Première partie

Optimisation en dimension finie

Chapitre 1

Généralités

L'optimisation (c'est-à-dire les techniques permettant de chercher les minima ou les maxima de fonctions ou de fonctionnelles) intervient dans pratiquement tous les processus de modélisation actuels. Qu'il s'agisse de problèmes directs (ajustement de données, contrôle optimal, résolution de systèmes linéaires par moindres carrés, etc ...) ou inverses (identification de paramètres, contrôle de frontières libres etc..), il est rare qu'un problème d'optimisation plus ou moins complexe n'intervienne pas à un stade donné de la modélisation et/ou de la simulation. Avant de donner les définitions et principes de base de la théorie de l'optimisation, nous allons présenter quelques exemples simples permettant d'introduire et d'illustrer par anticipation notre propos.

1.1 Quelques exemples

1.1.1 Détermination de coefficients en combustion

On considère un mélange de gaz et on appelle T le taux d'introduction de chaleur dans ce mélange. On appelle α la variable (par exemple la proportion d'un des gaz dans le mélange). La loi donnant T en fonction de α est de la forme

$$T(\alpha) = f[1 - e^{-a_1 e^{\alpha b_1}}] + (1-f)[1 - e^{-a_2 e^{\alpha b_2}}] \tag{1.1.1}$$

où les paramètres à déterminer sont $(f, a_1, b_1, a_2, b_2) \in \mathbb{R}^5$. Comme indiqué précédemment, on fait n mesures de T pour différentes valeurs de α de sorte que $T(\alpha_i) \simeq T_i,\ i = 1, \cdots, n$. On obtient alors la formulation au sens des moindres carrés du problème :

$$\min \sum_{i=1}^{n} [f(1 - e^{-a_1 e^{\alpha_i b_1}}) + (1-f)(1 - e^{-a_2 e^{\alpha_i b_2}}) - T_i]^2\ ,\ (f, a_1, b_1, a_2, b_2) \in \mathbb{R}^5\ . \tag{1.1.2}$$

1.1.2 Un exemple en hydrologie

En hydrologie dans des problèmes de corrélation hydropluviométrique, le débit d'un bassin $\mathcal{Y}$ est une variable aléatoire dépendant linéairement de n variables aléatoires $\mathcal{X}_1$,...,$\mathcal{X}_n$ (pluviosité

moyenne sur le bassin, différents indices de répartition de précipitation dans le temps etc...), selon la relation :

$$\mathcal{Y} = b_0 + b_1\mathcal{X}_1 + b_2\mathcal{X}_2 + \cdots + b_n\mathcal{X}_n \ . \tag{1.1.3}$$

On doit en général, chercher les coefficients $b_0,\ b_1, \cdots,\ b_n$ (dits de régression). Pour cela on effectue p observations (ou mesures) portant sur les variables $\mathcal{X}_i$ et $\mathcal{Y}$; on note $Y = [y_1, y_2, ..., y_p]$ le vecteur des valeurs observées de $\mathcal{Y}$ et $X = [x_{ij}],\ i = 1, \cdots, n,\ j = 1, \cdots, p$, la matrice observée : x_{ij} est j-ième observation de la variable $\mathcal{X}_i$. Le problème d'identification des paramètres b_i se formule alors de la manière suivante

$$\min \sum_{j=1}^{p} \left[y_j - \left(b_0 + \sum_{i=1}^{n} x_{ij}\, b_i \right) \right]^2 \ ,\ (b_0,\ b_1, \cdots,\ b_n) \in \mathbb{R}^{n+1} \ . \tag{1.1.4}$$

Toutefois, si on résout le problème de minimisation ci-dessus on risque de trouver des débits négatifs ce qui n'a pas de sens ! On impose donc une condition (ou contrainte) supplémentaire sur les débits qui doivent être positifs. Le problème se modélise alors comme suit

$$\begin{cases} \min \sum_{j=1}^{p} \left[y_j - \left(b_0 + \sum_{i=1}^{n} x_{ij}\, b_i \right) \right]^2 \\ (b_0,\ b_1, \cdots,\ b_n) \in \mathbb{R}^{n+1} \ , \\ b_0 + \sum_{i=1}^{n} b_i x_{ij} \geq 0 \ . \end{cases} \tag{1.1.5}$$

1.1.3 Un exemple en chimie : problème de l'équilibre chimique

On considère un mélange de m éléments chimiques. On a prédéterminé que les m atomes différents peuvent se combiner pour produire n composés. Soit x_j le nombre de moles du composé j (i.e. Nx_j est le nombre de molécules du composé j, où N est le nombre d'Avogadro), a_{ij} est le nombre d'atomes d'un élément i dans une molécule de composé j et Nb_i le nombre d'atomes de l'élément i dans le mélange. On veut identifier x_j c'est-à-dire la composition exacte du mélange. L'équation de bilan des masses donne un premier type de contraintes

$$\sum_{j=1}^{n} a_{ij}x_j = b_i,\ i = 1, \cdots, m$$

auxquelles s'ajoutent des contraintes naturelles

$$x_j \geq 0\ ,\ j = 1, \cdots, n\ .$$

D'autre part, le second principe de la thermodynamique nous apprend qu'un mélange de composés chimiques à température et à pression constantes atteint son état d'équilibre lorsque l'énergie libre du système est minimale (Principe de GIBBS). Cette énergie libre est donnée par

$$f(x_1, \cdots, x_n) = \sum_{j=1}^{n} x_j \left[c_j + \ln \left(\frac{x_j}{\sum_{j=1}^{n} x_j} \right) \right] ,$$

avec $c_j = \dfrac{F_j^0}{R_T} + \ln P$, F_j^0 désignant l'énergie libre de GIBBS par mole du composé j à la température T et à la pression d'une atmosphère, P est la pression totale et R la constante des gaz parfaits.
Le problème revient donc à minimiser f sous contraintes

$$\left\{ \begin{array}{l} \min \displaystyle\sum_{j=1}^{n} x_j \left[c_j + \ln \left(\frac{x_j}{\sum_{j=1}^{n} x_j} \right) \right] \\ \displaystyle\sum_{j=1}^{n} a_{ij} x_j = b_i,\ i = 1, \cdots, m\ , \\ x_j \geq 0\ ,\ j = 1, \cdots, n\ . . \end{array} \right. \tag{1.1.6}$$

Remarquons que ce problème n'est pas un problème formulé au sens des moindres carrés comme dans les exemples précédents. Toutefois, c'est encore un problème de minimisation d'une fonction de plusieurs variables.

1.2 Formulation mathématique

Les exemples précédents peuvent tous s'écrire sous la forme générale suivante

$$(\mathcal{P}) \qquad \left\{ \begin{array}{l} \min\ J(x) \\ g(x) \leq 0\ , \\ h(x) = 0\ , \\ x \in \mathbb{R}^n \end{array} \right.$$

où

- $J : \mathbb{R}^n \to \mathbb{R}$ est une fonction de plusieurs variables ($x = (x_1, \cdots, x_n)$) **à valeurs réelles**. Cette fonction (que l'on minimise) est appelée indifféremment fonction **coût**, **objectif** ou **critère**.
- $g : \mathbb{R}^n \to \mathbb{R}^p$ est une fonction de plusieurs variables $x \in \mathbb{R}^n$ à valeurs dans $\mathbb{R}^p$: elle a p composantes et on peut écrire

  $$g(x) = (g_1(x), \cdots, g_p(x))\ ,$$

  chaque fonction g_i étant définie sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$. La fonction g représente les **contraintes en inégalité**. La notation $g(x) \leq 0$ signifie qu'on considère les inégalités composante par composante : $g(x) \leq 0 \overset{\text{def}}{\Longleftrightarrow} \forall i = 1, \ldots, p \qquad g_i(x) \leq 0$.
- $h : \mathbb{R}^n \to \mathbb{R}^q$ est une fonction de plusieurs variables $x \in \mathbb{R}^n$ à valeurs dans $\mathbb{R}^q$: elle a q composantes et on peut écrire

  $$h(x) = (h_1(x), \cdots, h_q(x))\ ,$$

  chaque fonction h_i étant définie sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$. La fonction h représente les **contraintes en égalité** .

Remarque 1.2.1 *Plus généralement, on peut remplacer l'espace de dimension finie $\mathbb{R}^n$ par un espace vectoriel topologique sur $\mathbb{R}$ (de dimension a priori infinie). Nous nous bornerons à étendre (quand ce n'est pas trop compliqué) le cadre fonctionnel aux espaces de Hilbert réels.*

Précisons maintenant que qu'on entend par minimisation (ou maximisation) d'une fonction. Soit $\mathcal{C}$ l'ensemble des contraintes, c'est-à-dire par exemple dans le cas précédent

$$\mathcal{C} = \{\ x \in \mathbb{R}^n \mid g(x) \leq 0\ ,\ h(x) = 0\ \}\ .$$

On suppose que $\mathcal{C}$ est non vide ; un élément x de $\mathcal{C}$ sera dit **réalisable**.

Définition 1.2.1 (Minimum (maximum) local)
*Soient $\mathcal{C}$ un ensemble non vide d'un espace de Hilbert réel $\mathbb{H}$ et f une fonction de $\mathcal{C}$ dans $\mathbb{R}$. On dit que $x^* \in \mathcal{C}$ réalise un **minimum local** de f sur $\mathcal{C}$ si on peut trouver une boule de $\mathbb{H}$ centrée en x^* : $\mathcal{B}(x^*)$ telle que*

$$\forall x \in \mathcal{B}(x^*) \cap \mathcal{C} \quad f(x^*) \leq f(x)\ .$$

*On dit que $x^* \in \mathcal{C}$ réalise un **maximum local** de f sur $\mathcal{C}$ si on peut trouver une boule de $\mathbb{H}$ centrée en x^* : $\mathcal{B}(x^*)$ telle que*

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

On rappelle qu'une boule de $\mathbb{H}$ centrée en x^* de rayon $\rho > 0$ est l'ensemble

$$\mathcal{B}(x^*, \rho) = \{\ x \in \mathbb{H} \mid \|x - x^*\| \leq \rho\ \},$$

où $\|\cdot\|$ désigne la norme de $\mathbb{H}$.

Définition 1.2.2 (Minimum (maximum) global)
*On dit que $x^* \in \mathcal{C}$ réalise un minimum **global** de f sur $\mathcal{C}$ si $\forall x \in \mathcal{C} \quad f(x^*) \leq f(x)$.*
*On dit que $x^* \in \mathcal{C}$ réalise un maximum **global** de f sur $\mathcal{C}$ si $\forall x \in \mathcal{C} \quad f(x^*) \geq f(x)$.*

Nous donnons ci-dessous une illustration des différents cas de figure.

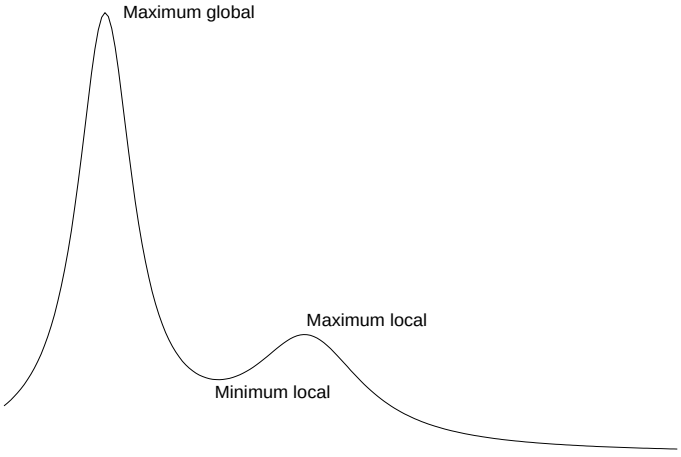

Figure 1.1a : Exemples de minima et de maxima locaux et globaux

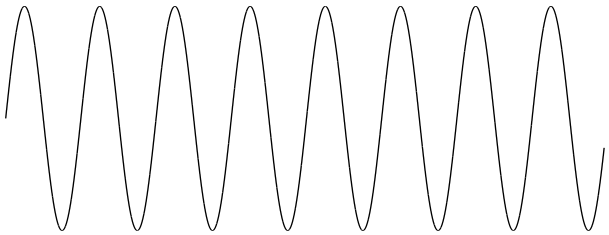

Figure 1.1b : Infinité de maxima et minima globaux

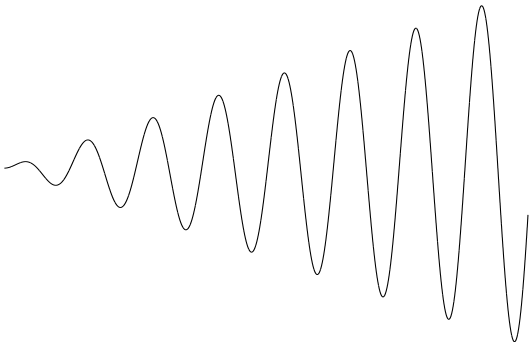

Figure 1.1c : Pas de maximum global - Infinité de maxima et minima locaux

Les minima et maxima sont dits **stricts** si les inégalités dans les définitions précédentes sont strictes. On s'intéressera essentiellement à la recherche des points réalisant des minima car la recherche des maxima peut se ramener à celle des minima comme le montre la proposition suivante :

Proposition 1.2.1 *Si x^* réalise un maximum (local ou global) de f sur $\mathcal{C}$, x^* réalise un minimum (local ou global) de $-f$ sur $\mathcal{C}$. Plus précisément*

$$\max \{ f(x) , x \in \mathcal{C} \} = -\min \{ -f(x) , x \in \mathcal{C} \} .$$

Démonstration - Donnons la preuve pour un maximum global : c'est exactement la même pour un maximum local.

Soit x^* tel que $f(x^*) = \max\{ f(x) , x \in \mathcal{C} \}$. On a donc

$$\begin{aligned} &\forall x \in \mathcal{C} \quad f(x) \leq f(x^*) \\ \Leftrightarrow &\forall x \in \mathcal{C} \quad -f(x) \geq -f(x^*) \\ \Leftrightarrow &-f(x^*) = \min \{ -f(x) , x \in \mathcal{C} \} \\ \Leftrightarrow &f(x^*) = -\min \{ -f(x) , x \in \mathcal{C} \} . \end{aligned}$$

□

Remarque 1.2.2 *Par abus de langage, on dit souvent que x^* est un minimum pour la fonction J ou de la fonction J : il faudrait dire que x^* réalise un minimum pour J ou que $J(x^*)$ est une valeur minimale de J.*

On ne peut évidemment résoudre le problème général sans quelques hypothèses sur J, g et h qui permettront au moins d'assurer l'existence de solutions. Nous préciserons ces hypothèses dans le prochain chapitre, mais nous aurons besoin des définitions suivantes.

1.3 Notion de convexité

1.3.1 Définitions

Le cas où les données sont convexes est un cas très important car les problèmes quadratiques sont à la base de nombreux algorithmes non linéaires. Nous rappelons quelques définitions et propriétés. Dans tout ce qui suit $\mathbb{H}$ désigne un espace de Hilbert réel. On désigne par $\|\cdot\|$ sa norme et $(\cdot,\cdot)$ son produit scalaire.

Définition 1.3.1 (Ensemble convexe)
*On dit que l'ensemble $\mathcal{C} \subset \mathbb{H}$ est **convexe** si*

$$\forall (x,y) \in \mathcal{C}\times\mathcal{C},\ \forall t \in [0,1] \quad tx+(1-t)y \in \mathcal{C}\ .$$

Autrement dit, $\mathcal{C}$ est convexe s'il contient tout "segment" reliant deux quelconques de ses points.

Exemple 1.3.1 (Ensembles convexes)

1. *Un intervalle $[a,b]$ est convexe dans $\mathbb{R}$.*
2. *Une réunion disjointe d'intervalles de $\mathbb{R}$ n'est pas convexe. ($\mathbb{R}^*$ par exemple).*

Définition 1.3.2 (Fonction convexe)
*On dit que la fonction $J : \mathcal{C} \subset \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$ est **convexe** si $\mathcal{C}$ est convexe et si*

$$\forall (x,y) \in \mathcal{C}\times\mathcal{C},\ \forall t \in [0,1] \quad J(tx+(1-t)y) \leq tJ(x)+(1-t)J(y)\ .$$

Définition 1.3.3 (Domaine d'une fonction convexe)
*Soit $J : \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$ une fonction convexe. On appelle **domaine** de J l'ensemble*

$$dom\ J = \{\ x \in \mathbb{H}\ \mid\ J(x) < +\infty\ \}\ .$$

Cet ensemble est convexe.
*Lorsque le domaine de J est non vide J est dite **propre**.*

Définition 1.3.4 (Fonction strictement convexe)
*On dit que la fonction $J : \mathcal{C} \to \mathbb{R} \cup \{+\infty\}$ est **strictement convexe** si $\mathcal{C}$ est convexe et si*

$$\forall (x,y) \in \mathcal{C}\times\mathcal{C}\ avec\ x \neq y,\ \forall t \in]0,1[\qquad J(tx+(1-t)y) < tJ(x)+(1-t)J(y)\ .$$

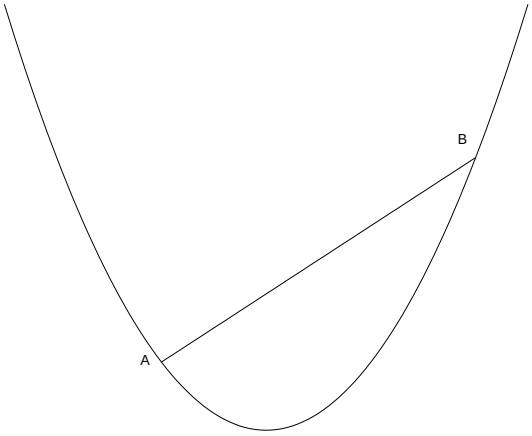

Figure 1.2 : Exemple de fonction convexe
(la corde AB est au-dessus de l'arc AB)

Exemples 1.3.1 ***Fonctions convexes et strictement convexes***

- $J(x) = \|x\|^2$ est strictement convexe.
- Toute application affine, c'est-à-dire de la forme

$$J(x) = (b, x) + \beta \ ,$$

où b et x sont des éléments de $\mathbb{H}$ et $\beta \in \mathbb{R}$ est convexe mais pas strictement.

- Soit A une matrice carrée symétrique d'ordre n semi-définie positive et b un vecteur de $\mathbb{R}^n$. Alors J définie par

$$J(x) = \frac{1}{2} \left(Ax, x\right)_n - \left(b, x\right)_n \ ,$$

est convexe. Si de plus A est définie positive, J est strictement convexe.
$(\ ,\ )_n$ désigne le produit scalaire de $\mathbb{R}^n$.
Plus généralement si $\mathcal{A}$ est un opérateur linéaire de $\mathbb{H}$ (espace de Hilbert) dans $\mathbb{H}$, auto-adjoint et **monotone** c'est-à-dire

$$\forall (x, y) \in \mathbb{H} \times \mathbb{H} \qquad (\mathcal{A}(x) - \mathcal{A}(y), x - y) \geq 0 \ ,$$

et $b \in \mathbb{H}$, alors J définie par

$$J(x) = \frac{1}{2} \left(\mathcal{A}x, x\right) - \left(b, x\right) \ ,$$

est convexe.

Les figures suivantes donnent quelques exemples de fonctions non convexes.

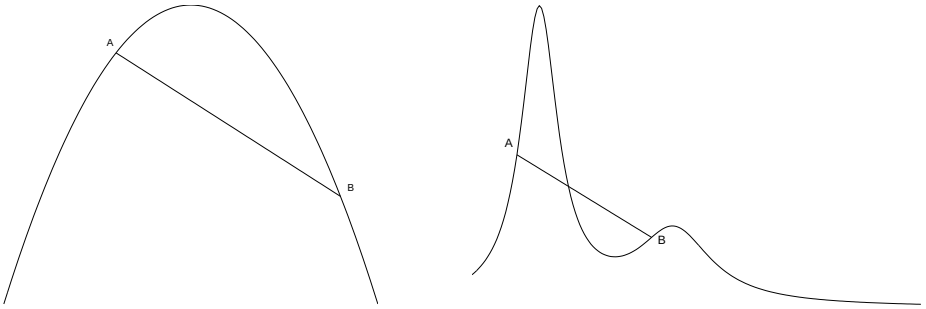

Figure 1.3 : Exemple de fonctions non convexes
(la corde AB n'est pas au-dessus de l'arc)

La fonction f apparaissant à gauche dans la figure 1.3 est telle que son opposée $-f$ est convexe : une telle fonction est dite **concave**. Le graphe de droite montre qu'une fonction peut n'être ni convexe, ni concave.

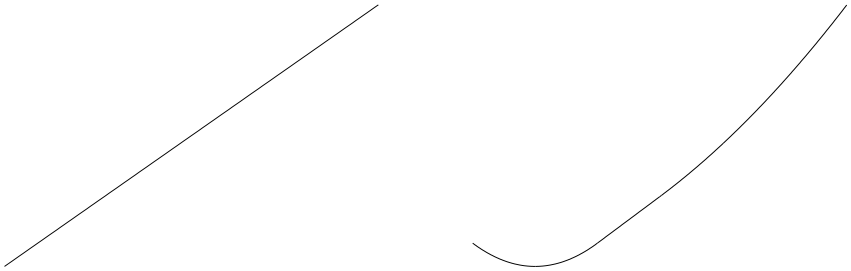

Figure 1.4 : Exemples de fonctions convexes, non strictement convexes

Définition 1.3.5 (Programmation convexe)
*On dit que le problème $(\mathcal{P})$ est un problème de **programmation convexe** quand les fonctions $J, g_i, i = 1, \cdots, p$ sont convexes et les fonctions $h_j, j = 1, \cdots, q$ sont affines.*

Définition 1.3.6 (Programmation linéaire)
*On dit que le problème $(\mathcal{P})$ est un problème de **programmation linéaire** quand les fonctions $J, g_i, i = 1, \cdots, p,\ h_j, j = 1, \cdots, q$ sont affines.*

Le cas de la programmation linéaire est certes un cas particulier de la programmation convexe mais il se présente plutôt comme un cas singulier pour lequel on n'est pas toujours sûr de trouver des solutions. De ce fait, les méthodes employées pour la résolution de ces problèmes sont des méthodes très spécifiques et non pas des cas particuliers des méthodes de programmation non linéaire que nous allons présenter. La résolution des problèmes de programmation linéaire relève de la Recherche Opérationnelle dont nous ne parlerons pas ici. On pourra par exemple consulter [10] à ce sujet.

1.3.2 Continuité des fonctions convexes

Donnons maintenant quelques propriétés (topologiques) importantes des fonctions convexes. Dans tout ce qui suit J est une fonction de $\mathbb{H}$ vers $\mathbb{R} \cup \{+\infty\}$. On suppose que le domaine de J est non vide.

Théorème 1.3.1 (Continuité)
Soit $J : \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$ convexe. Il est équivalent de dire
i) Il existe un ouvert non vide Ω sur lequel J est majorée par une constante a réelle et ne vaut pas constamment $-\infty$
ii) J est propre, dom J est d'intérieur non vide et J est continue sur l'intérieur de son domaine.

Démonstration - Il est clair que (ii) entraîne (i).
Pour démontrer le résultat réciproque nous aurons besoin du lemme suivant

Lemme 1.3.1 *Si, au voisinage d'un point $u_o \in \mathbb{H}$, une fonction convexe J est majorée par une constante finie, alors J est continue en u_o.*

Nous démontrerons ce lemme ensuite.
Supposons donc (i) vérifié : Ω est donc inclus dans l'intérieur du domaine de J qui est en particulier non vide. Donc J est propre. Soit $u \in \Omega$ tel que $J(u) > -\infty$. D'après le lemme (1.3.1), J sera continue en u, donc finie sur un voisinage de u. Pour tout $v \in$ int(dom$J$), il existe $\rho > 1$ tel que $w = u + \rho(v - u)$ appartienne encore à int(dom J) car l'intérieur d'un convexe est convexe (ce que nous admettrons). L'homothétie h de centre w et de rapport $1 - \dfrac{1}{\rho}$ transforme u en v et Ω en un ouvert $h(\Omega)$ contenant v. Pout tout v' de $h(\Omega)$, on a par convexité

$$J(v') \leq \frac{\rho - 1}{\rho} J(h^{-1}(v')) + \frac{1}{\rho} J(w) \leq \frac{\rho - 1}{\rho} a + \frac{1}{\rho} J(w) \, .$$

Par conséquent : tout point v de int(dom$J$) possède un voisinage $h(\Omega)$ sur lequel J est majorée par une constante finie. D'après le lemme (1.3.1), J est continue en v. □

Démonstration du lemme (1.3.1)
On se ramène par translation au cas où $u_o = 0$ et $J(0) = 0$. Soit $\mathcal{V}$ un voisinage de l'origine tel que $J(u) \leq a < +\infty$, pour tout u de $\mathcal{V}$. Posons $\mathcal{W} = \mathcal{V} \cap -\mathcal{V}$ et donnons nous $\varepsilon \in]0, 1[$. Si $u \in \varepsilon\mathcal{W}$, on par convexité

$$\frac{u}{\varepsilon} \in \mathcal{W}, \text{ donc } J(u) \leq (1 - \varepsilon) J(0) + \varepsilon J(\frac{u}{\varepsilon}) \leq \varepsilon \, a \, ,$$

$$-\frac{u}{\varepsilon} \in \mathcal{W}, \text{ donc } J(u) \geq (1 + \varepsilon) J(0) - \varepsilon J(-\frac{u}{\varepsilon}) \geq \varepsilon \, a \, .$$

Finalement

$$\forall u \in \varepsilon\mathcal{W} \qquad |J(u)| \leq \varepsilon a \, ,$$

d'où la continuité. □

Corollaire 1.3.1 *Toute fonction convexe propre sur un espace de dimension finie ($\mathbb{H} = \mathbb{R}^n$) est continue sur l'intérieur de son domaine.*

Démonstration - Si l'intérieur du domaine de J est non vide, il contient $n+1$ points affinement indépendants u_i, $1 \leq i \leq n+1$. D'après l'inégalité de convexité, J est majorée par $\max_{1\leq i\leq n+1} J(u_i)$ sur l'ouvert

$$\left\{ \sum_{i=1}^{n+1} \lambda_i u_i \ \mid \ \sum_{i=1}^{n+1} \lambda_i = 1 \ \text{ et } \lambda_i > 0 \ \forall i \right\} .$$

□

Pour plus de résultats sur les fonctions convexes on peut se référer à [11].

1.3.3 Différentiabilité des fonctions convexes

Donnons maintenant quelques propriétés de différentiabilité.

Définition 1.3.7 *Soit J une fonction de $\mathbb{H}$ dans $\mathbb{R} \cup \{+\infty\}$. On dit que J est **Gâteaux-différentiable** en $u \in$ dom (J) si la dérivée directionnelle*

$$J'(u;v) = \lim_{t\to 0^+} \frac{J(u+tv) - J(u)}{t} ,$$

existe dans toute direction v de $\mathbb{H}$ et si l'application

$$v \mapsto J'(u;v)$$

est linéaire continue.
D'après le théorème de représentation de Riesz (voir [5] par exemple) on identifie $\mathbb{H}$ et son dual ; on note alors

$$J'(u;v) = (\nabla J(u), v) ,$$

*où $(\ ,\ )$ désigne le produit scalaire de $\mathbb{H}$. L'élément $\nabla J(u)$ de $\mathbb{H}$ est le **gradient** de J en u.*

Il est clair que si J est différentiable au sens classique en u (on dit alors **Fréchet** - différentiable), alors J est Gâteaux-différentiable en u, et la dérivée classique et la dérivée au sens de Gâteaux coïncident.

La réciproque est fausse comme le montre le contre-exemple suivant : soit f de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par :

$$f(x,y) = \begin{cases} y & \text{si } x = y^2 , \\ 0 & \text{sinon} \end{cases}$$

La fonction f est continue en (0,0) et Gâteaux-différentiable en (0,0) mais pas Fréchet - différentiable en (0,0).

Théorème 1.3.2 *Soit $J : \mathcal{C} \subset \mathbb{H} \to \mathbb{R}$, Gâteaux différentiable sur $\mathcal{C}$, avec $\mathcal{C}$ convexe. J est convexe si et seulement si*

$$\forall (u,v) \in \mathcal{C} \times \mathcal{C} \qquad J(v) \geq J(u) + (\nabla J(u), v-u) \tag{1.3.1}$$

Démonstration - Supposons J convexe. Soient u et v dans $\mathcal{C}$. Par convexité de J on a

$$\forall t \in]0,1[\;\; J(u+t(v-u)) - J(u) \leq t(J(v) - J(u)) \; .$$

En divisant par $t > 0$ et en passant à la limite lorsque $t \to 0^+$ on obtient (1.3.1).
Réciproquement : on applique (1.3.1) à $u + t(v-u)$ ($t \in [0,1]$) et u, puis à $u + t(v-u)$ et v pour obtenir

$$J(u) \geq J(u+t(v-u)) - t\,(\nabla J(u+t(v-u)), v-u) \quad \text{et}$$

$$J(v) \geq J(u+t(v-u)) + (1-t)\,(\nabla J(u+t(v-u)), v-u) \; .$$

En faisant la combinaison convexe de ces deux inégalités on obtient

$$(1-t)J(u) + tJ(v) \geq (1-t+t)J(u+t(v-u)) \; ,$$

et la convexité de J. □

Théorème 1.3.3 *Soit $J : \mathcal{C} \subset \mathbb{H} \to \mathbb{R}$, Gâteaux différentiable sur $\mathcal{C}$, avec $\mathcal{C}$ convexe. J est convexe si et seulement si ∇J est un opérateur* ***monotone*** *, c'est-à-dire*

$$\forall (u,v) \in \mathcal{C} \times \mathcal{C} \qquad (\nabla J(u) - \nabla J(v), u - v) \geq 0 \; . \tag{1.3.2}$$

Démonstration - Soient (u,v) dans $\mathcal{C} \times \mathcal{C}$. D'après le théorème précédent, si J est convexe, alors

$$J(v) \geq J(u) + (\nabla J(u), v-u)$$

et

$$J(u) \geq J(v) + (\nabla J(v), u-v) \; .$$

En sommant on obtient (1.3.2).
Réciproquement : soient (u,v) dans $\mathcal{C} \times \mathcal{C}$, $u \neq v$; on définit φ de $[0,1]$ dans $\mathbb{R}$ de la manière suivante :

$$\varphi : t \mapsto \varphi(t) = (1-t)J(u) + tJ(v) - J(u+t(v-u)) \; .$$

Il est facile de voir que φ est dérivable et que

$$\forall t_1, t_2 \in [0,1] \;\; (\varphi'(t_1) - \varphi'(t_2))(t_1 - t_2) \leq 0 \; ,$$

grâce à (1.3.2). Donc φ' est décroissante sur $[0,1]$. De plus $\varphi(0) = \varphi(1) = 0$ et d'après le théorème de Rolle, il existe $a \in]0,1[$ tel que $\varphi'(a) = 0$. On a donc le tableau suivant qui montre que $\varphi \geq 0$ sur $[0,1]$. La convexité de J s'en déduit.

t	0	a	1
φ'	↘	0	↘
φ'	+	0	-
φ	0↗		↘0

□

Remarque 1.3.1 *Supposons que ∇J soit un opérateur **strictement** monotone :*

$$\forall (u,v) \in \mathcal{C} \times \mathcal{C},\ u \neq v, \qquad (\nabla J(u) - \nabla J(v), u - v) > 0\ . \tag{1.3.3}$$

*alors J est **strictement** convexe. En effet, on peut reprendre la deuxième partie de la démonstration du théorème précédent : φ' est **strictement** décroissante sur $[0,1]$ et donc ne s'annule qu'en un point a au plus. φ est alors **strictement** croissante sur $]\,0,a[$ et **strictement** décroissante sur $]\,a,1[$ donc strictement positive sur $]\,0,1[$.*

On définit de manière analogue la (Gâteaux) dérivée seconde de J en u, comme étant la dérivée de la fonction (vectorielle) $u \to \nabla J(u)$. On la note $D^2J(u)$ et on l'appellera aussi Hessien par abus de langage ; ce Hessien est identifiable à une matrice carrée $n \times n$ lorsque $\mathbb{H} = \mathbb{R}^n$.
Nous allons maintenant étudier le problème de minimisation $(\mathcal{P})$ c'est-à-dire

- donner des résultats d'existence de solutions (locales ou globales)
- étudier l'unicité éventuelle des solutions
- donner une caractérisation de ces solutions en terme de **conditions nécessaires et/ou suffisantes d'optimalité**
- donner des algorithmes de calcul de ces solutions.

[Exercices]

[Convexité]

1. Montrer qu'une norme est convexe.

   ---

2. Montrer que la fonction indicatrice d'un ensemble K définie par

   $$1_K = \begin{cases} 0 & \text{if } x \in K \\ +\infty & \text{sinon}\ , \end{cases}$$

   est convexe si et seulement si K est convexe.

   ---

3. Soit U une partie convexe d'un espace vectoriel V. Montrer que $f : U \subset V \to \mathbb{R}$ est convexe si et seulement si l'ensemble suivant :

   $$\text{épi}(f) = \{(v,\alpha) \in V \times \mathbb{R} \mid v \in U,\ \alpha \geq f(v)\}$$

   est une partie convexe de $V \times \mathbb{R}$.

   ---

4. Soit $(f_i)_{i \in I}$ une famille quelconque de fonctions convexes de $U \subset V \to \mathbb{R}$.
   Démontrer que la fonction $\sup_{i \in I} f_i$ est convexe .

   ---

5. Montrer l'inégalité de **Young** : $\forall a,b > 0\ ,\ \forall p,q \in \mathbb{N}$ tels que $\dfrac{1}{p} + \dfrac{1}{q} = 1\ ,\ \ ab \leq \dfrac{a^p}{p} + \dfrac{b^q}{q}$.

   ---

6. Soit f une fonction convexe de $\mathbb{R}^n$ dans $\mathbb{R}$. Montrer que :

$$\forall(\lambda_i)_{1\leq i\leq p} \in (\mathbb{R}^+)^p \text{ t.q. } \sum_{i=1}^{p}\lambda_i = 1, \forall(x_i)_{1\leq i\leq p} \in (\mathbb{R}^n)^p, \quad f(\sum_{i=1}^{p}\lambda_i x_i) \leq \sum_{i=1}^{p}\lambda_i f(x_i)$$

---

7. Donner une condition suffisante sur les fonctions g_i pour que l'ensemble $C = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\ , i = 1, \cdots, p\}$ soit convexe et fermé, puis donner une condition suffisante sur les fonctions $g_i,\ i = 1, \ldots, p$ et $h_j,\ j = 1, \ldots, q$ pour que l'ensemble

$$\mathcal{C} = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\ , i = 1, \cdots, p,\ h_j(x) = 0,\ j = 1, \ldots, q\ \}$$

soit convexe fermé.

---

8. Soit F une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$. Pour u et v fixés dans $\mathbb{R}^n$ on définit la fonction de $\mathbb{R}^{*+}$ vers $\mathbb{R}$ suivante :

$$\forall\lambda > 0 \qquad \Phi(\lambda) = \frac{F(u+\lambda v) - F(u)}{\lambda}\ .$$

Montrer que si F est convexe alors Φ est croissante.

---

9. **Fonctions conjuguées.-** Soit f une fonction quelconque de $\mathbb{R}^n$ dans $\bar{\mathbb{R}} = \mathbb{R} \cup \{\ \pm\infty\ \}$. On définit la fonction f* de $\mathbb{R}^n$ dans $\bar{\mathbb{R}}$ de la manière suivante :

$$\forall y \in \mathbb{R}^n \qquad f^*(y) = \sup_{x\in\mathbb{R}^n} \langle x, y\rangle - f(x)\ ,$$

où $\langle x, y\rangle$ désigne le produit scalaire usuel de $\mathbb{R}^n$. On note dom(f) l'ensemble :$\{x \in \mathbb{R}^n \mid f(x) \in \mathbb{R}\ \}$.

   (a) Montrer que : $\forall y \in \mathbb{R}^n \qquad f^*(y) = \sup_{x\in\text{dom}(f)} \langle x, y\rangle - f(x)$.

   (b) Montrer que f^* est convexe.

   (c) Montrer que :
   - $f^*(0) = -\inf_{x\in\mathbb{R}^n} f(x)$.
   - $\forall\lambda > 0\ ,\ \forall y \in \mathbb{R}^n \qquad (\lambda f)^*(y) = \lambda f^*(\frac{y}{\lambda})$.
   - $\forall\alpha \in \mathbb{R} \qquad (f+\alpha)^* = f^* - \alpha$.

   (d) Calculer explicitement f^* dans les cas suivants :
   - $\forall x \in \mathbb{R}^n \qquad f(x) = \langle b, x\rangle$, où b est un élément fixé de $\mathbb{R}^n$.
   - $\forall x \in \mathbb{R}^n \qquad f(x) = k$, où k est un réel.
   - $\forall x \in \mathbb{R}^n \quad f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$, où b est un élément fixé de $\mathbb{R}^n$ et A une matrice carrée, définie positive d'ordre n.

[Différentiabilité]

10. Soit f une fonction de $\mathbb{R}$ dans $\mathbb{R}$ dérivable sur l'intervalle]0,1]. On suppose que f' n'est pas bornée sur]0,1]. Montrer que f n'est pas lipschitzienne sur [0,1].

---

11. Les fonctions de $\mathbb{R}^n$ dans $\mathbb{R}$ suivantes sont-elles différentiables ?
Si oui , quelle est la différentielle ? Si non , sont-elles Gâteaux-différentiables ?
– $f_p(x) = \|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$, pour $p = 1, 2$.
– $f(x) = (Ax, x) = x^t Ax$, où (,) désigne le produit scalaire de $\mathbb{R}^n$.

---

12. Soit a une forme bilinéaire symétrique de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}$.
    (a) Montrer que l'on peut trouver une matrice symétrique A d'ordre n telle que :
    $$\forall u, v \in \mathbb{R}^n \qquad a(u, v) = (Au, v) \ .$$
    (b) Calculer le gradient et la dérivée seconde (hessien) de la fonctionnelle J définie sur $\mathbb{R}^n$ par : $J(v) = \frac{1}{2} (Av, v) - (b, v)$, où $b \in \mathbb{R}^n$ est fixé.
    (c) A quelle condition sur A, la fonction J est-elle convexe ? strictement convexe ?

---

13. Montrer que si f est (Fréchet) différentiable de différentielle Df alors :
    $$\forall v \in V \qquad Df(y).v = f'(y).v = (\nabla f(y), v)$$
    Réciproquement, montrer que si f est Gâteaux-différentiable de Gâteaux-différentielle continue, alors f est Fréchet-différentiable.

---

14. Soit $\mathcal{Y} = \mathcal{C}[a, b]$. On définit la fonctionnelle J de la manière suivante :
    $$\forall y \in \mathcal{Y} \quad J(y) = \int_a^b [\sin^3 x + y(x)^2] \, dx$$
    Pour y dans $\mathcal{Y}$ calculer la Gâteaux-différentielle $J'(y$ de J en y.

Chapitre 2

Minimisation sans contraintes

Dans ce chapitre nous allons étudier les problèmes d'optimisation évoqués dans le chapitre précédent dans le cas où $\mathbb{H} = \mathbb{R}^n$ muni du produit scalaire usuel et lorsqu'il n'y a pas de contraintes : on effectue la minimisation de la fonction J sur tout l'espace. Nous considérons donc le problème formulé de la façon suivante

$$(\mathcal{P}) \qquad \left\{ \begin{array}{l} \min\ J(x) \\ x \in \mathbb{R}^n \end{array} \right.$$

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

2.1 Résultats d'existence et d'unicité

Avant d'étudier les propriétés de la solution (ou des solutions) de $(\mathcal{P})$ il faut s'assurer de leur existence. Nous donnerons ensuite des résultats d'unicité.

Définition 2.1.1 *On dit que $J : \mathbb{H} \to \mathbb{R}$ est **coercive** si*

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

Ici $\|\cdot\|$ désigne la norme de l'espace de Hilbert $\mathbb{H}$. Dans le cas où $\mathbb{H} = \mathbb{R}^n$ les normes sont toutes équivalentes et $\|\cdot\|$ désigne une norme quelconque de $\mathbb{R}^n$. On notera $\|\cdot\|_p$ ($p \in \mathbb{N}$) la norme ℓ_p de $\mathbb{R}^n$:

$$\forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n \qquad \|x\|_p = \left[\sum_{i=1}^{n} |x_i|^p \right]^{\frac{1}{p}} \ .$$

La norme infinie de $\mathbb{R}^n$ est

$$\forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n \qquad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i| \ .$$

Rappelons que parmi les normes ci-dessus, seule la norme ℓ_2 munit $\mathbb{R}^n$ d'une structure d'espace de Hilbert. Nous noterons $(\cdot, \cdot)$ le produit scalaire associé.

Exemples 2.1.1

- $J(x) = \|x\|$ est coercive.
- Pour $n = 2$ et $x = (x_1, x_2)$, J définie par $J(x) = x_1^2 + x_2^2 - a\,x_1 - b\,x_2 - c$, est coercive pour tous réels a et b.
  J définie par $J(x) = x_1^2 - x_2^2$ n'est pas coercive. En effet la suite $x_n = (0, n)$ est telle que $\|x_n\| \to +\infty$ (on peut prendre par exemple $\|x_n\| = \|x_n\|_1 = n$) et pourtant $J(x_n) = -n^2$ ne converge pas vers $+\infty$.
  De la même manière, (toujours dans le cas n=2), la fonction J définie par $J(x) = x_1^2$ n'est pas coercive. On peut encore choisir pour le montrer la suite $(0, n)$ car $J(0, n) = 0$.
- Soit A une matrice carrée d'ordre n symétrique, définie positive et b un vecteur de $\mathbb{R}^n$. Alors J définie par
$$J(x) = \frac{1}{2}\,(Ax, x) - (b, x)\ ,$$
est coercive.
- Une fonction affine J définie par $J(x) = (\alpha, x) + \beta$, $\alpha \in \mathbb{R}^n,\ \beta \in \mathbb{R}$, n'est **jamais** coercive.

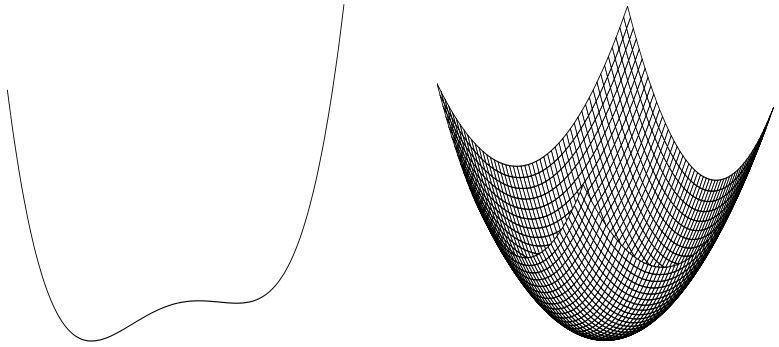

Figure 2.1 : exemples de fonctions coercives

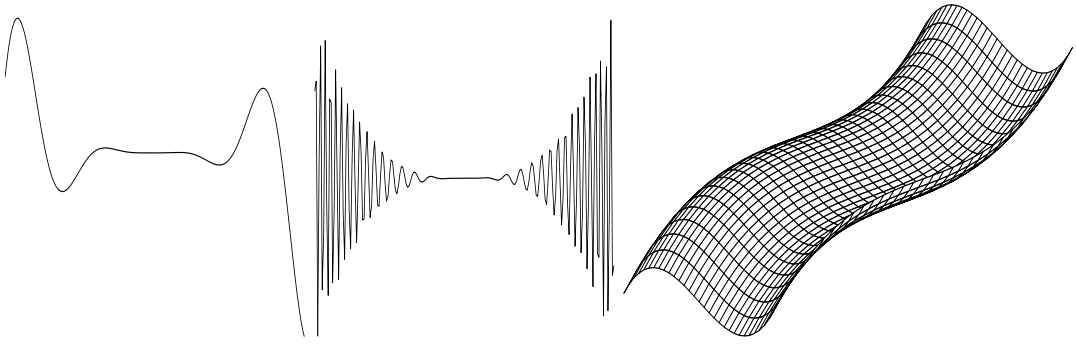

Figure 2.2 : exemples de fonctions non coercives

Théorème 2.1.1 (Existence)
Soit $J : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ propre, continue et coercive. Alors $(\mathcal{P})$ admet au moins une solution.

