Cours 1 : Optimisation sans contraintes
Méthodes du gradient

Optimisation déterministe

EISTI

Ingénieurs 2ème année GI

Nisrine Fortin Camdavant

- Un problème d'optimisation est défini par

$$\begin{cases} \text{minimiser sur } \mathbb{R}^n & J(x) \\ \text{avec} & g_i(x) \leq 0; \quad 1 \leq i \leq p \\ & h_j(x) = 0; \quad 1 \leq j \leq q \end{cases}$$

J une fonction de $\mathbb{R}^n$ dans $\mathbb{R} \cup \{+\infty\}$.

- Vocabulaire :
  - J (à valeur dans $\mathbb{R} \cup \{+\infty\}$) est la fonction de coût, la fonction objectif ou encore le critère
  - les g_i sont les contraintes d'inégalité
  - les h_j sont les contraintes d'égalité
  - l'ensemble des contraintes est

$$\mathcal{C} = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0; \quad 1 \leq i \leq p \quad \text{et} \quad h_j(x) = 0; \quad 1 \leq j \leq q\}$$

ensemble des points admissibles ou réalisables.

Optima locaux et globaux

- $x^* \in \mathcal{C}$ réalise un minimum **local** de J sur $\mathcal{C}$ ssi $\exists$ une boule ouverte B centrée en x^* telle que

$$\forall x \in B \cap \mathcal{C}, \quad J(x) \geq J(x^*)$$

- $x^* \in C$ réalise un minimum **global** de J sur $\mathcal{C}$ ssi

$$\forall x \in \mathcal{C}, \quad J(x) \geq J(x^*)$$

Etudier le problème d'optimisation **sans contraintes** où on effectue la minimisation de la fonction $J : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ sur tout l'espace.

Formulation du problème

$$(\mathcal{P}) \quad \begin{cases} \min J(x) \\ \\ x \in \mathbb{R}^n \end{cases}$$

J une fonction de $\mathbb{R}^n$ dans $\mathbb{R} \cup \{+\infty\}$.

Objectif

Présenter quelques algorithmes permettant de calculer (*de manière approchée*) la ou les solutions du problème $(\mathcal{P})$ sans contraintes.

Principe des méthodes déterministes

- Elles conduisent, pour une solution initiale donnée, toujours au même résultat final.
- Elles s'appuient sur une direction de recherche qui peut être fournie par les dérivées de la fonction objectif.
- La plupart de ces algorithmes exploitent les conditions d'optimalité.
- La question de la détermination de minima globaux est difficile.
- Ces méthodes ont la réputation d'être efficaces lorsque la solution initiale est proche de l'optimum recherché.
- Invonvénients :
  - Le cas d'une fonction objectif possédant plusieurs optimums,
  - peuvent converger vers un optimum local
  - faire l'hypothèse de différentiabilité de la fonction J.

Algorithme

Définition

Les principaux **algorithmes** d'optimisation sont définis par une application $\mathcal{A}$ de $\mathbb{R}^n$ dans $\mathbb{R}^n$ permettant la génération d'une suite d'éléments de $\mathbb{R}^n$ par la formule

$$\begin{cases} x \in \mathbb{R}^n \text{ donné}, \quad k = 0 & \text{Etape d'initialisation} \\ x_{k+1} = \mathcal{A}(x_k), \quad k = k+1 & \text{Itération k} \end{cases}$$

Ecrire un algorithme n'est ni plus ni moins que se donner une suite $(x_k)_{k\in\mathbb{N}}$ de $\mathbb{R}^n$; étudier la convergence de l'algorithme, c'est étudier la convergence de la suite $(x_k)_{k\in\mathbb{N}}$.

Définition (Convergence d'un algorithme)

On dit que l'algorithme $\mathcal{A}$ converge si la suite $(x_k)_{k\in\mathbb{N}}$ engendrée par l'algorithme converge vers une limite x^*.

Vitesse de convergence

- La vitesse de convergence et la complexité sont des facteurs à prendre en compte lors de l'utilisation (ou de la génération) d'un algorithme ;
- But : la méthode soit la plus rapide possible tout en restant précise et stable.
- critère de mesure de la vitesse (ou taux) de convergence : l'évolution de l'erreur commise ($e_k = \|x_k - x^*\|$)

Taux de convergence d'un algorithme

Soit $(x_k)_{k\in\mathbb{N}}$ une suite de limite x^* définie par la donnée d'un algorithme convergent $\mathcal{A}$. On dit que la convergence de $\mathcal{A}$ est

- **linéaire** si l'erreur $(e_k = \|x_k - x^*\|)$ décroit linéairement

$$\exists C \in [0,1[,\ \exists k_0,\quad \forall k \geq k_0 \qquad e_{k+1} \leq C e_k$$

- **super-linéaire** si l'erreur e_k décroit de la manière suivante :

$$e_{k+1} \leq \alpha_k e_k,$$

où α_k est une suite positive convergente vers 0.

