0$$

Théorème de Convergence Globale

Hypothèses: Supposons que

- $X \subset R^n$, un ensemble non vide fermé, et X^* l'ensemble non vide des solutions.
- A un algorithme défini sur X qui, partant d'un point $x^0 \in X$, génère une suite de points $\{x^k\}$ comme suit :
  - Si $x^k \in X^*$: l'algorithme s'arrête
  - Sinon : soit $x^{k+1} \in A(x^k) \in X$, Remplacer k par $(k+1)$
  - Répéter le processus;
- La suite de point $\{x^k\}$ est contenu dans un sous ensemble compact de X ;
- Il existe une fonction continue z par rapport à laquelle A est **un algorithme de descente**.

Sous ces conditions, si la multi application est fermée aux points n'appartenant pas à X, **alors** :

- soit que l'algorithme s'arrête en un nombre fini d'itérations à un point de X^*
- soit qu'il génère une suite infinie de points $\{x^k\}$ telle que le point limite de toute sous-suite convergente de $\{x^k\}$ appartient à X^*.