Démonstration - Soit $d = \inf(\mathcal{P})$; $d < +\infty$ car J est propre. Soit $(x_p)_{p\in\mathbb{N}} \in \mathbb{R}^n$ une suite minimisante, c'est-à-dire telle que

$$\lim_{p\to+\infty} J(x_p) = d \, .$$

Montrons que (x_p) est bornée.
Si ce n'était pas le cas on pourrait extraire de cette suite une sous-suite (encore notée (x_p)) telle $\lim_{p\to+\infty} \|x_p\| = +\infty$. Par coercivité de J on aurait $\lim_{p\to+\infty} J(x_p) = +\infty$, ce qui contredit le fait que $\lim_{p\to+\infty} J(x_p) = d < +\infty$.
Comme (x_p) est bornée, on peut alors en extraire une sous-suite (encore notée (x_p)) qui converge vers $\bar{x} \in \mathbb{R}^n$. Par continuité de J, on a alors

$$d = \lim_{p\to+\infty} J(x_p) = J(\bar{x}) \, .$$

En particulier $d > -\infty$ et $\bar{x}$ est une solution du problème $(\mathcal{P})$. □

Remarque 2.1.1 *Ce résultat est encore vrai si on remplace $\mathbb{R}^n$ par un espace de Hilbert $\mathbb{H}$ (ou plus généralement par un espace de Banach réflexif). La continuité de la fonctionnelle J est alors remplacée par de la semi-continuité inférieure pour la topologie faible. Nous ne voulons pas aborder les notions de topologie faible, aussi nous limitons nous au cas où l'espace de référence est de dimension finie.*

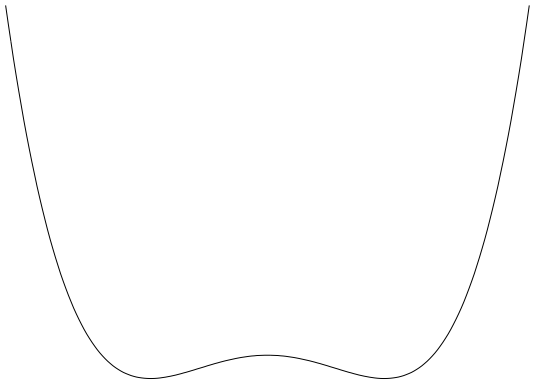

Figure 2.3 : minima (globaux) d'une fonction coercive (non convexe)

Toutefois, on n'a pas forcément unicité. . .Nous donnons ci-dessous un critère pour l'unicité.

Théorème 2.1.2 (Unicité)
Soit $J : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ strictement convexe. Alors le problème $(\mathcal{P})$ admet au plus une solution.

Démonstration - Supposons que J admette au moins un minimum m et soient $x_1 \neq x_2$ (dans $\mathbb{R}^n$) réalisant ce minimum : $J(x_1) = J(x_2) = m$. Par stricte convexité de la fonction J on a alors :

$$J\left(\frac{x_1 + x_2}{2}\right) < \frac{1}{2}\left(J(x_1) + J(x_2)\right) = m \ ;$$

ceci contredit le fait que m est le minimum. Donc $x_1 = x_2$. □
Donnons pour terminer un critère pour qu'une fonction soit strictement convexe et coercive :

Théorème 2.1.3 *Soit J une fonction $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$. On suppose qu'il existe $\alpha > 0$ tel que*

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \quad (\nabla J(x) - \nabla J(y), x - y) \geq \alpha \, \|x - y\|^2 \ . \tag{2.1.1}$$

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

Démonstration - D'après le théorème 1.3.3 la condition (2.1.1) implique que ∇J est monotone et que J est convexe. De plus la remarque 1.3.1 nous donne la stricte convexité de J. Enfin J est coercive : en effet, appliquons la formule de Taylor avec reste intégral

$$J(y) = J(x) + \int_0^1 \frac{d}{dt} J(x + t(y - x)) \, dt = J(x) + \int_0^1 (\nabla J(x + t(y - x)), y - x) \ dt \ .$$

Donc

$$J(y) = J(x) + (\nabla J(x), y - x) + \int_0^1 (\nabla J(x + t(y - x)) - \nabla J(x), y - x) \ dt \ . \tag{2.1.2}$$

D'après (2.1.1) on obtient

$$J(y) \geq J(x) + (\nabla J(x), y - x) + \int_0^1 t\alpha \|x - y\|^2 \, dt \ .$$

Finalement

$$J(y) \geq J(x) - \|\nabla J(x)\| \, \|y - x\| + \frac{\alpha}{2} \|x - y\|^2 \ . \tag{2.1.3}$$

Fixons $x = 0$ par exemple ; il est alors clair que J est coercive.
Par conséquent, J admet un minimum unique x^* sur $\mathbb{R}^n$ caractérisé par $\nabla J(x^*) = 0$.

□

La condition (2.1.1) nous amène à la définition suivante :

Définition 2.1.2 (Fonction elliptique)
*On dit que $J : \mathbb{H} \to \mathbb{R}$ est **elliptique** si la condition (2.1.1) est vérifiée, c'est-à-dire*

$$\exists \alpha > 0 \text{ tel que } \quad \forall (x, y) \in \mathbb{H} \times \mathbb{H} \quad (\nabla J(x) - \nabla J(y), x - y) \geq \alpha \, \|x - y\|^2 \ .$$

α est la constante d'ellipticité.

Donnons un critère d'ellipticité

Proposition 2.1.1 *Une fonction $J : \mathbb{H} \to \mathbb{R}$ deux fois différentiable sur $\mathbb{H}$ est elliptique si et seulement si*

$$\forall (x, y) \in \mathbb{H} \times \mathbb{H} \quad \left(D^2 J(x)\, y, y\right) \geq \alpha \, \|y\|^2 \, .$$

Démonstration - On utilise de nouveau la formule de Taylor appliquée à la fonction $\varphi : t \mapsto \varphi(t) = J(x + ty)$. La démonstration est laissée au lecteur. □

Il faut maintenant donner des conditions pour pouvoir calculer la (ou les) solutions. On va chercher à montrer que cette solution est solution de certaines équations, de sorte qu'il sera plus facile de la calculer.

2.2 Conditions d'optimalité

2.2.1 Conditions nécessaires du premier ordre

Les conditions que nous allons donner sont des conditions différentielles qui portent sur la dérivée de la fonction à minimiser. On va donc se restreindre au cas des fonctions Gâteaux-différentiables.

Théorème 2.2.1 (Condition nécessaire d'optimalité du premier ordre)
Soit $\mathbb{H}$ un espace de Hilbert réel et $J : \mathbb{H} \to \mathbb{R}$ une fonctionnelle Gâteaux-différentiable sur $\mathbb{H}$. Si x^* réalise un minimum (global ou local) de J sur $\mathbb{H}$ alors

$$\nabla J(x^*) = 0 \, . \tag{2.2.1}$$

Démonstration - Si x^* réalise un minimum de J sur $\mathbb{H}$ alors

$$\forall x \in \mathcal{B}(x^*, \rho) \qquad J(x^*) \leq J(x) \, ,$$

où $\mathcal{B}(x^*, \rho)$ est une boule de rayon $\rho > 0$ centrée en x^*.
Soit $h \in \mathbb{H},\ h \neq 0$; on peut trouver $t_h = \dfrac{\rho}{\|h\|} > 0$ tel que

$$\forall t \in]\, 0, t_h[\qquad x^* + t\, h \in \mathcal{B}(x^*, \rho) \, ,$$

et donc

$$\forall t \in]\, 0, t_h[\qquad J(x^*) \leq J(x^* + th) \, ,$$

Or J est Gâteaux-différentiable en x^*, donc

$$\lim_{t \to 0^+} \frac{J(x^* + th) - J(x^*)}{t} = (\nabla J(x^*), h) \ .$$

Donc

$$\forall h \in \mathbb{H} \qquad (\nabla J(x^*), h) \ \geq 0 \, , \tag{2.2.2}$$

c'est-à-dire $\nabla J(x^*) = 0$. □

Définition 2.2.1 *Un point x^* de $\mathbb{H}$ vérifiant $\nabla J(x^*) = 0$ est appelé* ***point critique*** *ou* ***point stationnaire*** .

La relation $\nabla J(x^*) = 0$ est aussi appelée **équation d'Euler** .
Ce théorème n'a pas de sens si la fonction J n'est pas différentiable comme le montre la figure 2.4. suivante :

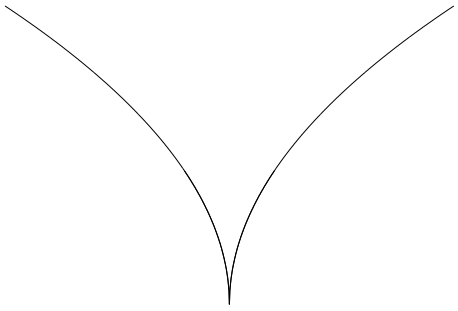

Figure 2.4 : Minimum d'une fonction non différentiable

On constate que cette condition nécessaire du premier ordre permet de sélectionner un certain nombre de "candidats" à être des minima (globaux) ; toutefois, la réciproque du Théorème 2.2.1 est **fausse** ; un point critique n'est pas nécessairement un minimum global. Ce peut-être un minimum local (cf. figure 2.5 : point B), un maximum local (cf. figure 2.5 : point C) ou ni l'un, ni l'autre (cf. figure 2.5 : point A). Le résultat du théorème est une condition **nécessaire** qui n'est en général pas suffisante (cf. la fonction f de $\mathbb{R}$ dans $\mathbb{R}$ définie par $f(x) = x^3$). C'est une condition du **premier ordre** car elle ne fait intervenir que la dérivée première de la fonction J.

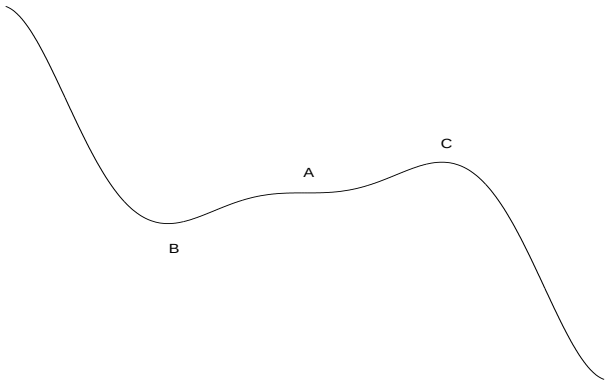

Figure 2.5 : Points critiques non optimaux

Il existe cependant des situations où la relation (2.2.1) est une condition **nécessaire et suffisante** : en **programmation convexe**.

Théorème 2.2.2 (CNS du premier ordre dans le cas convexe)
Soit $J : \mathbb{H} \to \mathbb{R}$ Gâteaux- différentiable et convexe sur $\mathbb{H}$. Un point x^* réalise un minimum global de J sur $\mathbb{H}$ si et seulement si $\nabla J(x^*) = 0$.

Démonstration - On a vu que la condition est toujours nécessaire. Montrons qu'elle est suffisante. Soit $x^* \in \mathbb{H}$ tel que $\nabla J(x^*) = 0$. Comme J est convexe on peut utiliser le théorème (1.3.1) du chapitre 1 et on obtient :

$$\forall x \in \mathbb{H} \qquad J(x) \geq J(x^*) + (\nabla J(x^*), x - x^*) = J(x^*) \; .$$

On a donc immédiatement le fait que x^* réalise un minimum de J sur $\mathbb{H}$. □

Remarque 2.2.1 *On peut raffiner le résultat précédent en ne supposant que la locale convexité de J au voisinage de x^*, c'est-à-dire en supposant que J est convexe sur une boule centrée en x^*. A ce moment là, nous pouvons affirmer que x^* est un minimum **local** de J.*

Le cas où J est convexe est fréquent dans la pratique mais pas systématique. Nous allons donc donner maintenant des conditions suffisantes pour qu'un point critique réalise un minimum (ou un maximum). Ces conditions vont faire intervenir la dérivée seconde de J : ce sont des conditions du **second ordre**.

2.2.2 Conditions du deuxième ordre

Nous commençons par une condition nécessaire permettant de préciser encore les éventuels mi- nima.

Théorème 2.2.3 (Condition nécessaire du second ordre)
On suppose que x^* est un minimum (local) de J et que J est deux fois dérivable sur $\mathbb{H}$. Alors
i. $\nabla J(x^*) = 0$ et
ii.$\forall x \in \mathbb{H} \qquad \left(D^2 J(x^*)\, x, x)\right) \geq 0.$

Démonstration - (i) a déjà été vue : montrons (ii).
Soit $x \in \mathbb{H}$. Appliquons la formule de Taylor à la fonction $\varphi : t \mapsto \varphi(t) = J(x^* + tx)$. Comme $\nabla J(x^*) = 0$ on obtient

$$0 \leq J(x^* + tx) - J(x^*) = \frac{t^2}{2} \left(D^2 J(x^*)\, x, x\right) + o(t^2) \; .$$

Après division par t^2, on fait tendre t vers 0 et on a le résultat voulu. □

Remarque 2.2.2 *Dans le cas où $\mathbb{H} = \mathbb{R}^n$, (ii) est équivalent à dire que la matrice Hessienne de J en x^* : $D^2 J(x^*)$ est semi-définie positive.*
Rappelons qu'un critère pour que $D^2 J(x^*)$ (qui est une matrice symétrique) soit semi-définie positive est que toutes ses valeurs propres soient positives ou nulles.

La réciproque du théorème précédent est fausse (il suffit de penser à la fonction $t \mapsto t^3$ pour s'en convaincre). Nous pouvons toutefois donner une réciproque sous forme de condition suffisante du second ordre plus forte (pour un résultat plus faible) de ce qui précède.

Théorème 2.2.4 (Condition suffisante du second ordre)
Soit J deux fois dérivable sur $\mathbb{H}$ vérifiant $\nabla J(x^*) = 0$ et

$$\exists \alpha > 0 \ , \ \forall x \in \mathbb{H} \qquad \left(D^2 J(x^*)\, x, x\right) \geq \alpha \|x\|^2 \ . \tag{2.2.3}$$

*Alors la fonction J admet un minimum **local** strict en x^*.*

Démonstration - Soit x dans $\mathbb{H}$. On utilise de nouveau la formule de Taylor appliquée à la fonction $\varphi : t \mapsto \varphi(t) = J(x^* + tx)$. Nous avons

$$J(x^* + tx) - J(x^*) = \frac{t^2}{2}\left(D^2 J(x^*)\, x, x\right) + o(t^2) \geq \frac{t^2}{2}\alpha\|x\|^2 + o(t^2) \ .$$

Ceci montre que x^* réalise un minimum local strict de J □
La condition (2.2.3) est une condition d'**ellipticité** locale. .

Remarque 2.2.3 *Si $\mathbb{H} = \mathbb{R}^n$ la condition (2.2.3) revient à dire que la matrice Hessienne $D^2 J(x^*)$ est définie positive, un choix possible pour α étant alors la plus petite valeur propre. C'est une condition de convexité (locale) stricte au voisinage de x^*.*

Exemple 2.2.1 *Soit $J : \mathbb{R}^2 \to \mathbb{R}$ définie par*

$$J(x_1, x_2) = 3x_1^4 - 4x_1^2 x_2 + x_2^2 \ .$$

Il est facile de voir que le point $(0,0)$ est un point critique mais on ne peut pas conclure immédiatement car la matrice Hessienne $D^2 J(0,0) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$, est seulement semi-définie positive puisque ses valeurs propres sont 0 et 2. On peut toutefois vérifier que ce point critique n'est ni un minimum ni un maximum en remarquant que $J(0,1) = 1 > J(0,0) = 0 > J(0.5, 0.5) = -0.0625$.
On peut aussi remarquer que J n'est pas coercive puisque $J(n, n^2) = 0$.

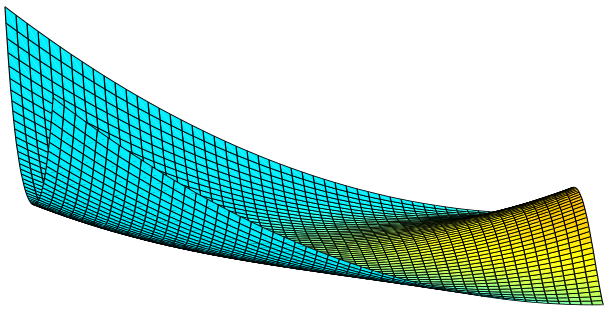

Figure 2.6 : fonction précédente au voisinage de (0,0)

Exemple 2.2.2 (Cas d'une fonction quadratique)
Nous allons détailler le cas où J est une fonction quadratique c'est-à-dire

$$J(x) = \frac{1}{2}\,(Ax, x) - (b, x)\ ,$$

où A est une matrice carrée symétrique et $b \in \mathbb{R}^n$.

Les points critiques de J sont solutions de $\nabla J(x) = 0$, c'est-à-dire sont solutions du système linéaire : $Ax = b$. Si de plus A est définie positive alors la solution de $Ax = b$ est le minimum (global) de J.

Le cadre de l'optimisation va donc permettre de donner des méthodes de résolution de systèmes linéaires grâce à des méthodes issues de l'optimisation. Nous y reviendrons dans la section de ce chapitre consacrée aux Algorithmes.

2.3 Exemple : régression linéaire

Nous allons illustrer les résultats de la section précédente par l'exemple très important de la **régression linéaire**.

Considérons un nuage de n points de $\mathbb{R}^2$: $M_i = (t_i, x_i), 1 \leq i \leq n$. Ces données sont souvent le résultat de mesures et on cherche à décrire le comportement global de ce nuage. En général ces points ne sont pas alignés, mais on décide de chercher une droite les "approchant" au mieux...

On utilise pour cela la méthode des moindres carrés : comme on n'a pas $x_i = a\ t_i + b$ pour tout i, on cherche à minimiser le carré des différences. On veut donc trouver un couple de réels (a, b) solution de

$$\min J(a, b)\ ,\ (a, b) \in \mathbb{R}^2\ ,$$

où $J(a, b) = \displaystyle\sum_{i=1}^{n} (x_i - at_i - b)^2$.

Calculons le gradient de J en un point quelconque (a, b) de $\mathbb{R}^2$.

$$\frac{\partial J}{\partial a}(a, b) = \sum_{i=1}^{n} 2\,(at_i + b - x_i)\ t_i = 2\,a \sum_{i=1}^{n} t_i^2 + 2\,b \sum_{i=1}^{n} t_i - 2 \sum_{i=1}^{n} t_i\,x_i\ ,$$

$$\frac{\partial J}{\partial b}(a, b) = \sum_{i=1}^{n} 2\,(at_i + b - x_i) = 2\,a \sum_{i=1}^{n} t_i + 2n\,b - 2 \sum_{i=1}^{n} x_i\ .$$

Notons

$$S_t = \sum_{i=1}^{n} t_i\ ,\ S_x = \sum_{i=1}^{n} x_i\ ,\ S_{xt} = \sum_{i=1}^{n} x_i\,t_i \text{ et } S_{t^2} = \sum_{i=1}^{n} t_i^2\ .$$

Ecrire que $\nabla J(a, b) = 0$ revient à écrire le système

$$\left\{ \begin{array}{lcl} S_{t^2}\ a + S_t\ b & = & S_{xt} \\ S_t\ a + n\ b & = & S_x \end{array} \right.$$

La résolution de ce système donne une solution unique si $(S_t)^2 - nS_{t^2} \neq 0$. On obtient

$$a = \frac{S_x S_t - nS_{xt}}{(S_t)^2 - nS_{t^2}} \text{ et } b = \frac{S_{xt} S_t - S_x S_{t^2}}{(S_t)^2 - nS_{t^2}} .$$

Il faut encore vérifier que le couple obtenu est bien un minimum. Calculons pour cela la matrice Hessienne de J en (a, b) :

$$D^2 J(a, b) = 2 \begin{bmatrix} S_{t^2} & S_t \\ S_t & n \end{bmatrix} ;$$

cette matrice est toujours (semi-définie) positive. Elle est définie positive si elle est inversible (car alors aucune valeur propre n'est nulle) c'est-à-dire lorsque que son déterminant $(S_t)^2 - nS_{t^2}$ est non nul. Dans ce cas, le couple obtenu est bien (l'unique) minimum strict de J.
La droite ainsi obtenue est appelée **droite de régression**.

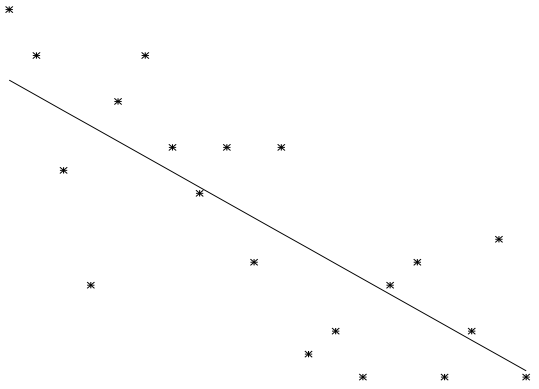

Figure 2.7 : Droite de régression d'un nuage de points

2.4 Algorithmes (déterministes)

Dans cette section, nous allons présenter quelques algorithmes permettant de calculer (de manière approchée) la ou les solutions du problème $(\mathcal{P})$ de départ. Bien entendu, nous ne pouvons pas être exhaustifs ; nous présentons les méthodes "de base" les plus classiques. Toutefois, la plupart de ces algorithmes exploitent les conditions d'optimalité dont on a vu qu'elles permettaient (au mieux) de déterminer des minima **locaux**. La question de la détermination de minima globaux est difficile et dépasse le cadre que nous nous sommes fixés. Néanmoins, nous décrirons dans la section suivante, un algorithme probabiliste permettant de "déterminer" un minimum global.

Remarquons aussi que nous avons fait l'hypothèse de différentiabilité de la fonction J. Il existe des méthodes permettant de traiter le cas non différentiable (ou non régulier). Nous n'en parlerons pas ici. Le lecteur intéressé peut se référer à [4].

Nous commencerons par quelques définitions :

Définition 2.4.1 (Algorithme)
*Un **algorithme** est défini 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 :

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

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 2.4.2 (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^*.*

Il est bien entendu très important d'assurer la convergence d'un algorithme via les hypothèses ad-hoc, mais la vitesse de convergence et la complexité sont aussi des facteurs à prendre en compte lors de l'utilisation (ou de la génération) d'un algorithme ; on a en effet "intérêt" à ce que la méthode soit la plus rapide possible tout en restant précise et stable. Un critère de mesure de la vitesse (ou taux) de convergence est l'évolution de l'erreur commise ($e_k = \|x_k - x^*\|$).

Définition 2.4.3 (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 :*

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

- ***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. Si α_k est une suite géométrique, la convergence de l'algorithme est dite **géométrique**.*
- ***d'ordre p** si l'erreur e_k décroit de la manière suivante :*

$$\exists C \geq 0 \ , \exists k_0, \ \ \forall k \geq k_0 \qquad e_{k+1} \leq C \, [e_k]^p \ .$$

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

*Enfin, la convergence est dite **locale** si elle n'a lieu que pour des points de départ x_0 dans un voisinage de x^*. Dans le cas contraire la convergence est globale.*

Remarque 2.4.1 *La "classification" précédente des vitesses de convergence renvoie à la notion de comparaison des fonctions au voisinage de $+\infty$. En effet, si on suppose que l'erreur e_k ne s'annule pas, une convergence linéaire revient à dire que $\dfrac{e_{k+1}}{e_k} = O(1)$, alors qu'une convergence super-linéaire est équivalente à $\dfrac{e_{k+1}}{e_k} = o(1)$. De manière analogue, un algorithme d'ordre $p \geq 2$ est tel que $\dfrac{e_{k+1}}{e_k} = o(e_k^{p-2})$. On a bien entendu intérêt à ce que la vitesse de convergence d'un algorithme soit la plus élevée possible (afin d'obtenir la solution avec un minimum d'itérations pour une précision donnée).*

2.4.1 Méthode du Gradient

La méthode (ou algorithme) du **Gradient** fait partie d'une classe plus grande de méthodes numéri- ques appelées **méthodes de descente**. Expliquons rapidement l'idée directrice de ces méthodes.

On veut minimiser une fonction J. Pour cela on se donne un point de départ arbitraire x_o. Pour construire l'itéré suivant x_1 il faut penser qu'on veut se rapprocher du minimum de J ; on veut donc que $J(x_1) < J(x_o)$. On cherche alors x_1 sous la forme $x_1 = x_o + \rho_1 d_1$ où d_1 est un vecteur non nul de $\mathbb{R}^n$ et ρ_1 un réel strictement positif. En pratique donc, on cherche d_1 et ρ_1 pour que $J(x_o + \rho_1 d_1) < J(x_o)$. On ne peut pas toujours trouver d_1. Quand d_1 existe on dit que c'est une **direction de descente** et ρ_1 est le **pas de descente**. La direction et le pas de descente peuvent être fixes ou changer à chaque itération. Le schéma général d'une méthode de descente est le suivant :

$$\begin{cases} x_0 \in \mathbb{R}^n \text{ donné} \\ x_{k+1} = x_k + \rho_k\, d_k,\ d_k \in \mathbb{R}^n - \{0\},\ \rho_k \in \mathbb{R}^{+*}\ , \end{cases}$$

où ρ_k et d_k sont choisis de telle sorte que $J(x_k + \rho_k\, d_k) \leq J(x_k)$.

Une idée naturelle pour trouver une direction de descente est de faire un développement de Taylor (formel) à l'ordre 2 de la fonction J entre deux itérés x_k et $x_{k+1} = x_k + \rho_k\, d_k$:

$$J(x_k + \rho_k\, d_k) = J(x_k) + \rho_k(\nabla J(x_k), d_k) + o(\rho_k d_k).$$

Comme on veut $J(x_k + \rho_k\, d_k) < J(x_k)$, on peut choisir en première approximation $d_k = -\nabla J(x_k)$. La méthode ainsi obtenue s'appelle l'algorithme du **Gradient**. Le pas ρ_k est choisi constant ou variable.

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 $\|x_{k+1} - x_k\| < \varepsilon$, 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.

Cette méthode a pour avantage d'être très facile à mettre en œuvre. Malheureusement, les conditions de convergence sont assez lourdes (c'est essentiellement de la stricte convexité) et la méthode est en général assez lente. Nous donnons ci-dessous un critère de convergence :

Théorème 2.4.1 *Soit J une fonction $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$, coercive et strictement convexe. On suppose qu'il existe une constante M strictement positive telle que*

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \quad \|\nabla J(x) - \nabla J(y)\| \leq M\, \|x - y\|\ . \tag{2.4.1}$$

Alors, si on choisit le pas ρ_k dans un intervalle $[\beta_1, \beta_2]$ tel que $0 < \beta_1 < \beta_2 < \dfrac{2}{M}$, la méthode du gradient converge vers le minimum de J.

Démonstration - J admet un minimum unique sur $\mathbb{R}^n$ et ce minimum x^* est caractérisé par $\nabla J(x^*) = 0$, puisque J est strictement convexe. Montrons que la suite x_k engendrée par l'algorithme converge vers x^*. Appliquons la relation (2.1.2) à $y = x_{k+1}$ et $x = x_k$:

$$J(x_{k+1}) = J(x_k) + (\nabla J(x_k), x_{k+1} - x_k) + \int_0^1 (\nabla J(x_k + t(x_{k+1} - x_k)) - \nabla J(x_k), x_{k+1} - x_k) \; dt.$$

Comme $x_{k+1} = x_k - \rho_k \nabla J(x_k)$ on obtient (avec (2.4.1))

$$J(x_{k+1}) - J(x_k) \leq -\frac{1}{\rho_k}\|x_{k+1} - x_k\|^2 + \int_0^1 \|\nabla J(x_k + t(x_{k+1} - x_k)) - \nabla J(x_k)\| \; \|x_{k+1} - x_k\| \; dt$$

$$J(x_{k+1}) - J(x_k) \leq -\frac{1}{\rho_k}\|x_{k+1} - x_k\|^2 + \frac{M}{2}\|x_{k+1} - x_k\|^2 = \left[\frac{M}{2} - \frac{1}{\rho_k}\right] \|x_{k+1} - x_k\|^2 .$$

Si on choisit ρ_k dans un intervalle $[\beta_1, \beta_2]$ tel que $0 < \beta_1 < \beta_2 < \dfrac{2}{M}$, nous obtenons alors

$$J(x_{k+1}) - J(x_k) \leq \left[\frac{M}{2} - \frac{1}{\beta_2}\right] \|x_{k+1} - x_k\|^2 .$$

La suite $J(x_k)$ est alors strictement décroissante ; comme elle est minorée elle converge. Cela entraîne d'une part que $J(x_{k+1}) - J(x_k)$ tend vers 0 et d'autre part que la suite (x_k) est bornée (par coercivité). On peut donc extraire de (x_k) une sous-suite convergente vers $\bar{x}$. De plus comme

$$\|x_{k+1} - x_k\|^2 \leq \left[\frac{1}{\beta_2} - \frac{M}{2}\right]^{-1} [J(x_{k+1}) - J(x_k)] \; ,$$

la suite $(x_{k+1} - x_k)$ tend également vers 0. Par conséquent $\nabla J(x_k) = \dfrac{x_{k+1} - x_k}{\rho_k}$ tend vers 0. Par continuité de ∇J, on obtient $\nabla J(\bar{x}) = 0$. Donc $\bar{x}$ est l'unique minimum x^* de J. Ceci étant vrai pour toute valeur d'adhérence de la suite (x_k) cela prouve que toute la suite (x_k) converge vers x^*. □

Définition 2.4.4 *On dit qu'une fonction F de $\mathbb{R}^n$ dans $\mathbb{R}^n$ est **Lipschitzienne** de rapport $M > 0$ si*

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \quad \|F(x) - F(y)\| \leq M \; \|x - y\| \; .$$

La condition (2.4.1) du théorème précédent signifie donc que ∇J est lipschitzienne.

Corollaire 2.4.1 *Soit J, une fonction $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$, elliptique et de dérivée lipschitzienne (c'est-à-dire vérifiant (2.1.1) et (2.4.1)). Alors, si on choisit le pas ρ_k dans un intervalle $[\beta_1, \beta_2]$ tel que $0 < \beta_1 < \beta_2 < \dfrac{2}{M}$, la méthode du gradient converge vers le minimum de J.*

Démonstration - Il suffit de coupler les résultats des théorèmes 2.1.3 et 2.4.1. □

Remarque 2.4.2 *Lorsque J vérifie (2.1.1) et (2.4.1), on peut aussi interpréter l'algorithme du gradient à pas constant comme la méthode des approximations successives appliquée à la recherche du point fixe de la fonction*

$$S_\rho(x) = x - \rho \nabla J(x) \ ,$$

où $\rho \neq 0$. On peut en effet montrer que S_ρ est lipschitzienne de rapport (1- $2\rho\alpha + \rho^2 M^2$). C'est donc une contraction stricte si $\rho \in]0,\ 2\alpha/M^2\ [$; elle possède alors un unique point fixe. Pour $\rho = \alpha/M^2$, le taux de contraction est (1- α/M^2) : c'est le meilleur possible. La convergence est alors celle d'une série géométrique de raison (1- α/M^2).

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**.

La méthode du gradient à **pas optimal** propose un choix du pas qui rend la fonction coût minimale le long de la direction de descente choisie. Plus précisément, l'étape 2. devient
2'.- **Itération** k

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

où ρ_k réalise le minimum sur $\mathbb{R}^+$ de la fonction Φ_k définie par

$$\Phi_k(\rho) = J(x_k - \rho \ \nabla J(x_k)) \ .$$

En pratique, on ne cherche pas le minimum de Φ_k et on détermine ρ_k en effectuant une **recherche linéaire** de pas optimal suivant une règle de la forme suivante par exemple :

Règle de recherche linéaire de Wolfe

1. Initialisation $\rho = 1$ (par exemple), $\rho_- = \rho_+ = 0$. On se donne $0 < \beta_1 < \beta_2 < 1$.
2. Si $\Phi_k(\rho) \leq \Phi_k(0) + \beta_1 \, \rho \, \Phi_k'(0)$ et $\Phi_k(\rho) \geq \beta_2 \, \Phi_k'(0)$, STOP : $\rho_k = \rho$.
3. Sinon
   – Si $\Phi_k(\rho) > \Phi_k(0) + \beta_1 \, \rho \, \Phi_k'(0)$, on pose $\rho_+ = \rho$
   – Si $\Phi_k(\rho) \leq \Phi_k(0) + \beta_1 \, \rho \, \Phi_k'(0)$ et $\Phi_k(\rho) < \beta_2 \, \Phi_k'(0)$, on pose $\rho_- = \rho$
   et on va à 4.
4. Choix d'un nouveau ρ :
   – Si $\rho_+ = 0$, on cherche $\rho > \rho_-$ (par exemple $\rho = 2 \, \rho_-$).
   – Si $\rho_+ > 0$, on cherche $\rho \in] \, \rho_-, \rho_+ \, [$ (par exemple $\rho = \dfrac{\rho_- + \rho_+}{2}$).
   Retour à 2.

La règle apparaissant à l'étape 2. est connue sous le nom générique de règle d'**Armijo**. Il existe beaucoup d'autres règles de recherche linéaire. Pour plus de détails sur ces variantes, on pourra se référer à [7, 12, 4] ou aux exercices en fin de chapitre.

Exemple 2.4.1 *Les conditions du théorème peuvent paraître compliquées, aussi nous donnons un exemple. Soit J la fonction de $\mathbb{R}^n$ vers $\mathbb{R}$ déjà évoquée plusieurs fois (car elle joue un rôle important) définie par*

$$J(x) = \frac{1}{2}(Ax, x) - (b, x) \ ,$$

où A est une matrice carrée, symétrique et définie positive et $b \in \mathbb{R}^n$ (voir l'exemple (2.1.1)). Cette fonction J vérifie les hypothèses du théorème ci-dessus avec pour constantes α et M la plus petite et la plus grande valeur propre de A (respectivement).

Remarque 2.4.3 *La notion d'ellipticité est très importante, car elle conditionne la convergence de la plupart des algorithmes qui vont être décrits par la suite. Toutefois, les conditions de convergence que nous donnons sont toujours des conditions **suffisantes**. L'algorithme converge si elles sont vérifiées mais **il peut éventuellement converger, même si elles ne le sont pas** ....*

En pratique, on ne calcule pas α et M. Pour trouver l'intervalle de convergence de ρ, on fait plusieurs tests pour différentes valeurs. La non convergence se traduit en général, soit par une explosion de la solution (elle va clairement vers $+\infty$) soit par des oscillations (périodiques ou non) qui empêchent la suite des itérés de converger vers une valeur.

2.4.2 Méthode de Newton

La méthode de NEWTON n'est pas une méthode d'optimisation à proprement parler. C'est en réalité une méthode utilisée pour résoudre des équations non linéaires de la forme $F(x) = 0$ où F est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$. Nous allons d'abord la décrire puis montrer comment on peut l'appliquer à la recherche de minimum.

Méthode de Newton

Présentons d'abord **formellement** cette méthode dans $\mathbb{R}$, pour résoudre $f(x) = 0$ où f est une fonction $\mathcal{C}^1$ de $\mathbb{R}$ dans $\mathbb{R}$. Nous préciserons ensuite les conditions d'utilisation de la méthode et justifierons l'existence des itérés successifs dans un théorème général de convergence.

Algorithme de Newton dans $\mathbb{R}$

1. **Initialisation**
   $k = 0$: choix de $x_0 \in \mathbb{R}$ dans un voisinage de x^*.
2. **Itération** k
   $x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)}$;
3. **Critère d'arrêt**
   Si $|x_{k+1} - x_k| < \varepsilon$, STOP
   Sinon, on pose $k = k + 1$ et on retourne à 2.

Remarquons qu'il faut non seulement assurer la convergence de la suite x_k vers la solution x^*, mais aussi montrer que cette suite est bien définie, c'est-à-dire montrer que $f'(x_k) \neq 0$ à l'étape 2.

Cette méthode est aussi appelée **méthode de la tangente**. En effet chaque itéré x_{k+1} est obtenu à partir du précédent en traçant la tangente à la courbe de f au point $(x_k, f(x_k))$ et en prenant son intersection avec l'axe des abscisses.

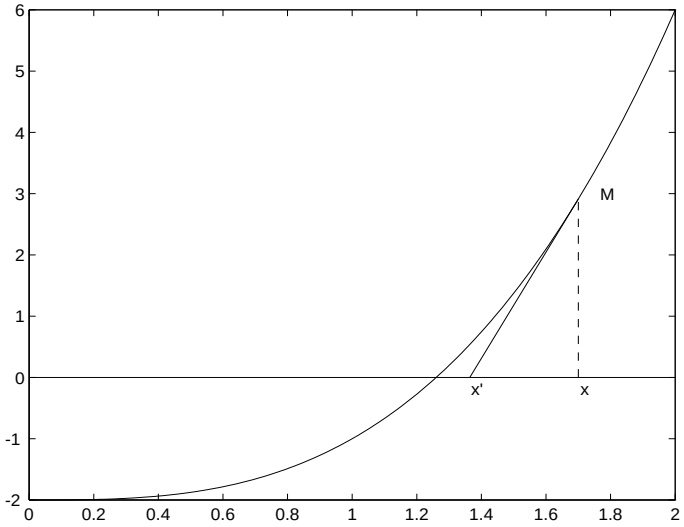

Figure 2.8 : Interprétation géométrique de la méthode de Newton

Nous pouvons maintenant généraliser à $\mathbb{R}^n$. Soit F une fonction de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}^n$. On suppose que l'équation

$$F(x) = 0\,, \tag{2.4.2}$$

possède au moins une solution notée x^* et que la matrice jacobienne $DF(x^*)$ est une matrice inversible. Nous verrons que la continuité de DF permet alors d'assurer l'inversibilité de $DF(x_k)$ pour tout x_k dans un voisinage de x^* et de fait, l'existence de x_{k+1} à l'étape 2. de l'algorithme décrit ci-dessous.

Algorithme de Newton dans $\mathbb{R}^n$

1. **Initialisation**
   $k = 0$: choix de x_0 dans un voisinage de x^*.
2. **Itération** k
   $x_{k+1} = x_k - [DF(x_k)]^{-1}\, F(x_k)$;
3. **Critère d'arrêt**
   Si $\|x_{k+1} - x_k\| < \varepsilon$, STOP
   Sinon, on pose $k = k + 1$ et on retourne à 2.

L'étape 2. de la méthode revient à résoudre le système linéaire suivant :

$$[DF(x_k)]\ \delta_k = \ F(x_k)$$

puis à poser $x_{k+1} = x_k - \delta_k$.

Nous avons le résultat de convergence suivant :

Théorème 2.4.2 *Soit F est une fonction de classe $\mathcal{C}^2$ de $\mathbb{R}^n$ dans $\mathbb{R}^n$ et x^* un zéro de F (c'est-à-dire $F(x^*) = 0$). On suppose en outre que ce zéro est isolé et que $DF(x^*)$ est inversible (DF désigne la dérivée première de F).*
Alors il existe une boule fermée $\mathcal{B}$ centrée en x^* telle que, pour tout point $x_0 \in \mathcal{B}$, la suite (x_k) définie par la méthode de Newton est entièrement contenue dans $\mathcal{B}$ et converge vers x^* qui est le seul zéro de F dans $\mathcal{B}$.
Enfin la convergence est géométrique : il existe $\beta \in]0,1[$ tel que

$$\forall k \geq 0 \qquad \|x_k - x^*\| \leq \beta^k \|x_0 - x^*\| .$$

En d'autres termes, si on choisit le point de départ x_0 "assez près" de x^*, alors l'algorithme converge vers x^*.

Démonstration - Comme F est $\mathcal{C}^1$ et $DF(x^*)$ est inversible, il existe une boule centrée en x^* : $\mathcal{B}(x^*, r_0)$ sur laquelle $DF(\cdot)$ est inversible et $DF(\cdot)^{-1}$ est uniformément bornée par m.
Appliquons la formule de Taylor avec reste intégral entre x^* et un itéré x_k en supposant que $x_k \in \mathcal{B}(x^*, r_0)$:

$$F(x^*) - F(x_k) = DF(x_k)(x^* - x_k) + \int_0^1 D^2F(x^* + t\,(x_k - x^*)) \cdot (x^* - x_k)^2\, t\, dt .$$

Comme F est $\mathcal{C}^2$, D^2F est continue et uniformément bornée sur $\mathcal{B}(x^*, r_0)$ par un réel $M > 0$, et nous avons

$$\left| \int_0^1 D^2F(x^* + t\,(x_k - x^*)) \cdot (x^* - x_k)^2\, t\, dt \right| \leq \frac{M}{2} \|x^* - x_k\|^2 .$$

Par conséquent, comme

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - DF(x_k)^{-1}[F(x_k) - F(x^*)] \\ &= DF(x_k)^{-1}[F(x^*) - F(x_k) - DF(x_k)(x^* - x_k)] \\ &= DF(x_k)^{-1} \left[\int_0^1 D^2F(x^* + t\,(x_k - x^*)) \cdot (x^* - x_k)^2\, t\, dt \right] \end{aligned}$$

on obtient

$$\|x_{k+1} - x^*\| \leq m \frac{M}{2} \|x^* - x_k\|^2 . \tag{2.4.3}$$

Posons $r = \min\left(r_0, \frac{2}{mM}, 1\right)$; il est alors facile de voir par récurrence que si $x_0 \in \mathcal{B}(x^*, r)$ alors pour tout k, $x_k \in \mathcal{B}(x^*, r)$: la suite des itérés est bien définie et reste dans la boule. Posons alors $e_k = \frac{mM}{2}\|x_k - x^*\|$. La relation (2.4.3) donne

$$e_{k+1} \leq {e_k}^2.$$

La suite e_k converge donc vers 0 si $e_0 = \dfrac{mM}{2}\|x_0 - x^*\| < 1$, cad si x_0 est dans la boule $\mathcal{B}(x^*, r^*)$ où $r^* = \min(r, \dfrac{1}{mM})$ par exemple. □

L'inconvénient majeur de la méthode est sa sensibilité au choix point de départ : la convergence est **locale**. Si ce point est mal choisi ("trop loin" de la solution) la méthode peut soit diverger, soit converger vers une autre solution. Il peut paraître surprenant de devoir choisir le point de départ x_0 "assez près" de x^* puisqu'on ne connaît pas x^* ! ! En pratique on essaie de s'approcher de x^* par une méthode de type gradient par exemple, puis on applique la méthode de Newton. L'avantage de cette méthode est sa grande rapidité. La convergence est quadratique d'après la relation (2.4.3), c'est-à-dire que l'erreur $e_k = \|x_{k+1} - x_k\|$ est élevée au carré à chaque itération. Concrètement, si elle vaut 10^{-2} à l'étape k elle vaudra 10^{-4} à l'étape $k+1$ et 10^{-8} à l'étape $k+2$.

Actuellement on développe surtout des méthodes dites **quasi-Newton** qui gardent la rapidité de la méthode de Newton, évitent le calcul (coûteux) de la matrice $[DF(x_k)]$ à chaque itération et sont plus robustes par rapport au du point de départ. On y trouve les méthodes dites de "région de confiance" qui s'attachent à rendre la méthode robuste (i.e. peu sensible) par rapport au choix du point de départ. La méthode BFGS propose un calcul de H_{k+1} (approximation de $[DF(x_k)]$) en fonction de H_k grâce à des formules algébriques. On peut aussi décider de "garder" la matrice H_k pendant plusieurs itérations et de l'actualiser périodiquement ...Pour plus de détails, on peut consulter par exemple [18].
Nous présentons ici une méthode quasi-Newton, très utilisée qui est la méthode de **Levenberg-Marquardt** couplée avec une règle de recherche linéaire de type Armijo, qui a l'avantage d'avoir une convergence quadratique globale sous des hypothèses standard (voir [14]).

Algorithme de Levenberg-Marquardt (avec recherche linéaire)

1. **Initialisation**
   On se donne α, β, $\gamma \in]\,0, 1\,[$.
   Choix de $x_0 \in \mathbb{R}^n$ et calcul de $\mu_0 = \|F(x_0)\|^2$, $k = 0$.
2. **Itération** k : trouver la solution d_k du système

$$[\mu_k Id + DF(x_k)^t DF(x_k)]\, d = -DF(x_k)^t\, F(x_k)\ ;$$

   Si d_k vérifie : $\|F(x_k + d_k)\| \leq \gamma \|F(x_k)\|$ alors : $x_{k+1} = x_k + d_k$ et on va à 4.
   Sinon, on va à 3.
3. **Etape de recherche linéaire du pas** : Soit m_k le plus petit entier positif tel que

$$\|\Phi(x_k + \beta^m d_k)\|^2 - \|\Phi(x_k)\|^2 \leq \alpha\, \beta^m \nabla\Phi(x_k)^t d_k\ ,$$

   où $\Phi(x) = \dfrac{1}{2}\|F(x)\|^2$.
   On pose $x_{k+1} = x_k + \beta^{m_k} d_k$ et on va à 4.
4. **Critère d'arrêt**
   Si $\|x_{k+1} - x_k\| < \varepsilon$, STOP
   Sinon, on pose $\mu_{k+1} = \|F(x_{k+1})\|^2$, $k = k+1$, et on retourne à 2.

Si M est une matrice (réelle), M^t désigne sa matrice transposée.