- **d'ordre p** si l'erreur e_k décroit de la manière suivante :

$$\exists C \geq 0,\ \exists k_0,\quad \forall k \geq k_0 \qquad e_{k+1} \leq C\,(e_k)^p$$

Si $p = 2$, la convergence de l'algorithme est dite **quadratique**.

Rappels

Définition

On dit que $J : H \to \mathbb{R}$ est **coercive** si

$$\lim_{\|x\| \to +\infty} J(x) = +\infty.$$

Ici $\|.\|$ désigne la norme de l'espace de Hilbert H.

Exemples

- J définie par $J(x) = \|x\|$ est coercive.
- Pour $n = 2$ et $x = (x_1, x_2)$, J définie par $J(x) = x_1^2 - x_2^2$ n'est pas coercive.
- Soit A une matrice carrée d'ordre n symétrique ($A^T = A$), définie positive ($\forall x \in \mathbb{R}^n,\ x^T Ax > 0$) et b un vecteur de $\mathbb{R}^n$, alors J définie par $J(x) = \frac{1}{2}(Ax, x) - (b, x)$ est coercive.

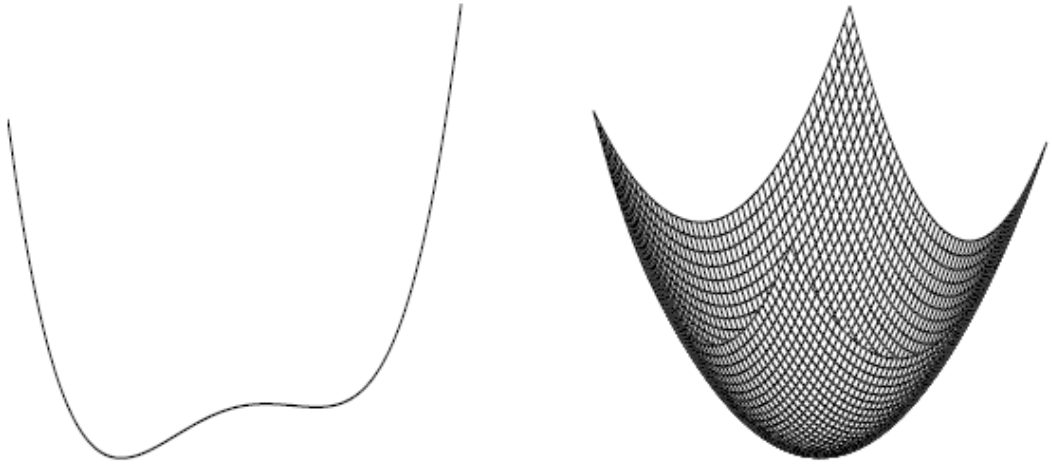

Figure 1 : exemples de fonctions coercives

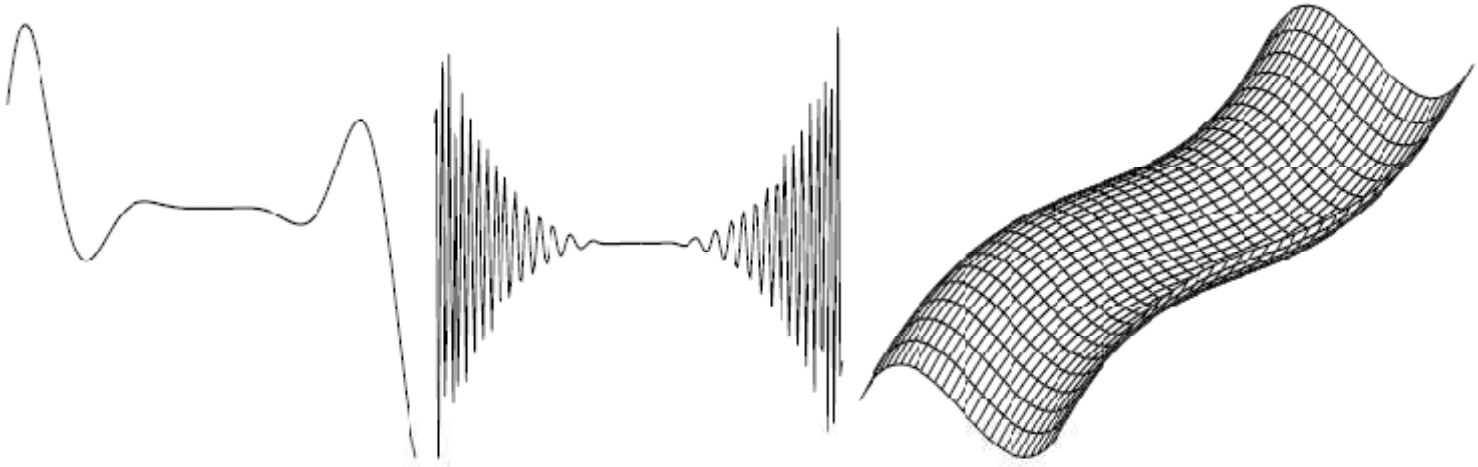

Figure 2 : exemples de fonctions non coercives

Rappels

Soient $J : \mathbb{R}^n \to \mathbb{R}$, $(x_1, x_2, \ldots, x_n) \mapsto J(x_1, x_2, \ldots, x_n)$ et $a \in \mathbb{R}^n$

- **Gradient d'une fonction :** Si J admet des dérivées partielles d'ordre 1 en a. On appelle **gradient** de J en a, le vecteur défini par

$$\overrightarrow{\nabla}^T J(a) = \left(\frac{\partial J}{\partial x_1}(a), \frac{\partial J}{\partial x_2}(a), \ldots, \frac{\partial J}{\partial x_n}(a) \right)$$

- **Hessien d'une fonction :** Si J admet des dérivées partielles d'ordre 2 en a. On appelle **hessien** de J en a la matrice $n \times n$ donnée par

$$\nabla^2 J(a) = \begin{pmatrix} \frac{\partial^2 J}{\partial x_1^2}(a) & \frac{\partial^2 J}{\partial x_1 \partial x_2}(a) & \ldots\ldots & \frac{\partial^2 J}{\partial x_1 \partial x_n}(a) \\ \frac{\partial^2 J}{\partial x_2 \partial x_1}(a) & \frac{\partial^2 J}{\partial x_2^2}(a) & \ldots\ldots & \frac{\partial^2 J}{\partial x_1 \partial x_2}(a) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 J}{\partial x_n \partial x_1}(a) & \frac{\partial^2 J}{\partial x_n \partial x_2}(a) & \ldots\ldots & \frac{\partial^2 J}{\partial x_n^2}(a) \end{pmatrix}$$

Rappels

- Un **sous-gradient** de J en x est un vecteur v tel que

$$\forall y,\ J(y) - J(x) \geq \langle v, y - x \rangle$$

- J est **M-lipschitzienne** si

$$\forall \mathbf{x},\ \mathbf{y}\quad \|J(\mathbf{x}) - J(\mathbf{y})\| \leq M\|\mathbf{x}\text{-}\mathbf{y}\|$$

- J est dite **fortement convexe** de rapport $a > 0$ (ou a-convexe) si $\forall t \in [0,1]$,

$$\forall x, y,\ \ J\left(tx + (1-t)y\right) \leq tJ(x) + (1-t)J(y) - \frac{a}{2}t(1-t)\|x-y\|^2$$

Résultats d'existence et d'unicité

Soit $J : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$

- **Existence :**

  Si J est **propre** (non identiquement infinie), **continue** et **coercive**. Alors $(\mathcal{P})$ admet *au moins une solution*.

- **Unicité**

  Si J est **strictement convexe**. Alors le problème $(\mathcal{P})$ admet *au plus une solution*.

- **Théorème :**

  Soit J une fonction $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$. On suppose qu'il existe $\alpha > 0$ tel que

  $$(\nabla J(x) - \nabla J(y), x - y) \geq \alpha \|x - y\|^2$$

  Alors J est strictement convexe et coercive ; en particulier le problème $(\mathcal{P})$ **admet une solution unique**.

Conditions nécessaires du premier ordre

Soit $J : \mathbb{R}^n \to \mathbb{R}$ une fonctionnelle **différentiable** sur $\mathbb{R}^n$.

- Si x^* réalise un minimum (global ou local) de J sur $\mathbb{R}^n$ alors

$$\nabla J(x^*) = 0.$$

La réciproque du théorème est fausse ; un point critique n'est pas nécessairement un minimum. **Exemple :** $J(x) = x^3$, $x^* = 0$ est un point critique de J mais pas un minimum.