Application à la recherche d'extrema

Nous avons vu dans la section 3.1.2 (Théorème 2.2.1) qu'une condition nécessaire d'optimalité est $\nabla J(x^*) = 0$. Ceci est une équation non linéaire (ou plutôt un système d'équations non linéaires) dans $\mathbb{R}^n$ et nous allons utiliser la méthode de Newton pour la résoudre. **Toutefois nous n'obtiendrons que les points critiques de J : il faudra ensuite vérifier que ce sont bien des minima.**

Ici $F = \nabla J$ est bien une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$. La dérivée de F n'est autre que la matrice hessienne de J (voir Annexe A) : $H(x) = D^2J(x)$. La méthode de Newton s'écrit alors :

Algorithme de Newton pour la recherche de points critiques

1. **Initialisation**
   $k = 0$: choix de x_0 dans un voisinage de x^*.
2. **Itération** k
   $x_{k+1} = x_k - [H(x_k)]^{-1} \nabla J(x_k)$;
3. **Critère d'arrêt**
   Si $\|x_{k+1} - x_k\| < \varepsilon$, STOP
   Sinon, on pose $k = k + 1$ et on retourne à 2.

L'étape 2. de la méthode revient à résoudre le système linéaire suivant :

$$H_k \delta_k = \nabla J(x_k) \ ,$$

où $H_k = H(x_k)$ puis à poser $x_{k+1} = x_k - \delta_k$.

Le calcul de H_k est l'étape la plus coûteuse de la méthode. Là encore des méthodes quasi-Newton peuvent être utilisées pour rendre le calcul moins "cher".

2.4.3 Méthode du Gradient conjugué

Méthode du Gradient conjugué : cas linéaire

La présentation des deux algorithmes précédents montrent qu'en réalité on ne calcule pas les extrema d'une fonction mais les points stationnaires (ou points critiques) qui vérifient la condition d'optimalité du premier ordre. Dans le cas particulier où J est quadratique nous avons vu que cela revient à résoudre un système linéaire : $Ax = b$. Nous allons donc présenter ici une méthode de résolution d'un système linéaire issue de la théorie de l'optimisation et convergente dans le cas des matrices symétriques définies positives. Dans ce cas l'application qui au couple (x, y) associe le produit $(A\,x, y)$ est un produit scalaire sur $\mathbb{R}^n$ qu'on note $(x, y)_A$. La méthode qui suit est une méthode de descente inspirée de la méthode du gradient. La direction de descente w_k n'est plus égale au gradient $g_k = Ax_k - b$: le gradient g_k est "corrigé" de façon que toutes les directions w_k obtenues soient orthogonales (ou conjuguées) pour le produit scalaire $(\cdot, \cdot)_A$. Plus précisément on pose :

$$w_k = g_k + \alpha_k w_{k-1} \ ,$$

où α_k est tel que :$(w_k, w_{k-1})_A = 0$.

Algorithme du Gradient conjugué

1. **Initialisation**
   $k = 0$: choix de $x_0 \in \mathbb{R}^n$ et calcul de $g_0 = Ax_0 - b$.
2. **Itération** k
   (a) Si $g_k = 0$ STOP ;
   (b) Sinon :
   - $w_k = \begin{cases} g_0 & \text{si} \quad k = 0 \\ g_k + \alpha_k\, w_{k-1} & \text{si} \quad k \geq 1 \end{cases}$ avec $\alpha_k = -\dfrac{(g_k, Aw_{k-1})}{(Aw_{k-1}, w_{k-1})}$.
   - $\rho_k = \dfrac{(g_k, w_k)}{(A\, w_k, w_k)}$
   - $x_{k+1} = x_k - \rho_k w_k$,
   - $g_{k+1} = A\, x_{k+1} - b$.

   (c) $k = k + 1$.

Une fois de plus, outre la convergence de la suite des itérés, nous devons assurer son existence c'est-à-dire $w_k \neq 0$ à l'étape 2b. de l'algorithme, ce que nous allons faire en démontrant le résultat suivant.

Théorème 2.4.3 *La méthode du gradient conjugué trouve le minimum d'une fonction quadratique J, où A est symétrique, définie positive, en au plus n itérations où n est l'ordre de A.*

Démonstration - Si $g_k = 0$, alors $x_k = \bar{x}$ est la solution de $Ax = b$.
Pour $k = 1$, nous avons $w_0 = g_0$; donc

$$(g_1, w_0) = (Ax_1 - b, w_0) = (Ax_0 - b, w_0) - \rho_0\,(A\, w_0, w_0) = (g_0, w_0) - \rho_0\,(A\, w_0, w_0) = 0\ ,$$

d'après la définition de ρ_0 ; ceci entraîne aussi

$$(g_1, g_0) = (g_1, w_0) = 0\ ,$$

et

$$(w_1, Aw_0) = (g_1, Aw_0) + \alpha_0\,(w_0, Aw_0) = 0\ ,$$

d'après la définition de α_0.
Nous faisons maintenant l'hypothèse de récurrence suivante :

$$(HR) \qquad \begin{cases} (g_k, g_j) = 0 & \text{pour } 0 \leq j < k\ , \\ (g_k, w_j) = 0 & \text{pour } 0 \leq j < k\ , \\ (w_k, Aw_j) = 0 & \text{pour } 0 \leq j < k\ . \end{cases}$$

Si $g_k \neq 0$, on peut construire l'algorithme au rang $k+1$ c'est-à-dire obtenir $(x_{k+1}, g_{k+1}, w_{k+1})$.
- Par construction, on a vu que $(g_{k+1}, w_k) = 0$.

– Pour $j < k$:

$$(g_{k+1}, w_j) = (g_{k+1}, w_j) - \underbrace{(g_k, w_j)}_{=0} = (g_{k+1} - g_k, w_j) = -\underbrace{\rho_k \, (Aw_k, w_j)}_{HR} = 0 \, .$$

Pour $j \leq k$,

$$(g_{k+1}, g_j) = (g_{k+1}, w_j) - \alpha_j \, (g_{k+1}, w_{j-1}) = 0 \, ,$$

car $g_j = w_j - \alpha_j w_{j-1}$.

– Maintenant : $w_{k+1} = g_{k+1} + \alpha_{k+1} w_k$. Pour $j < k$

$$(w_{k+1}, Aw_j) = (g_{k+1}, Aw_j) + \alpha_{k+1} \underbrace{(w_k, Aw_j)}_{=0} = (g_{k+1}, Aw_j) \ .$$

Comme $g_{j+1} = g_j - \rho_j A w_j$, on obtient

$$(g_{k+1}, Aw_j) = \frac{1}{\rho_j} \, (g_{k+1}, g_j - g_{j+1}) = 0 \text{ si } \rho_j \neq 0 \, .$$

Donc si $\rho_j \neq 0$, $(w_{k+1}, Aw_j) = 0$ pour $j < k$.

– D'autre part $(w_{k+1}, Aw_k) = 0$ par construction. Donc $(w_{k+1}, Aw_j) = 0$ pour $j < k + 1$.

La récurrence est démontrée tant qu'on a $\rho_j \neq 0$ et $g_j \neq 0$.
Mais on a

$$(g_k, w_k) = (g_k, g_k) + \alpha_k \, (g_k, w_{k-1}) = \|g_k\|^2 \, ,$$

et $\rho_k = \dfrac{(g_k, w_k)}{(Aw_k, w_k)}$. Donc ρ_k ne peut s'annuler que si $g_k = 0$, mais alors $x_k = \bar{x}$.
D'autre part

$$\|w_k\|^2 = \|g_k\|^2 + \alpha_k^2 \|w_{k-1}\|^2 \, .$$

Donc si $g_k \neq 0$ alors $w_k \neq 0$. Par conséquent, si les vecteurs $g_0, \ldots, g_k$ sont non nuls, il en est de même pour les vecteurs $w_0, \ldots, w_k$. Ceux-ci forment une famille orthogonale pour le produit scalaire $(\cdot, \cdot)_A$ et les $k + 1$- directions $g_0, \ldots, g_k$ forment une famille orthogonale pour le produit scalaire $(\cdot, \cdot)$. Ces directions sont donc indépendantes. Par suite si $g_0, \ldots, g_{n-1}$ ne sont pas nuls on a nécessairement $w_n = g_n = 0$, ce qui entraîne la convergence de l'algorithme à la n ième itération au plus. □

Remarque 2.4.4 *1. Montrons que*

$$\alpha_k = -\frac{(g_k, Aw_{k-1})}{(Aw_{k-1}, w_{k-1})} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \, .$$

En effet : $Aw_{k-1} = \dfrac{g_{k-1} - g_k}{\rho_{k-1}}$ *et donc* $(g_k, Aw_{k-1}) = -\dfrac{\|g_k\|^2}{\rho_{k-1}}$. *De même :*

$$(w_{k-1}, Aw_{k-1}) = \frac{(w_{k-1}, g_{k-1})}{\rho_{k-1}} = \frac{(g_{k-1} + \alpha_{k-1} w_{k-2}, g_{k-1})}{\rho_{k-1}} = \frac{\|g_{k-1}\|^2}{\rho_{k-1}} \, .$$

2. *Cette méthode est très stable même pour des matrices mal conditionnées (cf Annexe A.) Elle demande $2n^3$ opérations dans le cas d'une matrice pleine et de n itérations. Pour une matrice creuse, le nombre d'opérations diminue beaucoup.*

En pratique, la convergence n'est pas finie à cause des erreurs de précision dues à la machine. On ajoute donc en général une étape 3. "Critère d'arrêt" analogue à celle des algorithmes précédents.

Méthode du Gradient conjugué : cas général

On peut maintenant (indépendamment de la philosophie initiale de la méthode du gradient conjugué) étendre cette méthode au cas de la minimisation d'une fonction J quelconque (Gâteaux-différentiable). Nous présentons l'algorithme ci-dessous, mais cette méthode n'est pas des plus utilisées. Pour un résultat de convergence et plus de détails nous renvoyons à [4].

Algorithme du Gradient conjugué dans le cas général

1. **Initialisation**
   $k = 0$: choix de x_0 dans $\mathbb{R}^n$, de $\varepsilon > 0$ et calcul de $g_0 = \nabla J(x_0)$.
2. **Itération** k
   (a) Si $\|g_k\| \leq \varepsilon$, STOP ;
   (b) Sinon :
   $$w_k = \begin{cases} g_0 & \text{si} \quad k = 0 \\ g_k + \alpha_k\, w_{k-1} & \text{si} \quad k \geq 1 \end{cases} \quad \text{avec } \alpha_k = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}.$$
3. Si $(w_k, g_k) < 0$ aller en 4. Sinon on pose $w_k = g_k$
4. Recherche d'un pas ρ_k approchant le minimum de $J(x_k - \rho w_k)$, par
   $$(\nabla J(x_k - \rho w_k), w_k) = 0\ .$$
5. $x_{k+1} = x_k - \rho_k w_k$, $k = k + 1$.

Dans le cas non linéaire, il n'est pas évident que la direction w_k que l'on construit par "approximations linéaires" soit une direction de descente. Si ce n'est pas le cas, on impose $w_k = g_k$ (on ré-initialise en quelque sorte) : c'est l'objet de l'étape 3. du précédent algorithme.

2.4.4 Méthode de Relaxation

La dernière méthode que nous présentons permet de ramener un problème de minimisation dans $\mathbb{R}^n$ à la résolution successive de n problèmes de minimisation dans $\mathbb{R}$ (à chaque itération).

On cherche à minimiser $J : \mathbb{R}^n \to \mathbb{R}$; posons $X = (x_1, \ldots, x_n)$. Le principe de la méthode est le suivant : étant donné un itéré X^k de coordonnées $(x_1^k, \ldots, x_n^k)$, on fixe toutes les composantes sauf la première et on minimise sur la première :

$$\min J(x, x_2^k, x_3^k, \ldots, x_n^k)\ ,\ x \in \mathbb{R}\ .$$

On obtient ainsi la première coordonnée de l'itéré suivant X^{k+1} que l'on note x_1^{k+1} ; on peut, pour effectuer cette minimisation **dans** $\mathbb{R}$, utiliser par exemple la méthode de Newton dans $\mathbb{R}$. On recommence ensuite en fixant la première coordonnée à x_1^{k+1} et les $n-2$ dernières comme précédemment. On minimise sur la deuxième coordonnée et ainsi de suite. L'algorithme obtenu est le suivant :

Méthode de relaxation successive

1. **Initialisation**
   $k = 0$: choix de $X^o \in \mathbb{R}^n$.
2. **Itération** k ;
   pour i variant de 1 à n, on calcule la solution x_i^{k+1} de
   $$\min J(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x, x_{i+1}^k, \ldots, x_n^k) \ , \ x \in \mathbb{R} \ .$$
3. **Critère d'arrêt**
   Si $\|x_{k+1} - x_k\| < \varepsilon$, STOP
   Sinon, on pose $k = k+1$ et on retourne à 2.

Nous ne détaillerons pas les conditions de convergence de cette méthode. Elles sont analogues aux conditions de convergence de la méthode du gradient et de Newton. Pour plus de détails et pour des exemples d'utilisation on peut se référer à [13].

2.5 Une méthode probabiliste

Cette section est consacrée à la présentation d'un algorithme **stochastique** , c'est-à-dire un algorithme qui fait intervenir des variables aléatoires. La plupart des algorithmes décrit dans les sections précédentes fournissent des minima **locaux**, sauf dans des cas très particuliers (cas convexe par exemple). L'algorithme du **recuit simulé** permet d'obtenir les minimums globaux d'une fonction. On ne parlera que des algorithmes sur des ensembles finis. On suppose en effet qu'on a discrétisé l'ensemble des contraintes.

2.5.1 Dynamique de Métropolis

Soit E un espace **finie**. On considère une fonction V[1] de E dans $\mathbb{R}$ appelée **fonction d'énergie** ou **potentiel** que nous souhaitons minimiser. L'algorithme de **Métropolis** est un algorithme de recherche des minima de V. L'idée heuristique de cette méthode est la suivante : si à l'étape n l'itéré vaut $X_n = x$, on regarde la valeur de V pour un point y **voisin** de x **choisi aléatoirement**. Si $V(y) < V(x)$ on sait que x n'est pas bon et on prend $X_{n+1} = y$. Dans le cas contraire, on aurait envie de prendre $X_{n+1} = x$. Mais on veut éviter de rester piégé en un éventuel minimum local $X_n = x$. Donc on posera $X_{n+1} = y$ si $V(y) - V(x)$ est inférieur à une **variable aléatoire positive simulée**, et $X_{n+1} = x$ dans le cas contraire. L'algorithme s'écrit alors :

[1] Nous changeons de notations et de terminologie : la fonction V n'est autre que la fonction coût J des chapitres précédents restreinte au sous-ensemble fini E.

Dynamique de Métropolis

1. **Initialisation**
   $n = 0$: choix de $X_0 \in E$ déterministe arbitraire.
2. **Itération** n : on observe $X_n = x$.
   On simule une variable aléatoire Y_{n+1} de loi $Q(x, \cdot)$;
   puis on génère un **nombre au hasard** U_{n+1} (de loi uniforme sur $[0, 1]$)
   indépendant de Y_{n+1} et on pose

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{si } V(Y_{n+1}) \leq -\tau \log U_{n+1} + V(x) \\ x & \text{sinon.} \end{cases}$$

τ est un réel positif : c'est la **température**. Q est une matrice (de transition markovienne) sur E, symétrique : c'est la **règle de sélection des voisins**. Cette matrice exprime en général une relation de voisinage : si chaque point x a $r + 1$ voisins, la relation de voisinage étant symétrique, on peut prendre $Q(x, y) = \dfrac{1}{r}$ si x et y sont distincts et voisins.

Nous avons un résultat de convergence mais il nécessite pour être complètement détaillé la théorie des chaînes de Markov. Toutefois nous allons utiliser cette simulation pour décrire l'algorithme du **recuit simulé**. Pour plus de détails (en particulier sur les démonstrations de convergence) on peut se référer à [9].

2.5.2 Recuit simulé sur un ensemble fini

Pour fabriquer un acier de bonne qualité, on le recuit plusieurs fois en effectuant des fusions à des températures décroissantes. Le **recuit simulé** reprend cette idée.

On va utiliser la dynamique de Métropolis à des températures τ_n qui seront choisies décroissantes et on regarde si la suite de variables aléatoires ainsi engendrées converge en probabilité. Précisons :

Recuit simulé

1. **Initialisation**
   $n = 0$: choix de $X_0 \in E$ déterministe arbitraire et de $\tau_{-1} > 0$.
2. **Itération** n : on observe $X_n = x$ et on choisit $\tau_n < \tau_{n-1}$
   On simule une variable aléatoire Y_{n+1} de loi $Q(x, \cdot)$,
   puis on génère un nombre au hasard U_{n+1} (de loi uniforme sur $[0, 1]$)
   indépendant de Y_{n+1} et on pose

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{si } V(Y_{n+1}) \leq -\tau_n \log U_{n+1} + V(x) \\ x & \text{sinon.} \end{cases}$$

On a alors le résultat de convergence suivant ([9]) :

Théorème 2.5.1 *Si la suite τ_n est de la forme $\tau_n = \dfrac{T}{\log n}$ avec $T \geq T_0$ où T_0 est une constante liée à la structure géométrique de V, alors*

$$\lim_{n \to +\infty} P(X_n \in S_V) = 1 \, .$$

S_V désigne l'ensemble des minima (globaux) de V sur E.

Nous avons donc convergence en probabilité de X_n.

Exemple du voyageur de commerce

Soit $\{ 1, \ldots, N \}$ un ensemble de N villes. Le voyageur doit passer dans toutes ces villes en partant de 1 et revenant à 1, en ne passant qu'une fois dans chaque ville ; E est l'ensemble de tous les itinéraires possibles (donc des $(N-1)!$ permutations de $\{ 2, \ldots, N \}$; un **itinéraire** ou un **chemin** est un point de $\mathbb{R}^N$, $x = (x(1), \ldots, x(N))$. On pose $x(1) = 1$ (on part de la ville 1) et $[x(2), \ldots, x(N)]$ est une permutation des autres entiers. On notera $x(N+1) = x(1) = 1$ (on revient à la ville 1). Le coût à minimiser est la somme des distances entre deux points consécutifs d'un itinéraire :

$$V(x) = \sum_{j=1}^{N} d(x(j), x(j+1)) \, . \tag{2.5.1}$$

Les voisins sont classiquement choisis par la relation d'échange de deux étapes :

"x est voisin de y", $x \neq y$, si pour un couple (i, j) avec $2 \leq i < j \leq N$,
$x(i) = y(j)$, $x(j) = y(i)$, les autres coordonnées étant les mêmes (on a échangé deux étapes).

Par exemple : les chemins x="1-2-3-**4-5-6-7**-8-1" et y="1-2-3-**4-6-5-7**-8-1" sont voisins.

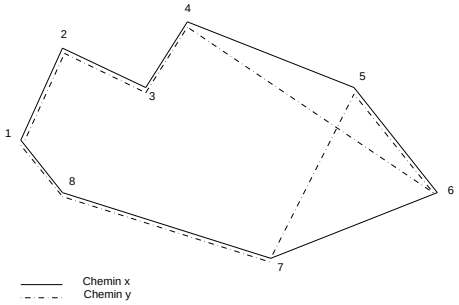

Figure 2.9 : les chemins x et y sont voisins

L'algorithme du recuit simulé s'écrit :

Etant donné $X_n = x$, choisir au hasard un voisin de x, $Y_{n+1} \neq x$, puis simuler une nouvelle variable aléatoire U_{n+1} uniforme sur $[0,1]$:

$$X_{n+1} = \begin{cases} Y_{n+1} & \textit{si } V(Y_{n+1}) \leq \tau_n \log U_{n+1} + V(X_n) \\ X_n & \textit{sinon.} \end{cases}$$

On limite le nombre d'itérations à N.
Du fait du caractère aléatoire de l'algorithme on peut en partant du même point de départ obtenir des résultats différents. Il faut donc en général faire "tourner" un certain nombre de fois la méthode et choisir ainsi le chemin correspondant à la valeur de V la plus petite.

Nous avons fait tourner plusieurs fois la méthode pour différentes valeurs du nombre maximal d'itérations N et une température initiale τ_0 fixée à 100 ; nous avons pris un exemple de 20 villes disposées suivant la figure suivante :

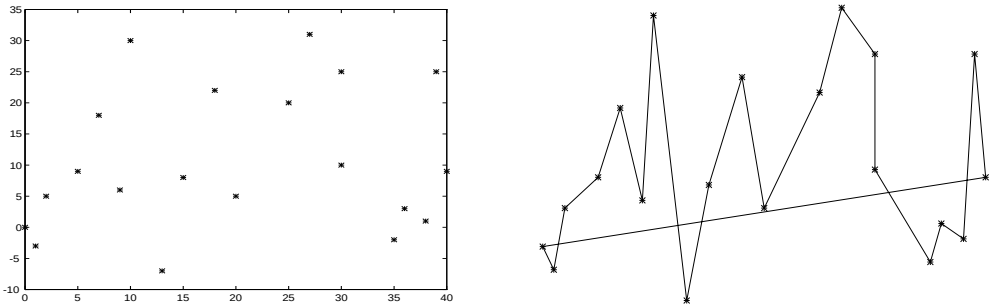

Figure 2.10 : Disposition des villes et chemin initial

Le chemin d'initialisation est le chemin reliant toutes les villes dans l'ordre croissant des abscisses et le valeur de V correspondante est $V = 68.53$. Voici les résultats obtenus sur plusieurs passages avec les valeurs de N et les chemins correspondants :

Passage	1	2	3	4
N	200	500	1000	1000
Valeur de V	57.51	53.90	50.55	49.07

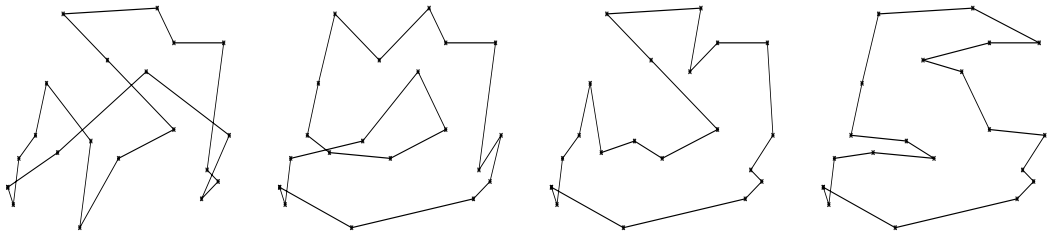

Figure 2.11 : Chemins des passages 1 à 4

Le meilleur résultat est obtenu pour une valeur élevée de N après plusieurs essais : il est consigné au passage 4. Mais il n'est pas certain qu'il n'y en ait pas de meilleur encore ! ! !

A titre d'information nous avons fait tourner le programme 50 fois avec $N = 5000$. Le meilleur résultat obtenu est donné par le chemin de la figure 2.12. Il correspond à une valeur de V égale à 48.125.

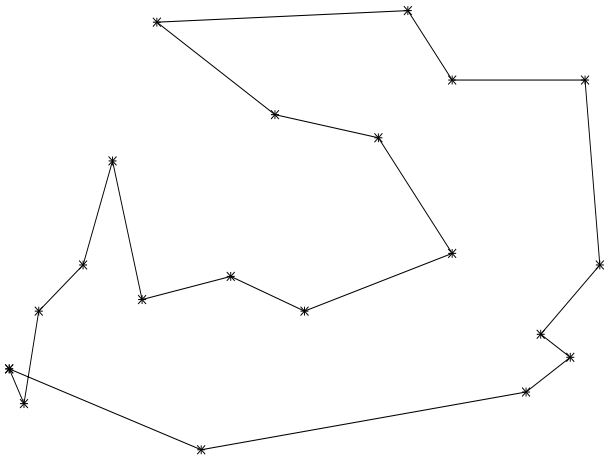

Figure 2.12 : Meilleur chemin obtenu au cours de 50 passages

Voici enfin un exemple plus ludique où on a effectivement essayé de calculer l'itinéraire d'un voyageur de commerce devant passer par 58 villes et partant d'Orléans avant d'y revenir. . .

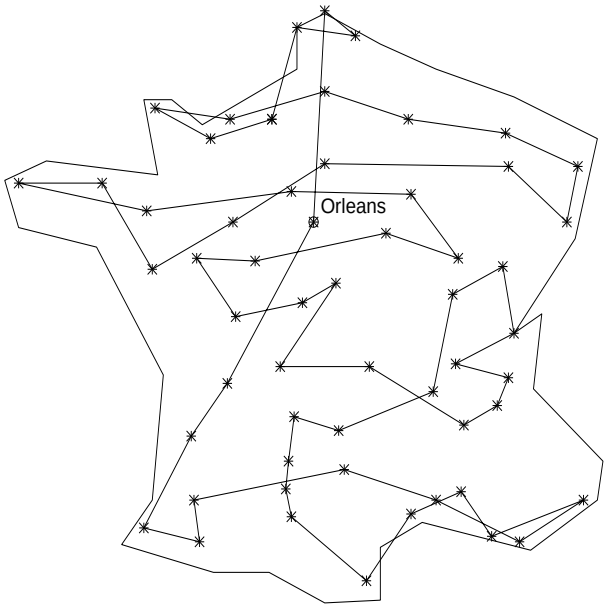

Figure 2.13 : Meilleur chemin obtenu au cours de 50 passages pour N= 10 000

[Travaux Pratiques]

[Algorithmes de minimisation de base]

Programmer les algorithmes de ce chapitre et les tester.
- Gradient à pas constant
- Algorithme de Newton
- Relaxation avec sous-programme Newton (pour $\mathbb{R}$)

Pour chacun d'entre eux, une étude de sensibilité sur le point de départ (initialisation) et le pas éventuel sera menée le plus rigoureusement possible. On fera une comparaison numérique des trois méthodes surtout en termes de
- vitesse de convergence - nombre d'itérations - temps CPU
- Robustesse et domaine de validité

en particulier sur les exemples suivants (f de $\mathbb{R}^2$ dans $\mathbb{R}$)

1. $f(x,y) = x^2 - 5x\,y + y^4 - 25\,x - 8\,y$
2. $f(x,y) = 5x^2 + 5y^2 - x\,y - 11\,x + 11\,y + 11$
3. $f(x,y) = (x^4 - 3)^2 + y^4$
4. $f(x) = x_1^4 - 4x_1^3 + 6(x_1^2 + x_2^2) - 4(x_1 + x_2)$

5. $f(x) = \frac{1}{2}\, x^t G x + c^t x$ avec

$$G = \begin{bmatrix} 2 & -1 & 0 & 0 & .. & .. \\ -1 & 2 & -1 & 0 & .. & .. \\ 0 & -1 & 2 & -1 & 0 & .. \\ .. & \ddots & \ddots & \ddots & .. & .. \\ .. & .. & 0 & -1 & 2 & -1 \\ .. & .. & .. & 0 & -1 & 2 \end{bmatrix} \quad c = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \\ 1 \end{bmatrix}$$

[Gradient conjugué]

Programmer l'algorithme du gradient conjugué pour résoudre Ax=b ; comparer avec les procédures de SCILAB (\ et `linsolve`) en terme de précision et de temps CPU.
On fera des tests sur des matrices définies positives de grande taille et sur l'exemple suivant

$$A = \begin{pmatrix} 0.78 & -0.02 & -0.12 & -0.14 \\ -0.02 & 0.86 & -0.04 & 0.06 \\ -0.12 & -0.04 & 0.72 & -0.08 \\ -0.14 & 0.06 & -0.08 & 0.74 \end{pmatrix} \text{ et } b = \begin{pmatrix} 0.76 \\ 0.08 \\ 1.12 \\ 0.68 \end{pmatrix}.$$

[Méthode de quasi-Newton avec recherche linéaire du pas]

On considère le problème de minimisation sans contraintes sur $\mathbb{R}^n$: $(\mathcal{P})$ $\begin{cases} \min f(x) \\ x \in \mathbb{R}^n \end{cases}$

1. Initialisation : i = 0
   $z_0 \in \mathbb{R}^n$ est tel que l'ensemble $C_{z_0} = \{z \in \mathbb{R}^n |\ f(z) \leq f(z_0)\}$ est borné et le Hessien $H(z) = \dfrac{\partial^2 f(z)}{\partial z^2}$ est défini positif sur cet ensemble.
   Soit $\alpha \in [0, \frac{1}{2}]$.
2. Itération i
   - Calcul de $\nabla f(z_i)$: Si $\nabla f(z_i)$ = 0 , STOP
     SINON : calcul de $H(z_i)$
   - On pose : $h(z_i) = -H(z_i)^{-1} \nabla f(z_i)$
   - Calcul de λ_i par la procédure suivante :
     Soit $\theta_1(\mu, z) = (f(z + \mu h(z)) - f(z)) - \mu(1-\alpha) < \nabla f(z), h(z) >$ et
     $\theta_2(\mu, z) = (f(z + \mu h(z)) - f(z)) - \mu\alpha < \nabla f(z), h(z) >$
     (a) $\mu = 1$
     (b) Calcul de $\theta_1(\mu, z_i)$
     (c) – Si $\theta_1(\mu, z_i)$ = 0 , on pose $\lambda_i = \mu$ et STOP
     – Si $\theta_1(\mu, z_i) < 0$, on pose $\mu = \mu + 1$ et RETOUR à b.
     – Si $\theta_1(\mu, z_i) > 0$ on continue à d.
     (d) Calcul de Calcul de $\theta_2(1, z_i)$

(e) – Si $\theta_2(1, z_i) \leq 0$, on pose $\lambda_i = \mu$ et STOP
– Sinon on pose $t_0 = \mu - 1$, $r_0 = \mu$ et on continue à f.

(f) *(on a* $\lambda_i \in [t_0, r_0]$*)*
On pose j = 0

(g) Calcul de $v_j = \dfrac{t_j + r_j}{2}$, de $\theta_1(v_j, z_i)$ et de $\theta_2(v_j, z_i)$

(h) Si $\theta_1(v_j, z_i) \geq 0$ et $\theta_2(v_j, z_i) \leq 0$ on pose $\lambda_j = v_j$ et STOP
SINON on va à i.

(i) Si $\theta_1(v_j, z_i) > 0$ alors $t_{j+1} = t_j$ et $r_{j+1} = v_j$, j = j +1 et on va à g.
SINON $t_{j+1} = v_j$ et $r_{j+1} = r_j$, j = j +1 et on va à g.

3. $z_{i+1} = z_i + \lambda_i h(z_i)$, i = i +1

[Exercices]

[Existence et conditions d'optimalité]

1. les fonctions J suivantes sont-elles coercives ?
   (a) $J : \mathbb{R} \to \mathbb{R}$ définie par $J(x) = x^3 + x^2 + 1$.
   (b) $J : \mathbb{R}^n \to \mathbb{R}$ définie par $J(x) = (a, x) + b$ avec $a \in \mathbb{R}^n$ et $b \in \mathbb{R}$.
   (c) $J : \mathbb{R}^2 \to \mathbb{R}$ définie par $J(x) = 2x_1^2 + x_2 - 1$.
   (d) $J : \mathbb{R}^2 \to \mathbb{R}$ définie par $J(x) = 2x_1^2 + x_2^3 + 2x_2^2$.
   (e) $J : \mathbb{R}^2 \to \mathbb{R}$ définie par $J(x) = x_1^2 + x_2^2 - 1000\, x_1 - 5000$.

---

2. Soit A une matrice symétrique définie positive à coefficients réels. Montrer qu'il existe une constante $\alpha > 0$ telle que

$$\forall v \in \mathbb{R}^n \qquad (Av, v) \geq \alpha \|v\|^2 ,$$

où $(\cdot, \cdot)$ est le produit scalaire de $\mathbb{R}^n$ et $\| \cdot \|$ la norme euclidienne associée.

---

3. Montrer par un exemple que la condition $\nabla f = 0$ est une condition **nécessaire** d'optimalité et **pas suffisante**.

---

4. Trouver les minima et les maxima sur $\mathbb{R}^2$ de la fonction f définie sur $\mathbb{R}^2$ par :
   (a) $f(x_1, x_2) = x_1^2 - x_1 x_2 + \dfrac{1}{6} x_2^3$
   (b) $f(x_1, x_2) = x_1^2 - 2\, x_1 x_2 + 1$
   (c) $f(x, y) = x^3 + y^3 - 9xy + 27$.

---

5. Soit $J(v) = \frac{1}{2}(Av, v) - (b, v)$, où A est une matrice symétrique de $\mathbb{R}^n$ dans $\mathbb{R}^n$ et $v \in \mathbb{R}^n$, une fonctionnelle quadratique de $\mathbb{R}^n$ dans $\mathbb{R}$. Démontrer les propositions suivantes :
   (a) J est convexe si et seulement si A est semi-définie positive.
   (b) J est strictement convexe si et seulement si A est définie positive.
   (c) $\exists u \in \mathbb{R}^n$ tel que : $\forall v \in \mathbb{R}^n - \{u\}\ \ J(u) < J(v)$ si et seulement si A est définie positive.
   (d) $\exists u \in \mathbb{R}^n$ tel que : $\forall v \in \mathbb{R}^n\ \ J(u) \leq J(v)$ si et seulement si A est semi-définie positive et l'ensemble $\{w \in \mathbb{R}^n \mid Aw = b\}$ n'est pas vide.
   (e) Si la matrice A est semi-définie positive et si l'ensemble $\{w \in \mathbb{R}^n \mid Aw = b\}$ est vide, alors $\inf_{v \in \mathbb{R}^n} J(v) = -\infty$.

---

6. Chercher les dimensions d'un wagon rectangulaire non couvert (ou d'une caisse sans couvercle) telles que pour un volume donné V, la somme des aires des côtés et du plancher soit minimale.

---

7. **Lissage.-** On se propose d'approcher un nuage de points donnés par les couples de réels $(t_i, x_i), i \in \{1, \ldots, N\}$ par une parabole d'équation $x(t) = a\,t^2 + b\,t + c$ où a, b et c sont trois réels à déterminer. Autrement dit, on fait une régression "parabolique".
   (a) Exprimer le problème ci-dessus sous forme de problème de minimisation au sens des moindres carrés. On précisera en particulier la fonction coût, les inconnues et l'ensemble des contraintes.
   (b) Ce problème de minimisation a-t'il une solution ? Pourquoi ? Est-elle unique ?
   (c) Ecrire le système d'optimalité permettant de trouver le minimum.

   On notera S_k la quantité $S_k = \sum_{i=1}^{i=N} t_i^k$.

---

8. **Problèmes de gestion de stocks**

   (a) Soient p_1 = 52 et p_2 = 44 les prix respectifs de deux produits . Soient q_1 et q_2 les quantités respectives de ces produits. Le revenu issu de la vente est donc : $R = p_1\, q_1 + p_2\, q_2$. La fonction coût est : $C = q_1^2 + q_1\, q_2 + q_2^2$ et le bénéfice réalisé est : $\Pi = R - C$. Trouver les quantités q_1 et q_2 maximisant le bénéfice.

   (b) Même problème avec des prix adaptatifs , i.e. variant en fonction de la quantité de produits :

$$\left\{ \begin{array}{rcl} p_1 & = & 256 - 3\, q_1 - q_2 \\ p_2 & = & 222 + q_1 - 5\, q_2 \end{array} \right.$$

[Algorithmes]

9. **Algorithme du gradient à pas constant**

   On veut résoudre le système $Ax = b,\ x \in \mathbb{R}^n$ (avec A symétrique, définie, positive) par une méthode de gradient à pas constant. Soit $\bar{x}$ la solution de ce système. On propose l'algorithme suivant :

$$\left\{ \begin{array}{l} x_0, r_0 = b - Ax_0 \\ x_{k+1} = x_k + \alpha r_k \\ \text{où } r_k = b - Ax_k \end{array} \right.$$

   α est un réel constant .

   (a) Soit $e_k = x_k - \bar{x}$ (pour $k \geq 0$) ; montrer que $e_k = (I - \alpha A)^k e_0$, (pour $k \geq 0$).

   (b) Soient $0 < \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1$ les valeurs propres de A. Montrer que l'algorithme converge si et seulement si $0 < \alpha < \dfrac{2}{\lambda_1}$.

   (c) Montrer que le meilleur choix de α est : $\alpha_{opt} = \dfrac{2}{\lambda_1 + \lambda_n}$ et qu'alors :

$$\rho(I - \alpha_{opt} A) = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}$$

   où $\rho(M)$ désigne le rayon spectral de la matrice M.

---

10. **Algorithme du gradient à pas optimal**

    On veut résoudre le système suivant par une méthode de gradient à paramètre optimal :

$$\left\{ \begin{array}{l} \dfrac{1}{2} x = 0 \\ \dfrac{c}{2} y = 0 \end{array} \right. \qquad \text{où} c \geq 1.$$

    (a) Ecrire le système sous la forme $Ax = b$ et calculer les valeurs propres de A.

    (b) Soit r le résidu : $b - Ax$. Calculer r et le paramètre α correspondant à la minimisation sur $\mathbb{R}$ de la fonction qui à α associe $J(x_k + \alpha r_k)$.

(c) Soit P_k le point de coordonnées x_k et y_k. Exprimer x_{k+1} et y_{k+1} en fonction de x_k et y_k.

(d) Soit $t_k = \dfrac{y_k}{x_k}$ la pente de la droite (OP$_k$). Exprimer t_{k+2} en fonction de t_k. Interprétation géométrique. Conclusion ?

(e) Soit $t \in \{t_k, t_{k+1}\}$. (k donné)
On appelle τ le facteur moyen de réduction de l'erreur : $\tau^2 = \dfrac{y_{k+2}}{y_k} = \dfrac{x_{k+2}}{x_k}$.
Montrer que :

$$\tau^2 = [\frac{c-1}{c+1}]^2 \frac{1}{1 + \frac{c}{(c+1)^2}(ct - \frac{1}{ct})^2}$$

Pour quelle valeur de t , τ est-il maximum ?

---

11. **Algorithme du gradient conjugué - 1**
On note (x, y) le produit scalaire euclidien de $\mathbb{R}^n$, $x^t y$ sous forme matricielle, u_i le vecteur propre associé à une valeur propre λ_i et W_k le sous-espace engendré par les k vecteurs propres $(u_i)_{1 \leq i \leq k}$. A est une matrice symétrique, définie positive dont les valeurs propres λ_i sont rangées par ordre décroissant. On appelle **quotient de Rayleigh** de la matrice A l'application de $\mathbb{R}^n - \{0\}$ vers $\mathbb{R}$ définie par :

$$R_A(x) = \frac{(Ax, x)}{(x, x)}$$

Montrer que :

(a) $\lambda_k = R_A(u_k)$.

(b) $\lambda_k = \min\limits_{x \in W_k} R_A(x)$.

(c) $\lambda_k = \max\limits_{x \in W_{k-1}^{\perp}} R_A(x)$.

(d) Pour $x \neq 0$ et λ scalaire quelconque, on définit $\eta = Ax - \lambda x$.
Montrer que

$$\min_{i \in \{1, \cdots, n\}} |\lambda - \lambda_i| \leq \frac{\|\eta\|_2}{\|x\|_2}$$

et que η est minimum au sens de la norme euclidienne pour $\lambda = R_A(x)$.

---

12. **Algorithme du gradient conjugué - 2**
Soit A une matrice carrée d'ordre N symétrique , définie positive . Deux vecteurs $u \neq 0$ et $v \neq 0$ sont dits A-conjugués si $(Av, u) = 0$.

(a) Montrer que si les vecteurs $v_0, v_1, \cdots, v_{N-1}$ sont A-conjugués deux à deux, ils forment une base de $\mathbb{R}^N$.

(b) On définit les deux suites de matrices suivantes :

$$C_k = \sum_{i=0}^{k-1} \frac{v_i v_i^t}{(Av_i, v_i)} , \qquad D_k = I - C_k A .$$

Montrer que pour $0 \leq j \leq k-1$: $\left\{ \begin{array}{lcl} C_k A v_j & = & v_j \\ D_k v_j & = & 0 \\ D_k^t A v_j & = & 0 \end{array} \right.$.

Que valent alors D_N et C_N ?

(c) Supposons que $v_0, v_1, \cdots, v_{k-1}$ soient connus .
Si D_k = 0 , que peut-on conclure ?
Sinon , soit $v \in \mathbb{R}^n$ tel que $D_k v \neq 0$. Montrer que $v_k = D_k v$ est A-conjugué par rapport à $v_0, v_1, \cdots, v_{k-1}$.

(d) Ecrire un algorithme qui construit la suite $v_1, \cdots, v_{N-1}$ à partir de $v_0 \in \mathbb{R}^N - \{0\}$, donné.
Pour cela on pourra considérer, tant que $D_k \neq 0$, un vecteur de la forme $D_k e_i$ où e_i est le ième vecteur de la base canonique.
En déduire un algorithme pour calculer A^{-1} .

---

13. **Méthode de Newton - 1**
Vérifier que le calcul de l'inverse d'un scalaire α par la méthode de Newton correspond à la méthode itérative :

$$x_{k+1} = x_k(2 - \alpha x_k) \ , \ k \geq 0$$

Construire , par analogie , une méthode itérative d'approximation de l'inverse d'une matrice inversible A , de la forme :

$$\begin{array}{c} B_0 \text{ matrice arbitraire} \\ B_{k+1} = \text{ fonction}(B_k, A), \ k \geq 0 \end{array}$$

Démontrer qu'une CNS de convergence de cette méthode est : $\rho(I - AB_0) < 1$ où $\rho(M)$ désigne le rayon spectral de la matrice M.
Supposant la matrice A symétrique , définie , positive et supposant connu son rayon spectral, comment choisir simplement la matrice B_0 pour vérifier la condition précédente ?

---

14. **Méthode de Newton relaxée**
Une variante de la méthode de Newton pour la résolution des systèmes d'équations non

linéaires est la méthode de *Gauss-Seidel* qui se présente sous la forme suivante :

$$\begin{aligned}
x_1^{(k+1)} &= x_1^{(k)} - \frac{f_1(x_1^{(k)}, x_2^{(k)}, \cdots, x_n^{(k)})}{\partial_1 f_1(x_1^{(k)}, x_2^{(k)}, \cdots, x_n^{(k)})} \\
x_2^{(k+1)} &= x_2^{(k)} - \frac{f_2(x_1^{(k+1)}, x_2^{(k)}, \cdots, x_n^{(k)})}{\partial_2 f_2(x_1^{(k+1)}, x_2^{(k)}, \cdots, x_n^{(k)})} \\
&\vdots \\
x_n^{(k+1)} &= x_n^{(k)} - \frac{f_n(x_1^{(k+1)}, x_2^{(k+1)}, \cdots, x_{n-1}^{(k+1)}, x_n^{(k)})}{\partial_n f_n(x_1^{(k+1)}, x_2^{(k+1)}, \cdots, x_{n-1}^{(k+1)}, x_n^{(k)})}
\end{aligned}$$

où $\partial_i = \dfrac{\partial}{\partial x_i}$.

Montrer que si les fonctions f_i sont affines : $f_i(x) = \displaystyle\sum_{j=1}^{n} a_{ij}x_j - b_i$, cette méthode n'est autre que la méthode itérative de Gauss-Seidel pour le résolution des systèmes linéaires.

Chapitre 3

Minimisation avec contraintes

Dans ce chapitre on s'intéresse au cas où le problème de minimisation comporte des contraintes. Plus précisément on se donne un sous-ensemble C non vide, fermé de $\mathbb{R}^n$ et on étudie le problème

$$(\mathcal{P}) \qquad \min\ J(x)\ ,\ x \in C\ .$$

C est l'ensemble des contraintes.
Dans tout le chapitre $\|\cdot\|$ désigne la norme euclidienne de $\mathbb{R}^n$ et $(\cdot,\cdot)$ le produit scalaire associé.

3.1 Résultats d'existence et d'unicité

Commençons par donner un résultat d'existence.

Théorème 3.1.1 *Supposons que J est continue, que C est un sous-ensemble fermé non vide de $\mathbb{R}^n$ et que l'une des conditions suivantes est réalisée :*

1. *soit C est borné,*
2. *soit J est coercive.*

Alors le problème $(\mathcal{P})$ admet au moins une solution.

Démonstration - La démonstration est la même que dans le cas sans contraintes. On montre qu'une suite minimisante est bornée soit parce qu'elle est dans C qui est borné, soit parce que la fonctionnelle J est coercive. La limite de la sous-suite extraite est alors dans C puisque cet ensemble est fermé. C'est donc une solution de $(\mathcal{P})$. □

Exemple 3.1.1 *Nous traiterons plus particulièrement le cas où C est défini par des égalités et des inégalités :*

$$C = \{\ x \in \mathbb{R}^n\ \mid\ h(x) = 0\ ,\ g(x) \leq 0\ \}\ , \tag{3.1.1}$$

où

- *$h : \mathbb{R}^n \to \mathbb{R}^p$ représente p contraintes en égalité avec $h(x) = (h_1(x), \ldots, h_p(x))$; h est supposée continue.*

– $g : \mathbb{R}^n \to \mathbb{R}^q$ *représente* q *contraintes en inégalité avec* $g(x) = (g_1(x), \ldots, g_q(x))$ *;* g *est supposée continue.*

Dans ce cas C *est un ensemble fermé.*

Exemple 3.1.2

$$C = \{\, x = (x_1, x_2) \in \mathbb{R}^2 \mid\ x_1^2 + x_2^2 \leq 1\} \,,$$

est un ensemble fermé borné. En revanche

$$\tilde{C} = \{\, x = (x_1, x_2) \in \mathbb{R}^2 \mid\ x_1^2 + x_2^2 < 1\} \,,$$

est borné mais pas fermé.

Les solutions de $(\mathcal{P})$ ne sont en général pas uniques mais nous pouvons donner un cas important où nous avons unicité :

Théorème 3.1.2 *Sous les hypothèses du théorème (3.1.1), si* J *est strictement convexe et si* C *est convexe, alors le problème* $(\mathcal{P})$ *admet une solution unique.*

Démonstration - La démonstration est exactement celle du théorème 2.1.2. □

Exemple 3.1.3 *C'est le cas si* J *est quadratique avec* A *définie positive et si* C *est convexe.*

De plus, l'ensemble C *défini par la relation (3.1.1) est convexe si les fonctions* h_i *sont affines et les* g_j *sont convexes.*

3.2 Conditions d'optimalité du premier ordre

Tout comme dans le cas sans contraintes, nous allons établir des conditions d'optimalité du premier et du second ordre permettant de calculer les éventuels minima de J.

3.2.1 Condition d'optimalité du premier ordre générale

La condition nécessaire suivante est l'analogue de celle que nous avons dans le cas sans contraintes. Elle fait bien sûr intervenir l'ensemble des contraintes.

Théorème 3.2.1 (Condition nécessaire du premier ordre)
Si J *est une fonction (Gâteaux) différentiable et si* C *est un convexe fermé, alors toute solution* x^* *de* $(\mathcal{P})$ *vérifie la condition nécessaire d'optimalité du premier ordre :*

$$\forall x \in C \qquad (\nabla J(x^*), x - x^*) \geq 0 \,. \tag{3.2.1}$$

Démonstration - Soit x^* une solution de $(\mathcal{P})$ et $x \in C$. Par convexité de C, $x^* + t(x - x^*)$ est un élément de C pour tout $t \in [0, 1]$ et donc

$$J(x^* + t(x - x^*)) - J(x^*) \geq 0 \,.$$

On divise ensuite par t (> 0) et on fait tendre t vers 0^+. □

Nous avons un résultat intéressant dans le cas convexe puisqu'il donne une condition nécessaire et suffisante.

Théorème 3.2.2 (CNS du premier ordre dans le cas convexe)
Supposons J convexe, Gâteaux-différentiable et C convexe fermé ; soit x^* un élément de C. La condition (3.2.1) est nécessaire et suffisante pour que x^* soit solution de $(\mathcal{P})$.
Elle caractérise donc les minima de J sur C.

Démonstration - Il reste à montrer que la condition est suffisante. Soit x^* un élément de C. Grâce à la convexité de J, nous savons (théorème 1.3.2 et relation (1.3.1)) que

$$\forall x \in C \qquad J(x) \geq J(x^*) + (\nabla J(x^*), x - x^*) \ ;$$

Il est alors immédiat de voir que la condition (3.2.1) implique que $J(x) \geq J(x^*)$ pour tout $x \in C$. □

On peut remarquer que si $C = \mathbb{R}^n$ (cas sans contraintes) la condition (3.2.1) est équivalente à $\nabla J(x^*) = 0$. On retrouve ainsi la condition (3.1.1).

Nous allons maintenant détailler la condition (3.2.1) qui reste tout de même très abstraite, dans le cas où C est donné par la relation (3.1.1) de l'exemple 3.1.1. Nous donnons d'abord à titre d'exemple le cas où il n'y a que des contraintes en égalité avant de présenter le cas général.

3.2.2 Contraintes en égalité

Le problème $(\mathcal{P})$ se réduit à

$$\min \ J(x) \ , \ x \in \mathbb{R}^n, \ h(x) = 0 \ ,$$

avec $h(x) = (h_1(x), \ldots, h_p(x))$ et h est continue de $\mathbb{R}^n$ dans $\mathbb{R}^p$.

Théorème 3.2.3 (CN du premier ordre-contraintes en égalité)
On suppose que
- *J et h sont de classe $\mathcal{C}^1$ sur $\mathbb{R}^n$,*
- *le problème $(\mathcal{P})$ a une solution x^*,*
- *les p vecteurs de $\mathbb{R}^n$: $(\nabla h_1(x^*), \ldots, \nabla h_p(x^*))$ sont linéairement indépendants (et donc $p \leq n$) ;*

alors il existe p réels $(\lambda_1^*, \ldots, \lambda_p^*)$ tels que

$$\nabla J(x^*) + \sum_{j=1}^{p} \lambda_j^* \, \nabla h_j(x^*) = 0. \tag{3.2.2}$$

Démonstration - Ce résultat est un cas particulier du theorème 3.2.5. Nous renvoyons donc à la démonstration de ce théorème. □

Définition 3.2.1 *Les réels λ_j^* obtenus par le théorème précédent sont appelés des* ***multiplicateurs de Lagrange*** .

3.2.3 Contraintes en égalité et en inégalité

Le problème $(\mathcal{P})$ est de la forme

$$\min\ J(x)\ ,\ x \in C$$

où

$$C = \{\ x \in \mathbb{R}^n \mid\ h(x) = 0\ ,\ g(x) \leq 0\ \}\ .$$

Théorème 3.2.4 (Conditions d'optimalité non qualifiées)
On suppose que $J,\ h$ et g sont de classe $\mathcal{C}^1$. Soit x^* une solution du problème $(\mathcal{P})$. Alors il existe $\lambda^* = (\lambda_1^*, \ldots, \lambda_p^*) \in \mathbb{R}^p,\ \mu^* = (\mu_1^*, \ldots, \mu_q^*) \in \mathbb{R}^{+,q}$ et $\mu_0^* \in \mathbb{R}^+$ tels que

$$\forall j \in \{\ 0, \ldots, q\ \} \qquad \mu_i^* \geq 0\ , \tag{3.2.3a}$$

$$h(x^*) = 0\ ,\ g(x^*) \leq 0\ , \tag{3.2.3b}$$

$$\forall j \in \{\ 1, \ldots, q\ \} \qquad \mu_j^*\ g_j(x^*) = 0\ , \tag{3.2.3c}$$

$$\mu_0^*\ \nabla J(x^*) + \sum_{i=1}^{p}\ \lambda_i^*\ \nabla h_i(x^*) + \sum_{j=1}^{q}\ \mu_j^*\ \nabla g_j(x^*) = 0. \tag{3.2.3d}$$

Démonstration - Nous allons démontrer ce résultat par une méthode de pénalisation comme dans [15]. Pour tout entier k on considère le problème :

$$(\mathcal{P}_k) \qquad\qquad \min\ J_k(x)\ \ ,\ \ x \in \mathcal{B}(x^*, \rho)\ ,$$

où

$$J_k(x) = J(x) + \frac{k}{2}\sum_{i=1}^{p}[h_i(x)]^2 + \frac{k}{2}\sum_{j=1}^{q}[g_j^+(x)]^2 + \|x - x^*\|^2\ , \tag{3.2.4}$$

où $g_j^+(x) = \max(0, g_j(x))$ et $\mathcal{B}(x^*, \rho)$ est une boule fermée (compacte) non vide ($\rho > 0$), centrée en x^*. Il est facile de voir que si g_j est de classe $\mathcal{C}^1$ alors $(g_j^+)^2$ est différentiable et que $\dfrac{d(g_j^+)^2}{dx}(x) = 2\,\dfrac{dg_j}{dx}(x) \cdot g_j^+(x)$ (cf. exercices en fin de chapitre).

• **Le problème $(\mathcal{P}_k)$ a au moins une solution x_k.**
En effet J_k est continue (et même différentiable). Elle atteint donc son minimum sur le compact $\mathcal{B}(x^*, \rho)$.

• **La suite (x_k) converge vers (x^*).**
En effet, la suite (x_k) est dans le compact $\mathcal{B}(x^*, \rho)$ et on peut donc en extraire une sous-suite (que nous noterons de la même façon) qui converge vers $\tilde{x} \in \mathcal{B}(x^*, \rho)$. Comme

$$J_k(x_k) \leq J_k(x^*) = J(x^*) < +\infty\ ,$$

nous avons

$$\sum_{i=1}^{p}[h_i(x_k)]^2 + \sum_{j=1}^{q}[g_j^+(x_k)]^2 \leq \frac{2}{k}\left[J(x^*) - J(x_k) - \|x_k - x^*\|^2\right]\ . \tag{3.2.5}$$

L'expression du membre de droite entre crochets est bornée : donc

$$\forall i = 1, \dots, p \quad \lim_{k \to +\infty} h_i(x_k) = h_i(\tilde{x}) = 0 \text{ et } \forall j = 1, \dots, q \quad \lim_{k \to +\infty} g_i^+(x_k) = g_i^+(\tilde{x}) = 0 \,.$$

Par conséquent $\tilde{x}$ est réalisable. D'autre part

$$J(x_k) + \|x_k - x^*\|^2 \leq J_k(x_k) \leq J(x^*) \,.$$

En passant à la limite on obtient

$$J(\tilde{x}) + \|\tilde{x} - x^*\|^2 \leq J(x^*);$$

comme x^* est une solution de $(\mathcal{P})$ nous avons finalement

$$J(\tilde{x}) + \|\tilde{x} - x^*\|^2 \leq J(x^*) \leq J(\tilde{x}) \,,$$

ce qui signifie : $\tilde{x} = x^*$. Ce raisonnement peut être fait pour toute les valeurs d'adhérence $\tilde{x}$ de la suite (x_k). Par conséquent **toute** la suite converge vers x^*.

• **Ecrivons les conditions d'optimalité pour $(\mathcal{P}_k)$.**

Comme la suite (x_k) converge vers x^* elle est dans l'intérieur de $\mathcal{B}(x^*, \rho)$ à partir d'un certain rang et donc x^k est un minimum local (sans contraintes) de J_k. Par conséquent $\nabla J_k(x_k) = 0$ pour k assez grand, c'est-à-dire

$$\nabla J(x_k) + k \sum_{i=1}^{p} h_i(x_k) \nabla h_i(x_k) + k \sum_{j=1}^{q} g_j^+(x_k) \nabla g_j(x_k) + 2(x_k - x^*) = 0 \,. \tag{3.2.6}$$

Posons

$$s_k = \left(1 + k^2 \sum_{i=1}^{p} [h_i(x_k)]^2 + k^2 \sum_{j=1}^{q} [g_j^+(x_k)]^2 \right)^{\frac{1}{2}} \text{ et}$$

$$\mu_0^k = \frac{1}{s_k} \,,\ \lambda_i^k = \frac{k\, h_i(x_k)}{s_k} \,, i = 1, \dots, p \text{ et } \mu_j^k = \frac{k\, g_j^+(x_k)}{s_k} \,, j = 1, \dots, q \,.$$

La relation (3.2.6) devient

$$\mu_0^k \nabla J(x_k) + \sum_{i=1}^{p} \lambda_i^k \nabla h_i(x_k) + \sum_{j=1}^{q} \mu_j^k \nabla g_j(x_k) + \frac{2}{s_k}(x_k - x^*) = 0 \,. \tag{3.2.7}$$

Comme le vecteur $(\lambda_1^k, \dots, \lambda_p^k, \mu_0^k, \mu_1^k, \dots, \mu_q^k)$ de $\mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^q$ est de norme 1 on peut extraire une sous-suite convergente vers $(\lambda_1^*, \dots, \lambda_p^*, \mu_0^*, \mu_1^*, \dots, \mu_q^*) \neq 0$ et passer à la limite dans (3.2.7). On obtient alors

$$\mu_0^* \nabla J(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{q} \mu_j^* \nabla g_j(x^*) + = 0 \,, \tag{3.2.8}$$

car toutes les fonctions considérées sont continues et $s_k \geq 1$.
D'autre part il est clair que $\mu_j^* \geq 0, j = 0, \ldots, q$.
• **Relation 3.2.3c**
Enfin si $g_j(x^*) < 0$, alors $g_j(x_k) < 0$ à partir d'un certain rang et $\mu_j^k = 0$. par conséquent à la limite $\mu_j^* = 0$. □

Remarque 3.2.1 *Si on remplace l'ensemble C par $\mathcal{O} \cap C$ où $\mathcal{O}$ est un ouvert de $\mathbb{R}^n$ dans la démonstration précédente, on obtient le même résultat pour un minimum* ***local*** *de J sur C.*

Les réels λ_j^* et μ_i^* sont des multiplicateurs de Lagrange. La relation (3.2.3b) est une relation de **réalisabilité**. La relation (3.2.3c) est une relation de **complémentarité**.

Les conditions obtenues ci-dessus sont dites **non qualifiées.** Dans ce cas, en effet ces relations peuvent être "dégénérées" ; en effet le réel μ_0^* peut être nul et on n'a donc aucun renseignement sur le minimum de la fonction J puisqu'elle n'apparaît alors nulle part dans les relations d'optimalité. Il est donc important de donner des conditions qui permettent d'assurer que μ_0^* est non nul. Dans ce cas, quitte à tout diviser par μ_0^* on pourra le supposer égal à 1. De telles conditions sont appelées conditions de **qualification** ou de régularité. Lorsqu'elles sont vérifiées le problème est dit **qualifié** .

Définition 3.2.2 (Contrainte active)
On dit qu'une contrainte en inégalité g_j est ***active*** *(ou* ***saturée****) au point x^* de $\mathbb{R}^n$ si $g_j(x^*) = 0$.*
Une contrainte qui n'est pas active est ***inactive*** *.*
On note $I(x^*)$ l'ensemble des indices j correspondant aux contraintes actives en x^*.

Définition 3.2.3 (Point régulier)
On dit qu'un élément x^* de $\mathbb{R}^n$ est ***régulier*** *pour les contraintes h et g,*
• *s'il est réalisable : $h(x^*) = 0,\ g(x^*) \leq 0$,*
• *les vecteurs $\nabla h_i(x^*),\ i = 1, \ldots, p$ sont indépendants.*
• *et si on peut trouver $d \neq 0 \in \mathbb{R}^n$ tel que*

$$(\nabla h_i(x^*), d) = 0,\ i = 1, \ldots, p \ \text{ et } \ (\nabla g_j(x^*), d) < 0,\ \ \forall j \in I(x^*)\ .$$

On dit ausi que x^* vérifie la contrainte de qualification de MANGASARIAN-FROMOWITZ*, que nous noterons ici (CQ1).*

Il existe de nombreux critères de régularité pour lesquels les théorèmes suivants sont vrais. La condition (CQ2) suivante est plus forte que la condition définie ci-dessus (voir exercice).

Définition 3.2.4 *On dit qu'un élément x^* de $\mathbb{R}^n$ vérifie la condition de qualification (CQ2) pour les contraintes h et g,*
- s'il est réalisable ($h(x^*) = 0,\ g(x^*) \leq 0$)
- et si les vecteurs $\nabla h_i(x^*), \nabla g_j(x^*),\ 1 \leq i \leq p,\ j \in I(x^*)$, sont indépendants.

Théorème 3.2.5 (Conditions de Karush-Kuhn-Tucker)
On suppose que $J,\ h$ et g sont de classe $\mathcal{C}^1$. Soit x^* une solution du problème $(\mathcal{P})$. On suppose

que x^* est régulier pour les contraintes h et g. Alors il existe $\lambda^* = (\lambda_1^*, \ldots, \lambda_p^*) \in \mathbb{R}^p$ et $\mu^* = (\mu_1^*, \ldots, \mu_q^*) \in \mathbb{R}^q$ tels que

$$\forall j \in \{\, 1, \ldots, q \,\} \qquad \mu_j^* \geq 0 \,, \tag{3.2.9a}$$

$$h(x^*) = 0 \,,\ g(x^*) \leq 0 \,, \tag{3.2.9b}$$

$$\forall j \in \{\, 1, \ldots, q \,\} \qquad \mu_j^*\, g_j(x^*) = 0 \,, \tag{3.2.9c}$$

$$\nabla J(x^*) + \sum_{i=1}^{p} \lambda_i^*\, \nabla h_i(x^*) + \sum_{j=1}^{q} \mu_j^*\, \nabla g_j(x^*) = 0. \tag{3.2.9d}$$

Démonstration - Il faut montrer que sous l'hypothèse de régularité (CQ1), le réel μ_0^* donné par le thórème 3.2.4 est non nul. Supposons donc que $\mu_0^* = 0$.
Supposons que tous les $\mu_j^*, j \in I(x^*)$ sont tous nuls. Alors le vecteur $(\lambda_1^*, \ldots, \lambda_p^*)$ n'est pas nul et nous avons alors

$$\sum_{j=1}^{p} \lambda_j^*\, \nabla h_j(x^*) = 0 \,,$$

ce qui contredit l'indépendance linéaire des $(\nabla h_j(x^*))$.
Par conséquent, il existe $j_0 \in I(x^*)$ tel que $\mu_{j_0}^* \neq 0$. Nous avons alors en prenant la direction d donnée par (CQ1)

$$0 = \sum_{i=1}^{p} \lambda_i^*\, (\nabla h_i(x^*), d) + \sum_{j=1}^{q} \mu_j^*\, (\nabla g_j(x^*), d) \leq \mu_{j_0}^*\, (\nabla g_{j_0}(x^*), d) < 0 \,;$$

d'où la contradiction. □
L'ensemble des équations (3.2.9) du théorème précédent sont appelées les conditions de KARUSH-KUHN-TUCKER en abrégé (KKT).

Définition 3.2.5 *On appelle **Lagrangien** du problème $(\mathcal{P})$ la fonction définie sur $\mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q$ par*

$$\mathcal{L}(x, \lambda, \mu) = J(x) + \sum_{i=1}^{p} \lambda_i\, h_i(x) + \sum_{j=1}^{q} \mu_j\, g_j(x) \,. \tag{3.2.10}$$

La relation (3.2.9d) s'écrit alors

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0 \,,$$

où ∇_x désigne le gradient par rapport à la première variable.
Le cas convexe est une fois de plus un cas particulier important :

Théorème 3.2.6 (CNS dans le cas convexe)
On suppose que $J,\ h$ et g sont $\mathcal{C}^1$, que $J,\ g$ sont convexes, h est affine et que x^* est régulier pour les contraintes h et g. Alors x^* une solution du problème $(\mathcal{P})$ si et seulement si les conditions (3.2.9) sont satisfaites.

Démonstration - Il faut prouver que les conditions (3.2.9) sont suffisantes pour que le point x^* soit une solution de $(\mathcal{P})$. Grâce à (3.2.9b), x^* est bien réalisable. Comme toutes les fonctions sont convexes le Lagrangien est convexe par rapport à la variable x et la condition (3.2.9d) est équivalente à dire que x^* est un minimum de $x \mapsto \mathcal{L}(x, \lambda^*, \mu^*)$. On obtient donc

$$\forall x \in \mathbb{R}^n \qquad \mathcal{L}(x^*, \lambda^*, \mu^*) \leq \mathcal{L}(x, \lambda^*, \mu^*) \; .$$

Si $x \in C$ alors $h(x) = 0$ et $g(x) \leq 0$ de sorte que

$$\sum_{i=1}^{p} \lambda_i^* \; h_i(x) + \sum_{j=1}^{q} \mu_j^* \; g_j(x) = \sum_{j=1}^{q} \mu_j^* \; g_j(x) \leq 0 \; ,$$

puisque $\mu_j^* \geq 0$ d'après (3.2.9a). De plus avec la relation de complémentarité (3.2.9c), on voit que $\mathcal{L}(x^*, \lambda^*, \mu^*) = J(x^*)$. Finalement, nous obtenons

$$\forall x \in C \qquad J(x^*) = \mathcal{L}(x^*, \lambda^*, \mu^*) \leq \mathcal{L}(x, \lambda^*, \mu^*) \leq J(x) \; ,$$

ce qui prouve que x^* est solution de $(\mathcal{P})$. □

3.3 Conditions d'optimalité du deuxième ordre

3.3.1 Conditions d'optimalité nécessaires du deuxième ordre

Les conditions données dans la section précédente sont nécessaires. Elles permettent de déterminer les bons "candidats" à la solution de $(\mathcal{P})$, c'est-à-dire les points critiques du Lagrangien. Il faut maintenant avoir des critères permettant de conclure et de savoir si le point obtenu est un minimum ou non. Comme dans le cas sans contraintes, nous pouvons dans un premier temps restreindre le nombre de "candidats" grâce à une condition nécessaire du second ordre.

Théorème 3.3.1 *On suppose que J, h et g sont de classe $\mathcal{C}^2$, que x^* est un minimum (local) de J sur C et que la condition la condition (CQ2) de la définition 3.2.4 est vérifiée.*
Alors il existe $\lambda^* = (\lambda_1^*, \ldots, \lambda_p^*) \in \mathbb{R}^p$ et $\mu^* = (\mu_1^*, \ldots, \mu_q^*) \in \mathbb{R}^q$ tels que
• *Les relations de KKT (3.2.9) sont satisfaites et*
• *Pour toute direction $d \in \mathbb{R}^n$ vérifiant*

$$\left\{ \begin{array}{ll} (\nabla h_i(x^*), d\,) = 0 & \textit{pour } i = 1, \ldots, p \\ (\nabla g_j(x^*), d\,) = 0 & \textit{pour } j \in I^+(x^*) \\ (\nabla g_j(x^*), d\,) \leq 0 & \textit{pour } j \in I(x^*) \backslash I^+(x^*) \end{array} \right. \tag{3.3.1}$$

où

$$I^+(x^*) = \{\; j,\; 1 \leq j \leq q \;\mid\; g_j(x^*) = 0 \textit{ et } \mu_j^* > 0 \;\} \; ,$$

on a

$$\left(\nabla^2_{xx} \mathcal{L}(x^*, \lambda^*, \mu^*) d, d\right) \geq 0 \; , \tag{3.3.2}$$

$\nabla^2_{xx} \mathcal{L}(x, \lambda, \mu)$ *désignant la dérivée seconde de $\mathcal{L}$ au point (x, λ, μ).*

Définition 3.3.1 *L'ensemble $I^+(x^*)$ est l'ensemble des contraintes **fortement actives**.*
Lorsque $I^+(x^*) = I(x^*)$ c'est-à-dire

$$g_j(x^*) = 0 \Longleftrightarrow \mu_j^* > 0 \ ,$$

*on dit qu'il y a **stricte complémentarité**.*

Nous pouvons aussi donner une condition suffisante du second ordre :

Théorème 3.3.2 *On suppose que J, h et g sont de classe $\mathcal{C}^2$; soit x^* un point de $\mathbb{R}^n$ vérifiant les conditions (nécessaires) de KKT (3.2.9) avec les multiplicateurs (λ^*, μ^*).*
Si la matrice hessienne du Lagrangien au point (x^*, λ^*, μ^*) :

$$H(x^*) = \nabla^2_{xx}\mathcal{L}(x^*, \lambda^*, \mu^*) = \nabla^2 J(x^*) + \sum_{i=1}^{p} \lambda_i^* \, \nabla^2 h_i(x^*) + \sum_{j=1}^{q} \mu_j^* \, \nabla^2 g_j(x^*) \ ,$$

est définie positive sur le sous-espace

$$\mathcal{T} = \{ \, d \in \mathbb{R}^n \ \mid \ (\nabla h_i(x^*), d\,) = 0, \ i = 1, \ldots, p, \ (\nabla g_j(x^*), d\,) = 0, \ \textit{pour tout } j \in I^+(x^*) \, \} \ ,$$

alors x^* est un minimum (local) strict de J sur C.

Les démonstrations des deux théorèmes précédents nécessitent d'introduire la notion de cône tangent et d'expliciter ce cône via la condition de MANGASARIAN-FROMOWITZ. Nous renvoyons à [15] pour une preuve complète de ces deux résultats.
Nous allons maintenant illustrer ces conditions par différents exemples.

3.4 Applications et Exemples

3.4.1 Projection sur un convexe fermé

On se donne un sous-ensemble non vide de $\mathbb{R}^n$ et un point x n'appartenant pas à ce sous-ensemble. On veut définir la "distance" de ce point à l'ensemble. Les questions qui se posent alors sont les suivantes :

- comment définir cette distance pour qu'elle soit finie ?
- peut-on trouver un point x^* de l'ensemble considéré qui réalise cette distance ?

La réponse n'est pas évidente a priori. On peut toutefois résoudre complètement ce problème lorsque l'ensemble considéré est **convexe** et **fermé**.

Théorème 3.4.1 *Etant donnés C un sous-ensemble convexe, fermé et non vide de $\mathbb{R}^n$ et x un élément quelconque de $\mathbb{R}^n$. Alors le problème*

$$\min \ \|x - y\|^2 = \sum_{i=1}^{n} (x_i - y_i)^2 \ , \ y \in C$$

a une solution unique $x^* \in C$. De plus $x^* \in C$ est caractérisé par :

$$\forall y \in C \qquad (x - x^*, y - x^*) \leq 0 \ . \tag{3.4.1}$$

Démonstration - La démonstration est immédiate. Nous avons affaire à un problème de moindres carrés. La fonction coût est continue, coercive et strictement convexe et l'ensemble C est convexe fermé. On peut donc appliquer le théorème 3.1.2. La caractérisation de x^* s'obtient par application du théorème 3.2.2 □

Nous avons une seconde caractérisation de x^* de manière immédiate :

Corollaire 3.4.1 *Sous les hypothèses du théorème (3.4.1) on peut caractériser le projeté $x^* \in C$ de x par :*

$$\forall y \in C \qquad (x^* - y, y - x) \leq 0 \; . \tag{3.4.2}$$

Démonstration - Si x^* est le projeté de x sur C, (3.4.1) donne :

$$\forall y \in C \qquad (x - x^*, y - x^*) \leq 0 \; .$$

Donc

$$\forall y \in C \qquad (x^* - y, y - x) = (x^* - y, x^* - x) - \|y - x^*\|^2 \leq 0 \; .$$

Réciproquement : Soit $y \in C$ et $z = x^* + t(y - x^*) \in C$, pour $t \in]0, 1[$. La relation (3.4.2) implique

$$\forall t \in]\,0, 1[\qquad (x^* - z, z - x) = -t\,(y - x^*, x^* - x + t(y - x^*)) \leq 0$$

$$\forall t \in]\,0, 1[\qquad (y - x^*, x^* - x + t(y - x^*)) \geq 0 \; .$$

On fait ensuite tendre t vers 0^+ pour obtenir $(3.4.1)$. □

Le point x^* est le **projeté** de x sur C. L'application $\pi_C = \mathbb{R}^n \to C$ qui à x associe son projeté x^* est la **projection** sur C. Le projeté $\pi_C(x)$ est donc le point de C qui est le "plus près" de x. On définit de manière standard la fonction **distance** d'un point x à l'ensemble C par

$$d(x, C) = \inf_{y \in C} \|x - y\| \; . \tag{3.4.3}$$

Dans le cas où C est un convexe fermé, on vient donc de démontrer que

$$d(x, C) = \|x - \pi_C(x)\| \; .$$

Proposition 3.4.1 *La projection π_C est continue. Plus précisément on a*

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \qquad \|\pi_C(x) - \pi_C(y)\| \leq \|x - y\| \; ,$$

c'est-à-dire π_C est une contraction.

Démonstration - Soient x_1 et x_2 deux éléments quelconques de $\mathbb{R}^n$. Appliquons la relation (3.4.1) à $x = x_1, x^* = \pi_C(x_1)$ et $y = \pi_C(x_2) \in C$ puis à $x = x_2, x^* = \pi_C(x_2)$ et $y = \pi_C(x_1) \in C$: on obtient

$$\begin{aligned} &(x_1 - \pi_C(x_1), \pi_C(x_2) - \pi_C(x_1)) \leq 0 \\ &(x_2 - \pi_C(x_2), \pi_C(x_1) - \pi_C(x_2)) \leq 0 \; ; \end{aligned}$$

La somme des deux inégalités donne

$$(x_1 - x_2, \pi_C(x_2) - \pi_C(x_1)) + \|\pi_C(x_2) - \pi_C(x_1)\|^2 \leq 0 \,,$$

c'est-à-dire avec l'inégalité de Cauchy-Schwarzrtz

$$\|\pi_C(x_2) - \pi_C(x_1)\|^2 \leq (x_2 - x_1, \pi_C(x_2) - \pi_C(x_1)) \leq \|x_2 - x_1\| \, \|\pi_C(x_2) - \pi_C(x_1)\| \,.$$

Si $\pi_C(x_2) = \pi_C(x_1)$ la relation que l'on cherche est évidente.
Sinon on divise par $\pi_C(x_2) - \pi_C(x_1)$ et on obtient le résultat souhaité. □

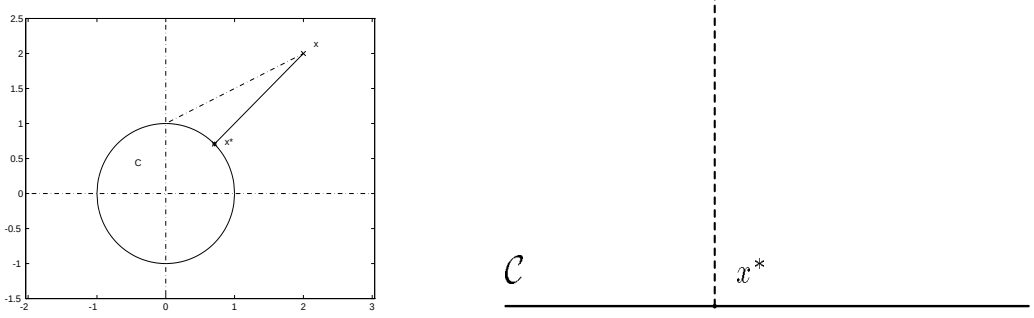

Figure 3.1 : Exemples de projection sur un convexe C

Remarque 3.4.1 *1. Si $x \in C$ alors $\pi_C(x) = x$. Plus généralement si $C = \mathbb{R}^n$ alors $\pi_C = Id_{\mathbb{R}^n}$.*
2. Le théorème 3.4.1 est faux si C n'est pas convexe.
3. La projection π_C n'est pas différentiable en général, mais l'application $x \mapsto \|x - \pi_C(x)\|^2$ l'est (cf. exercice 3.6.2.2.)

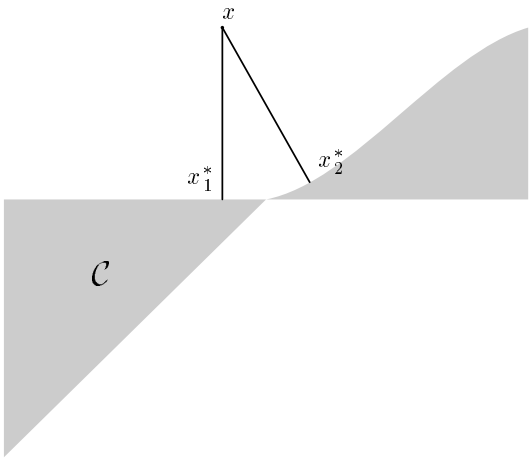

Figure 3.2 : cas où C n'est pas convexe

Le théorème est également faux si C n'est pas fermé : prenons par exemple C égal au disque ouvert de centre $(0,0)$ et de rayon 1 :

$$C = \{\ (x,y) \in \mathbb{R}^2 \ \mid \ x^2 + y^2 < 1\ \}\ .$$

Il n'y a pas de point de C réalisant la distance du point $(1,1)$ à C. En effet le seul point possible se situerait sur le cercle de centre $(0,0)$ et de rayon 1 (voir Figure 3.3) ; mais dans ce cas ce point n'appartient pas à C.

Exemple 3.4.1 (Projection sur un sous-espace vectoriel) *Dans le cas où C est un sous-espace vectoriel de $\mathbb{R}^n$, c'est bien sûr un convexe fermé non vide. L'opérateur de projection est dans ce cas **linéaire** (c'est **faux** dans le cas général). Le projeté x^* d'un élément x, sur C, est caractérisé par (3.4.1) qui dans ce cas est équivalente à*

$$\forall y \in C \qquad (x - x^*, y) = 0\ .$$

Cela signifie que $x - x^* \in C^\perp$ (l'orthogonal de C). On retrouve ainsi la classique projection orthogonale sur un sous-espace vectoriel.

Exemple 3.4.2 *Donnons l'expression de la projection π_C quand C est un ensemble "simple".*

1. *Cas où C est l'orthant positif de $\mathbb{R}^n$:*

$$C = \{\ x = (x_1, \ldots, x_n) \in \mathbb{R}^n \ \mid \ x_i \geq 0 \text{ pour tout } i\ \}\ .$$

Alors la projection sur C est définie par

$$\pi_C(x) = (x_1^+, \ldots, x_n^+)\ ,$$

où $x_i^+ = \max(0, x_i)$ est la partie positive du réel x_i.

2. *Cas d'un pavé borné :*

$$C = \{\, x = (x_1, \ldots, x_n \in \mathbb{R}^n) \;\mid\; a_i \leq x_i \leq b_i \;\; \textit{pour tout } i \,\}\,,$$

où $a_i \leq b_i$ pour tout i compris entre 1 et n.
Alors $\pi_C(x) = (\tilde{x}_1, \ldots, \tilde{x}_n)$, où $\tilde{x}_i$ est défini de la manière suivante :

$$\tilde{x}_i = \begin{cases} x_i & \textit{si} \quad a_i \leq x_i \leq b_i \\ a_i & \textit{si} \quad x_i < a_i \\ b_i & \textit{si} \quad x_i > b_i \end{cases}$$

La démonstration est laissée en exercice au lecteur. On commencera par le cas $n = 1$.

Le Théorème 3.4.1 permet aussi de montrer un résultat de séparation entre $x \notin C$ et C. Ce résultat est un cas particulier du Théorème de Hahn-Banach (cf Annexe A - Section A.4), mais peut se montrer directement ici :

Théorème 3.4.2 (Séparation d'un point et d'un convexe fermé) *Soit C un sous-ensemble convexe, fermé de $\mathbb{R}^n$ et $x \notin C$. Alors, il existe un hyperplan fermé séparant x et C* ***strictement****. Plus précisément, on peut trouver $\alpha \in \mathbb{R}^n$ non nul, $\beta \in \mathbb{R}$ et $\varepsilon > 0$ tels que*

$$(\alpha, x) + \beta \geq \varepsilon \quad \textit{et} \quad \forall y \in C \qquad (\alpha, y) + \beta \leq -\varepsilon\,.$$

Démonstration - Pour la définition précise de la notion de séparation, nous renvoyons à l'annexe A. Soit $x^* = \pi_C(x)$. Comme $x \notin C$, $\alpha = x - x^* \neq 0$. Posons alors

$$\beta = -\frac{1}{2}\,(x - x^*, x + x^*) \in \mathbb{R} \;\; \text{et} \;\; \varepsilon = \frac{1}{4}\|x - x^*\|^2 > 0.$$

On constate que

$$(\alpha, x) + \beta = \frac{1}{2}\,(x - x^*, x - x^*) = \frac{1}{2}\|x - x^*\|^2 > \varepsilon\,,$$

et pour tout $y \in C$

$$(\alpha, y) + \beta = \frac{1}{2}\,(x - x^*, 2\,y - x - x^*) = \frac{1}{2}\,(x - x^*, y - x) + \frac{1}{2}\,(x - x^*, y - x^*)$$

$$= (x - x^*, y - x) + \frac{1}{2}\,(x - x^*, x - x^*) = (x - x^*, y - x) + 2\varepsilon\,.$$

D'après le corollaire 3.4.1, $(x - x^*, y - x) \leq 0$, donc

$$\forall y \in C \qquad (\alpha, y) + \beta \leq -2\varepsilon < -\varepsilon\,.$$

□

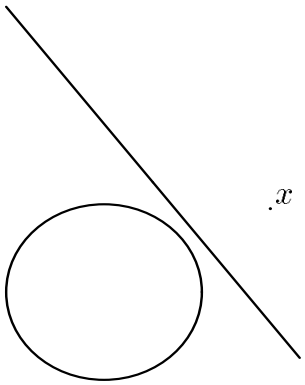

Figure 3.3. : séparation stricte d'un convexe fermé et d'un point x

Remarque 3.4.2 *Nous avons un résultat similaire (avec séparation au sens large) dans le cas où x est un élément de la frontière de C. Cependant, ce n'est pas une conséquence directe de ce qui précède ou du théorème de projection. La démonstration n'est pas simple. C'est en fait un corollaire du Théorème de Hahn-Banach cité en Annexe A - Section A.4.*

3.4.2 Régression linéaire avec contraintes

Reprenons l'exemple de la régression linéaire, en ajoutant des contraintes. Nous gardons les notations du chapitre 3, Section 3.2.

Considérons un nuage de n points de $\mathbb{R}^2$: $M_i = (t_i, x_i), 1 \leq i \leq n$. On cherche la droite de régression $x = at + b$ mais on impose (par exemple) $b \geq 0$. Cela signifie par exemple qu'on ne connaît pas la donnée à l'origine (ou qu'on n'a pas fait de mesure) mais qu'on sait qu'elle est positive (si c'est la concentration d'un composant dans un mélange par exemple). Le problème de régression se formule maintenant de la façon suivante : trouver un couple de réels (a, b) solution de

$$\min J(a,b)\ ,\ (a,b) \in \mathbb{R}^2\ ,\ b \geq 0$$

où $J(a,b) = \sum_{i=1}^{n} (x_i - at_i - b)^2$.

La fonction J est (strictement) convexe, coercive et l'ensemble $C = \{\ (a,b) \in \mathbb{R}^2\ \mid\ b \geq 0\ \}$ est convexe ; ce problème a donc une solution unique. On pose $g(a,b) = -b \leq 0$. Comme $\nabla g(a,b) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$ n'est pas nul, tout point (a,b) est régulier. Ecrivons les équations de KKT : il existe $\lambda \geq 0$ tel que

$$\lambda \cdot b = 0\ ,\ b \geq 0\ ,\ \nabla J(a,b) + \lambda \begin{bmatrix} 0 \\ -1 \end{bmatrix} = 0\ .$$

Cela donne donc $b \geq 0\ ,\ \lambda \geq 0$, et

$$\left\{ \begin{array}{lcl} S_{t^2} a + S_t b & = & S_{xt} \\ S_t a + n\ b - \lambda & = & S_x \\ \lambda \cdot b & = & 0 \end{array} \right.$$

Si $b > 0$ alors $\lambda = 0$. On résout le système ci-dessus. Si le réel b ainsi obtenu est strictement positif on a trouvé la solution. Sinon, c'est que $b = 0$ et on termine alors le calcul : $a = \dfrac{S_{xt}}{S_{t^2}}$.

3.4.3 Cas de la programmation linéaire

On considère le problème

$$(\mathcal{P}) \qquad \begin{cases} \min \ (a, v)_n \\ v \geq 0 \\ Cv \leq d \end{cases}$$

où $(\cdot, \cdot)_n$ désigne le produit scalaire dans $\mathbb{R}^n$, $a, v \in \mathbb{R}^n$, $d \in \mathbb{R}^m$ et C est une matrice réelle $m \times n$. (Les inégalités sont à comprendre composante par composante).
Le Lagrangien s'écrit dans ce cas :

$$\forall (v, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^n \qquad \mathcal{L}(v, \lambda, \mu) = \ (a, v)_n + (\lambda, Cv - d)_p - \ (\mu, v)_n \ .$$

On suppose que ce problème possède au moins une solution v^*. Ecrivons (au moins formellement) les conditions de KKT : il existe $\lambda^* = (\lambda_1^*, \ldots, \lambda_p^*) \in \mathbb{R}^p$ et $\mu^* = (\mu_1^*, \ldots, \mu_n^*) \in \mathbb{R}^n$ tels que

$$\mu^* \geq 0 \ , \ \lambda^* \geq 0 \ , \tag{3.4.4a}$$

$$v^* \geq 0 \ , \ Cv^* \leq d \ , \tag{3.4.4b}$$

$$(\lambda^*, Cv^* - d)_p = 0 \ , \ (\mu^*, v^*)_n = 0 \ , \tag{3.4.4c}$$

$$a + C^t \lambda^* - \mu^* = 0 \ , \tag{3.4.4d}$$

où C^t désigne la matrice transposée de C. Les équations précédentes sont équivalentes à

$$v^* \geq 0 \ , \ w^* = d - Cv^* \geq 0 \ , \tag{3.4.5a}$$

$$\lambda^* \geq 0 \ , \ \mu^* = C^t \lambda^* + a \geq 0 \ , \tag{3.4.5b}$$

$$(\lambda^*, Cv^* - d)_p = (\lambda^*, w^*)_p = 0 \ , \ (\mu^*, v^*)_n = \ \left(v^*, C^t \lambda^* + a\right)_n = 0 \ , \tag{3.4.5c}$$

$$Cv^* + w^* = d \ . \tag{3.4.5d}$$

On reconnaît alors les conditions de KKT du problème suivant

$$(\mathcal{D}) \qquad \begin{cases} \max \ -(d, \lambda)_p = -\min(d, \lambda)_p \\ \lambda \geq 0 \\ C^t \lambda + a \geq 0 \end{cases}$$

λ^* étant la solution et $v^*, \ w^*$ jouant le rôle des multiplicateurs de Lagrange. Ce problème est le problème **dual** de $(\mathcal{P})$.

3.4.4 Un exemple

On considère le problème suivant : $\left\{ \begin{array}{ll} \min & 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\ \text{avec} & x_1^2 + x_2^2 \leq 5 \\ & 3x_1 + x_2 \leq 6 \, . \end{array} \right.$

Dans ce cas $x = (x_1, x_2)$, il n'y pas de contraintes en égalité ($p = 0$) et

$$J(x) = 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \, ,$$

$$q = 2, \; g = (g_1, g_2) \text{ avec } g_1(x) = x_1^2 + x_2^2 - 5 \text{ et } g_2(x) = 3x_1 + x_2 - 6 \, .$$

On peut remarquer que $J, \; h$ et g sont convexes. J est même strictement convexe.

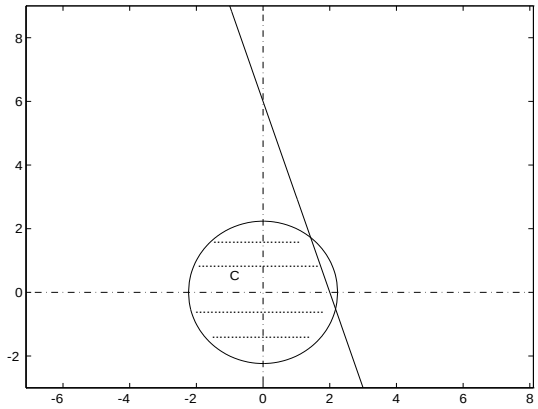

Figure 3.4 : Ensemble C des contraintes

Ecrivons les conditions de KKT a priori. On vérifiera la régularité du point obtenu a posteriori. On obtient

$$\mu_1 \geq 0 \, , \; \mu_2 \geq 0 \, , \tag{3.4.6a}$$

$$x_1^2 + x_2^2 \leq 5, \;\; 3x_1 + x_2 \leq 6 \, , \tag{3.4.6b}$$

$$\mu_1(x_1^2 + x_2^2 - 5) = 0 \, , \; \mu_2(\, 3x_1 + x_2 - 6) = 0 \, , \tag{3.4.6c}$$

$$4x_1 + 2x_2 - 10 + 2\mu_1 x_1 + 3\mu_2 = 0 \, , \tag{3.4.6d}$$

$$2x_1 + 2x_2 - 10 + 2\mu_1 x_2 + \mu_2 = 0 \, . \tag{3.4.6e}$$

Comme il y a deux contraintes en inégalité il y a quatre possibilités :

1. les deux contraintes sont inactives : $x_1^2 + x_2^2 < 5, \;\; 3x_1 + x_2 < 6$, ce qui donne : $\mu_1 = 0$ et $\mu_2 = 0$.
2. g_1 est active et g_2 ne l'est pas, ce qui donne : $x_1^2 + x_2^2 = 5$ et $\mu_2 = 0$.
3. g_1 est inactive et g_2 est active , ce qui donne : $\mu_1 = 0$ et $3x_1 + x_2 = 6$.
4. Les deux contraintes sont actives : $x_1^2 + x_2^2 = 5$ et $3x_1 + x_2 = 6$.

Il faut donc tester chacun de ces cas et résoudre les équations de KKT à chaque fois.

1. Premier cas : les deux contraintes sont inactives : $\mu_1 = 0$ et $\mu_2 = 0$. Les équations (3.4.6d) et (3.4.6e) deviennent :

$$4x_1 + 2x_2 = 10 \text{ et } 2x_1 + 2x_2 = 10,$$

ce qui donne $x_1 = 0$ et $x_2 = 5$. Or ce point ne vérifie pas les contraintes (condition de réalisabilité (3.4.6b)). Par conséquent **ce cas est impossible**.

Remarquons toutefois que le point $(0, 5)$ obtenu satisfait $\nabla J(x) = 0$. On vérifie facilement que c'est le **minimum (global) sans contraintes**.

2. Deuxième cas : g_1 est active et g_2 ne l'est pas : on obtient alors

$$x_1^2 + x_2^2 = 5,\ \mu_2 = 0$$

$$4x_1 + 2x_2 - 10 + 2\mu_1 x_1 = 0\ ,\ 2x_1 + 2x_2 - 10 + 2\mu_1 x_2 = 0\ .$$

Ce système a pour solution $x_1 = 1,\ x_2 = 2$ et $\mu_1 = 1$. **Cette solution est admissible** : le point trouvé vérifie les conditions de KKT.