- Si J est **convexe** sur $\mathbb{R}^n$. Un point x^* réalise un minimum global de J sur $\mathbb{R}^n$ ssi $\nabla J(x^*) = 0$.

Un point x^* de $\mathbb{R}^n$ vérifiant $\nabla J(x^*) = 0$ est appelé point critique ou point stationnaire.

Conditions d'optimalité de $2^{\text{ème}}$ ordre

- Si J est $\mathcal{C}^2$ et si x^* réalise un minimum de J alors :
  - $\nabla J(x^*) = 0$
  - $\langle \nabla^2 J(x^*)h, h\rangle \geq 0$ pour tout h.
- Si J est $\mathcal{C}^2$ et si
  - $\nabla J(x^*) = 0$
  - il existe $\alpha > 0$ tel que $\langle \nabla^2 J(x^*)h, h\rangle \geq \alpha \|h^2\|$ ($\nabla^2 J(x^*)$ est défini positif)

  alors x^* réalise un minimum local (strict) de J.

Méthodes de descente

- Principe général :
  1. point initial $\mathbf{x}_0$
  2. pour $k \geq 1$ croissant
     - 2.1 choisir une **direction de descente** $\mathbf{d}_k \neq 0$
     - 2.2 choisir un **pas de descente** $\rho_k > 0$
     - 2.3 poser $\mathbf{x}_{k+1} = \mathbf{x}_k + \rho_k \mathbf{d}_k$
     - 2.4 tester la convergence
- il faut qu'on puisse descendre :
  - on doit pouvoir trouver ρ_k tel que $J(\mathbf{x}_{k+1}) < J(\mathbf{x}_k)$
  - si J est convexe, $J(\mathbf{x}_{k+1}) \geq J(\mathbf{x}_k) + \rho_k \langle v, \mathbf{d}_k \rangle$ pour tout sous-gradient v : on doit donc avoir $\langle v, \mathbf{d}_k \rangle < 0$ pour au moins un sous-gradient v

Direction de descente

Définition

Soit $J : \mathbb{R}^n \to \mathbb{R}$, une fonction continûment différentiable, et x un vecteur de $\mathbb{R}^n$. Le vecteur $\mathbf{d} \in \mathbb{R}^n$ est appelé **direction de descente** de J en x ssi

$$\langle \nabla J(x), \mathbf{d} \rangle < 0.$$

Par définition de la dérivée, si $\mathbf{d}$ est une direction de descente alors pour tout $\alpha > 0$ suffisamment petit, on a :

$$J(x + \alpha \mathbf{d}) < J(x)$$

Pas de descente

Objectif : diminuer la fonction J en effectuant un déplacement dans la direction $\mathbf{d}$ en construisant une suite d'itérés $(x_k)_{k\geq 1}$ approchant la solution x^* du problème $\mathcal{P}$ de la façon suivante :

$$x_{k+1} = x_k + \rho_k \mathbf{d}_k.$$

Le paramètre ρ_k est **le pas** à effectuer le long de la direction de descente d_k au point courant x_k. Un algorithme à directions de descente est donc déterminé par les paramètres d et ρ : la façon dont la direction de descente d est calculée donne son nom à l'algorithme et la façon dont le pas ρ est déterminé est appelée recherche linéaire et peut se déterminer de différentes façon à préciser.

Choix de pas de descente

La recherche linéaire consiste à déterminer le pas ρ_k à effectuer le long d'une direction de descente d_k. Des valeurs de ce pas peuvent être obtenues par différentes méthodes :

- recherche linéaire exacte : déterminer le pas optimal, c'est à dire le pas qui *minimise* la fonction J le long de la direction de descente d_k. Le pas ρ_k est donc solution du problème :

$$\rho_k = \arg\min_{\rho \geq 0} J(x_k + \rho_k d_k)$$

- méthode d'Armijo : Choisir un pas qui diminue suffisamment la valeur de la fonction, c'est une méthode plus simple qui consiste à délimiter un intervalle convenable pour ρ et choisir le plus grand ρ_k possible sur cet intervalle. Conditions d'Armijo-Goldstein

$$J(x_k + \rho_k d_k) \leq J(x_k) + \rho_k \beta_1 \langle \nabla J(x_k), d_k \rangle, \qquad \beta_1 \in]0, 1[$$

La méthode (ou algorithme) du **Gradient** fait partie de la classe des **méthodes de descente**.