3. Troisième cas : g_1 est inactive et g_2 est active ; on obtient

$$4x_1 + 2x_2 - 10 + 3\mu_2 = 0\ , 2x_1 + 2x_2 - 10 + \mu_2 = 0 \text{ et } x_2 = 6 - 3x_1.$$

C'est un système linéaire de trois équations à trois inconnues dont la solution est $x_1 = 2/5,\ x_2 = 24/5$ et $\mu_2 = -2/5$. On voit que μ_2 n'est pas positif, donc **ce cas est impossible**.

4. Le cas où les deux contraintes sont actives correspond aux points situés à l'intersection de la droite et du cercle. Un rapide calcul montre que ce sont les points $x_1 = 9/5 + \sqrt{14}/10 \simeq 2.17$ ou $x_1 = 9/5 - \sqrt{14}/10 \simeq 1.42$ avec $x_2 = 6 - 3x_1$ et que $\mu_1 = (4 - 10x_1)/(20x_1 - 36)$. On voit donc que μ_1 est négatif pour $x_1 = 9/5 + \sqrt{14}/10 \simeq 2.17$. Dans l'autre cas $\mu_1 \simeq 1.34$. mais $\mu_2 \simeq -0.98$ est négatif. Finalement **ce cas est impossible**.

Finalement la seule solution possible des équations de KKT est $x^* = (1, 2)$. On peut vérifier a posteriori que ce point est régulier. La seule contrainte active est g_1. Comme $\nabla g_1(x^*) = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ n'est pas nul, la condition de régularité est donc satisfaite. On pourrait aussi vérifier la condition du second ordre. Mais, nous savons que le problème considéré a une solution unique d'après les théorèmes généraux (C est borné et J est strictement convexe). Comme il n'y a qu'un point possible ce point est nécessairement la solution.

Nous pouvons constater sur cet exemple (pourtant simple) que la résolution des équations de KKT peut-être longue et fastidieuse, surtout lorsqu'il y a beaucoup de contraintes. C'est pour cela qu'on développe des algorithmes pour calculer les solutions de ces équations.

3.5 Algorithmes

3.5.1 Méthode du Gradient projeté

La méthode du gradient projeté s'inspire de la méthode du gradient décrite dans le chapitre précédent. Comme cette dernière est une méthode de descente, nous avons de manière très générale

dans le cas sans contraintes, la formulation suivante :

$$\begin{cases} x_0 \in \mathbb{R}^n \text{ donné} \\ x_{k+1} = x_k + \rho_k\, d_k,\ d_k \in \mathbb{R}^n - \{0\},\ \rho_k \in \mathbb{R}^{+*}\ , \end{cases}$$

où ρ_k et d_k sont choisis de telle sorte que $J(x_k + \rho_k\, d_k) \leq J(x_k)$. Toutefois, lorsqu'on minimise sur un ensemble de contraintes C et que $x_k \in C$ on n'est pas sûr avec la formulation précédente que l'itéré $x_{k+1} = x_k + \rho_k\, d_k$ appartienne à C. Il faut donc le "ramener" dans C, ce qu'on fait grâce à une projection sur C.

Algorithme du Gradient projeté

1. **Initialisation**
   $k = 0$: choix de x_0 et de $\rho_0 > 0$
2. **Itération** k
   $x_{k+1} = \pi_C(x_k - \rho_k\, \nabla J(x_k))$;
3. **Critère d'arrêt**
   Si $\|x_{k+1} - x_k\| < \varepsilon$, STOP
   Sinon, on pose $k = k + 1$ et on retourne à 2.

Nous avons un résultat de convergence :

Théorème 3.5.1 *Soit J une fonction $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$. On suppose que J est elliptique de dérivée lipschitzienne (c'est-à-dire J vérifie (2.1.1) et (2.4.1)).*
Alors, si on choisit le pas ρ_k dans un intervalle $[\beta_1, \beta_2]$ tel que $0 < \beta_1 < \beta_2 < \dfrac{2\alpha}{M}$, la suite x_n définie par la méthode du gradient projeté converge vers la solution du problème $(\mathcal{P})$.

Démonstration - Nous avons vu au chapitre 2 , (théorème 2.1.3) que J est strictement convexe, coercive et que le problème $(\mathcal{P})$ admet une solution unique x^*. Remarquons ensuite que

$$x^* = \pi_C(x^* - \rho \nabla J(x^*)) \ \text{ pour tout } \rho > 0\ . \tag{3.5.1}$$

En effet, grâce au théorème 3.2.1, nous avons

$$\forall x \in C \qquad (\nabla J(x^*), x - x^*) \geq 0\ ,$$

c'est-à-dire

$$\forall x \in C \qquad ((x^* - \rho\, \nabla J(x^*)) - x^*, x - x^*) \leq 0\ ;$$

on reconnaît la caractérisation du projeté de x^* (avec (3.4.1)). Comme

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

en soustrayant (3.5.1), nous obtenons

$$x_{k+1} - x^* = \pi_C(x_k - \rho_k\, \nabla J(x_k)) - \pi_C(x^* - \rho_k \nabla J(x^*))\ .$$

Comme la projection est contractante nous avons

$$\|x_{k+1} - x^*\|^2 \leq \|(x_k - \rho_k \nabla J(x_k)) - (x^* - \rho_k \nabla J(x^*))\|^2 ,$$

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + \rho_k^2 \|\nabla J(x_k) - \nabla J(x^*)\|^2 - 2\rho_k \left(x_k - x^*, \nabla J(x_k) - \nabla J(x^*) \right) ;$$

Avec (2.1.1) et (2.4.1), on a finalement

$$\|x_{k+1} - x^*\|^2 \leq (1 + M\,\rho_k^2 - 2\,\rho_k \alpha)\, \|x_k - x^*\|^2 .$$

Si on choisit le pas ρ_k dans un intervalle $[\beta_1, \beta_2]$ tel que $0 < \beta_1 < \beta_2 < \dfrac{2\alpha}{M}$, on obtient

$$\|x_{k+1} - x^*\| \leq \kappa\, \|x_k - x^*\| ,$$

où κ est une constante de $]\,0, 1\,[$ indépendante de k. La suite x_k converge donc vers x^*. □

Malgré son apparente simplicité, cette méthode est souvent difficile à mettre en oeuvre. En effet, il faut à chaque étape calculer le projeté d'un vecteur de $\mathbb{R}^n$ sur C. Lorsque C est simple (pavé borné par exemple), c'est faisable. Dès que les contraintes ne sont pas des contraintes de borne, le calcul de la projection devient délicat.

3.5.2 Méthode de Lagrange-Newton pour des contraintes en égalité

Plaçons nous pour commencer dans le cas particulier d'un problème quadratique avec contraintes en égalité affines :

$$\min \frac{1}{2} x^t Q x - c^t x \ , Ax = b \ ,$$

où Q est une matrice carrée $n \times n$, x et c des vecteurs de $\mathbb{R}^n$, A une matrice $p \times n$ et b un vecteur de $\mathbb{R}^p$. Ecrivons (au moins formellement) les conditions de KKT, c'est-à-dire le système d'optimalité du premier ordre. On obtient

$$\left\{ \begin{array}{l} \nabla_x \left(\dfrac{1}{2} x^t Q x - c^t x + \lambda^t (Ax - b) \right) = 0 \\ Ax = b \end{array} \right.$$

où $\lambda \in \mathbb{R}^p$ est le multiplicateur de Lagrange associé à la contrainte $A\,x - b = 0$. Par conséquent, le couple optimal (x^*, λ^*) est solution du système linéaire :

$$\left[\begin{array}{cc} Q & A^t \\ A & 0 \end{array} \right] \left[\begin{array}{c} x^* \\ \lambda^* \end{array} \right] = \left[\begin{array}{c} c \\ b \end{array} \right] .$$

Si la matrice $M = \left[\begin{array}{cc} Q & A^t \\ A & 0 \end{array} \right]$ est inversible, ce système admet une solution que l'on peut calculer par n'importe quelle méthode de résolution de systèmes linéaires.

Supposons maintenant que le problème n'est plus quadratique. On va utiliser la méthode de Newton pour résoudre le système d'optimalité. Plus précisément, on considère le problème en égalité suivant

$$\min J(x)\ ,\ \ h(x)=0\ ,$$

où $J:\mathbb{R}^n\to\mathbb{R}$ et $h:\mathbb{R}^n\to\mathbb{R}^p$ sont suffisamment régulières. Les conditions du premier ordre s'écrivent (au moins formellement)

$$\left\{\begin{array}{l}\nabla_x\mathcal{L}(x,\lambda)=0\\ h(x)=0\ ,\end{array}\right.$$

où $\lambda\in\mathbb{R}^p$ et $\mathcal{L}(x,\lambda)=J(x)+\lambda^t h(x)$ est le Lagrangien du problème. On peut résoudre ce système d'équations non linéaires par la méthode de Newton ce qui donne :

Algorithme de Lagrange-Newton

1. **Initialisation**
   $k=1$: choix de $(x_0,\lambda_0)\in\mathbb{R}^n\times\mathbb{R}^p$
2. **Itération** k : on connaît (x_k,λ_k).
   Résoudre

$$\left[\begin{array}{cc}D^2_{xx}\mathcal{L}(x_k,\lambda_k) & Dh(x_k)^t\\ Dh(x_k) & 0\end{array}\right]\left[\begin{array}{c}d_k\\ y_k\end{array}\right]=-\left[\begin{array}{c}\nabla_x\mathcal{L}(x_k,\lambda_k)\\ h(x_k)\end{array}\right]\ , \qquad (3.5.2)$$

   où $D^2_{xx}\mathcal{L}(x_k,\lambda_k)$ est le Hessien (par rapport à x) de $\mathcal{L}$. Puis

$$x_{k+1}=x_k+d_k\ ,\ \lambda_{k+1}=\lambda_k+y_k\ .$$

3. **Critère d'arrêt**
   Si x_k est "satisfaisant" STOP.
   Sinon, on pose $k=k+1$, et on retourne à 2.

On remarque que si on ajoute $\nabla h(x_k)^t\lambda_k$ à la première ligne de (3.5.2), celle-ci est alors équivalente à

$$\left[\begin{array}{cc}D^2_{xx}\mathcal{L}(x_k,\lambda_k) & Dh(x_k)^t\\ Dh(x_k) & 0\end{array}\right]\left[\begin{array}{c}d_k\\ \lambda_{k+1}\end{array}\right]=-\left[\begin{array}{c}\nabla_x J(x_k)\\ h(x_k)\end{array}\right]\ ,$$

3.5.3 Méthode de Newton projetée pour des contraintes de borne

La méthode de Newton projetée (essentiellement due à Bertsekas [2]) repose sur une idée analogue à celle du gradient projeté : puisque les itérés successifs n'appartiennent pas nécessairement à l'ensemble des contraintes C, on les projette sur C. Nous ne présenterons que le cas d'un problème de minimisation avec contraintes de bornes (pour des variantes on peut se référer à [2]), de la forme

$$(\mathcal{P})\qquad\qquad \min f(x)\ ,\qquad a\le x\le b\ .$$

On peut toutefois remarquer que beaucoup de problèmes duaux (dont les variables sont les multiplicateurs de Lagrange) sont de cette forme là.
Commençons par un problème encore plus simple :

$$(\mathcal{P})_0 \qquad \min f(x)\ , \qquad x \geq 0\ ,$$

où f est $\mathcal{C}^2$ sur $\mathbb{R}^n$. Rappelons que si H_k désigne la matrice hessienne $[D^2 f(x_k)]$, une itération de la méthode de Newton est de la forme :

$$x_{k+1} = x_k - H_k^{-1} \nabla f(x_k)\ .$$

On peut raffiner un peu la méthode en introduisant un pas $\alpha_k > 0$, ce qui donne

$$x_{k+1} = x_k - \alpha_k H_k^{-1} \nabla f(x_k)\ .$$

Enfin , on peut projeter sur l'ensemble des contraintes et on obtient

$$x_{k+1} = \big(x_k - \alpha_k H_k^{-1} \nabla f(x_k)\big)^+\ ;$$

nous obtenons ainsi la méthode de Newton projetée.

En fait, nous allons présenter une méthode un peu plus générale dont les critères de convergence sont voisins et qui présente l'avantage d'être plus souple d'utilisation (c'est une méthode quasi-Newton). Dans ce qui suit x^i désigne la composante numéro i de x.

Algorithme de Newton projeté : cas unilatéral

1. **Initialisation**
   $k = 0$, choix de $x_0 \in \mathbb{R}^n,\ x_0 \geq 0$. On se donne une tolérance $\varepsilon > 0$.
2. **Itération** k : on connaît x_k.
   – Utilisation d' une règle **anti zig-zag** :

$$\omega_k = \|x_k - (x_k - \nabla f(x_k))^+\| \quad \text{et} \quad \varepsilon_k = \min(\varepsilon, \omega_k)\ .$$

$$I_k^+ = \{ i \ \mid \ 0 \leq x_k^i \leq \varepsilon_k \text{ et } \frac{\partial f(x_k)}{\partial x^i} > 0 \}\ .$$

$$D_k = H_k^{-1} \text{ où } H_k^{ij} = \begin{cases} 0 \text{ si } i \neq j \text{ et } (i \in I_k^+ \text{ ou } j \in I_k^+)\ , \\ \dfrac{\partial^2 f}{\partial x^i \partial x^j}(x_k) \text{ sinon.} \end{cases}$$

   – Choix de la direction de descente : $p_k = D_k \nabla f(x_k)$.
   – Choix du pas par une recherche linéaire :

$$x_k(\alpha) = (x_k - \alpha p_k)^+\ .$$

   α_k est choisi avec une règle du type **Armijo** comme $(3.5.3)$ ci-dessous.
   $x_{k+1} = x_k(\alpha_k)$

3. **Critère d'arrêt**
   Si x_{k+1} est "satisfaisant" STOP.
   Sinon, on pose $k = k + 1$, et on retourne à 2.

Comme au chapitre 2, tout résultat de convergence inclut le fait que la suite des itérés est bien définie, c'est-à-dire ici, que les matrices H_k sont inversibles.

Règle de choix du pas

On se donne $\beta \in]0,1[$ et $\sigma \in]0, \frac{1}{2}[$. On pose $\alpha_k = \beta^{m_k}$ où m_k est le premier entier tel que

$$f(x_k) - f(x_k(\beta^m)) \geq \sigma \left(\beta^m \sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} \, p_k^i + \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} \, (x_k^i - x_k(\beta^m)^i) \right) . \quad (3.5.3)$$

Remarque 3.5.1 *1. I_k^+ est l'ensemble des indices des contraintes "presque" actives (c'est-à-dire à ε_k près). La règle **anti zig-zag** permet d'éviter des oscillations de l'algorithme.*

2. D_k^{-1} est une approximation de la matrice hessienne plus "facile" à calculer.

*3. La règle de choix du pas (3.5.3) est une règle du type **Armijo** qui permet de choisir un pas "optimal" à moindre coût.*

Nous avons alors un résultat de convergence (dont la démonstration se trouve dans [2]) :

Théorème 3.5.2 *On suppose que la fonction f est convexe et $\mathcal{C}^2$ et que le problème $(\mathcal{P})_0$ a une solution unique x^*vérifiant $\dfrac{\partial f(x^*)}{\partial x^i} > 0$ pour tout $i \in I(x^*)$ (ensemble actif).*

On suppose également qu'on peut trouver m_1 et m_2 deux réels strictement positifs tels que

$$m_1 \|z\|^2 \leq \left(D^2 f(x)\, z, z \right) \leq m_2 \|z\|^2 ,$$

dans chacun des cas suivants :
• pour tout $z \in \{x \mid f(x) \leq f(x_0)\}$ d'une part
• pour tout x dans une boule centrée en x^* et $z \neq 0$ tel que $z^i = 0$ pour $i \in I(x^*)$ d'autre part.

Alors, la suite x_k engendrée par l'algorithme converge vers x^* et le taux de convergence est super-linéaire (au moins quadratique si $D^2 f$ est lipschitzienne au voisinage de x^*).

On peut généraliser cet algorithme au problème $(\mathcal{P})$; nous obtenons

Algorithme de Newton projeté : cas bilatéral

1. **Initialisation**
   $k = 1$: choix de $x_0 \in \mathbb{R}^n,\ x_0 \geq 0$. On se donne une tolérance $\varepsilon > 0$.
2. **Itération** k : on connaît x_k.

– On utilise une règle **anti zig-zag** :

$$\omega_k = \|x_k - [x_k - \nabla f(x_k)]^\sharp\| \text{ et } \varepsilon_k = \min(\varepsilon, \omega_k) \ .$$

$$I_k^\sharp = \{i \mid a^i \leq x_k^i \leq a^i + \varepsilon_k \text{ et } \frac{\partial f(x_k)}{\partial x^i} > 0\} \cup \{i \mid b^i - \varepsilon_k \leq x_k^i \leq b^i \text{ et } \frac{\partial f(x_k)}{\partial x^i} < 0\} \ .$$

– D_k est définie positive et diagonale par rapport à $I_k^\sharp$.

–
$$x_k(\alpha) = [x_k - \alpha D_k \nabla f(x_k)]^\sharp \ ,$$
où pour tout z de $\mathbb{R}^n$, $z^\sharp = \pi_C(z)$ désigne le vecteur de coordonnées
$$[z]^{\sharp i} = \begin{cases} b^i & \text{si } b^i \leq z^i \\ z^i & \text{si } a^i \leq z^i \leq b^i \\ a^i & \text{si } z^i \leq a^i \ . \end{cases}$$
α_k est choisi avec une règle du type **Armijo** comme $(3.5.3)$ ci-dessus où $\sharp$ remplace $+$.

$x_{k+1} = x_k(\alpha_k)$

3. **Critère d'arrêt**
Si x_{k+1} est "satisfaisant" STOP.
Sinon, on pose $k = k + 1$, et on retourne à 2.

3.5.4 Méthodes de pénalisation

Grâce à leur facilité de mise en oeuvre les méthodes de pénalisation sont très souvent utilisées en pratique. Elles relèvent toutes du principe suivant. On remplace le problème

$$(\mathcal{P}) \qquad \min \ J(x) \ , \ x \in C \subset \mathbb{R}^n \ ,$$

par un problème **sans contraintes**

$$(\mathcal{P}_r) \qquad \min \ J(x) + r\alpha(x) \ , \ x \in \mathbb{R}^n \ ,$$

où $\alpha : \mathbb{R}^n \to \mathbb{R}$ est une fonction de **pénalisation des contraintes** et $r > 0$. Le but est de trouver des fonctions α telles que les problèmes $(\mathcal{P})$ et $(\mathcal{P}_r)$ soient équivalents c'est-à-dire qu'ils aient les mêmes solutions. Dans ce cas on dit que la pénalisation est **exacte** . On peut par exemple choisir

$$\alpha(x) = \begin{cases} 0 & \text{si} \quad x \in C \\ +\infty & \text{si} \quad x \notin C \, . \end{cases}$$

Cette fonction un peu "sommaire" n'a pas de bonnes propriétés mathématiques (en particulier de dérivabilité) pour qu'on puisse appliquer les résultats de la section précédente. Il faut donc trouver d'autres fonctions.

En général on effectue une pénalisation **inexacte** c'est-à-dire telle que le problème $(\mathcal{P})$ a des solutions qui ne sont pas solutions de $(\mathcal{P}_r)$; l'ensemble des solutions de $(\mathcal{P}_r)$ ne couvre pas tout l'ensemble des solutions de $(\mathcal{P})$. En revanche dans ce cas, on peut trouver des fonctions α qui sont dérivables ce qui permet d'utiliser les résultats de minimisation sans contraintes. Nous donnons ici un exemple de fonctions de pénalisation où la pénalisation est dite **extérieure** .

On prend $\alpha : \mathbb{R}^n \to \mathbb{R}$ vérifiant les conditions suivantes :

$$\begin{array}{ll} (i) & \alpha \text{ est continue sur } \mathbb{R}^n \ , \\ (ii) & \forall x \in \mathbb{R}^n \qquad \alpha(x) \geq 0 \ , \\ (iii) & \alpha(x) = 0 \Leftrightarrow x \in C \ . \end{array} \tag{3.5.4}$$

Nous donnons ci-dessous un exemple de fonction de pénalisation α pour différentes contraintes :

Contrainte	$x \leq 0$	$h(x) = 0$	$g(x) \leq 0$
Fonction α	$\|x^+\|^2$	$\|h(x)\|^2$	$\|g(x)^+\|^2$

où $\|\cdot\|$ désigne toujours la norme euclidienne de $\mathbb{R}^n$ et $x^+ = (x_1^+, \ldots, x_n^+)$.
Nous avons alors un résultat de convergence :

Théorème 3.5.3 *Supposons que J soit continue et coercive. Soit C un ensemble fermé non vide. On suppose que α vérifie les conditions (3.5.4). Alors on a :*

- *$\forall r > 0$, $(\mathcal{P}_r)$ a au moins une solution x_r,*
- *La famille $(x_r)_{r>0}$ est bornée.*
- *Toute sous-suite convergente extraite de $(x_r)_{r>0}$ converge vers une solution de $(\mathcal{P})$ lorsque r tend vers $+\infty$.*

Démonstration - On ne considère que des réels $r > 0$. On note x^* une solution de $(\mathcal{P})$. Comme

$$J_r(x) \stackrel{\text{def}}{\equiv} J(x) + r\,\alpha(x) \geq J(x) \ ,$$

la coercivité de J entraîne celle de J_r. De plus, la continuité de J et de α impliquent celle de J_r. Par conséquent, $(\mathcal{P}_r)$ a au moins une solution x_r.
D'autre part

$$J(x_r) \leq J_r(x_r) \leq J_r(x^*) = J(x^*) = j^* < +\infty \ ; \tag{3.5.5}$$

donc $J(x_r)$ est uniformément bornée par rapport à r et suivant un raisonnement désormais standard, la coercivité de J entraîne que la famille $(x_r)_{r>0}$ est bornée.
On peut donc trouver une sous-suite de $(x_r)_{r>0}$, notée $(x_{r_k})_k$ qui converge vers $\tilde{x}$. Comme $J(x^*) - J(x_{r_k})$ est bornée (car J est continue) et

$$\alpha(x_{r_k}) \leq \frac{1}{r_k}\,[J(x^*) - J(x_{r_k})]$$

on en déduit que

$$0 \leq \alpha(\tilde{x}) = \lim_{r_k \to +\infty} \alpha(x_{r_k}) \leq 0 \, ,$$

grâce à la continuité de α ; donc $\tilde{x} \in C$. On peut maintenant passer à la limite dans (3.5.5) pour obtenir

$$J(\tilde{x}) \leq J(x^*) \, .$$

Ceci prouve que $\tilde{x}$ est solution de $(\mathcal{P})$ (mais pas nécessairement égale à x^*). □
On peut alors proposer le schéma suivant pour approcher les solutions de $(\mathcal{P})$:

Algorithme de pénalisation extérieure

1. **Initialisation**
   $k = 1$: choix de $x_0 \in \mathbb{R}^n$ et de $r_1 > 0$
2. **Itération** k :
   Résoudre le sous-problème

$$(\mathcal{P}_{r_k}) \qquad \min \; J(x) + r_k \, \alpha(x) \; , \; x \in \mathbb{R}^n \; ,$$

   en prenant x_{k-1} comme point d'initialisation.
3. **Critère d'arrêt**
   Si x_k est "satisfaisant" STOP.
   Sinon, on pose $k = k + 1$, on choisit $r_{k+1} > r_k$ et on retourne à 2.

Cet algorithme est simple mais assez délicat à mettre en oeuvre. Tout d'abord il faut se donner un critère d'arrêt et définir ce qu'on entend par "satisfaisant". D'autre part il faut augmenter le facteur r_k progressivement ($r_{k+1} \geq r_k + 1$ par exemple) car il ne faut pas que le terme $r_k \, \alpha(x)$ soit trop grand par rapport à $J(x)$ faute de quoi, on construit un itéré qui vérifie les contraintes ($\alpha(x) = 0$) mais qui "néglige " de minimiser J. Enfin, si la solution du problème n'est pas unique, la suite x_k peut osciller entre deux valeurs d'adhérence.

3.5.5 Méthodes de Programmation Quadratique Successive (SQP)

Nous allons présenter maintenant une classe de méthodes directement fondées sur le système d'optimalité établi dans les sections 3.2 et 3.3. Ce sont les méthodes de Programmation Quadratique Successive dites **SQP** (de l'anglais *Sequential Quadratic Programming*).

Cas de contraintes en égalité

On considère le problème

$$\min \; J(x) \; , \; x \in C$$

où

$$C = \{ \, x \in \mathbb{R}^n \mid \; h_i(x) = 0 \, , \; i = 1, \ldots, p \, \} \, .$$

Nous avons vu dans la section 3.2 qu'une solution x^* de ce problème est un point critique du Lagrangien $\mathcal{L}$, mais ce n'est pas en général un minimum de ce Lagrangien. Nous allons établir une méthode de descente qui exploite le système d'optimalité établi dans le théorème 3.2.3. L'idée est de résoudre une succession de problèmes quadratiques avec contraintes linéaires (on a vu que ces problèmes étaient relativement "faciles" à résoudre) qui sont des approximations du problème de départ.

Etant donné un itéré x_k on cherche

$$x_{k+1} = x_k + \rho_k d_k \ ,$$

où $d_k \in \mathbb{R}^n$ est une direction de descente et $\rho_k > 0$ le pas. Commençons par faire une approximation des contraintes h grâce à la formule de Taylor :

$$h_i(x_k + d) = h_i(x_k) + \nabla h_i(x_k) \cdot d + O(\|d\|^2) \ ;$$

si on néglige les termes d'ordre supérieur ou égal à 2, on définit la direction d_k comme étant la direction permettant d'assurer que $h_i(x_k + d) \simeq 0$. Plus précisément, on pose

$$h_i(x_k) + \nabla h_i(x_k) \cdot d_k = 0, \qquad \forall i = 1, \ldots, q \ ,$$

c'est-à-dire

$$Dh(x_k)\, d_k = -h(x_k) \ , \tag{3.5.6}$$

où $Dh(x_k)$ est la matrice jacobienne de h en x_k. Cette relation correspond à une linéarisation des contraintes au voisinage de x_k : c'est un système linéaire.

D'autre part, il faudrait que x_{k+1} diminue la valeur du Lagrangien (puisque que c'est le Lagrangien qui joue le rôle de la fonction objectif, quand on a des contraintes). De manière similaire on va faire une approximation du Lagrangien $\mathcal{L}(x, \lambda) = J(x) + \lambda^t h(x)$: elle sera quadratique cette fois, puisque le point cherché est un point critique et qu'on ne peut se contenter d'une approximation du premier ordre.

$$\mathcal{L}(x_k + d, \lambda) = \mathcal{L}(x_k, \lambda) + (\nabla_x \mathcal{L}(x_k, \lambda), d) + \frac{1}{2}\left(D^2_{xx}\mathcal{L}(x_k, \lambda)\, d, d\right) + O((\|d\|^3) \ .$$

Si on néglige les termes d'ordre supérieur ou égal à 3, on voit qu'il faut minimiser

$$(\nabla_x \mathcal{L}(x_k, \lambda), d) + \frac{1}{2}\left(D^2_{xx}\mathcal{L}(x_k, \lambda)\, d, d\right)$$

pour espérer minimiser le Lagrangien. On cherche donc finalement d_k comme solution du problème

$$(QP)_e \qquad \begin{cases} \min\, (\nabla J(x_k), d) + \frac{1}{2}\left(D^2_{xx}\mathcal{L}(x_k, \lambda)\, d, d\right) \\ Dh(x_k)\, d_k + h(x_k) = 0 \ . \end{cases}$$

En effet, avec (3.5.6)

$$(\nabla_x \mathcal{L}(x_k, \lambda), d) = (\nabla_x J(x_k), d) + \lambda^t Dh(x_k)\, d = (\nabla_x J(x_k), d) + \underbrace{\lambda^t h(x_k)}_{\text{= cste}} \ .$$

Il reste à déterminer le pas ρ_k et le multiplicateur λ_k à chaque itération. Il y a bien sûr beaucoup de possibilités qui donnent lieu à autant de variantes de la méthode. Nous présentons l'algorithme "de base" où $\rho_k = 1$:

Méthode SQP pour des contraintes en égalité

1. **Initialisation**
   $k = 1$: choix de $x_0 \in \mathbb{R}^n$ et de $\lambda_0 \in \mathbb{R}^q$.
2. **Itération** k :
   Résoudre le sous-problème quadratique

$$(QP)_e \qquad \begin{cases} \min \left(\nabla J(x_k), d\right) + \frac{1}{2}\left(D^2_{xx}\mathcal{L}(x_k, \lambda_k)\, d, d\right) \\ D(x_k)\, d + h(x_k) = 0\ . \end{cases}$$

3. $\lambda_{k+1} \in \mathbb{R}^p$ est le multiplicateur associé à la contrainte (en égalité) de (QP_e) et $x_{k+1} = x_k + d_k$.
4. **Critère d'arrêt**
   Si x_{k+1} est "satisfaisant" STOP.
   Sinon, on pose $k = k + 1$, et on retourne à 2.

Cas de contraintes générales

Dans le cas de contraintes en égalité et inégalité le principe est le même : seule la fonction Lagrangienne change. Pour le problème

$$(\mathcal{P}) \qquad \min\ J(x)\ ,\ x \in C$$

où

$$C = \{\ x \in \mathbb{R}^n \mid\ h(x) = 0,\ g(x) \leq 0\}\ ,\ h = (h_i)_{1 \leq i \leq p},\ \ g = (g_j)_{1 \leq j \leq q}$$

elle vaut $\mathcal{L}(x, \lambda, \mu) = J(x) + \lambda^t h(x) + \mu^t g(x)$, où $\lambda \in \mathbb{R}^p$ et $\mu \in \mathbb{R}^q$. La méthode SQP s'écrit de la même façon : on linéarise les contraintes et on fait une approximation quadratique de $\mathcal{L}$. Cela donne

Méthode SQP pour des contraintes générales

1. **Initialisation**
   $k = 1$: choix de $x_0 \in \mathbb{R}^n$ et de $(\lambda_0, \mu_0) \in \mathbb{R}^p \times \mathbb{R}^{q,+}$.
2. **Itération** k :
   Résoudre le sous-problème quadratique

$$(QP) \qquad \begin{cases} \min \left(\nabla J(x_k), d\right) + \dfrac{1}{2}\left(D^2_{xx}\mathcal{L}(x_k, \lambda_k, \mu_k)\, d, d\right) \\ Dh(x_k)\, d + h(x_k) = 0\ , \\ Dg(x_k)\, d + g(x_k) \leq 0\ . \end{cases}$$

3. $\lambda_{k+1} \in \mathbb{R}^p$ est le multiplicateur associé à la contrainte en égalité de (QP) et $\mu_{k+1} \in \mathbb{R}^{q,+}$ le multiplicateur (positif) associé à la contrainte en inégalité .
   $x_{k+1} = x_k + d_k$.

4. **Critère d'arrêt**

   Si x_{k+1} est "satisfaisant" STOP.

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

Nous ne détaillons pas les résultats de convergence. Nous donnons juste un résultat que l'on peut trouver dans [4] p.153.

Théorème 3.5.4 *Supposons que f, h et g soient $\mathcal{C}^2$ à dérivées secondes bornées dans un voisinage d'une solution x^* de $(\mathcal{P})$. On suppose que les contraintes en x^*vérifient la condition (CQ1) de Mangasarian-Fromowitz de la définition 3.2.3 et on note (λ^*, μ^*) les multiplicateurs associés à x^*. On suppose enfin que les conditions suffisantes du second ordre sont vérifiées.*
Considérons la méthode SQP dans laquelle la solution de (QP) d_k est de norme minimale (s'il y en a plusieurs). Alors, il existe un voisinage $\mathcal{V}$ de (x^*, λ^*, μ^*) tel que si $(x_0, \lambda_0, \mu_0) \in \mathcal{V}$, la méthode est bien définie et la suite (x_k, λ_k, μ_k) converge quadratiquement vers (x^*, λ^*, μ^*).

3.5.6 Méthode de dualité : Méthode d'Uzawa

La méthode que nous allons présenter est issue de la **théorie de la dualité convexe**, théorie puissante que nous ne pouvons détailler ici. L'idée générale est de considérer le Lagrangien $\mathcal{L}$ au lieu de la fonction J ; ce choix est motivé (au moins) par deux raisons : la fonction Lagrangienne $\mathcal{L}$ englobe à la fois la fonction J et les contraintes h et g et représente bien le problème. Ensuite, nous avons vu qu'une condition nécessaire du premier ordre pour que x^* soit un minimum de J avec contraintes est que x^* (associé aux multiplicateurs de Lagrange) soit un point critique de $\mathcal{L}$. Nous rappelons que le Lagrangien du problème est

$$\mathcal{L}(x, \lambda, \mu) = J(x) + \sum_{i=1}^{p} \lambda_i \, h_i(x) + \sum_{j=1}^{q} \mu_j \, g_j(x) \, .$$

Nous allons avoir besoin de la notion de **point selle** :

Définition 3.5.1 *On appelle **point selle** de $\mathcal{L}$ sur $\mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$ tout triplet $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$ vérifiant l'équation*

$$\mathcal{L}(x^*, \lambda, \mu) \leq \mathcal{L}(x^*, \lambda^*, \mu^*) \leq \mathcal{L}(x, \lambda^*, \mu^*) \tag{3.5.7}$$

pour tous $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$.

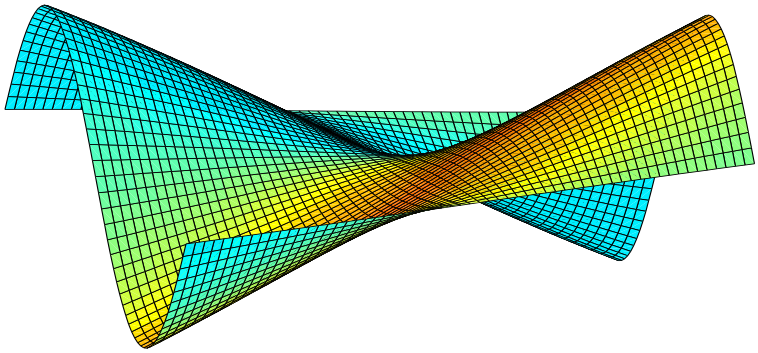

Figure 3.5 : un point selle

La méthode qui va suivre utilise des résultats corollaires du théorème 3.2.5.

Théorème 3.5.5 *Supposons que J, g et h sont $\mathcal{C}^1$ et que le triplet $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$ est un point selle de $\mathcal{L}$ sur $\mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$. Alors ce triplet vérifie les conditions de Karush-Kuhn-Tucker (3.2.9).*

Démonstration - Soit $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$ un point selle de $\mathcal{L}$ sur $\mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$.
La condition de signe des multiplicateurs (3.2.9a) est manifestement vérifiée.
Comme x^* est un minimum (global) de la fonction $x \mapsto \mathcal{L}(x, \lambda^*, \mu^*)$ sur $\mathbb{R}^n$ nous savons (chapitre 1. théorème 2.2.1) que

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0 \ ,$$

c'est-à-dire que la relation (3.2.9d) est satisfaite.
D'autre part la première inégalité de (3.5.7) s'écrit

$$\sum_{i=1}^{p} (\lambda_i - \lambda_i^*)\, h_i(x^*) + \sum_{j=1}^{q} (\mu_j - \mu_j^*)\, g_j(x^*) \leq 0, \text{ pour tous } (\lambda, \mu) \in \mathbb{R}^p \times (\mathbb{R}^+)^q \ .$$

En choisissant successivement

$$(\lambda_1, \ldots, \lambda_{i-1}, \lambda_i, \lambda_{i+1} \ldots, \lambda_p, \mu) = (\lambda_1^*, \ldots, \lambda_{i-1}^*, \lambda_i^* + t, \lambda_{i+1}^*, \ldots, \lambda_p^*, \mu^*)$$

avec $t \in \mathbb{R}$ on obtient $h_i(x^*) = 0, i = 1, \ldots, p$. De même, en choisissant successivement

$$(\lambda, \mu_1, \ldots, \mu_{j-1}, \mu_j, \mu_{j+1} \ldots, \mu_q) = (\lambda^*, \mu_1^*, \ldots, \mu_{j-1}^*, \mu_j^* + s, \mu_{j+1}^* \ldots, \mu_q^*)$$

avec $s \in \mathbb{R}^+$, on obtient $g_j(x^*) \leq 0$. La relation (3.2.9b) est donc aussi satisfaite.
Il reste à montrer la relation de complémentarité : si $j \in \{1, \ldots, q\}$ est tel que $\mu_j^* > 0$ on peut prendre dans ce qui précède $s = \mu_j^*$ puis $s = -\dfrac{\mu_j^*}{2}$: on obtient alors $\mu_j^* g_j(x^*) = 0$. □

Dans la cas **convexe** nous avons **caractérisation** des points selles grâce aux conditions de KKT :

Théorème 3.5.6 *Supposons que J, g et h sont convexes et $\mathcal{C}^1$. Alors le triplet $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$ est point selle de $\mathcal{L}$ sur $\mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$ si et seulement si il vérifie les conditions de KKT (3.2.9).*

Démonstration - Nous devons montrer que si le triplet $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$ vérifie les conditions de KKT (3.2.9), alors il est point selle de $\mathcal{L}$ sur $\mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$.
Comme $\mathcal{L}$ est convexe par rappport à x, la relation (3.2.9d) est équivalente à

$$\mathcal{L}(x^*, \lambda^*, \mu^*) \leq \mathcal{L}(x, \lambda^*, \mu^*) \text{ pour tout } x \in \mathbb{R}^n \ .$$

Montrons l'autre inégalité : tout d'abord

$$\mathcal{L}(x^*, \lambda^*, \mu^*) = J(x^*) \ ,$$

d'après la condition de réalisabilité (3.2.9b) et de complémentarité (3.2.9c) et

$$\mathcal{L}(x^*, \lambda, \mu) = J(x^*) + \sum_{j=1}^{q} \mu_j \, g_j(x^*) \ .$$

Or $\mu_j \geq 0$ et $g_j(x^*) \leq 0$, donc

$$\mathcal{L}(x^*, \lambda^*, \mu^*) \leq \mathcal{L}(x^*, \lambda, \mu) \ , \text{ pour tous } (\lambda, \mu) \in \mathbb{R}^p \times (\mathbb{R}^+)^q \ .$$

□

Le théorème précédent indique que nous allons chercher un triplet $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$ vérifiant les conditions de KKT de la façon suivante :

1. Pour (λ^*, μ^*) fixés dans $\mathbb{R}^p \times (\mathbb{R}^+)^q$, nous allons chercher le minimum **sans contraintes** (sur tout l'espace $\mathbb{R}^n$) de la fonction $x \mapsto \mathcal{L}(x, \lambda^*, \mu^*)$; nous traduisons ainsi ce que signifie le terme de gauche de la relation (3.5.7).
2. Pour x^* fixé dans $\mathbb{R}^n$, on cherche le maximum sur $\mathbb{R}^p \times (\mathbb{R}^+)^q$ (c'est-à-dire avec des contraintes de bornes simples) de la fonction $(\lambda, \mu) \mapsto \mathcal{L}(x^*, \lambda, \mu)$; c'est ce que signifie le terme de droite de la relation (3.5.7).

Bien sûr on ne va pas faire ces deux calculs simultanément ; on va résoudre successivement les étapes 1. et 2. ci-dessus. On obtient l'algorithme d' UZAWA .

Algorithme d'Uzawa

1. **Initialisation**
   $k = 0$: choix de $\lambda^o \in \mathbb{R}^p$ et de $\mu^o \in (\mathbb{R}^+)^q$

2. **Itération** k :
   $\lambda^k = (\lambda_1^k, \ldots, \lambda_p^k) \in \mathbb{R}^p$ et $\mu^k = (\mu_1^k, \ldots, \mu_q^k) \in (\mathbb{R}^+)^q$ sont connus ; puis

   (a) Calcul de $x^k \in \mathbb{R}^n$ solution de

   $$(\mathcal{P}^k) \qquad \min\ \mathcal{L}(x, \lambda^k, \mu^k)\ ,\ x \in \mathbb{R}^n\ .$$

   (b) Calcul de λ^{k+1} et μ^{k+1} avec :

   $$\lambda_i^{k+1} = \lambda_i^k + \rho\ h_i(x^k) \qquad i = 1, \ldots, p$$

   $$\mu_j^{k+1} = \max(0, \mu_j^k + \rho\ g_j(x^k)) \quad j = 1, \ldots, q\ .$$

   où $\rho > 0$ est un réel fixé (choisi par l'utilisateur).

3. **Critère d'arrêt**
   Si $\|x^{k+1} - x^k\| < \varepsilon$, STOP
   Sinon, on pose $k = k + 1$ et on retourne à 2.

L'étape 2a. de cet algorithme revient à résoudre

$$\nabla_x \mathcal{L}(x, \lambda^k, \mu^k) = \nabla J(x) + \sum_{j=1}^{p} \lambda_j^k\ \nabla h_j(x) + \sum_{i=1}^{q} \mu_i^k\ \nabla g_i(x)\ = 0\ .$$

L'étape 2b. est très facile à réaliser.

Théorème 3.5.7 *On suppose que J est $\mathcal{C}^1$ elliptique, (relation (2.1.1) que h et g sont convexes (h est affine), $\mathcal{C}^1$ et lipschitziennes. On suppose de plus que le Lagrangien $\mathcal{L}$ possède un point selle (x^*, λ^*, μ^*) sur $\mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}^+)^q$. Alors, il existe ρ_1, ρ_2 avec $0 < \rho_1 < \rho_2$ tels que pour tout $\rho \in [\rho_1, \rho_2]$ la suite x^k générée par l'algorithme d'Uzawa converge vers x^*.*

Démonstration - Nous avons vu dans le théorème 3.5.5 que si (x^*, λ^*, μ^*) est point selle de $\mathcal{L}$, alors ce triplet vérifie les conditions de KKT. En particulier

$$\lambda^* = \lambda^* + \rho\ h(x^*) \quad \text{et} \quad \mu^* = \max(0, \mu^* + \rho\ g(x^*))\ .$$

Justifions la seconde relation : si $\mu_j^* + \rho\ g_j(x^*) \leq 0$, comme $\mu_j^* \geq 0$ et par complémentarité,

$$(\mu_j^*)^2 + \rho\ \mu_j^*\ g_j(x^*) = (\mu_j^*)^2 \leq 0\ ,$$

et donc $\mu_j^* = 0$. Si $\mu_j^* + \rho\ g_j(x^*) \geq 0$, comme $g_j(x^*) \leq 0$ et par complémentarité,

$$\mu_j^*\ g_j(x^*) + \rho\ (g_j(x^*))^2 = \rho\ (g_j(x^*))^2 \leq 0\ ;$$

donc $g_j(x^*) = 0$. Dans les deux cas : $\mu_j^* = \max(0, \mu_j^* + \rho\ g_j(x^*))$. Nous noterons π_+ la projection de $\mathbb{R}^q$ sur $(\mathbb{R}^+)^q$ de sorte que

$$\mu^* = \pi_+(\mu^* + \rho\ g(x^*))\ .$$

D'autre part,

$$\lambda^{k+1} = \lambda^k + \rho\, h(x^k) \quad \text{et} \quad \mu^{k+1} = \pi_+(\mu^k + \rho\, g(x^k))\,.$$

Par différence on obtient

$$\lambda^{k+1}-\lambda^* = \lambda^k-\lambda^*+\rho\,[h(x^k)-h(x^*)] \quad \text{et} \quad \mu^{k+1}-\mu^* = \pi_+(\mu^k+\rho\, g(x^k))-\pi_+(\mu^*+\rho\, g(x^*))\,.$$

Donc

$$\|\lambda^{k+1}-\lambda^*\|^2 \leq \|\lambda^k-\lambda^*+\rho\,(h(x^k)-h(x^*))\|^2 \text{ et } \|\mu^{k+1}-\mu^*\|^2 \leq \|\mu^k-\mu^*+\rho\,(g(x^k)-g(x^*))\|^2\,,$$

car π_+ est contractante. Cela donne

$$\begin{aligned} \|\lambda^{k+1}-\lambda^*\|^2 \leq \;& \|\lambda^k-\lambda^*\|^2 + \rho^2\|h(x^k)-h(x^*)\|^2 \\ & +2\rho\left(\lambda^k-\lambda^*, h(x^k)-h(x^*)\right)\,, \\ \|\mu^{k+1}-\mu^*\|^2 \leq \;& \|\mu^k-\mu^*\|^2 + \rho^2\|g(x^k)-g(x^*)\|^2 \\ & +2\rho\left(\mu^k-\mu^*, g(x^k)-g(x^*)\right)\,. \end{aligned} \tag{3.5.8}$$

D'autre part nous savons que pour tout $y \in \mathbb{R}^n$

$$J(y) - J(x^*) + \sum_{i=1}^{p} \lambda_i^*\,(h_i(y)-h_i(x^*)) + \sum_{j=1}^{q} \mu_j^*\,(g_j(y)-g_j(x^*)) \geq 0\,.$$

Prenons $y = x^* + t(x-x^*)$ avec $t \in]\,0,1\,[$ et $x \in \mathbb{R}^n$. Par convexité

$$h_i(x^*+t(x-x^*)) - h_i(x^*) \leq t\,(h_i(x)-h_i(x^*))\,, i = 1,\ldots,p, \text{ et}$$
$$g_j(x^*+t(x-x^*)) - g_j(x^*) \leq t\,(g_j(x)-g_j(x^*))\,, j = 1,\ldots,q\,.$$

Donc

$$\frac{J(x^*+t(x-x^*))-J(x^*)}{t} + \sum_{i=1}^{p} \lambda_i^*\,(h_i(x)-h_i(x^*)) + \sum_{j=1}^{q} \mu_j^*\,(g_j(x)-g_j(x^*)) \geq 0\,. \tag{3.5.9}$$

En passant à la limite pour $t \to 0^+$ on obtient :

$$(\nabla J(x^*), x-x^*) + (\lambda^*, h(x)-h(x^*)) + (\mu^*, g(x)-g(x^*)) \geq 0 \qquad \forall x \in \mathbb{R}^n\,. \tag{3.5.10}$$

Le même raisonnement sur l'itération k de l'algorithme donne

$$\left(\nabla J(x^k), x-x^k\right) + \left(\lambda^k, h(x)-h(x^k)\right) + \left(\mu^k, g(x)-g(x^k)\right) \geq 0 \quad \forall x \in \mathbb{R}^n\,. \tag{3.5.11}$$

On prend $x = x^k$ dans (3.5.10) et $x = x^*$ dans (3.5.11), on somme et on utilise l'ellipticité de J :

$$\begin{aligned} &\left(\lambda^k-\lambda^*, h(x^k)-h(x^*)\right) + \left(\mu^k-\mu^*, g(x^k)-g(x^*)\right) \\ &\qquad \leq -\left(\nabla J(x^*) - \nabla J(x^k), x^*-x^k\right) \\ &\qquad \leq -\alpha\|x^*-x^k\|^2. \end{aligned} \tag{3.5.12}$$

Finalement (3.5.8) devient

$$\nu_{k+1} \leq \nu_k + \rho^2 \left(\|h(x^k) - h(x^*)\|^2 + \|g(x^k) - g(x^*)\|^2 \right) - 2\,\rho\,\alpha \|x^* - x^k\|^2. \qquad (3.5.13)$$

où on a posé

$$\nu_k = \|\lambda^{k+1} - \lambda^*\|^2 + \|\mu^{k+1} - \mu^*\|^2 .$$

Comme h et g sont lipschitziennes

$$\|h(x^k) - h(x^*)\| \leq M_h \|x^* - x^k\| \quad \text{et} \quad \|g(x^k) - g(x^*)\| \leq M_g \|x^* - x^k\| .$$

Donc

$$\nu_{k+1} - \nu_k \leq \left[\rho^2 (M_h^2 + M_g^2) - 2\,\rho\,\alpha\right] \|x^* - x^k\|^2. \qquad (3.5.14)$$

Choisissons ρ dans l'intervalle $]\,0, \rho_2\,[$ avec $\rho_2 = \dfrac{\alpha}{2\,(M_h^2 + M_g^2)}$. La suite ν_k est alors une suite décroissante, positive. Par conséquent, elle est convergente et la différence $\nu_{k+1} - \nu_k$ tend vers 0. La relation (3.5.14) permet de conclure que la suite x_k converge vers x^*.
Remarquons qu'on ne peut rien dire (sans hypothèse supplémentaire) sur la convergence des multiplicateurs.
Remarquons enfin que nous sommes dans le cas convexe avec J elliptique. Donc tout point selle du Lagrangien est aussi solution (unique) du problème $(\mathcal{P})$. □

[Travaux Pratiques]

[Méthodes classiques]

Programmer les algorithmes du cours sur la minimisation avec contraintes dans $\mathbb{R}^n$ et les tester.
- Gradient projeté et Newton projeté pour des problèmes avec contraintes de borne
- Méthode de Lagrange - Newton pour des problèmes avec contraintes en égalité
- Méthode de pénalisation (inexacte)
- Méthode d'Uzawa avec le Lagrangien ordinaire

Pour chacun d'entre eux, une étude de sensibilité sur le point de départ (initialisation) et les différents paramètres (pas, paramètre d'augmentation) sera menée le plus rigoureusement possible.
On fera une comparaison numérique des trois méthodes surtout en termes de
- vitesse de convergence - nombre d'itérations - temps CPU
- Robustesse et domaine de validité

[Lagrangien / Lagrangien augmenté sur un problème quadratique]

Le problème est le suivant

$$(\mathcal{P}) \quad \left\{ \begin{array}{c} \min \frac{1}{2}\, x^t G x + c^t x \\ Ax = b \end{array} \right.$$

où $x \in \mathbb{R}^n,\ b \in \mathbb{R}^p$, A est une matrice p $\times$ n à coefficients réels et G est une matrice $n \times n$ symétrique à coefficients réels , (éventuellement définie positive)

On rappelle que , dans la méthode d'Uzawa le pas ρ est le réel permettant d'incrémenter la suite des multiplicateurs .

Lagrangien - Méthode d'Uzawa

- Ecrire le Lagrangien $\mathcal{L}$ associé à ce problème .
- Résoudre $(\mathcal{P})$ par la méthode d' Uzawa avec un pas ρ constant (mais choisi judicieusement)

Lagrangien augmenté - Pas constant

- Soit $\mathcal{L}_r$ défini de la manière suivante :
$$\forall (x,q) \in \mathbb{R}^n \times \mathbb{R}^p \qquad \mathcal{L}_r(x,q) = \mathcal{L}(x,q) + \frac{r}{2} \|Ax-b\|^2$$
où r est un réel positif et $\|\ \|$ désigne la norme euclidienne de $\mathbb{R}^p$.
- Programmer la méthode d'Uzawa avec $\mathcal{L}_r$. Etudier plus particulièrement la rapidité de la convergence (à précision donnée) pour différentes valeurs du paramètre r.
Que se passe-t'il si r = 0 ?
On prendra le pas ρ constant égal à r .

[Lagrangien augmenté - Variantes]

Les données sont les mêmes que dans la section précédente mais cette fois le pas ρ n'est pas choisi constant .

Algorithme du résidu minimal

On le détermine à chaque itération grâce à l' algorithme dit du **résidu minimal** ce qui donne , combiné à la méthode d'Uzawa :

1. q_0 donné , $x_0 = -G_r^{-1}[c + A^t(q_0 - rb)]$ où $G_r = G + r\ A^tA$.
$g_0 = Ax_0 - b$
2. $q_n\ ,\ x_n\ ,\ g_n$ connus :

$$(\mathcal{RM}) \quad \left\{ \begin{array}{lcl} * \ z_n & = & -G_r^{-1}A^t g_n \\ * \ \rho_n & = & -\dfrac{< g_n, Az_n >}{\|Az_n\|^2} \\ * \ q_{n+1} & = & q_n + \rho_n\ g_n \\ * \ x_{n+1} & = & x_n + \rho_n\ z_n \\ * \ g_{n+1} & = & g_n + \rho_n\ Az_n \end{array} \right.$$

$\langle\ ,\ \rangle$ désigne le produit scalaire usuel de $\mathbb{R}^n$ et $\|\ \|$ la norme euclidienne.

Algorithme de descente optimale

On détermine le pas ρ à chaque itération grâce à l' algorithme de **descente optimale** ce qui donne, combiné à la méthode d'Uzawa :

1. q_0 donné , $x_0 = -G_r^{-1}[c + A^t(q_0 - rb)]$ où $G_r = G + r\ A^tA$.
   $g_0 = Ax_0 - b$
2. $q_n\ ,\ x_n\ ,\ g_n$ connus :

$$(\mathcal{DO}) \quad \begin{cases} * \ z_n & = & -G_r^{-1}A^t g_n \\ * \ \rho_n & = & -\dfrac{\| g_n \|^2}{< g_n, Az_n >} \\ * \ q_{n+1} & = & q_n + \rho_n\ g_n \\ * \ x_{n+1} & = & x_n + \rho_n\ z_n \\ * \ g_{n+1} & = & g_n + \rho_n\ Az_n \end{cases}$$

Algorithme du gradient conjugué

On détermine le pas ρ à chaque itération grâce à l' algorithme du **gradient conjugué** ce qui donne, combiné à la méthode d'Uzawa :

1. q_0 donné , $x_0 = -G_r^{-1}[c + A^t(q_0 - rb)]$ où $G_r = G + r\ A^tA$.
   $g_0 = Ax_0 - b$ et $w_0 = g_0$
2. Itération n : $q_n\ ,\ x_n\ ,\ g_n\ ,\ w_n$ connus :

$$(\mathcal{GC}) \quad \begin{cases} * \ z_n = -G_r^{-1}A^t w_n \\ * \ \rho_n = -\dfrac{\| g_n \|^2}{< Az_n, w_n >} \\ * \ q_{n+1} = q_n + \rho_n\ w_n \\ * \ x_{n+1} = x_n + \rho_n\ z_n \\ * \ g_{n+1} = g_n + \rho_n Az_n \\ * \ \lambda_n = \dfrac{\| g_{n+1} \|^2}{\| g_n \|^2} \\ * \ w_{n+1} = g_{n+1} + \lambda_n w_n \end{cases}$$

Méthode d'Arrow-Hurwicz (Relaxation)

Pour améliorer la convergence on peut parfois combiner la méthode précédente avec une méthode de relaxation . Plus précisément :

Soit ω un réel compris entre 0 et 1 . On considère l'algorithme suivant :

1. q_0 et x_0 donnés

2. q_n , x_n connus :

$$(\mathcal{AH}) \quad \begin{cases} * \ \text{ calcul de } x_{n+\frac{1}{2}} \text{ par l'algorithme } (\mathcal{GC}) \\ * \ x_{n+1} = x_n + \omega \, (x_{n+\frac{1}{2}} - x_n) \\ * \ q_{n+1} = q_n + \rho \, (A x_{n+1} - b) \end{cases}$$

où ρ = r par exemple .
On testera plusieurs valeurs de ω.

Tests numériques

1. p= 3

$$\text{G} = \begin{bmatrix} 2 & -1 & 0 & 0 & .. & .. \\ -1 & 2 & -1 & 0 & .. & .. \\ 0 & -1 & 2 & -1 & 0 & .. \\ .. & \ddots & \ddots & \ddots & .. & .. \\ .. & .. & 0 & -1 & 2 & -1 \\ .. & .. & .. & 0 & -1 & 2 \end{bmatrix} \quad \text{c} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \\ 1 \end{bmatrix}$$

$$\text{A} = \begin{bmatrix} 3 & 1 & 0 & -1 & 0 & 0 \\ -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 \end{bmatrix} \qquad \text{b} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

2.

$$\begin{cases} \min \frac{1}{2} \, (x_1^2 + 2x_1 \, x_2 + x_3^2 + x_4^2) - 4x_1 \ + 5x_2 \\ \\ x_1 + 2 \, x_4 = 5 \\ x_2 - \ x_3 = 1 \end{cases}$$

[Exercices]

[Conditions de qualification]

1. Montrer que la condition (CQ2) donnée par la définition 3.2.4 implique la condition de régularité de MANGASARIAN-FROMOWITZ (CQ1) de la définition 3.2.3.

---

2. Dans $\mathbb{R}^2$ on considère les contraintes

$$x_1 \geq 0 \ , \ x_2 \geq 0 \ , \ x_2 - (x_1 - 1)^2 \leq 0 \ .$$

Dessiner l'ensemble C des points satisfaisant ces contraintes.
Montrer que le point $x_1 = 1$, $x_2 = 0$ est réalisable mais pas régulier.

[Le théorème de projection]

3. Soit H un sous-espace vectoriel de V Hilbert . Soit f un élément de V.
Montrer que g ($\in$ H) est le projeté de f sur H si et seulement si $f - g \in H^{\perp}$.

---

4. Soit C un convexe fermé non vide de $\mathbb{R}^n$ et P la projection de $\mathbb{R}^n$ sur C.
$(\cdot, \cdot)$ désigne le produit scalaire usuel de $\mathbb{R}^n$ et $\| \cdot \|$ la norme euclidienne.
   (a) Soit $x \in \mathbb{R}^n$ et $P(x)$ son projeté sur C. Rappeler la caractérisation de $P(x)$ en termes de produit scalaire. Montrer qu'une formulation équivalente de cette caractérisation est
   $$P(x) \in C \text{ et } \forall y \in C \qquad (x - y, P(x) - y) \geq 0 \, .$$
   (b) Montrer que tous x et y de $\mathbb{R}^n$,
   $$0 \leq (x - P(x), P(x) - P(y)) \leq \ (x - y, P(x) - P(y)) - \|P(x) - P(y)\|^2 \, .$$
   (c) Montrer que pour tous x et y de $\mathbb{R}^n$,
   $$0 \leq \|y - P(y)\|^2 - \|x - P(x)\|^2 - 2 \ (x - P(x), y - x) \leq \|y - x\|^2 - \|P(y) - P(x)\|^2 \, .$$
   (d) En déduire que l'application $f : \mathbb{R}^n \to \mathbb{R}$ définie par $f(x) = \|x - P(x)\|^2$ est différentiable en tout point. Quelle est sa différentielle ?
   (e) Soit $C = \{ x = (x_1, x_2) \in \mathbb{R}^2 \ \mid \ x_1 \geq 0, \ x_2 \geq 0 \ \}$. Déterminer l'expression de l'application $g : \mathbb{R}^2 \to \mathbb{R}$ définie par $g(x) = \|x - P(x)\|$.
   Est-elle différentiable ?

[Conditions d'optimalité]

5. Soient les points $(t_i, x_i), i = 1, \cdots, 10$, donnés par le tableau suivant :

1	2	3	4	5	6	7	8	9	10
0	-3	6	-3	6	3.8	5	-2	1.4	8

   (a) Quelle est l'équation de la droite de régression de ce nuage ? (on l'appelle D)
   (b) Quel est le point le plus "éloigné" de D ?
   (c) Quelle est la droite de régression D' obtenue en enlevant ce point ?
   (d) Calculer la droite de régression Δ obtenue pour le nuage en imposant $b \geq 0$ puis $b \geq 1$ (l'équation de Δ est de la forme $x = a\, t + b$.)

---

6. Soient A et b définis par
$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \quad \text{et} \quad b = \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix} ,$$
et $J : \mathbb{R}^3 \to \mathbb{R}$ définie par
$$J(y) = \frac{1}{2} \, (Ay, y) + (b, y) \ .$$

(a) Le problème

$$(\mathcal{P}) \begin{cases} \min\ J(y) \\ y_1 \geq 0 \\ y_2 + y_3 = 0 \\ y \in \mathbb{R}^3 \end{cases}$$

admet-il une solution ? Est-elle unique ?

(b) Ecrire les conditions d'optimalité que doit vérifier la solution de $(\mathcal{P})$. Résoudre

(c) Ecrire les algorithmes que vous connaissez dans ce cas précis

---

7. Maximiser la fonction $J(x, y) = 14x - x^2 + 6y - y^2 + 7$ sous les contraintes

$$x + y \leq 2,\ x + 2y \leq 3\ .$$

On précisera l'existence et l'unicité et on dessinera l'ensemble réalisable C.

---

8. Résoudre

$$\begin{array}{ll} \min & \dfrac{1}{2}\,(Ax, x) - (b, x) \\ & x_1 \geq 1 \\ & x_2 - 2x_3 = 1 \end{array}$$

avec

$$A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix} \text{ et } b = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}\ .$$

Comparer avec la solution sans contraintes.

---

9. On veut maximiser la fonction f de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par :

$$f(x_1, x_2) = 2x_1 - x_1^2 + x_2$$

sur l'ensemble K défini par :

$$K = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 + x_2 \leq 1\ ,\ x_1 x_2 \geq 0\}$$

(a) Résoudre ce problème graphiquement.

(b) Ecrire les conditions de KKT et résoudre.

(c) Peut-on résoudre par la méthode d'Uzawa ?

(d) Minimiser la fonction g sur l'ensemble K avec

$$g(x_1, x_2) = 2x_1^2 - 3x_1 + x_2^2 + 3x_2\ .$$

---

10. L'évolution de la concentration c d'une espèce dans un mélange est donnée par une loi linéaire en fonction du temps :
$$c(t) = at + b\ .$$
Un expérimentateur fait une série de mesures pour déterminer les paramètres inconnus, mesures résumées dans le tableau suivant.

0	1	2	3	4	5	6	7	8	9	10	11
1.54	3.06	4.97	7.43	10.65	14.92	20.6	28.2	38.42	52.15	70.65	96.6

Déterminer les paramètres a et b. En réalité la concentration initiale est positive. Combien valent les paramètres a et b ?
Pensez-vous que la loi est vraiment linéaire ? Expliquer sommairement (sans faire de calculs, en posant juste le problème) la démarche à suivre pour trouver les paramètres d'une loi de la forme
$$c(t) = ae^{t} + be^{-t}\ ,$$
sachant que la concentration initiale est positive.

---

11. On se place dans $\mathbb{R}^2$, et on note $x = (x_1, x_2)$. On considère la fonction f de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par :
$$f(x) = \frac{1}{2}\,(Bx, x) + (b, x) = \frac{1}{2}(x_1^2 + \alpha x_2^2) + x_1$$
où B est une matrice symétrique 2-2, b un vecteur de $\mathbb{R}^2$ et $\alpha \in \mathbb{R}$.
   (a) Préciser B et b ; calculer $\nabla f(x)$.
   (b) Donner une condition nécessaire pour que x soit un minimum (local) sans contraintes de f.
      i. Si $\alpha = 0$, montrer que f possède un minimum et qu'il y a une infinité de x réalisant ce minimum. Si $\alpha \neq 0$, quel est l'élément x^* pouvant éventuellement réaliser le minimum ?
      ii. Si $\alpha > 0$, x^* réalise-t'il le minimum de f ? Pourquoi ?
      iii. Si $\alpha < 0$, montrer que f ne possède pas de minimum.
      iv. Ecrire les deux premières itérations de l'algorithme du gradient dans le cas $\alpha = 2$.
   (c) On suppose maintenant que $\alpha = 2$; on veut minimiser la fonction f avec la contrainte supplémentaire :
$$\sqrt{x_1^2 + x_2^2} \leq \frac{1}{2}\ .$$
      i. La fonction f admet-elle un minimum ?
      ii. Ecrire les conditions de KARUSH-KUHN-TUCKER permettant de calculer le minimum éventuel. Résoudre.
      iii. Décrire sur cet exemple les deux premières itérations de l'algorithme d' UZAWA.

---

12. **Méthode de pénalisation**
Dans tout ce qui suit $\mathbb{R}^n$ est muni du produit scalaire $(\cdot,\cdot)$ et de la norme $\|\ \|$.

(a) Soit g une fonction convexe de $\mathbb{R}^n$ dans $\mathbb{R}$ et soit g^+ la fonction définie par

$$g^+(x) = \max(g(x), 0).$$

– Montrer que g^+ est convexe puis que $(g^+)^2$ est convexe.
– Montrer que la fonction $s \mapsto (s^+)^2$ de $\mathbb{R}$ dans $\mathbb{R}$ est dérivable. En déduire que si g est Gâteaux (respectivement Fréchet-différentiable) , $(g^+)^2$ est Gâteaux-différentiable (respectivement Fréchet-différentiable) et calculer sa différentielle en fonction de celle de g.

(b) Pour $i = 1, ..., m$ on se donne $c_i \in \mathbb{R}^n$ et $\gamma_i \in \mathbb{R}$, puis on définit pour $y \in \mathbb{R}^n$

$$g_i(y) = (c_i, y) - \gamma_i\ ,$$

et

$$K = \{\ y \in \mathbb{R}^n\ \mid\ g_i(y) \leq 0,\ \text{pour } i = 1, \ldots, m\ \}.$$

On suppose que $K \neq \emptyset$. Soient A une matrice $n \times n$, symétrique, définie positive et $b \in \mathbb{R}^n$. Montrer qu'il existe une solution unique $\bar{x}$ au problème

$$(\mathcal{P}) \qquad x \in K \quad ; \quad J(x) = \min_{y \in K} J(y),$$

où

$$J(y) = \frac{1}{2}\,(A\,y, y) - (b, y)\,.$$

(c) On pose pour $\varepsilon > 0$, et pour $y \in \mathbb{R}^n$,

$$J_\varepsilon(y) = \frac{1}{2}\,(A\,y, y) - (b, y) + \frac{1}{2\varepsilon}\sum_{i=1}^{m}(g_i^+(y))^2.$$

Montrer que pour $\varepsilon > 0$ fixé, il existe une solution unique $x_\varepsilon \in \mathbb{R}^n$ au problème

$$(\mathcal{P}_\varepsilon) \qquad x_\varepsilon \in \mathbb{R}^n \quad ; \quad J_\varepsilon(x_\varepsilon) = \min_{y \in \mathbb{R}^n} J_\varepsilon(y).$$

(d) Montrer que x_ε est solution de $(\mathcal{P}_\varepsilon)$ si et seulement si x_ε est solution d'une équation de la forme

$$(\mathcal{E}_\varepsilon) \qquad A\,x_\varepsilon - b + \sum_{i=1}^{m} h_\varepsilon^i c_i = 0,$$

où

$$h_\varepsilon^i = \frac{g_i^+(x_\varepsilon)}{\varepsilon}.$$

(e) Montrer que x_ε reste borné dans $\mathbb{R}^n$ indépendamment de ε et que $\dfrac{g_i^+(x_\varepsilon)}{\sqrt{\varepsilon}}$ reste borné dans $\mathbb{R}$ indépendamment de ε pour $i = 1, \ldots, m$.

(f) Montrer que x_ε converge vers $\bar{x}$ lorsque $\varepsilon \to 0$.

(g) On suppose que $c_1, ..., c_m$ sont linéairement indépendants. Si $i \in \{\ 1, \ldots, m\ \}$, on note Π_i l'orthogonal dans $\mathbb{R}^n$ de $\{c_j\}_{j \neq i}$. Montrer qu'il existe $d_i \in \Pi_i$ de norme 1 (par exemple) tel que $(d_i, c_i) \neq 0$.
En utilisant la multiplication scalaire de $(\mathcal{E}_\varepsilon)$ par d_i et un résultat précédent, montrer que h_ε^i reste borné dans $\mathbb{R}$ indépendamment de ε.

(h) En déduire que si $\bar{x}$ est solution du problème $(\mathcal{P})$ il existe des nombres réels $h^i \geq 0$, $i = 1, \ldots, m$ tels que

$$(\mathcal{E}) \qquad A\,\bar{x} - b + \sum_{i=1}^{m} h^i c_i = 0 \quad ; \quad x \in K,$$

et $h^i\, g_i(\bar{x}) = 0$.

---

13. **Analyse de sensibilité**
Soient f et h deux fonctions de classe $\mathcal{C}^2$ de $\mathbb{R}$ vers $\mathbb{R}$ et $t \in \mathbb{R}$.
On considère le problème de minimisation sous contraintes suivant

$$(\mathcal{P}_t) \qquad \left\{ \begin{array}{ll} \min & f(x) \\ \text{tel que} & h(x) = t\ . \end{array} \right.$$

On suppose que le problème $(\mathcal{P}_0)$ (obtenu pour $t = 0$) admet au moins une solution x^* et que ce point est **régulier**.

(a) Rappeler ce que signifie "régulier" dans ce cas précis et écrire le système d'optimalité permettant de calculer x^* (on notera λ^* le multiplicateur associé).

(b) Quel est le système d'optimalité que doit satisfaire une éventuelle solution de $(\mathcal{P}_t)$

(c) On note $\mathcal{L}(x, \lambda)$ le Lagrangien du problème $(\mathcal{P}_0)$ et

$$H(x^*, \lambda^*) = \begin{pmatrix} \mathcal{L}''(x^*, \lambda^*) & h'(x^*) \\ h'(x^*) & 0 \end{pmatrix}$$

où la dérivation se fait par rapport à la variable x. Montrer que le système d'optimalité établi en 2. a une solution $(\mathbf{x}(t), \lambda(t))$ dans un voisinage de (x^*, λ^*) et que $\mathbf{x}(0) = x^*$, $\lambda^* = \lambda(0)$. Montrer que les applications $t \mapsto \mathbf{x}(t)$ et $t \mapsto \lambda(t)$ sont dérivables au voisinage de 0.
(On utilisera le théorème des fonctions implicites (cf. Annexe A)).

(d) La fonction valeur optimale du problème $(\mathcal{P}_t)$ est donc égale à $F(t) = f(\mathbf{x}(t))$ au voisinage de $t = 0$ et vérifie $F(0) = f(x^*)$.
Montrer que

$$F'(0) = \frac{d\,f}{dx}(x^*)\,\frac{d\,\mathbf{x}}{dt}(0) \quad \text{ et que } \quad \frac{d\,(h \circ \mathbf{x})}{dt}(0) = \frac{d\,h}{dx}(x^*)\,\frac{d\,\mathbf{x}}{dt}(0)\ .$$

En déduire que $F'(0) = -\lambda^*$, et donner un développement limité de F au voisinage de $t = 0$.

---

14. Soient U et V deux sous-ensembles convexes fermés non vides de $\mathbb{R}^p$ et $\mathbb{R}^q$ respectivement et $\mathcal{L}$ une application de $\mathbb{R}^p \times \mathbb{R}^q$ dans $\mathbb{R}$. On fait les hypothèses suivantes

(H1) pour tout $v \in V$, l'application $u \in U \mapsto \mathcal{L}(u, v)$ est convexe et continue,
(H2) pour tout $u \in U$, l'application $v \in V \mapsto \mathcal{L}(u, v)$ est concave et continue,
(H3) U et V sont bornés.

(a) Dans cette question, on suppose que pour tout $v \in V$, l'application $u \in U \mapsto \mathcal{L}(u, v)$ est strictement convexe.

i. Pour tout $v \in V$ on pose
$$G(v) = \inf_{u \in U} \mathcal{L}(u, v) \ .$$
Montrer que l'inf est atteint en un point unique que l'on notera $\varphi(v)$.

ii. Pour tout $u \in U$ on pose
$$F(u) = \sup_{v \in V} \mathcal{L}(u, v) \ .$$
Montrer que le sup est atteint.

iii. Montrer qu'il existe $v^* \in V$ tel que
$$G(v^*) = \max_{v \in V} G(v) \ .$$
Indication : On prendra une suite maximisante et on montrera qu'elle converge vers v^*.

iv. Etant donné $v \in V$, on considère les suites v_n et u_n définies par
$$v_n = (1 - \frac{1}{n})v^* + \frac{1}{n}v, u_n = \varphi(v_n) \ .$$
Montrer que (u_n) admet une seule valeur d'adhérence $\varphi(v^*)$ notée u^*, que (u^*, v^*) est un point-selle de $\mathcal{L}$ sur $U \times V$.
Indication : On pourra montrer que $G(v^*) \leq \mathcal{L}(u, v^*)$ pour tout $u \in U$ puis $G(v^*) \geq \mathcal{L}(u^*, v)$.

(b) On ne suppose plus la stricte convexité. Montrer l'existence d'un point-selle de $\mathcal{L}$ sur $U \times V$.
Indication : On pourra considérer $\mathcal{L}_n(u, v) = \mathcal{L}(u, v) + \frac{1}{n}\|u\|^2$.

[Algorithmes]

15. **Méthode d'Uzawa et d'Arrow-Hurwicz**
Soit A une matrice symétrique définie positive, et soit pour y dans $\mathbb{R}^n$:

$$J(v) = \frac{1}{2}(Av, v) - (b, v)$$

où b est un vecteur donné de $\mathbb{R}^n$.
On cherche à approcher la solution unique du problème : trouver u tel que

$$\begin{cases} u \in U = \{v \in \mathbb{R}^n \mid Cv = 0\} \\ J(u) = \inf\limits_{v \in U} J(v) \end{cases}$$

où C est une matrice m × n .

(a) Décrire la méthode d'Uzawa pour ce problème.

(b) On définit la méthode itérative suivante :

$$\begin{cases} \bullet \quad (u^o, \lambda^o) \text{ donnés dans } \mathbb{R}^n \times \mathbb{R}^m \\ \bullet \quad (u^k, \lambda^k) \text{ étant connus , calcul de } (u^{k+1}, \lambda^{k+1}) \\ \quad u^{k+1} = u^k - \rho_1(Au^k - b + C^t\lambda^k) \\ \quad \lambda^{k+1} = \lambda^k + \rho_1\rho_2 C u^{k+1} \end{cases}$$

où ρ_1 et ρ_2 sont des paramètres > 0.
Montrer que si $\rho_1 > 0$ est suffisamment petit ,

$$\beta \stackrel{def}{=} \|I - \rho_1 A\| < 1$$

(c) Soit λ un vecteur de $\mathbb{R}^m$ qui vérifie : $Au + C^t\lambda = b$.
Dire pourquoi il existe de tels vecteurs .
On choisit le paramètre ρ_1 pour que l'inégalité $\beta < 1$ ait lieu. Montrer que si le paramètre $\rho_2 > 0$ est suffisamment petit , il existe une constante $\gamma > 0$ indépendante de l'entier k telle que :

$$\gamma\|u^{k+1} - u\|_n^2 \leq (\frac{\|\lambda^k - \lambda\|_m^2}{\rho_2} + \beta\|u^k - u\|_n^2) - (\frac{\|\lambda^{k+1} - \lambda\|_m^2}{\rho_2} + \beta\|u^{k+1} - u\|_n^2)$$

(d) En déduire que pour de tels choix des paramètres ρ_1 et ρ_2 , on a $\lim\limits_{k \to +\infty} u_k = u$.

(e) Que peut-on dire de la suite (λ^k) lorsque rang(C) = m ?

La méthode ci-dessus s'appelle la méthode *d'Arrow-Hurwicz.*
Quel avantage présente -t'elle par rapport à la méthode d'Uzawa ?

---

16. On considère le problème d'optimisation suivant :

$$\min \ \{ \ J(v) \ \mid \ Bv = 0\} \ , \tag{1}$$

où J est une fonctionnelle quadratique définie sur $\mathbb{R}^N$ par

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle_N \ ,$$

A est une matrice réelle $N \times N$, symétrique, définie positive, B est une matrice réelle $M \times N$ dont les vecteurs lignes sont indépendants ($M \leq N$), et $\langle\ ,\ \rangle_N$ désigne le produit scalaire usuel sur $\mathbb{R}^N$.

(a) Montrer que le problème (1) admet une solution unique notée u.

(b) Montrer que la solution u de (1) est caractérisée par l'existence de $p \in \mathbb{R}^M$ tel que

$$\left\{ \begin{array}{rcl} Au + B^t p & = & b \\ Bu & = & 0 \end{array} \right. , \tag{2}$$

où B^t désigne la matrice transposée de B.

(c) Montrer que les relations (2) caractérisent les points-selles sur $\mathbb{R}^N \times \mathbb{R}^M$ du Lagrangien du problème :

$$\mathcal{L}(v, q) = J(v) + \langle q, Bv\rangle_M \ .$$

(d) On introduit le Lagrangien augmenté du problème $\mathcal{L}_r$ avec $r > 0$, défini par :

$$\mathcal{L}_r(v, q) = \mathcal{L}(v, q) + \frac{r}{2}\|Bv\|_M^2 \ ,$$

où $\|\ \|_M$ désigne la norme euclidienne usuelle sur $\mathbb{R}^M$.

Montrer que les points-selles de $\mathcal{L}_r$ sont et ne sont que les points-selles de $\mathcal{L}$.

Montrer, grâce aux questions précédentes qu'une caractérisation des points-selles de $\mathcal{L}_r$ est :

$$\left\{ \begin{array}{rcl} (A + rB^tB)u + B^t p & = & b \\ Bu & = & 0 \end{array} \right. , \tag{3}$$

(e) On considère maintenant l'algorithme suivant :

i. p^o donné arbitrairement dans $\mathbb{R}^M$

ii. p^n étant connu, on calcule u^n puis p^{n+1} par

$$\left\{ \begin{array}{l} \mathcal{L}_r(u^n, p^n) \leq \mathcal{L}_r(v, p^n) \quad \forall v \in \mathbb{R}^N \\ u^n \in \mathbb{R}^N \end{array} \right. \tag{4}$$

$$p^{n+1} = p^n + \rho_n B u^n \ , \ \rho_n > 0 \ . \tag{5}$$

Reconnaissez-vous cette méthode ?

Montrer que la relation (4) est équivalente à

$$(A + rB^tB)u^n + B^t p^n = b \ . \tag{6}$$

(f) On pose $\bar{u}^n = u^n - u$ et $\bar{p}^n = p^n - p$ où u est la solution de (1) et p le multiplicateur associé défini par (3). En utilisant les relations (3), (5) et (6), montrer successivement que

i. $\|\bar{p}^n\|_M^2 - \|\bar{p}^{n+1}\|_M^2 = -2\rho_n \langle \bar{p}^n, B\bar{u}^n \rangle_M - \rho_n^2 \|B\bar{u}^n\|_M^2$,

ii. $\langle A\bar{u}^n, \bar{u}^n \rangle_N + r\|B\bar{u}^n\|_M^2 = -\langle \bar{p}^n, B\bar{u}^n \rangle_M$,

iii. $\|\bar{p}^n\|_M^2 - \|\bar{p}^{n+1}\|_M^2 = 2\rho_n \langle A\bar{u}^n, \bar{u}^n \rangle_N + \rho_n(2r - \rho_n)\|B\bar{u}^n\|_M^2$.

(g) En déduire que si on prend ρ_n tel que $0 < \alpha_0 \leq \rho_n \leq 2r$, la suite $\|\bar{p}^n\|_M^2$ est décroissante.

(h) Montrer que la suite $\|\bar{p}^n\|_M^2$ est convergente et que $\lim_{n\to+\infty} \langle A\bar{u}^n, \bar{u}^n \rangle_N = 0$.

Que peut-on en déduire sur la suite u^n ?

(i) On considère à présent la fonctionnelle J_r^* définie sur $\mathbb{R}^M$ par :

$$J_r^*(q) = -\min_{v\in\mathbb{R}^N} \mathcal{L}_r(v, q) .$$

i. Montrer que

$$\forall q \in \mathbb{R}^M \qquad J_r^*(q) = \frac{1}{2}\left\langle BA_r^{-1}B^t q, q\right\rangle_M - \left\langle BA_r^{-1}b, q\right\rangle_M + \frac{1}{2}\left\langle A_r^{-1}b, b\right\rangle_N ,$$

où $A_r = A + rB^tB$.

ii. Montrer que par élimination de u^n, l'algorithme décrit par (4) et (5) peut s'écrire :
– p^o donné arbitrairement dans $\mathbb{R}^M$
– $p^{n+1} = p^n - \rho_n(BA_r^{-1}B^t p^n - BA_r^{-1}b)$.

iii. Montrer que l'algorithme décrit dans la partie 2. par (4) et (5) est en fait l'algorithme du gradient appliqué à la minimisation de la fonctionnelle J_r^* sur $\mathbb{R}^M$.

---

17. Soit A une matrice symétrique définie positive , et soit pour y dans $\mathbb{R}^n$:

$$J(y) = \frac{1}{2}(Ay, y)_n - (b, y)_n$$

où b est une vecteur donné de $\mathbb{R}^n$ et $(\cdot,\cdot)_n$ désigne le produit scalaire de $\mathbb{R}^n$. On considère le sous-ensemble convexe fermé de $\mathbb{R}^n$ défini par :

$$K = \{y \in \mathbb{R}^n \mid g_i(x) = 0, i = 1, \ldots, p\}$$

où les g_i sont des contraintes affines définies par :

$$g_i(y) = (c_i, y)_n - \gamma_i \quad , \ c_i \in \mathbb{R}^n \ , \ \gamma_i \in \mathbb{R}$$

Les vecteurs c_i seront supposés indépendants et l'ensemble K non vide. On considère le problème

$$\min\ J(y) \ , \ y \in K \qquad (\mathcal{P})$$

(a) Ecrire le Lagrangien du problème $\mathcal{L}(y,\mu)$ et pour $\mu \in \mathbb{R}^p$ calculer :

$$h(\mu) = \min_{y\in\mathbb{R}^n} \mathcal{L}(y,\mu) \ .$$

On écrira $h(\mu)$ sous la forme $h(\mu) = -\frac{1}{2}\left(B\mu,\mu\right)_p + (d,\mu)_p + \alpha$, avec B matrice $p\times p$, $d\in\mathbb{R}^p$ et $\alpha\in\mathbb{R}$.
On admettra que la matrice $p\times p$ de coefficients $\left(c_i, A^{-1}c_j\right)_n$ est inversible (c'est une matrice de Gram associée au produit scalaire défini par A).

(b) Montrer que $\lambda\in\mathbb{R}^p$ satisfait les conditions de KKT pour le problème $\mathcal{P}$, si et seulement si il est solution du problème suivant :

$$h(\lambda) = \max_{\mu\in\mathbb{R}^p} h(\mu) \qquad (\mathcal{D})$$

(c) Décrire en détail l'algorithme d'Uzawa pour le problème $(\mathcal{P})$.

(d) Décrire en détail l'algorithme du gradient pour le problème $(\mathcal{D})$.

(e) Quelle conclusion pouvez-vous en tirer ?

---

18. Si $x=(\xi_1,\cdots,\xi_n)$ et $y=(\eta_1,\cdots,\eta_n)$ sont deux vecteurs de $\mathbb{R}^n$, on note $\langle x,y\rangle = \sum_{i=1}^{n}\xi_i\eta_i$ le produit scalaire usuel, et $\|x\| = (\langle x,x\rangle)^{1/2}$ la norme euclidienne associée. On considère $\mathbb{R}^n$ muni de la topologie associée à cette norme.
On identifiera dans certaines questions, les éléments de $\mathbb{R}^n$ avec des matrices colonnes.
Si A est une matrice carrée d'ordre n, à coefficients réels, on note $\|A\|$ la norme de l'application linéaire associée à A, norme définie par

$$\|A\| = \sup\ \{\ \|Ax\|\ ,\ x\in\mathbb{R}^n\ ,\ \|x\|\le 1\ \}\ .$$

Dans tout le problème U désigne une partie **convexe fermée non vide** de $\mathbb{R}^n$.
Soit A une matrice carrée symétrique d'ordre n, à coefficients réels, telle que :

$$\forall v\in\mathbb{R}^n, \qquad \langle Av,v\rangle \ge 0\ .$$

Soit f un élément fixé de $\mathbb{R}^n$; on considère l'application J de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par

$$J(v) = \frac{1}{2}\langle Av,v\rangle - \langle f,v\rangle\ . \qquad (7)$$

On se propose d'étudier l'ensemble $\mathcal{P}$ des éléments u de U tels que

$$\forall v\in U,\quad J(u)\le J(v)\ , \qquad (8)$$

(a) Montrer que u est dans $\mathcal{P}$ si et seulement si u est dans U et $\langle Au-f, v-u\rangle\ge 0$ pour tout v de U.
Montrer que $\mathcal{P}$ est un convexe (éventuellement vide) de $\mathbb{R}^n$.

(b) On suppose dans cette partie que pour tout v non nul de $\mathbb{R}^n$, on a $\langle Av, v\rangle > 0$.
Montrer que J est strictement convexe et qu'il existe $\alpha > 0$ tel que pour tout v de $\mathbb{R}^n$, on ait $\langle Av, v\rangle \geq \alpha\|v\|^2$.
Montrer que $\mathcal{P}$ est non vide. Est-il réduit à un élément ?

(c) On suppose dans cette partie que U est bornée, et que pour tout v de $\mathbb{R}^n$, on a $\langle Av, v\rangle \geq 0$.

i. Montrer que $\mathcal{P}$ est non vide. Est-il réduit à un élément ?

ii. Soit u_0 dans U. On pose pour m dans $\mathbb{N}$,

$$u_{m+1} = \pi(u_m - \rho(Au_m - f)) \ ,$$

où ρ est un réel qui sera déterminé ultérieurement et π la projection sur U. Montrer que

$$J(u_{m+1}) - J(u_m) = \langle Au_m - f, u_{m+1} - u_m\rangle + \frac{1}{2}\langle A(u_{m+1} - u_m), u_{m+1} - u_m\rangle \ , \qquad (9)$$

$$\text{et} \qquad \rho\langle Au_m - f, u_{m+1} - u_m\rangle + \|u_{m+1} - u_m\|^2 \ \leq 0 \ .$$

Montrer qu'il existe $\rho_0 > 0$, tel que si on fixe ρ dans] 0 , ρ_0 [, la suite (J(u_m))$_{m\in\mathbb{N}}$ est décroissante. En déduire que $\lim_{m\to+\infty}(u_{m+1} - u_m) = 0$.
Comment peut-on utiliser la suite $(u_m)_{m\in\mathbb{N}}$ pour trouver des éléments de $\mathcal{P}$?

---

19. On considère le problème : trouver u tel que

$$\begin{cases} u \in U \stackrel{\text{def}}{=} \{v \in \mathbb{R}^n \mid \varphi_i(v) \leq 0 \ , i = 1, \cdots, m \ \} \\ J(u) = \inf\limits_{v\in U} J(v) \end{cases} \qquad (\mathcal{P})$$

et on fait **une fois pour toutes** les hypothèses suivantes :

– La fonctionnelle J est $\mathbb{R}^n$ - elliptique : elle est $\mathcal{C}^1$ dans $\mathbb{R}^n$ et il existe une constante $\alpha > 0$ telle que :

$$(\nabla J(v) - \nabla J(u), v - u)_n \geq \alpha\|v - u\|_n^2, \ \ \forall u, v \in \mathbb{R}^n \qquad (10)$$

où $(\ ,\ )_p$ et $\|\ \|_p$ désignent respectivement le produit scalaire et la norme euclidienne de $\mathbb{R}^p$.

– Il existe une constante M telle que :

$$\|\nabla J(v) - \nabla J(u)\|_n \leq \frac{M}{2}\|v - u\|_n, \ \ \forall u, v \in \mathbb{R}^n \qquad (11)$$

– L'ensemble U est non vide.

– Les fonctions $\varphi_i \ : \ \mathbb{R}^n \to \mathbb{R}$, $i = 1, \cdots, m$, sont convexes.

– Il existe une constante C telle que :

$$\|\Phi(v) - \Phi(u)\|_m \leq C\|v - u\|_n, \ \ \forall u, v \in \mathbb{R}^n$$

où Φ désigne l'application de $\mathbb{R}^n$ dans $\mathbb{R}^m$ de composantes φ_i.

(a) Etablir les inégalités : $\forall u, v \in \mathbb{R}^n$

$$(\nabla J(u), v-u)_n + \frac{\alpha}{2}\|v-u\|_n^2 \leq J(v) - J(u) \leq (\nabla J(u), v-u)_n + M\|v-u\|_n^2 \ .$$

(b) Démontrer que le problème $(\mathcal{P})$ possède une solution unique que l'on notera u .

(c) On définit le Lagrangien associé à ce problème :

$$\mathcal{L} : (v, \mu) \in \mathbb{R}^n \times \mathbb{R}_+^m \mapsto \mathcal{L}(v, \mu) \stackrel{\text{def}}{=} J(v) + (\mu, \Phi(v))_m \ .$$

Démontrer que si $(u, \lambda) \in \mathbb{R}^n \times \mathbb{R}_+^m$ est un point selle de $\mathcal{L}$ sur l'ensemble $\mathbb{R}^n \times \mathbb{R}_+^m$ alors le point u est solution du problème $(\mathcal{P})$.