La direction de descente $\mathbf{d}$ vérifie $\langle \nabla J(x), \mathbf{d} \rangle < 0$, un choix classique pour cette méthode est donné par

$$\mathbf{d}_k = -\nabla J(\mathbf{x}_k)$$

Méthodes du Gradient

Algorithme du Gradient

1. Initialisation

   $k = 0$: choix de x_0 et de $\rho_0 > 0$

2. Itération k

   $x_{k+1} = x_k - \rho_k \nabla J(x_k)$;

3. Critère d'arrêt

   Si $\|x_{k+1} - x_k\| < \epsilon$, STOP

   Sinon, on pose $k = k + 1$ et on retourne à 2.

Dans tout ce qui suit, ϵ est un réel positif (petit) donné qui représente la précision désirée.

- La convergence n'est pas toujours assurée donc une règle de base est de fixer un nombre maximum d'itérations k_{max}.
- Méthodes conceptuellement très simples et peuvent être programmées directement, mais elles sont souvent lentes dans la pratique.
- Elles convergent mais sous des conditions de convergence souvent complexes.

Convergence

- L'algorithme du gradient converge sous des conditions assez fortes :
  - J est C^1, coercive et strictement convexe
  - ∇J est M-Lipschitzienne
  - si $\rho_k \in [\beta_1, \beta_2]$ avec $0 < \beta_1 < \beta_2 < \frac{2}{M}$
- On peut obtenir une preuve plus simple en supposant que J est α-convexe, $\mathcal{C}^1$ et de gradient M-Lipschitzien :
  - on suppose alors que $\rho_k \in \left]0, \frac{2\alpha}{M^2}\right[$
  - on obtient une vitesse de convergence linéaire

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \sqrt{1 - 2\alpha\rho_k + M^2\rho_k^2}\ \|\mathbf{x}_k - \mathbf{x}^*\|,$$

optimale de valeur $\sqrt{1 - \frac{\alpha^2}{M^2}}$ (pour $\rho_k = \frac{\alpha}{M^2}$)

Choix du pas

On utilise le plus souvent la méthode du gradient à pas constant ($\rho_k \equiv \rho$ constant). Toutefois, on peut faire varier le pas à chaque itération : on obtient alors la méthode du gradient à pas variable.

- pas variable :
  - adaptation du pas au problème
  - pas optimal :
    - $\rho_k = \arg\min_{\rho>0} J(\mathbf{x}_k - \rho\nabla J(\mathbf{x}_k))$
    - avantage : meilleure réduction possible par itération
    - inconvénient : coût de la recherche
  - pas approximativement optimal :
    - rechercher un ρ_k qui réduit assez J
    - comme $J(\mathbf{x}_k - \rho\nabla J(\mathbf{x}_k)) = J(\mathbf{x}_k) - \rho\|\nabla J(\mathbf{x}_k)\|^2 + o(\rho\|\nabla J(\mathbf{x}_k)\|)$, on peut demander une réduction d'au moins $\alpha\rho|\nabla J(\mathbf{x}_k)\|^2$ (avec $\alpha < \frac{1}{2}$)

Le gradient à pas fixe

- La méthode du gradient à pas constant ($\rho_k \equiv \rho$ constant).
- Si ρ est trop grand, l'algorithme diverge ; si ρ est trop petit la convergence est très lente.
- Lorsqu'on dispose d'informations supplémentaires sur la fonctionnelle J (constante de Lipchitz L de la dérivée, constante de forte convexité a,...) on peut choisir $\rho = \frac{a}{L^2}$.

Le gradient à pas optimal pour une fonction quadratique

$$J(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$$

où A est **définie positive**.

- $d_k = -\nabla J(x_k) = -(Ax_k + b)$
- On calcule le pas ρ_k qui réalise le minimum de $f(\rho) = J(x_k + \rho d_k)$.

$$\frac{\partial f}{\partial \rho} = 0 \Longrightarrow \rho = \frac{\langle d_k, d_k\rangle}{\langle Ad_k, d_k\rangle}$$

Le gradient à pas optimal pour une fonction quadratique

Algorithme du Gradient à pas optimal pour une fonction quadratique

1. Initialisation
   $k = 0$: choix de x_0 et de $\epsilon > 0$
2. Itération k
   $d_k = -Ax_k - b$;
3. Critère d'arrêt
   Si $\|d_k\| < \epsilon$, STOP
   Sinon, $\rho_k = \frac{\|d_k\|^2}{Ad_k, d_k}$
   on pose $k = k + 1$ et on retourne à 2.

ϵ est un réel positif (petit) donné qui représente la précision désirée.