(d) On note P_+ l'opérateur de projection de $\mathbb{R}^m$ sur $\mathbb{R}_+^m$. Vérifier l'équivalence :

$$\lambda = P_+(\lambda + \rho\Phi(u)) \ , \ \rho > 0 \text{ fixé } \Leftrightarrow \begin{cases} \lambda \in \mathbb{R}_+^m \ , \ \Phi(u) \leq 0, \\ (\lambda, \Phi(u))_m = 0 \end{cases}$$

(e) Vérifier qu'un couple (u, λ) est un point-selle du Lagrangien $\mathcal{L}$ si et seulement si :

$$(\nabla J(u), v-u)_n + (\lambda, \Phi(v))_m \geq 0 \text{ pour tout v de } \mathbb{R}^n$$

$$\lambda = P_+(\lambda + \rho\Phi(u)) \ , \ \rho > 0 \text{ fixé (mais arbitraire) .}$$

(f) On définit une méthode itérative de la façon suivante : partant d'un couple $(u^o, \lambda^o) \in \mathbb{R}^n \times \mathbb{R}_+^m$ arbitraire , on définit une suite de couples $(u^k, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}_+^m$ par récurrence :

$$\begin{cases} u^{k+1} \text{ est solution de : } \inf\limits_{v \in \mathbb{R}^n} \dfrac{1}{2}\|v\|_n^2 + (\varepsilon \nabla J(u^k) - u^k, v)_n + \varepsilon(\lambda^k, \Phi(v))_m \\ \lambda^{k+1} = P_+(\lambda^k + \rho\Phi(u^{k+1}) \end{cases}$$

où ε et ρ sont deux nombres > 0 fixés.
Démontrer que le problème d'optimisation définissant le vecteur u^{k+1} à partir du couple (u^k, λ^k) admet une solution et une seule et que le vecteur u^{k+1} est solution de ce problème si et seulement si :

$$\forall v \in \mathbb{R}^n \ \ (u^{k+1} - u^k + \varepsilon \nabla J(u^k), v - u^{k+1})_n + \varepsilon(\lambda^k, \Phi(v) - \Phi(u^{k+1}))_m \geq 0 \ .$$

(g) Soit (u, λ) un point selle du Lagrangien L. Etablir les inégalités suivantes :

$$2\left(\lambda^k - \lambda, \Phi(u) - \Phi(u^{k+1})\right)_m \leq \frac{1}{\rho}(\|\lambda^k - \lambda\|_m^2 - \|\lambda^{k+1} - \lambda\|_m^2) + \rho C^2 \|u^{k+1} - u\|_n^2 \qquad (12)$$

$$\left(u^{k+1} - u^k, u - u^{k+1}\right)_n + \varepsilon\left(\nabla J(u^k) - \nabla J(u), u - u^{k+1}\right)_n + \varepsilon\left(\lambda^k - \lambda, \Phi(u) - \Phi(u^{k+1})\right)_m \geq 0 \qquad (13)$$

$$\left(\nabla J(u^k) - \nabla J(u), u - u^{k+1}\right)_n \leq \frac{M}{2}\|u^k - u^{k+1}\|_n^2 - \frac{\alpha}{2}(\|u^k - u\|_n^2 + \|u^{k+1} - u\|_n^2) \qquad (14)$$

$$\begin{aligned} &\frac{1}{2}(\varepsilon M - 1)\|u^k - u^{k+1}\|_n^2 + \frac{\varepsilon\alpha - 1}{2}(\|u^{k+1} - u\|_n^2 - \|u^k - u\|_n^2) + \\ &\varepsilon(\rho\frac{C^2}{2} - \alpha)\|u^{k+1} - u\|_n^2 + \frac{\varepsilon}{2\rho}(\|\lambda^k - \lambda\|_m^2 - \|\lambda^{k+1} - \lambda\|_m^2) \quad \geq 0 \end{aligned} \qquad (15)$$

(h) Déduire de la dernière inégalité de la question (g) que, si

$$0 < \varepsilon < \frac{1}{M} \text{ et } 0 < \rho < \frac{2\alpha}{C^2}$$

alors :
$\lim_{k \to +\infty} u^k = u$; la suite (λ^k) est bornée et la suite $(\|\lambda^k - \lambda\|_m)$ converge.

Deuxième partie

Contrôle des Systèmes Linéaires

Chapitre 4

Introduction à la théorie du contrôle

On considère un **système** (dynamique) dont l'état est décrit par une fonction inconnue x dite **fonction (ou variable) d'état** . Cette fonction dépend de variables réelles notées pour l'instant abstraitement t et vérifie des relations (souvent différentielles) plus ou moins compliquées appelées lois d'état (ou lois de comportement). En théorie du **contrôle** (ou **commande**) on veut agir sur le système en agissant sur l'état, via des fonctions qu'on appelle **contrôles** (ou **commandes**).

4.1 Quelques exemples

4.1.1 Exemple 1. Economie

L'économie d'un pays est un système constitué d'une population (consommateurs - producteurs), de compagnies, de bourses, etc. L'état de ce système est une collection importante de données (salaires, profits, coûts de production, taux d'inflation, chômage etc.) On peut agir sur ce système en agissant sur les taux, en débloquant des crédits etc. Ces systèmes sont toutefois très difficiles à modéliser du fait de leur complexité. Nous en resterons donc là pour cet exemple...

4.1.2 Exemple 2. Stockage de l'eau dans un réservoir

On considère un réservoir dont le schéma est donné par la figure 4.1. Le flotteur régule le niveau de l'eau. L'eau dans le réservoir est le **système** ; le **contrôle** est la position du flotteur. L'**état** à chaque instant est un vecteur constitué par la hauteur de l'eau dans le réservoir $h(t)$, le débit d'entrée et le débit de sortie de l'eau.

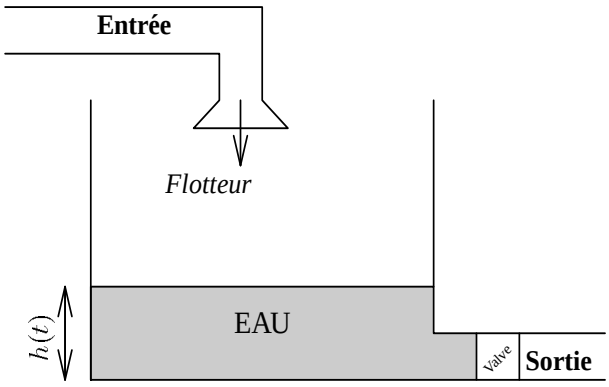

Figure 4.1. Réservoir

4.1.3 Exemple 3 . Stabilisation d'un véhicule

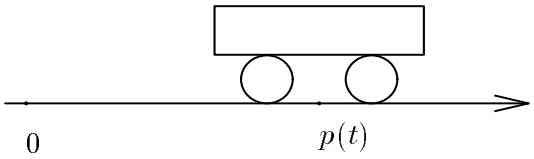

Figure 4.2. Véhicule

On considère un véhicule sur un rail. Il a deux moteurs, un à l'avant (pour accélérer) et l'autre à l'arrière (pour freiner). A l'instant $t = 0$, le véhicule est à la position p_0, à la vitesse v_0. On veut que le véhicule s'arrête à l'origine 0 (la vitesse y est donc nulle). Pour cela, il faut allumer les moteurs au "bon moment", d'une manière correcte et si possible optimale (en mettant par exemple le moins de temps possible et en consommant le moins d'énergie possible). Cet exemple peut modéliser de façon très simplifiée, une voiture lancée à 60 km/h qui doit s'arrêter à un feu situé à une certaine distance p_0.

Dans ce cas le système est le **véhicule**. La fonction d'état est donnée par le vecteur $x(t) = (p(t), \frac{dp}{dt}(t))$ (c'est-à-dire (position, vitesse)). L'état initial $x_0 = (p_0, v_0)$ est supposé connu. Le contrôle $u(t)$ est la force appliquée au véhicule, due à l'allumage des moteurs à l'instant t. Si le moteur arrière s'allume la force est négative. La dynamique du système est donnée par la loi fondamentale de la dynamique "force = masse $\times$ accélération", qu'on peut écrire ici $\frac{d^2p}{dt^2}(t) = u(t)$ (on normalise la masse supposée constante). On obtient

$$x(t) = \begin{bmatrix} p(t) \\ \frac{dp}{dt}(t) \end{bmatrix}, \text{ et } \frac{dx}{dt}(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t) .$$

L'**équation d'état** est donc une équation différentielle dans $\mathbb{R}^2$. De plus, il y a des contraintes sur u dues à la taille des moteurs et à la limitation de l'accélération. Une hypothèse mathématique

raisonnable est que u est mesurable et bornée. Plus généralement, on choisira u dans un ensemble de **contrôles admissibles** : U_{ad} ; on peut prendre par exemple : $|u(t)| \leq 1$, et/ou des fonctions constantes par morceaux (ce qui correspond à des allumages successifs de chacun des deux moteurs).

Si on intègre l'équation différentielle on obtient

$$\left\{ \begin{array}{rcl} \dfrac{dp}{dt}(t) & = & v_0 + \displaystyle\int_0^t u(s)\, ds\ . \\ p(t) & = & p_0 + v_0 t + \displaystyle\int_0^t \int_0^s u(\sigma)\, d\sigma \end{array} \right.$$

L'état dépend donc du contrôle via les intégrales ; on le note $x(\cdot) \stackrel{\text{def}}{=} x[u, x_0](\cdot)$.
Si on peut trouver un temps t_1 et une fonction u tels que $x[u, x_0](t_1) = (0, 0)$, on dit que u est solution et que le système est **contrôlable** ou **commandable**. Il peut y avoir aucune, plusieurs ou même une infinité de solutions.

Si on se fixe T et qu'on peut trouver u tel que $x[u, x_0](T) = (0, 0)$, on dit que que le système est **contrôlable en un temps** T.

On peut également, pour un système contrôlable, chercher le temps le plus petit t^* pour lequel on peut trouver u tel que $x[u, x_0](t^*) = (0, 0)$. On a alors affaire à un problème de **temps optimal.**

Enfin, on peut imaginer que le mobile ne pourra pas atteindre l'origine : on décide alors de s'en rapprocher le plus possible et de formuler le problème au sens des **moindres carrés**, de la manière suivante

Chercher un contrôle u à temps fixé T solution de

$$\min_{u \in U_{ad}} \|x[u, x_0](T)\|^2 = \|p[u, x_0](T)\|^2 + \|\frac{dp}{dt}[u, x_0](T)\|^2$$

Nous avons alors affaire à un problème de **contrôle optimal**. On voit que la fonction à minimiser dépend de la "variable" u par l'intermédiaire de la fonction d'état x. Cette dépendance a été explicitée dans cet exemple. En général la dépendance reste implicite et on préfère alors exprimer le problème de contrôle comme un problème à deux variables x et u en ajoutant une contrainte qui n'est autre que l'équation d'état.

Il y a bien sûr beaucoup d'autres exemples. Le principe de base est qu'on peut agir sur l'état d'un système grâce à l'action d'un contrôle. La dynamique du système (c'est-à-dire la manière dont l'état change sous l'influence du contrôle) peut être compliquée.

Il faut donc une modélisation raisonnable (c'est-à-dire qui décrit correctement et qui n'est pas trop compliquée). On peut en outre faire intervenir des termes aléatoires et on obtient alors du contrôle **stochastique**. Nous nous limiterons à des problèmes de contrôle **déterministe**.

4.2 Formulation mathématique d'un problème de contrôle

4.2.1 Définitions

Soient les ensembles de fonctions suivants :

$$\mathcal{U}_p[0,T] = \{\, u : \mathbb{R} \to \mathbb{R}^p \;\mid u \text{ est intégrable sur } [0,T] \,\},$$

$$\mathcal{U}_p = \bigcup_{T>0} \mathcal{U}_p[0,T] \;; \text{ (ensemble des contrôles).}$$

Pour chaque $t \geq 0$, on se donne un ensemble **cible** fermé, $\mathcal{T}(t) \subset \mathbb{R}^n$. En général on prend $\mathcal{T}(t) = \{\, 0 \,\}$.

On suppose que l'état du système $x : [0\,, +\infty[\to \mathbb{R}^n$ est donné par une équation différentielle ordinaire (EDO) de la forme

$$\begin{cases} \dfrac{dx}{dt}(t) = f(t, x(t), u(t)) \text{ sur }]\,0, +\infty\,[\\ x(t_0) = x_0 \end{cases} \tag{4.2.1}$$

On supposera que f vérifie les conditions du Théorème de Cauchy-Lipschitz (Annexe A - Théorème A.3.1) de sorte qu'on puisse assurer l'existence d'une solution unique de (4.2.1) notée : $x[u, x_0](\cdot)$.

Un problème de contrôle consiste en
- **une classe de contrôles admissibles :** $U_{ad} \subset \mathcal{U}_p$
- **une EDO de la forme (4.2.1) décrivant l'état du système**
- **une famille d'ensembles cibles :** $\mathcal{T}$**.**

Définition 4.2.1 *Soit x_0 dans $\mathbb{R}^n$; si on peut trouver $u \in U_{ad}$ et $t_1 > 0$ tels que $x[u, x_0](t_1) \in \mathcal{T}(t_1)$, on dit que u envoie x_0 à la cible ; x_0 est dit* ***contrôlable****.*

On peut alors se poser plusieurs questions :
- Trouver l'ensemble des états initiaux $x_0 \in \mathbb{R}^n$ que l'on peut envoyer à la cible, c'est à dire l'ensemble des états initiaux contrôlables. On a alors un problème de **contrôlabilité**.
- Si on connaît un état initial contrôlable x_0, un deuxième problème est un problème de **synthèse** : trouver et décrire au moins un contrôle qui réalise la jonction de x_0 à la cible et plus généralement décrire une méthode constructive de calcul d'un contrôle pour un x_0 contrôlable donné.
  Un contrôle qui dépend de l'état du système est un contrôle **feedback**.
- Quand il n'y a pas unicité du contrôle, on peut aussi chercher le "meilleur" relativement à un critère (ou coût) donné. C'est alors un problème de **contrôle optimal.**

En pratique, on se borne à l'étude sur un intervalle $[0, T]$ avec un ensemble cible fixe $\mathcal{T}$.

4.2.2 Contrôlabilité

On se donne un problème de contrôle :

$$\begin{cases} \dfrac{dx}{dt}(t) = f(t,x,u)\ ,\ x(t_0) = x_0\ , \\ u \in U_{ad} \\ \mathcal{T} \text{ donné.} \end{cases} \tag{4.2.2}$$

et on veut décrire l'ensemble des états initiaux (ou de départ) $x(0) = x_0$ pour lesquels on peut trouver un "bon" contrôle, c'est-à-dire un contrôle envoyant x_0 à la cible.

Définition 4.2.2 *L'ensemble **contrôlable** (ou **commandable**) **en un temps** τ est défini par*

$$\mathcal{C}(\tau) = \{\ x_0 \in \mathbb{R}^n\ \mid \exists u \in U_{ad} \textit{ tel que } x[u,x_0](\tau) \in \mathcal{T}\ \}\ . \tag{4.2.3}$$

*L'ensemble **contrôlable** est alors*

$$\mathcal{C} = \bigcup_{\tau\in[0,T]} \mathcal{C}(\tau) = \{\ x_0 \in \mathbb{R}^n\ \mid \exists u \in U_{ad}, \exists \tau \in [0,T] \textit{ tels que } x[u,x_0](\tau) \in \mathcal{T}\ \}\ . \tag{4.2.4}$$

Remarquons que T peut être infini : on a alors contrôlabilité en temps infini. Deux questions se posent alors

- Décrire $\mathcal{C}$
- Décrire les variations de $\mathcal{C}$ en fonction des variations de U_{ad}.

4.2.3 Observabilité

Bien souvent, l'état d'un système n'est pas directement mesurable ou observable. On peut, par exemple, mesurer plus facilement la pression d'un fluide que sa vitesse (qui elle même induit le gradient de pression). On introduit alors la notion d'équation de sortie

$$y = C(x)\ ,$$

qui correspond en fait à l'observation de l'état x. Un système dynamique sera décrit par l'équation d'état et par l'équation de sortie. On dit alors que le système est **observable** si on peut déterminer de manière unique l'état x à partir des mesures y sur l'intervalle de temps considéré.

4.2.4 Stabilité

La notion de stabilité est cruciale en automatique. Elle est très liée à la contrôlabilité. Nous donnons ici la définition et nous développerons cette notion par la suite.

Considérons un système dont la dynamique est décrite par l'équation d'état (non contrôlée) suivante :

$$\frac{dx}{dt}(t) = f(t,x)\ ,\ x(0) = x_0\ ; \tag{4.2.5}$$

une telle équation est dite aussi **autonome** . On appelle $\bar{x}$ le point d'équilibre de ce système (s'il existe) c'est-à-dire $\bar{x}$ tel que $f(t,\bar{x}) = 0$ pour tout t.

Définition 4.2.3 (Stabilité au sens de Lyapounov)
Le système (4.2.5) est stable au sens de Lyapounov si

$$\forall t_0, \forall \varepsilon > 0\ , \exists \eta > 0\ \textit{ tel que }\ \|x(t_0) - \bar{x}\| \leq \eta \Longrightarrow \forall t \geq t_0 \qquad \|x(t) - \bar{x}\| \leq \varepsilon\ ,$$

où $\bar{x}$ est l'état d'équilibre du système.

Cela revient à dire qu'une petite perturbation de l'état d'équilibre, à chaque instant t_0 n'affecte pas l'évolution vers cet état d'équilibre.

Définition 4.2.4 (Stabilité asymptotique)
Le système (4.2.5) est asymptotiquement stable au sens de Lyapounov si
• *il est stable*
• $\exists \alpha > 0$ *tel que*

$$\|x(t_0) - \bar{x}\| < \alpha \Longrightarrow \lim_{t \to +\infty} \|x(t) - \bar{x}\| = 0\ .$$

Nous verrons qu'un grand problème en automatique consiste à déterminer une **loi de commande**, c'est-à-dire une loi donnant le contrôle en fonction de l'état (réel ou observé), qui permette de stabiliser le système.

4.2.5 Contrôle Optimal

On se donne à présent un système dont l'état est décrit par l'équation d'état suivante :

$$\left\{ \begin{array}{l} \dfrac{dx}{dt}(t) = f(t, x(t), u(t)) \qquad t \in]\,0, +\infty[,\ x(0) = x_0\ , \\ u \in U_{ad} \end{array} \right. \tag{4.2.6}$$

L'équation (4.2.6) a une solution supposée unique (cf Théorème de Cauchy-Lipschitz A.3.1) notée $x[u, x_0](\cdot)$. **Désormais on fixe x_0.**
On se donne alors une fonctionnelle **coût** J (ou **objectif**), et on cherche une fonction de contrôle qui rend minimale cette fonctionnelle. On choisit J par exemple de la forme

$$J(x, u) = \int_0^T \Phi(x(t), u(t))\ dt\ ,$$

et on cherche à résoudre

$$(\mathcal{P}) \qquad \left\{ \begin{array}{ll} \min\ J(x, u) & \\ x = x[u, x_0](\cdot) & \text{Equation d'état} \\ u \in U_{ad} \subset \mathcal{U}_p & \text{Contraintes sur le contrôle .} \end{array} \right.$$

où $T > 0$.

Deux questions se posent alors
– Prouver l'existence d'un contrôle optimal
– Trouver un moyen de le calculer. Pour cela, on va écrire des conditions d'optimalité.

4.2.6 Boucle ouverte / Boucle fermée

Nous alllons voir qu'on peut avoir deux types de stratégies pour contrôler (ou commander) un système dynamique.

1. La première est une stratégie dite en **boucle ouverte**.
   On considère un système donné par son équation d'état et un ensemble de contrôles. On cherche un contrôle qui stabilise le système ou un contrôle optimal (ou tout autre type de contrôle) qui ne dépend pas a priori de l'état du système à un moment quelconque. Ce type de stratégie peut se résumer par le schéma suivant :

Figure 4.3. : Commande en boucle ouverte

2. La deuxième stratégie dite en **boucle fermée**.
   On considère toujours un système donné par son équation d'état et un ensemble de contrôles. On applique le contrôle au système. L'équation d'état retourne une fonction d'état (contrôlée). On introduit alors une **loi de commande** qui permet d'exprimer la fonction de contrôle précisément en fonction de cet état. L'état du système est pris en compte à chaque instant pour déterminer "en temps réel" la commande. Le contrôle est alors appelé contrôle **feedback**. Cette stratégie est schématisée par la figure suivante.

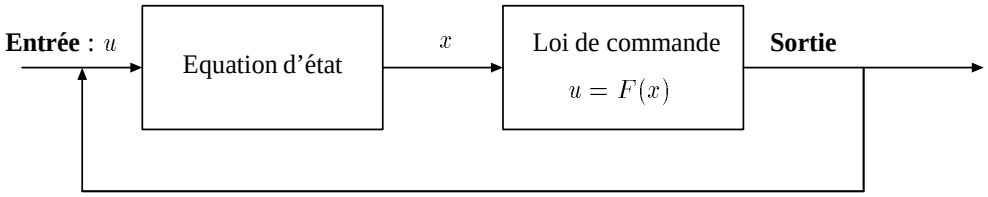

Figure 4.4. : Commande en boucle fermée

Il arrive souvent (nous le soulignerons au moment voulu) qu'une stratégie a priori en boucle ouverte donne un contrôle que l'on peut expliciter en fonction de l'état : nous retrouvons alors un contrôle feedback.

4.3 Encore quelques exemples ...

4.3.1 Stabilisation d'un véhicule

Reprenons l'exemple 3 de la première section, en changeant un peu les notations. La dynamique du système est décrite par

$$\begin{cases} \dfrac{dp}{dt}(t) = q(t),\ p(0) = 0 \\ \dfrac{dq}{dt}(t) = u(t),\ q(0) = q_0 \end{cases} , \tag{4.3.1}$$

avec $U_{ad} = \{\ u \text{ intégrable } \mid\ -1 \leq u \leq 1\ \}$.

Contrôlabilité

Le but est de trouver un instant T tel que $(p(T), q(T)) = (p^*, 0)$, et un contrôle ad-hoc. Ici $t_0 = 0$ et $\mathcal{T} = \mathcal{T}(T) = \{\ (p^*, 0)\ \}$ avec $p^* > 0$.

Pour simplifier les calculs, supposons que nous cherchons un contrôle bang-bang c'est-à-dire

$$u = 1 \text{ sur } [0, t_1[\text{(on commence par accélérer)},$$

$$u = -1 \text{ sur } [t_1, T] \text{(on freine ensuite)} .$$

Les inconnues sont donc t_1 et T. On peut alors calculer explicitement la solution $x(t) = (p(t), q(t))$ de (4.3.1).

$$q(t) = q_0 + v(t) \text{ et } p(t) = q_0 t + \int_0^t v(s)\, ds$$

avec

$$v(t) = \int_0^t u(s)\, ds = \begin{cases} t & \text{si } t \in [0, t_1] \\ 2t_1 - t & \text{si } t \in [t_1, T] \end{cases} .$$

On veut $p(T) = p^*$ et $q(T) = 0$ avec $t_1 \leq T$; on obtient

$$q(T) = q_0 + v(T) = q_0 + 2t_1 - T\ = 0 .$$

D'autre part

$$p(T) = q_0 T + \int_0^T v(t)\, dt = q_0 T - \frac{T^2}{2} + 2\, T t_1 - t_1^2 = p^* .$$

Comme $T = q_0 + 2t_1$ et $t_1 \leq T$, il faut finalement résoudre l'équation suivante

$$t_1^2 + 2q_0 t_1 + \frac{q_0^2}{2} - p^* = 0 \text{ avec } 0 \leq t_1 \leq T \quad \text{(c'est-à-dire } q_0 + t_1 \geq 0).$$

Le discriminant réduit est égal à $\dfrac{q_0^2}{2} + p^*$. Il est donc positif (puisque nous avons supposé $p^* > 0$) et on obtient alors a priori deux solutions :

$$\tau_1 = -q_0 + \sqrt{p^* + \frac{q_0^2}{2}} \qquad \text{et} \qquad \tau_2 = -\left(q_0 + \sqrt{p^* + \frac{q_0^2}{2}}\right) .$$

Pour que τ_1 soit solution il faut que

$$q_0 \leq \sqrt{p^* + \frac{q_0^2}{2}} \geq 0 ,$$

c'est-à-dire $|q_o| \leq \sqrt{2p^*}$. La vitesse de départ doit être positive et même supérieure à $\sqrt{2\,p^*}$. Remarquons au passage que si $q_0 = 0$ (le véhicule est à l'arrêt), t_1 sera solution si $p^* = 0$, c'est-à-dire dans le cas où le véhicule est déjà où il doit être (et ne démarre pas) .

Contrôle optimal

On va définir une fonctionnelle coût grâce aux critères suivants :

- le processus doit être fini en temps "raisonnable" ; on va donc minimiser $T = \displaystyle\int_0^T dt$.
- l'énergie cinétique autorisée doit être limitée pour être sûr qu'on ne cassera pas les moteurs. Cette énergie est donnée par
$$E_c = \int_0^T q(t)^2 \, dt .$$
- la dépense de carburant doit être aussi raisonnable ; en supposant qu'elle est proportionnelle à la force, nous avons
$$D_c = \int_0^T |u(t)| \, dt .$$

Nous allons rassembler toutes ces informations dans une fonctionnelle en attribuant à chaque terme un poids différent selon qu'on privilégie l'une des trois options. On obtient

$$J(p, q, u, T) = \int_0^T (\lambda_1 + \lambda_2 \, q(t)^2 + \lambda_3 |u(t)|) \, dt ,$$

avec $\lambda_i \geq 0,\ i = 1, 2, 3$ fixés et $\displaystyle\sum_{i=1}^{3} \lambda_i = 1$. On cherche ensuite à résoudre le problème

$$(\mathcal{P}) \qquad \begin{cases} \min\ J(p, q, u, T) \\ (p, q) = (p[u], q[u]) \text{ est solution de l'équation d'état (4.3.1)}, \\ u \in U_{ad},\ T > 0 . \end{cases}$$

4.3.2 Rendez-vous spatial

Cet exemple est inspiré par [12]. On considère une station orbitale dont la trajectoire est supposée circulaire (au voisinage de la Terre) et dont le vecteur position est $\vec{r_1}$ (dans un repère ayant pour origine le centre de la Terre). On veut amarrer à cette station une navette (dont le vecteur position est $\vec{r_2}$) par l'action de moteurs dont la poussée est $\vec{u}$. Les lois de la mécanique s'écrivent :

$$\left\{ \begin{array}{rcl} \dfrac{d^2\vec{r_1}}{dt^2} & = & -\mu \dfrac{\vec{r_1}}{r_1^3} \\ \dfrac{d^2\vec{r_2}}{dt^2} & = & -\mu \dfrac{\vec{r_2}}{r_2^3} + \vec{u}\,. \end{array} \right.$$

La position relative des deux objets est donnée par $\vec{R} = \vec{r_2} - \vec{r_1}$ dont les composantes sont $(x_1, x_2, x_3)^t$. La commande correspond à la poussée des moteurs $\vec{u} = (u_1, u_2, u_3)^t$ et l'état du système est décrit par $(\vec{R}, \dfrac{d\vec{R}}{dt})$.

Comme $\vec{R}$ est petit par rapport à $\vec{r_1}$, on peut linéariser ces équations et les projeter sur le repère de Frénet associé à la station (origine : la station, axes radial (selon $\vec{r_1}$) et tangentiel (perpendiculaire à $\vec{r_1}$)). Le mouvement dans le plan de l'orbite s'écrit alors

$$\frac{dX}{dt} = A\,X + B\,U\,, \tag{4.3.2}$$

avec

$$X = \begin{bmatrix} x_1 \\ x_1' \\ x_2 \\ x_2' \end{bmatrix}\,, \quad U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\,, \quad A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\omega^2 & 0 & 0 & 2\omega \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega & 0 & 0 \end{bmatrix}\,, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}\,.$$

Ici la notation x' désigne la dérivée par rapport à t : $x'(t) = \dfrac{dx}{dt}(t)$.
Le mouvement perpendiculaire au plan de l'orbite (avec x_3) vérifie une équation du type (4.3.2) avec

$$X = \begin{bmatrix} x_3 \\ x_3' \end{bmatrix}\,, \quad U = u_3\,, \quad A = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix}\,, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\,.$$

Dans ce cas particulier le paramètre ω correspond à la période de révolution $T = \frac{2\pi}{\omega}$ de la station. On peut alors se poser les mêmes questions que dans l'exemple précédent.

Contrôlabilité

Peut-on trouver un contrôle U pour que les deux objets se rencontrent (sans se percuter) c'est-à-dire pour que l'état du système soit amené à 0 ?

La question qui se pose a priori est celle du rendez-vous **en un temps fini**. On peut aussi exiger que le rendez-vous ait lieu en un temps T donné. On a alors affaire à un problème de contrôlabilité en temps T.

Temps minimum

Supposons que le système soit contrôlable, c'est-à-dire que le rendez-vous a lieu en temps fini sous l'effet d'un contrôle. On peut exiger que le temps mis pour obtenir la rencontre soit minimal. Un critère envisageable est le suivant :

$$J_1(U) = \int_0^T 1 \, ds \, ,$$

sous la contrainte d'état explicitée précédemment.

Consommation minimum

On peut aussi demander, non pas que le temps soit le plus court possible mais que la consommation de la navette soit la moins coûteuse possible. On peut alors minimiser un critère représentant la dépense d'énergie de la navette, qui est fonction de la poussée des moteurs U :

$$J_2(U) = \int_0^T f(U) \, ds \, ,$$

toujours sous la contrainte d'état précédente.

Enfin on peut minimiser toute combinaison pondérée de ces deux fonctionnelles, de façon à obtenir un "bon" compromis en un gain de temps et une dépense d'énergie raisonnable.

[Exercices/exemples]

Pour les exercices 1 à 3 donner une interprétation des mots *système, état* et *contrôle.* Décrire un ensemble de contraintes et d'états à atteindre raisonnable.

- 1. Le système de chauffage et de régulation thermique d'une maison.
- 2. Un avion avec un moteur, des ailerons à l'arrière des ailes, des élévateurs horizontaux et un stabilisateur vertical à l'arrière. Le pilote fournit des impulsions par le moteur et a le contrôle des ailerons, des élévateurs et du stabilisateur.
- 3. Une tumeur dans le corps humain. Elle se "nourrit" via le système sanguin. Elle est attaquée (et contrôlée) par des radiations et des médicaments véhiculés par le sang.
  (Indication : l'état pourrait être décrit par la masse, la densité ou le volume de la tumeur et son taux de croissance ou de réduction.)
- 4. ***Un modèle de pêche optimale***.
  La population $x(t)$ (mesurée par exemple par sa masse en tonnes) d'une espèce donnée de poissons est supposée croître continuement en l'absence de pêche, suivant la loi suivante :

$$\frac{dx}{dt} = rx(1 - \frac{x}{K}) \stackrel{\text{def}}{=} F(x) \, ,$$

où la constante $r > 0$ est le taux de croissance et la constante K est la capacité d'absorption de l'environnement, (si $x(t) > K$, $\dfrac{dx}{dt} < 0$).
Si on introduit une fonction "taux de pêche" $h(t)$ (en tonne par unité de temps par exemple), l'évolution du système est donnée par :

$$\frac{dx}{dt} = F(x) - h(t)\ ,\ x(0) = x_0\ .$$

Ici $h(t)$ est le contrôle. On peut supposer que $h(t)$ est de la forme $h(t) = E(t)x(t)$ où $E(t)$ est une moyenne entre l'effort des pêcheurs et les conditions de pêche.
Montrer que si

$$h(t) = E\,x(t)\ ,\quad \text{où } E \text{ est constante et } 0 < E < r\ , \tag{1}$$

alors le système atteint un état d'équilibre en $x_1 = \dfrac{K}{r}(r - E)$.
Ainsi on peut maintenir la population en x_1 en appliquant le taux de pêche $Y = Ex_1 = KE(1 - \dfrac{E}{r})$.
Montrer aussi que toute solution $x(t; x_0, h(.))$ converge vers x_1 quand t tend vers $+\infty$ (sous l'hypothèse (1)).
Une fonctionnelle *coût* pour ce modèle est donnée par

$$C[h(.)] = \int_0^{+\infty} e^{-\delta t}\,[p - c(x(t))\,x(t)]\,h(t)\,dt\ , \tag{2}$$

où p est le prix de vente (supposé constant ! !) du poisson par unité de masse et $c(x(t))$ est le prix de revient de la prise d'une unité de masse de poisson quand la population est $x(t)$.
La constante δ est un taux de réduction dû aux coûts de stockage, etc.

Chapitre 5

Contrôle optimal à horizon fini

Nous allons étudier dans ce chapitre le cas de problèmes de contrôle optimal "simples". En effet nous supposerons que **l'équation d'état est une équation différentielle ordinaire (EDO) linéaire et que le coût est une fonctionnelle quadratique**. Ce choix n'est toutefois pas très restrictif car de nombreux problèmes de contrôle optimal sont issus d'une minimisation au sens des moindres carrés et la fonctionnelle qui apparaît alors "naturellement" est quadratique.

5.1 Présentation du problème - Théorèmes d'existence

5.1.1 L'équation d'état

On considère un système défini par l'équation d'état (linéaire suivante) :

$$\left\{ \begin{array}{rcl} \dfrac{dx}{dt} & = & A\,x(t) + B\,v(t) + f(t)\ ,\ \text{sur }]0,T[\ , \\ x(0) & = & x_0\ . \end{array} \right. \tag{5.1.1}$$

- T est un réel positif fixé et $f \in L^2(0,T)^n$. On dit qu'on travaille à **horizon fini**.
- Le contrôle $v : [0,T] \to \mathbb{R}^p$ appartient à un sous-ensemble $\mathcal{U}$, convexe et fermé de l'espace des contrôles : $L^2(0,T)^p$. On peut choisir par exemple

$$\mathcal{U} = \{\ v \in L^2(0,T)^p \ \mid\ v(s) \in U \ \text{ pour presque tout } s \text{ de } [0,T]\ \} \tag{5.1.2}$$

  où U est un convexe fermé de $\mathbb{R}^p$.
  On munit l'ensemble des contrôles de la norme usuelle de $L^2(0,T)^p$

$$\|v\|_{\mathcal{U}} = \left[\int_0^T \|v(t)\|_p^2 \right]^{\frac{1}{2}}\ .$$

- La fonction d'état $x = (x_1, \ldots, x_n)$ est à valeurs dans $\mathbb{R}^n$ et la donnée initiale x_0 est dans $\mathbb{R}^n$.

– A est une matrice (réelle) carrée $n \times n$. Pour simplifier on la suppose constante (c'est-à-dire ne dépendant pas de la variable t).
De la même manière B est une matrice réelle $p \times n$ constante.

Remarque 5.1.1 *Le système précédent est en réalité un système d'équations différentielles. On rappelle que toute équation différentielle linéaire d'ordre n à coefficients constants peut se ramener à un système différentiel d'ordre 1 de n équations à n inconnues.*
On n'a aucune information sur la valeur finale $x(T)$ et on ne désire pas lui donner de valeur particulière (0 par exemple). On n'a donc pas affaire à un problème de contrôlabilité.

Commençons par rappeler un résultat sur les EDO linéaires.

Proposition 5.1.1 *Le système (5.1.1) a une solution unique $x = x[v]$. De plus l'application $v \mapsto x[v]$ est affine, continue de $L^2(0,T;\mathbb{R}^p)$ dans $L^\infty(0,T;\mathbb{R}^n)$.*

Démonstration - L'existence et l'unicité de la solution de (5.1.1) découlent des résultats généraux sur les EDO linéaires (voir Annexe A).
Montrons que $v \mapsto x[v]$ est affine. Soit $x^o = x[0]$ la solution de l'équation (5.1.1) correspondant à $v = 0$:

$$\left\{ \begin{array}{rcl} \dfrac{dx^o}{dt} & = & A\,x^o(t) + f(t)\ ,\ \text{sur }]0,T[\ , \\ x^o(0) & = & x_0\ . \end{array} \right.$$

et $x_H[v]$ la solution de l'EDO homogène :

$$\left\{ \begin{array}{rcl} \dfrac{dx_H}{dt} & = & A\,x_H(t) + B\,v\ ,\ \text{sur }]0,T[\ , \\ x_H(0) & = & 0\ . \end{array} \right.$$

Il est clair que l'application $v \mapsto x_H[v]$ est linéaire. Comme $x[v] = x_H[v] + x^o$, le résultat suit.
Pour montrer la continuité il suffit de montrer la continuité de l'application de $L^2(0,T)^p$ dans $L^2(0,T)^n$ qui à v associe $x_H[v]$. Cette solution peut s'écrire

$$x_H[v](t) = \int_0^t e^{A(t-s)}\ Bv(s)\ ds\ .$$

Comme $|||\, e^{At}\, ||| \leq e^{|||\, A\, |||t} \leq e^{|||\, A\, |||\, T}$ pour toute norme matricielle $|||\ \cdot\ |||$ induite on obtient

$$\|x_H[v](t)\|_n \leq e^{|||\, A\, |||\, T} |||\, B\, ||| \int_0^T \|v(s)\|_p\ ds\ ,$$

c'est-à-dire d'une part

$$\|x[v]_H\|_{L^2(0,T;\mathbb{R}^n)} \leq \sqrt{T}\, \|x_H[v]\|_{L^\infty(0,T;\mathbb{R}^n)} \leq C\|v\|_{L^\infty(0,T;\mathbb{R}^p)}\ ,\ \text{si } v \in L^\infty(0,T;\mathbb{R}^p)$$

et d'autre part par l'inégalité de Cauchy-Schwarz

$$\|x_H[v]\|_{L^\infty(0,T;\mathbb{R}^n)} \leq C\|v\|_{L^2(0,T;\mathbb{R}^p)}\ .$$

□

Remarque 5.1.2 *La proposition précédente montre que $x[v] \in L^\infty(0,T;\mathbb{R}^n)$ lorsqu'on choisit $v \in L^2(0,T;\mathbb{R}^p)$. En fait on peut montrer "mieux", à savoir que $x[v]$ est toujours continu (même si v ne l'est pas).*

Pour alléger les notations on écrira désormais $x(t,v) = x[v](t)$: la quantité x est fonction de la variable t et dépend aussi du paramètre v. On notera

$$\mathcal{X} = \{\ x \in L^2(0,T)^n \ \mid\ x(0) = x_0\ \}$$

l'espace d'état, c'est-à-dire l'espace (affine) auquel appartient la fonction d'état $x[v]$ solution de l'EDO, quand v appartient à $\mathcal{U}$. C'est une espace vectoriel si $x_0 = 0$.
On se donne, à présent une fonction **coût** (ou critère) **quadratique** de la forme suivante

$$\begin{aligned}\mathcal{J}(x,v) = \quad & \frac{1}{2}\int_0^T \langle x(t) - z_d(t), Q\,(x(t) - z_d(t))\rangle_n \ dt \\ & + \langle x(T) - z_d(T), D\,(x(T) - z_d(T))\rangle_n \\ & + \frac{1}{2}\int_0^T \langle v(t), R\,v(t)\rangle_p \ dt\ ,\end{aligned} \tag{5.1.3}$$

où R est une matrice $p \times p$ définie positive, Q et D sont des matrices $n \times n$ semi-définies positives, et symétriques. $\langle \cdot,\cdot\rangle_n$ désigne le produit scalaire de $\mathbb{R}^n$: $\langle x(t), z(t)\rangle_n = x(t)^t z(t)$. On pose ensuite

$$J(v) = \mathcal{J}(x[v], v)\ . \tag{5.1.4}$$

Remarque 5.1.3 *La fonctionnelle J est définie sur l'ensemble $\mathcal{U} \subset L^2(0,T)^p$. En pratique, v (et donc x) sera une fonction continue et il n'y aura pas de problème pour définir J.*
J est une norme et il est facile de voir que J ainsi définie est continue.
*Les termes "intégraux" de J sont des termes **distribués**, qui agissent sur tout l'intervalle $[0,T]$; le troisième terme est une terme d'observation finale, au temps T.*

5.1.2 Le problème de contrôle optimal

Nous allons considérer le problème d'optimisation suivant

$$(\mathcal{P}) \qquad \begin{cases} \min\ J(v) = \mathcal{J}(x(v), v) \\ v \in \mathcal{U}\ . \end{cases}$$

Le problème $(\mathcal{P})$ comporte une contrainte implicite qui est l'équation d'état. On peut le formuler de manière équivalente en faisant apparaître l'équation d'état comme une contrainte explicite.

$$(\mathcal{P}) \qquad \begin{cases} \min\ \mathcal{J}(x,v) \\ \\ \dfrac{dx}{dt} = A\,x(t) + B\,v(t) + f(t)\ ,\ \text{sur}\ ]0,T[\ ,\ x(0) = x_0 \\ v \in \mathcal{U}\ . \end{cases}$$

Définition 5.1.1 *Un contrôle u solution du problème de minimisation précédent s'appelle un* ***contrôle optimal*** *pour le critère J (et le point x_0).*

On peut choisir le contrôle u "à l'avance", c'est-à-dire choisir une fonction u qui servira de contrôle sur $[0,T]$ et le système est alors en **boucle ouverte** , ou bien choisir une application fixe Φ telle que pour tout t de $[0,T]$, $\Phi(t) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ et *utiliser* la sortie $x(t)$ pour construire un contrôle de la forme $\Phi(t)x(t)$: le système est alors en **boucle fermée** et le contrôle est **feedback**.

Divers types de contrôle

Si $u(t) \in \mathbb{R}$ ($p = 1$), le système est dit à **simple commande** ; si $p > 1$ le système est dit à **commandes multiples** .

5.1.3 Exemples

Fonctionnelle avec observation distribuée

Choisissons $D = O$ (matrice nulle) : il n'y a pas d'observation finale ;

$$J(v) = \frac{1}{2}\int_0^T \langle x(t,v) - z_d(t), Q\,(x(t,v) - z_d(t))\rangle_n \; dt + \frac{1}{2}\int_0^T \langle v(t), R\,v(t)\rangle_p \; dt \; .$$

Lorsque $Q = I_n$ (matrice identité d'ordre n) et $R = \alpha I_p$ avec $\alpha > 0$, on retrouve l'expression classique d'une fonctionnelle d'énergie :

$$J(v) = \frac{1}{2}\int_0^T \|x(t,v) - z_d(t)\|_n^2 \; dt + \frac{\alpha}{2}\int_0^T \|v(t)\|_p^2 \; dt \; .$$

Fonctionnelle avec observation finale

On choisit cette fois $Q = 0$:

$$J(v) = \frac{1}{2}\int_0^T \langle v(t), R\,v(t)\rangle_p \; dt + \langle x(T,v) - z_d(T), D\,(x(T,v) - z_d(T))\rangle_n \; .$$

Si on prend $D = I_n$ et $R = \alpha I_p$ avec $\alpha > 0$, on obtient

$$J(v) = \frac{1}{2}\|x(T,v) - z_d(T)\|_n^2 \; dt + \frac{\alpha}{2}\int_0^T \|v(t)\|_p^2 \; dt \; .$$

5.1.4 Etude de la fonction coût

Commençons par une définition

Définition 5.1.2 *J définie est dite λ-convexe si on peut trouver $\lambda > 0$ tel que*

$$J(\frac{u+v}{2}) \leq \frac{1}{2}\,(J(u) + J(v)) - \frac{\lambda}{8}\|u - v\|^2 \; .$$

Théorème 5.1.1 *La fonctionnelle J définie par (5.1.3) et (5.1.4) est λ-convexe, strictement convexe et coercive.*

Démonstration - J est convexe.
Il est clair que $\mathcal{J}$ est convexe car les matrices ont été choisies semi-définies positives. D'autre part, $v \to x[v]$ est affine. Par composition J est convexe.
J est coercive.
La matrice R est définie positive. Soit $\beta > 0$ sa plus petite valeur propre. On a donc pour presque tout t dans $[0, T]$

$$\langle v(t), R\, v(t)\rangle_p \geq \beta \|v(t)\|_p^2 \,,$$

$$\frac{1}{2}\int_0^T \langle v(t), R\, v(t)\rangle_p \, dt \geq \frac{\beta}{2}\int_0^T \|v(t)\|_p^2 \,.$$

Par conséquent

$$J(v) \geq \frac{\beta}{2}\|v\|_{\mathcal{U}}^2 \,,$$

ce qui entraîne la coercivité de J, i.e. $\lim_{\|v\|_{\mathcal{U}} \to +\infty} J(v) = +\infty$.
J est λ-convexe.
Soient u et v dans $\mathcal{U}$ et posons $w = \dfrac{u+v}{2}$.

$$\begin{aligned} J(w) = \frac{1}{2}\int_0^T & \langle x(t,w) - z_d(t), Q\,[x(t,w) - z_d(t)]\rangle_n \, dt + \langle x(T,w) - z_d(T), D\,[x(T,w) - z_d(T)]\rangle_n \\ & + \frac{1}{2}\int_0^T \langle w(t), R\, w(t)\rangle_p \, dt \,. \end{aligned}$$

On sait que $v \mapsto x[v]$ est affine, c'est-à-dire , en posant $z_0 = x[0]$, $x[v] = x_H[v] + z_0$ où x_H est linéaire. Pour tout t fixé on obtient :

$$\begin{aligned} x(t,w) - z_d(t) = & \quad x_H(t,w) - z_d(t) + z_0(t) \\ = & \quad \frac{x_H(t,u) + x_H(t,v)}{2} - z_d(t) + z_0(t) \\ = & \quad \frac{x(t,u) + x(t,v)}{2} - z_d(t) \,. \end{aligned}$$

Par conséquent, pour tout t et pour toute matrice P

$$\sigma(t) = \left(\frac{x(t,u) + x(t,v)}{2} - z_d(t)\right)^t P \left(\frac{x(t,u) + x(t,v)}{2} - z_d(t)\right).$$

Pour rendre la démonstration plus lisible, on utilise la notation matricielle $\langle \xi, \zeta \rangle_n = \xi^t \zeta$ pour tous $\xi,\ \zeta$ de $\mathbb{R}^n$, et on a posé $\sigma(t) \stackrel{\text{def}}{=} (x(t,w) - z_d(t))^t\, P\, (x(t,w) - z_d(t))$. Nous obtenons

$$\begin{aligned}
\sigma(t) = &\ \frac{1}{4}(x(t,u) - z_d(t) + x(t,v) - z_d(t))^t\, P\, (x(t,u) - z_d(t) + x(t,v) - z_d(t)) \\
= &\ \frac{1}{4}(x(t,u) - z_d(t))^t P\, (x(t,u) - z_d(t)) + \frac{1}{4}(x(t,v) - z_d(t))^t P\, (x(t,v) - z_d(t)) \\
&+ \frac{1}{2}(x(t,v) - z_d(t))^t P\, (x(t,u) - z_d(t))\ .
\end{aligned}$$

Par conséquent

$$\begin{aligned}
&\sigma(t) - \frac{1}{2}(x(t,u) - z_d(t))^t\, P\, (x(t,u) - z_d(t)) - \frac{1}{2}(x(t,v) - z_d(t))^t\, P\, (x(t,v) - z_d(t)) \\
&\quad = -\frac{1}{4}(x(t,u) - z_d(t))^t\, P\, (x(t,u) - z_d(t)) - \frac{1}{4}(x(t,v) - z_d(t))^t\, P\, (x(t,v) - z_d(t)) \\
&\qquad\qquad + \frac{1}{2}(x(t,v) - z_d(t))^t\, P\, (x(t,u) - z_d(t)) \\
&\quad = -\frac{1}{4}(x(t,u) - x(t,v))^t\, P\, (x(t,u) - x(t,v)) \\
&\quad \leq 0\ ,
\end{aligned}$$

c'est-à-dire $\sigma(t) \leq 0$. En intégrant et en remarquant que

$$-\frac{1}{4}\left\langle u(t) - v(t), R\,(u(t) - v(t))\right\rangle_p \leq -\frac{1}{4}\lambda \|u - v\|^2_{\mathcal{U}}\ ,$$

on obtient l'inégalité voulue :

$$J(w) = J\left(\frac{u+v}{2}\right) \leq \frac{1}{2}J(u) + \frac{1}{2}J(v) - \frac{\lambda}{8}\|u - v\|^2_{\mathcal{U}}\ .$$

Enfin la λ-convexité entraîne la stricte convexité. □

Théorème 5.1.2 *La fonctionnelle J définie par (5.1.3) est Gâteaux-différentiable sur $\mathcal{U}$.*

Démonstration - On rappelle que J est Gâteaux-différentiable en $v \in \mathcal{U}$ si

$$\forall w \in L^2(0,T;\mathbb{R}^p) \quad \lim_{t \to 0^+} \frac{J(v+tw) - J(v)}{t} \stackrel{def}{=} J'(v)\cdot w = \left\langle J'(v), w\right\rangle\ ,$$

où $w \mapsto J'(v)\cdot w$ est linéaire.
Soit $v \in \mathcal{U}$. Montrons tout d'abord que $\mathcal{J}$ est Gâteaux-différentiable en v et on conclura par composition : en effet

$$J'_v(v)\cdot w = \mathcal{J}'_{(x,v)}(x(v),v)(x'_v(v)\cdot w, w)\ .$$

Comme $v \mapsto x(v)$ est affine, $x'_v(v)\cdot w = x(w) - x(0)\ = x_H(w)$.

Soit $w \in L^2(0,T;\mathbb{R}^p)$ et z dans l'espace d'état associé. Comme l'expression de $\mathcal{J}$ est "symétrique" en v et en x on ne calcule la Gâteaux-dérivée que pour le terme en v.
Le terme correspondant dans $\dfrac{J(x+\tau z, v+tw)-J(x,v)}{\tau}$ est

$$\frac{1}{2\tau}\left[\int_0^T \langle v(t)+\tau w(t), R\,v(t)+\tau w(t)\rangle_p \; dt - \int_0^T \langle v(t), R\,v(t)\rangle_p \; dt\right]$$

$$= \frac{1}{2\tau}\left[2\tau\int_0^T \langle v(t), R\,w(t)\rangle_p \; dt + \tau^2 \int_0^T \langle w(t), R\,w(t)\rangle_p \; dt\right] .$$

Le passage à la limite quand τ tend vers 0 donne donc

$$\int_0^T \langle v(t), R\,w(t)\rangle_p \; dt \; .$$

Un calcul analogue permet d'établir la Gâteaux-dérivée de $\mathcal{J}$ en (x,v).

$$\mathcal{J}'_{x,v}(x,v)(z,w) = \int_0^T \Big[\langle x(t)-z_d(t), Qz(t)\rangle_n + \langle v(t), R\,w(t)\rangle_p\Big]\; dt + \langle x(T)-z_d(T), D\,z(T)\rangle_n \,. \quad (5.1.5)$$

On peut alors en déduire la Gâteaux-dérivée de J en v dans la direction w.

$$\begin{aligned} J'(v)w = & \int_0^T \Big[\langle x[v](t)-z_d(t), Q(x(t,w)-x[0](t))\rangle_n + \langle v(t), R\,w(t)\rangle_p\Big]\; dt \\ & + \langle x[v](T)-z_d(T), D(x[w](T)-x[0](T))\rangle_n \,. \end{aligned}$$

En particulier pour $w = u - v$ on obtient

$$\begin{aligned} J'(v)(u-v) = & \int_0^T \Big[\langle x[v](t)-z_d(t), Q(x[u](t)-x[v](t))\rangle_n + \langle v(t), R\,u(t)-v(t)\rangle_p\Big]\; dt \\ & + \langle x[v](T)-z_d(T), D(x[u](T)-x[v](T))\rangle_n \,. \end{aligned}$$

□

Pour éviter les confusions, nous adoptons volontairement deux notations pour désigner la dérivation : la notation J' désigne la Gâteaux-dérivée de la **fonctionnelle** J (par rapport à une fonction x ou v donc) et la notation $\dfrac{dx}{dt}$ désigne la dérivée usuelle de la fonction x par rapport à la variable t.

5.1.5 Existence et unicité de la solution de $(\mathcal{P})$

Théorème 5.1.3 *Le problème $(\mathcal{P})$ admet une solution unique.*

Démonstration - C'est un résultat classique d'existence dans les espaces de Hilbert car J est convexe, continue et coercive et $\mathcal{U}$ est (convexe) fermé. L'unicité provient de la stricte convexité de J et de la convexité de U.
On peut toutefois donner une démonstration directe. Comme J est positive, $m = \inf_{v\in\mathcal{U}} J(v) \geq 0$. Soit v_n une suite minimisante :

$$v_n \in \mathcal{U} \text{ et } J(v_n) \to m \ .$$

Grâce la λ-convexité, on a

$$J(\frac{v_n + v_q}{2}) \leq \frac{1}{2}\left[J(v_n) + J(v_q)\right] - \frac{\lambda}{8}\|v_n - v_q\|^2_{\mathcal{U}} \ .$$

D'où

$$\frac{\lambda}{8}\|v_n - v_q\|^2_{\mathcal{U}} \leq \frac{1}{2}\left[J(v_n) + J(v_q)\right] - m \ ,$$

et par passage à la limite on voit que (v_n) est une suite de Cauchy. L'espace des contrôles étant un espace de Hilbert, cette suite converge vers une limite $\bar{v}$ (dans $\mathcal{U}$, car $\mathcal{U}$ est fermé). De plus J est continue : on a donc bien $J(\bar{v}) = m$.
Supposons qu'on ait deux solutions u et v de $(\mathcal{P})$. L'inégalité de λ-convexité donne encore :

$$0 \leq \frac{\lambda}{8}\|u - v\|^2_{\mathcal{U}} \leq \frac{1}{2}\left[J(u) + J(v)\right] - m \ = 0 \ .$$

Par conséquent, la solution est unique. □

5.2 Conditions d'optimalité

Nous savons maintenant que le problème $(\mathcal{P})$ admet une solution unique $\bar{v}$. On notera $\bar{x} = x(\bar{v})$ l'état associé. On sait également que J est Gâteaux-différentiable. Par conséquent, nous avons la condition nécessaire d'optimalité du premier ordre (3.2.1) démontrée dans le chapitre 3 - section 3.2 .

$$\forall v \in \mathcal{U} \qquad J'(\bar{v})(v - \bar{v}) \geq 0 \ .$$

Nous savons d'autre part que cette condition est **nécessaire et suffisante** car J est convexe (ainsi que $\mathcal{U}$).

5.2.1 Un exemple.

Avant de calculer $J'(\bar{v})(v - \bar{v})$ dans le cas général, commençons par un cas simple : $n = p = 1,\ Q = 1,\ D = O$ et $R = \alpha > 0$. Dans ce cas x et v sont des fonctions scalaires à valeurs dans $\mathbb{R}$. La fonction coût s'écrit alors

$$J(v) = \frac{1}{2}\int_0^T \left[x_v(t) - z_d(t)\right]^2 dt + \frac{\alpha}{2}\int_0^T \left[v(t)\right]^2 dt \ ,$$

où x_v est solution de l'EDO

$$\frac{dx}{dt}(t) = a\,x(t) + b\,v(t) \ \text{ dans }]\,0, T[\text{ et } x(0) = 0 \ .$$

Supposons de plus qu'il n'y a pas de contraintes sur le contrôle : $\mathcal{U} = L^2(0,T)$; une condition nécessaire et suffisante d'optimalité est alors

$$J'(\bar{v}) = 0 \ ,$$

où $\bar{v}$ est la solution du problème de contrôle optimal. Cela donne, pour tout $v \in L^2(0,T)$:

$$\int_0^T [\bar{x}(t) - z_d(t)][x_v(t) - \bar{x}(t)]\, dt + \alpha \int_0^T [\bar{v}(t)][v(t) - \bar{v}(t)]\, dt = 0 \ ,$$

où $\bar{x} = x_{\bar{v}}$. Introduisons une variable auxiliaire solution de l'équation (dite adjointe) suivante :

$$-\frac{d\bar{p}}{dt}(t) = a\,\bar{p}(t) + \bar{x}(t) - z_d(t) \ \text{ dans }]\,0,T[\text{ et } \bar{p}(T) = 0 \ .$$

Par intégration par parties, nous obtenons

$$\int_0^T [b\,\bar{p}(t) + \alpha\bar{v}(t)][v(t) - \bar{v}(t)]\, dt = 0 \ ,$$

c'est-à-dire $b\,\bar{p}(t) + \alpha\bar{v}(t) = 0$. Finalement la solution optimale est caractérisée par le système d'optimalité suivant :

$$\left\{ \begin{array}{l} \dfrac{d\bar{x}}{dt}(t) = a\,\bar{x}(t) + b\,\bar{v}(t) \ \text{ dans }]\,0,T[\text{ et } \bar{x}(0) = 0 \ , \\ -\dfrac{d\bar{p}}{dt}(t) = a\,\bar{p}(t) + \bar{x}(t) - z_d(t) \ \text{ dans }]\,0,T[\text{ et } \bar{p}(T) = 0 \ , \\ \bar{v} = -\dfrac{b\,\bar{p}}{\alpha} \ . \end{array} \right.$$

c'est-à-dire

$$\left\{ \begin{array}{l} \dfrac{d\bar{x}}{dt}(t) = a\,\bar{x}(t) - \dfrac{b^2}{\alpha}\bar{p}(t) \ \text{ dans }]\,0,T[\text{ et } \bar{x}(0) = 0 \ , \\ -\dfrac{d\bar{p}}{dt}(t) = \bar{x}(t) + a\,\bar{p}(t) - z_d(t) \ \text{ dans }]\,0,T[\text{ et } \bar{p}(T) = 0 \ . \end{array} \right. \tag{5.2.1}$$

5.2.2 Cas général.

Calculons maintenant $J'(\bar{v})(v - \bar{v})$:

$$\begin{array}{rl} J'(\bar{v})(v-\bar{v}) = & \displaystyle\int_0^T \Big[\langle \bar{x}(t) - z_d(t), Q(x(t) - \bar{x}(t)) \rangle_n + \langle \bar{v}(t), R\,(v(t) - \bar{v}(t)) \rangle_p \Big]\, dt \\ & + \langle \bar{x}(T) - \ z_d(T), D\,(x(T) - \bar{x}(T)) \rangle_n \ , \end{array}$$

où on a posé $x = x(v)$. Les différentes matrices sont symétriques, donc

$$\begin{array}{rll} J'(\bar{v})(v-\bar{v}) = & \left. \begin{array}{l} \displaystyle\int_0^T \langle Q\,(\bar{x}(t) - z_d(t)), x(t) - \bar{x}(t) \rangle_n\, dt \\ + \langle D\,(\bar{x}(T) - z_d(T)), x(T) - \bar{x}(T) \rangle_n \end{array} \right\} & (a) \\ & \displaystyle + \int_0^T \langle R\,\bar{v}(t), v(t) - \bar{v}(t) \rangle_p\, dt \ . & (b) \end{array} \tag{5.2.2}$$

On va transformer cette expression par **intégration par parties**, pour se ramener à des données "connues". Par exemple, on ne connaît pas $x(T)$ mais $x(0)$ grâce à l'équation d'état. Tout d'abord définissons **l'état adjoint** de la manière suivante : $\bar{p}$ est la solution de l'EDO (rétrograde) dite **équation adjointe**

$$\left\{ \begin{array}{ll} \dfrac{d\bar{p}}{dt}(t) & = -A^t\,\bar{p}(t) - Q\,(\bar{x}(t) - z_d(t)),\ \text{sur}\ ]0,T[\\ \bar{p}(T) & = D\,[\bar{x}(T) - z_d]\ , \end{array} \right. \tag{5.2.3}$$

où A^t désigne la matrice transposée de A. L'équation (5.2.2) devient

$$\begin{aligned} J'(\bar{v})(v-\bar{v}) = & \int_0^T \left[\left\langle -A^t\,\bar{p}, x(t)-\bar{x}(t) \right\rangle_n - \left\langle \frac{d\bar{p}}{dt}(t), x(t)-\bar{x}(t) \right\rangle_n \right] dt \\ & + \int_0^T \langle R\,\bar{v}(t), v(t)-\bar{v}(t) \rangle_p\, dt \\ & + \langle \bar{p}(T), x(T)-\bar{x}(T) \rangle_n\,. \end{aligned}$$

Intégrons par parties le terme (a) de (5.2.2) :

$$\begin{aligned} & \int_0^T \left\langle \frac{d\bar{p}}{dt}(t), x(t)-\bar{x}(t) \right\rangle_n dt \\ = & \left[\langle \bar{p}(T), x(T)-\bar{x}(T) \rangle_n - \langle \bar{p}(0), x(0)-\bar{x}(0) \rangle_n \right] - \int_0^T \left\langle \bar{p}(t), \frac{dx}{dt}(t) - \frac{d\bar{x}}{dt}(t) \right\rangle_n dt \\ = & \langle \bar{p}(T), x(T)-\bar{x}(T) \rangle_n - \int_0^T \left\langle \bar{p}(t), \frac{dx}{dt}(t) - \frac{d\bar{x}}{dt}(t) \right\rangle_n dt \\ = & \langle \bar{p}(T), x(T)-\bar{x}(T) \rangle_n - \int_0^T \langle \bar{p}(t), A(x(t)-\bar{x}(t)) + B(v(t)-\bar{v}(t)) \rangle_n\, dt\,. \end{aligned}$$

(a) se réduit donc à :

$$\int_0^T \langle \bar{p}(t), B(v(t)-\bar{v}(t)) \rangle_n\, dt \ = \int_0^T \left\langle B^t \bar{p}(t), v(t)-\bar{v}(t) \right\rangle_n dt\,.$$

Finalement

$$J'(\bar{v})(v-\bar{v}) = \int_0^T \left\langle B^t\,\bar{p} + R\,\bar{v}(t), v(t)-\bar{v}(t) \right\rangle_p dt \ \geq 0. \tag{5.2.4}$$

On obtient donc le théorème suivant :

Théorème 5.2.1 *La solution $\bar{v}$ du problème $(\mathcal{P})$ est caractérisée par les conditions d'optimalité du premier ordre suivantes :*

$$\left\{ \begin{array}{l} \dfrac{d\bar{x}}{dt}(t) = A\,\bar{x}(t) + B\,\bar{v}(t) + f(t)\ \textit{sur}\ ]0,T[,\ \bar{x}(0) = x_0\ : \qquad \textit{Equation d'état} \\ \dfrac{d\bar{p}}{dt}(t) = -A^t\bar{p}(t) - Q\,(\bar{x}(t) - z_d(t))\ \textit{sur}\ ]0,T[, \bar{p}(T) = D(\bar{x}(T) - z_d(T))\ : \quad \textit{Equation adjointe} \\ \\ \forall v \in \mathcal{U} \qquad \displaystyle\int_0^T \left\langle B^t\,\bar{p}(t) + R\,\bar{v}(t), v(t)-\bar{v}(t) \right\rangle_p dt \ \geq 0\,. \end{array} \right.$$

5.2.3 Cas particulier fondamental

On se place dans le cas où

$$\mathcal{U} = \{\, v \in L^2(0,T)^p \;\mid\; v(t) \in U \text{ presque partout } \}$$

et où U est un convexe fermé de $\mathbb{R}^p$. Alors la relation (5.2.4) est équivalente à

$$\forall v \in U\ ,\ \text{pp.}t \in [0,T] \quad \left\langle B^t\,\bar{p} + R\,\bar{v}(t), v - \bar{v}(t)\right\rangle_p \geq 0\ . \tag{5.2.5}$$

On définit alors le **Hamiltonien** du système sur par $H : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^n \times [0,T] \to \mathbb{R}$

$$H(x,v,p,t) = \frac{1}{2}\left\langle x - z_d, Q\,(x - z_d)\right\rangle_n + \frac{1}{2}\left\langle v, R\,v\right\rangle_p + \left\langle p, Ax + Bv\right\rangle_n\ , \tag{5.2.6}$$

($z_d \in \mathbb{R}^n$). La relation (5.2.5) est équivalente à

$$\forall v \in U\ ,\ \ \text{pp. } t \in [0,T] \quad H'_v(\bar{x}(t),\bar{v}(t),\bar{p}(t),t)(v - \bar{v}(t)) \geq 0\ .$$

Autrement dit, pour presque tout t dans [0,T], le Hamiltonien est minimisé pour $v = \bar{v}(t)$.

5.3 Cas sans contraintes sur le contrôle : équation de Ricatti

Détaillons ce que deviennent les conditions d'optimalité quand il n'y a pas de contrainte sur le contrôle : $\mathcal{U} = L^2(0,T)^p$ (ou $U = \mathbb{R}^p$).

Dans ce cas (5.2.5) nous donne

$$\bar{v}(t) = -R^{-1}\,B^t\bar{p}(t) \qquad \text{presque partout}\ .$$

On peut remarquer que le contrôle ne dépend que de la sortie $\bar{x}$ via $\bar{p}$: c'est donc un contrôle **feedback**. Les conditions d'optimalité se réduisent à

$$\begin{cases} \dfrac{d\bar{x}}{dt}(t) = A\,\bar{x}(t) - B\,R^{-1}\,B^t\bar{p}(t) + f(t) \text{ sur }]0,T[,\ \bar{x}(0) = x_0 \\ \\ \dfrac{d\bar{p}}{dt}(t) = -A^t\,\bar{p}(t) - Q\,(\bar{x}(t) - z_d(t)) \text{ sur }]0,T[,\ \bar{p}(T) = D\,(\bar{x}(T) - z_d(T))\ . \end{cases} \tag{5.3.1}$$

On va simplifier encore ces équations pour obtenir plus précisément le résultat suivant :

Théorème 5.3.1 *On se place dans le cas sans contraintes sur le contrôle ($U = \mathbb{R}^p$) et on suppose en outre que $f \equiv 0$ et $z_d \equiv 0$.*
Alors, il existe une unique matrice $P(t)$, $n \times n$, $\mathcal{C}^1$ sur $[\,0,T\,]$, telle que la solution $\bar{v}$ du problème $(\mathcal{P})$, l'état optimal associé $\bar{x}$ et l'état adjoint $\bar{p}$ vérifient les relations suivantes :

$$\bar{v}(t) = -R^{-1}B^t\bar{p}(t)\ , \qquad \bar{p}(t) = P(t)\,\bar{x}(t) \quad \forall t \in\ [0,T]\ . \tag{5.3.2}$$

La matrice P est l'unique solution de l'équation différentielle suivante, appelée ***équation de Riccatti*** *:*

$$\frac{dP}{dt} = -A^t P - P\,A + P\,B\,R^{-1}B^t P - Q \text{ sur } [\,0,T], \;\; P(T) = D\,. \tag{5.3.3}$$

De plus, pour tout $t \in [\,0,T]$ la matrice $P(t)$ est symétrique semi-définie positive, et elle est définie positive si D l'est.

Démonstration - Dans le cas où $f \equiv 0$ et $z_d \equiv 0$ les conditions d'optimalité (5.3.1) deviennent

$$\begin{cases} \dfrac{d\bar{x}}{dt}(t) = A\,\bar{x}(t) - B\,R^{-1}\,B^t\bar{p}(t) \text{ sur }]\,0,T[,\; \bar{x}(0) = x_0 \\ \\ \dfrac{d\bar{p}}{dt}(t) = -A^t\,\bar{p}(t) - Q\,\bar{x}(t) \text{ sur }]\,0,T[,\; \bar{p}(T) = D\,\bar{x}(T)\,. \end{cases} \tag{5.3.4}$$

La solution $(\bar{x},\bar{p})$ de (5.3.4) dépend linéairement de x_0 : en effet $\bar{x}$ dépend linéairement de x_0 et $\bar{p}$ dépend linéairement de $\bar{x}$. Comme on est en dimension finie on peut donc écrire, pour tout $t \in [0,T]$:

$$\bar{x}(t) = X(t)\,x_0, \qquad \bar{p}(t) = \Pi(t)\,x_0\,,$$

où X et Π sont deux applications $\mathcal{C}^1$ de $[0,T]$ dans l'ensemble des matrices $n \times n$. D'autre part $X(0) = I_n$.

Montrons que $X(t)$ est inversible pour tout $t \in [0,T]$. Pour cela, on va montrer que le noyau de $X(t)$ est réduit à $\{0\}$, pour tout t. Soit donc t^* (quelconque) de $[0,T]$ et $x^* \in \ker X(t^*)$ c'est-à-dire $X(t^*)\,x^* = 0$.
Soit $t \in [\,0, T - t^*[$ et (x,p) définis par :

$$x(t) = X(t+t^*)\,x^* \;\text{ et }\; p(t) = \Pi(t+t^*)\,x^*.$$

On vérifie que (x,p) est solution du système (5.3.4) avec $T - t^*$ à la place de T et 0 à la place de x^*. Or le système (5.3.4) qui est linéaire admet aussi $(0,0)$ comme solution. D'après l'unicité de la solution, on en déduit que $x \equiv 0$ et par conséquent $x^* = 0$. On a donc prouvé que la matrice $X(t)$ est inversible pour tout $t \in [0,T\,]$.

On peut donc écrire $x_0 = X^{-1}(t)\,\bar{x}(t)$ pour tout t et

$$\bar{p}(t) = \Pi(t)\,x_0 = \Pi(t)\,X^{-1}(t)\bar{x}(t)\,.$$

Posons $P(t) = \Pi(t)\,X^{-1}(t)$. P est de classe $\mathcal{C}^1$ et on peut écrire la deuxième équation de (5.3.4) de la manière suivante :

$$\frac{d\bar{p}}{dt}(t) = -A^t\,P(t)\,\bar{x}(t) - Q\,\bar{x}(t) \text{ sur }]0,T\,[,\; \bar{p}(T) = D\,\bar{x}(T)\,.$$

Comme

$$\frac{d\bar{p}}{dt}(t) = \frac{dP}{dt}(t)\,\bar{x}(t) + P(t)\,\frac{d\bar{x}}{dt}(t) = \frac{dP}{dt}(t)\,\bar{x}(t) + P(t)\,\left[A\,\bar{x}(t) - B\,R^{-1}\,B^t\bar{p}(t)\right]$$

$$= \frac{dP}{dt}(t)\,\bar{x}(t) + P(t)\,A\,\bar{x}(t) - P(t)\,B\,R^{-1}\,B^t P(t)\,\bar{x}(t)\ ,$$

on obtient

$$\left(\frac{dP}{dt}(t) + P(t)\,A - P(t)\,B\,R^{-1}\,B^t P(t)\right)\bar{x}(t) = -\left(A^t\,P(t) - Q\right)\bar{x}(t)$$

c'est-à-dire pour tout $t \in [0,T]$

$$\left(\frac{dP}{dt}(t) + P(t)\,A - P(t)\,B\,R^{-1}\,B^t P(t) + A^t\,P(t) - Q\right)\bar{x}(t) = 0\ .$$

D'autre part, cette relation est vraie quel que soit le choix de x_0. Comme $\bar{x}(t)$ décrit $\mathbb{R}^n$ lorsque x_0 décrit $\mathbb{R}^n$, on obtient l'équation de Ricatti annoncée. La condition finale est donnée par

$$\bar{p}(T) = D\,\bar{x}(T) = P(T)\,\bar{x}(T)\ ,$$

qui entraîne $P(T) = D$.

Le théorème de Cauchy-Lipschitz (cf Annexe A.) appliqué à l'équation de Ricatti montre que celle-ci a une solution unique. Comme P^t est aussi solution de cette équation, on déduit par unicité que $P(t)$ est symétrique.
Il reste à montrer la positivité de P. Pour cela, nous remarquons que

$$\frac{d}{dt}\,\langle \bar{x}(t), \bar{p}(t)\rangle_n = \left\langle \frac{d\bar{x}}{dt}(t), \bar{p}(t)\right\rangle_n + \left\langle \bar{x}(t), \frac{d\bar{p}}{dt}(t)\right\rangle_n$$

$$= \left\langle A\,\bar{x}(t) - B\,R^{-1}\,B^t\bar{p}(t), \bar{p}(t)\right\rangle_n - \left\langle \bar{x}(t), A^t\,\bar{p}(t) + Q\,\bar{x}(t)\right\rangle_n$$

$$= -\left\langle B\,R^{-1}\,B^t\bar{p}(t), \bar{p}(t)\right\rangle_n - \left\langle \bar{x}(t), Q\,\bar{x}(t)\right\rangle_n\ .$$

En intégrant entre t et T on obtient

$$\langle \bar{x}(t), \bar{p}(t)\rangle_n = \langle D\bar{x}(T), \bar{x}(T)\rangle_n + \int_t^T \left\langle R^{-1}B^t\bar{p}(s), \bar{p}(s)\right\rangle_n\,ds + \int_t^T \langle Q\bar{x}(s), \bar{x}(s)\rangle_n\,ds\ ,$$

c'est-à-dire

$$\langle \bar{x}(t), P(t)\bar{x}(t)\rangle_n \geq \langle D\bar{x}(T), \bar{x}(T)\rangle_n \geq 0\ .$$

Comme $\bar{x}(t)$ décrit décrit $\mathbb{R}^n$ lorsque x_0 décrit $\mathbb{R}^n$, on obtient la positivité de $P(t)$ (définie si D l'est aussi).

5.4 Formulation en termes de Lagrangien

On reprend la formulation du problème, où l'équation d'état apparaît comme une contrainte explicite :

$$(\mathcal{P}) \qquad \begin{cases} \min\ \mathcal{J}(x,v) \\ \\ \dfrac{dx}{dt} = A\,x(t) + B\,v(t) + f(t)\ ,\ \text{sur } [0,T]\ ,\ x(0) = x_0 \\ v \in \mathcal{U}\ . \end{cases}$$

Définissons le Lagrangien associé au problème :

$$\mathcal{L}(x,v,q) = \mathcal{J}(x,v) + \int_0^T \left\langle q(t)\ ,\ \frac{dx}{dt}(t) - A\,x(t) - B\,v(t) - f(t) \right\rangle_n dt\ . \qquad (5.4.1)$$

Donnons d'abord quelques propriétés du Lagrangien :

Proposition 5.4.1 *$\mathcal{L}$ est convexe par rapport à (x,v) et linéaire par rapport à q. De plus, $\mathcal{L}$ est Gâteaux-différentiable.*

Démonstration - La convexité par rapport à (x,v) est immédiate car l'équation d'état est linéaire. La linéarité par rapport à q et la Gâteaux-différentiabilité sont tout aussi évidentes. □

On rappelle que $(\bar{x},\bar{v},\bar{q}) \in \mathcal{X}\times\mathcal{U}\times\mathcal{Q}$ est un **point-selle** du Lagrangien si

$$\forall (x,v,q) \in \mathcal{X}\times\mathcal{U}\times\mathcal{Q} \qquad \mathcal{L}(\bar{x},\bar{v},q) \le \mathcal{L}(\bar{x},\bar{v},\bar{q}) \le \mathcal{L}(x,v,\bar{q})\ .$$

Théorème 5.4.1 *$(\bar{x},\bar{v})$ est solution de $(\mathcal{P})$ si et seulement si $(\bar{x},\bar{v},-\bar{p})$ est point-selle de $\mathcal{L}$ sur $\mathcal{X}\times\mathcal{U}\times L^2(0,T)^n$. ($\bar{p}$ désigne l'état adjoint.)*

Démonstration - Montrons que la condition est nécessaire. Soit $(\bar{x},\bar{v})$ la solution de $(\mathcal{P})$ et $\bar{p}$ l'état adjoint associé. Dérivons le Lagrangien par rapport à (x,v) au point $(\bar{x},\bar{v},-\bar{p})$.

$$\mathcal{L}'_{x,v}(\bar{x},\bar{v},-\bar{p})(x-\bar{x},v-\bar{v}) = \mathcal{J}'(\bar{x},\bar{v})(x-\bar{x},v-\bar{v}) - \int_0^T \left\langle \bar{p}\,,\,\frac{d(x-\bar{x})}{dt} - A\,(x-\bar{x}) - B\,(v-\bar{v}) \right\rangle_n dt.$$

$$\begin{aligned}\mathcal{L}'_{x,v}(\bar{x},\bar{v},-\bar{p})(x-\bar{x},v-\bar{v}) = \quad & \int_0^T \langle \bar{x}-z_d\,,Q\,(x-\bar{x})\rangle_n\ dt\ + \int_0^T \langle \bar{v}\,,R\,(v-\bar{v})\rangle_p\ dt \\ & + \langle \bar{x}(T)-z_d(T)\,,D\,(x(T)-\bar{x}(T))\,\rangle_n \\ & -\int_0^T \left\langle \bar{p},\,\frac{d(x-\bar{x})}{dt}\right\rangle_n dt\ + \int_0^T \langle \bar{p}, A\,(x-\bar{x}) + B\,(v-\bar{v})\rangle_n\, dt\end{aligned}$$

$$\begin{aligned}\mathcal{L}'_{x,v}(\bar{x},\bar{v},-\bar{p})(x-\bar{x},\ v-\bar{v}) = \quad & \int_0^T \left\langle Q\,(\bar{x}-z_d) + A^t\bar{p}\ ,\ x-\bar{x}\ \right\rangle_n\ dt \\ & + \int_0^T \left\langle\ R\,\bar{v} + B^t\bar{p}\ ,\ v-\bar{v}\ \right\rangle_p\ dt \\ & + \left\langle\ D\,(\bar{x}(T)-z_d(T))\ ,\ x(T)-\bar{x}(T)\ \right\rangle_n \\ & - \int_0^T \left\langle\ \bar{p}\ ,\ \frac{d(x-\bar{x})}{dt}\ \right\rangle_n dt\ .\end{aligned}$$

Le calcul de $\displaystyle\int_0^T \left\langle \bar{p},\,\frac{d(x-\bar{x})}{dt}\right\rangle_n dt$ par intégration par parties a déjà été fait :

$$\begin{aligned}\int_0^T \left\langle\ \bar{p}\ ,\ \frac{d(x-\bar{x})}{dt}\ \right\rangle_n dt\ = \quad & -\int_0^T \left\langle\ A^t\bar{p} + Q\,(\bar{x}-z_d)\ ,\ x-\bar{x}\ \right\rangle_n dt \\ & + \langle \bar{p}(0),\,x(0)-\bar{x}(0)\rangle_n - \langle D\,(\bar{x}(T)-z_d(T)),\,x(T)-\bar{x}(T)\rangle_n\,.\end{aligned}$$

Finalement

$$\mathcal{L}'_{x,v}(\bar{x},\bar{v},-\bar{p})(x-\bar{x},v-\bar{v}) = \int_0^T \left\langle\, R\,\bar{v} + B^t\bar{p}\,,\, v-\bar{v}\,\right\rangle_p\, dt \;+ \left\langle\, \bar{p}(0)\,,\, x(0)-x_0\,\right\rangle_n .$$

Or d'après le principe du minimum

$$\forall v \in \mathcal{U} \qquad \int_0^T \left\langle\, R\,\bar{v} + B^t\bar{p}\,,\, v-\bar{v}\,\right\rangle_p\, dt \;\geq 0 .$$

Donc

$$\forall (x,v) \in \mathcal{X}\times\mathcal{U} \qquad \mathcal{L}'_{x,v}(\bar{x},\bar{v},-\bar{p})(x-\bar{x},v-\bar{v}) \geq 0 .$$

Comme $\mathcal{L}$ est convexe par rapport à (x,v) cela entraîne

$$\forall (x,v) \in \mathcal{X}\times\mathcal{U} \qquad \mathcal{L}(\bar{x},\bar{v},-\bar{p}) \leq \mathcal{L}(x,v,-\bar{p}) .$$

Comme $\mathcal{L}(\bar{x},\bar{v},q) = \mathcal{J}(\bar{x},\bar{v}) = \mathcal{L}(\bar{x},\bar{v},-\bar{p})$, nous avons donc bien un point-selle.

Réciproquement. Soit $\mathcal{L}(\bar{x},\bar{v},\bar{q})$ un point selle. En particulier pour tout q de $L^2(0,T)^n$ on a

$$\mathcal{L}(\bar{x},\bar{v},q) \leq \mathcal{L}(\bar{x},\bar{v},\bar{q}) ,$$

c'est-à-dire

$$\int_0^T \left\langle\, q(t)-\bar{q}(t)\,,\, \frac{d\bar{x}}{dt}(t) - A\,\bar{x}(t) - B\,\bar{v}(t) - f(t)\,\right\rangle_n\, dt \;\leq 0 .$$

Comme q varie dans un espace-vectoriel on en déduit l'égalité et donc

$$\forall q \in L^2(0,T)^n \qquad \int_0^T \left\langle\, q(t)\,,\, \frac{d\bar{x}}{dt}(t) - A\,\bar{x}(t) - B\,\bar{v}(t) - f(t)\,\right\rangle_n\, dt \;= 0 .$$

Cela implique $\dfrac{d\bar{x}}{dt}(t) - A\,\bar{x}(t) - B\,\bar{v}(t) - f(t) \;= 0$ sur $[0,T]$. Donc $(\bar{x},\bar{v})$ est un **point réalisable.**
D'autre part la deuxième inégalité de point-selle donne

$$\forall (x,v) \in \mathcal{X}\times\mathcal{U} \qquad \mathcal{L}(\bar{x},\bar{v},\bar{q}) = \mathcal{J}(\bar{x},\bar{v}) \leq \mathcal{L}(x,v,\bar{q}) = \mathcal{J}(x,v) ,$$

de sorte que $(\bar{x},\bar{v})$ est bien un couple optimal. □
Nous sommes donc ramenés à la recherche de points-selles.

5.5 Algorithmes de résolution

Nous allons appliquer quelques algorithmes de minimisation classiques au problème que nous venons d'étudier.

5.5.1 Résolution directe de $(\mathcal{P})$

On traite le problème directement en "ignorant" l'étude théorique effectuée précédemment. Celui-ci peut s'écrire

$$\min\{\ \mathcal{J}(x,v)\ \mid\ (x,v)\in\mathcal{D}\ \}\ ,$$

où le domaine réalisable $\mathcal{D}$ est donné par

$$\mathcal{D}=\{\ (x,v)\in\mathcal{X}\times\mathcal{U}\ \mid\ \frac{dx}{dt}-A\,x-B\,v-f=0\ \}\ .$$

L'EDO peut se discrétiser par le schéma d'Euler par exemple.
On peut essayer d'utiliser l'algorithme du **gradient projeté**, mais ici l'ensemble $\mathcal{D}$ est difficile à décrire et la projection impossible à mettre en oeuvre. Cette technique n'est donc pas à conseiller...

5.5.2 Pénalisation de la contrainte d'état

On peut aussi utiliser une méthode de pénalisation extérieure qui permet de ramener le problème avec contraintes à un problème sans contraintes. Ici nous allons pénaliser l'équation d'état. Le problème pénalisé s'écrit alors

$$(\mathcal{P}_r)\qquad\left\{\begin{array}{l}\min\ \mathcal{J}_r(x,v)=\mathcal{J}(x,v)+\dfrac{1}{2r}\|\dfrac{dx}{dt}-A\,x-B\,v-f\|^2_{\mathcal{X}}\\ (x,v)\in\mathcal{X}\times\mathcal{U}\ .\end{array}\right.$$

Ce problème admet une solution unique (x_r,v_r) qui converge (à une sous-suite près) vers la solution $(\bar{x},\bar{v})$ de $(\mathcal{P})$ quand r tend vers 0. On peut, en la calculant obtenir une "bonne" approximation de la solution. Le problème majeur est le choix du paramètre r. Habituellement on prend une suite r_n que l'on fait décroître vers 0. Toutefois cette méthode est très sensible au choix de ce paramètre et difficile à mettre en oeuvre.

Pour résoudre $(\mathcal{P}_r)$, on discrétise d'abord l'équation d'état, par exemple avec le schéma d'Euler et on calcule les normes et la fonction $\mathcal{J}$ avec une méthode d'intégration numérique (trapèzes par exemple). On est ramené à un problème de minimisation dans un espace de dimension finie $\mathbb{R}^N$, sans contraintes, que l'on peut résoudre par exemple avec les algorithmes présentés au chapitre 2.

5.5.3 Recherche d'un point-selle

On a vu que la solution du problème $(\mathcal{P})$ est aussi un point-selle du Lagrangien. On peut donc utiliser l'algorithme d'UZAWA, qui dans ce cas précis devient :

Algorithme d'Uzawa (dimension infinie)

1. **Initialisation**
   $k=0$: choix de $q_0\in L^2(0,T)^n$
2. **Itération** k :
   $q_k\in L^2(0,T)^n$ est connu ; puis

(a) Calcul de $(x_k, v_k) \in \mathcal{X} \times \mathcal{U}$ solution de

$$(\mathcal{P}_k) \qquad \min\ \mathcal{L}(x, v, q_k)\ ,\ (x, v) \in \mathcal{X} \times \mathcal{U}\ .$$

(b) Calcul de q_{k+1} avec :

$$q_{k+1} = q_k + \rho\ (\frac{dx_k}{dt} - A\,x_k - B\,v_k - f)\ ,$$

où $\rho > 0$ est un réel fixé (choisi par l'utilisateur).

3. **Critère d'arrêt**
Si $\|(x_{k+1}, v_{k+1}) - (x_k, v_k)\|_{\mathcal{X}\times\mathcal{U}} < \varepsilon$, STOP
Sinon, on pose $k = k + 1$ et on retourne à 2.

Cet algorithme est écrit dans sa version "dimension infinie". Pour une mise en oeuvre numérique, il convient bien sûr de discrétiser l'EDO et les intégrales par des méthodes numériques standard (voir par exemple [7, 8]).

La méthode d'Uzawa est considérablement améliorée si on utilise le **Lagrangien augmenté** $\mathcal{L}_r$ à la place de $\mathcal{L}$, où $\mathcal{L}_r$ est défini par

$$\mathcal{L}_r(x, v, q) = \mathcal{L}(x, v, q) + \frac{1}{2r}\,\|\frac{dx}{dt} - A\,x - B\,v - f\|^2_{\mathcal{X}}\ .$$

La contrainte a été pénalisée. Si on applique l'algorithme d'Uzawa au Lagrangien augmenté, on augmente la vitesse de convergence. Il faut bien sûr choisir un paramètre r convenable.

5.5.4 Méthode de point fixe

On peut aussi considérer le système d'optimalité comme une équation non linéaire (où la fonction considérée est non différentiable) et utiliser la méthode des approximations successives pour le résoudre. Plus précisément, dans le cas où

$$\mathcal{U} = \{\ v \in L^2(0,T)^p\ \mid\ v(t) \in U \text{ presque partout }\}$$

la relation (5.2.5)

$$\forall v \in U\ ,\ \text{pp.}t \in [0,T] \quad \left\langle B^t\,\bar{p} + R\,\bar{v}(t), v - \bar{v}(t)\right\rangle_p \geq 0$$

est équivalente à

$$\bar{v}(t) \text{ réalise } \min_{v\in U} \|v + R^{-1}B^t\bar{p}(t)\|^2,$$

c'est-à-dire

$$\bar{v}(t) = \Pi_U\left(-R^{-1}B^t\bar{p}(t)\right)\ , \tag{5.5.1}$$

où Π_U désigne la projection orthogonale sur U. Le système d'optimalité peut donc s'écrire

$$\Phi(\bar{x}, \bar{p}, \bar{p}(T), \bar{v}) = [0,0,0,0]^t\ ,$$

où $\Phi : \mathcal{X} \times L^2(0,T)^n \times \mathbb{R}^n \times \mathcal{U} \to \mathcal{X} \times L^2(0,T)^n \times \mathbb{R}^n \times \mathcal{U}$ est définie par

$$\Phi(x,p,p_T,v) = \begin{bmatrix} \dfrac{dx}{dt} - Ax - Bv - f \\ \dfrac{dp}{dt} - A^t p - Q(x - z_d) \\ p_T - D[x(T) - z_d(T)] \\ v - \Pi_U[-R^{-1}B^t p(\cdot)] \end{bmatrix} .$$

L'équation est équivalente à

$$\begin{bmatrix} \bar{x} \\ \bar{p} \\ \bar{p}(T) \\ \bar{v} \end{bmatrix} + \rho\, \Phi(\bar{x},\bar{p},\bar{p}(T),\bar{v}) = \begin{bmatrix} \bar{x} \\ \bar{p} \\ \bar{p}(T) \\ \bar{v} \end{bmatrix} ,$$

pour tout $\rho > 0$. Formellement (car il faut trouver ρ pour que $I_d + \rho\,\Phi$ soit contractante) on peut écrire la méthode des approximations successives appliquée à la recherche de points fixes de $I_d + \rho\,\Phi$. Ceci donne

Approximations successives

1. **Initialisation**
   $k = 1$: choix de $v_0 \in \mathcal{U}$
2. **Itération** k :
   $v_{k-1} \in \mathcal{U}$ est connu ; puis
   (a) Calcul de $x_k \in \mathcal{X}$ solution de
   $$\frac{dx_k}{dt} = Ax_k + Bv_k + f\ ,\ x_k(0) = x_0\ .$$
   (b) Calcul de p_k avec :
   $$\frac{dp_k}{dt} = A^t p_k - Q(x_k - z_d),\ p_k(T) = D[x_k(T) - z_d(T)]\ .$$
   (c) Calcul de v_k avec
   $$v_k(t) = \Pi_U\left(-R^{-1}B^t p_k(t)\right)\ ,\ \text{p.p. } t.$$
3. **Critère d'arrêt**
   Si $\|\Phi(x_k,p_k,p_k(T),v_k)\|_{\mathcal{X}\times L^2(0,T)^n\times\mathbb{R}^n\times\mathcal{U}} < \varepsilon$, STOP
   Sinon, on pose $k = k+1$ et on retourne à 2.

L'étude des conditions de convergence donne des conditions sur R, dont le rayon spectral doit être assez petit.

[Exercices]

1. On considère un mobile se déplaçant sur une droite (par exemple un véhicule sur une route droite). Soit $x(t)$ sa position à l'instant t et $v(t) = \dfrac{dx}{dt}(t)$ sa vitesse à l'instant t. La loi fondamentale de la dynamique nous donne

$$\frac{dv}{dt}(t) = \kappa \, ,$$

où κ est une force d'accélération (ou de freinage) liée à la puissance du moteur par exemple ; on suppose que cette force est de la forme $\kappa = \alpha\, t + \beta$ où α et β sont deux paramètres constants à déterminer. Comme on accélère à l'instant $t = 0$, **on impose en plus que** $\beta \geq 0$. La loi d'état est alors

$$\left\{ \begin{array}{rcl} \dfrac{dx}{dt} & = & v(t), \;\; t \in]0, T[\\ \dfrac{dv}{dt} & = & \alpha t + \beta, \;\; t \in]0, T[\\ x(0) = x_0 & , & v(0) = v_0 \, , \end{array} \right. \tag{1}$$

où x_0 est la position du mobile à l'instant initial et v_0 sa vitesse. Le repère est choisi de telle sorte que $x_0 = 0$. On veut stabiliser la voiture c'est-à-dire déterminer les paramètres α et β pour qu'à l'instant $T = 3$ mn, la voiture s'arrête à un endroit donné. On veut trouver α et β dans $\mathbb{R}$ pour que $x(3) = x^*$ (donné) et $v(3) = 0$. On va aborder ce problème de deux manières différentes.

   (a) **Première approche**
   - Résoudre le système (1) : exprimer x et v en fonction de t, v_0, α et β. En déduire $x(3)$ et $v(3)$ (en fonction de v_0, α et β.)
     Que vaudraient α et β si on n'imposait aucune contrainte sur β ?
   - Ecrire le problème de stabilisation sous forme d'un problème de minimisation au sens des moindres carrés où α et β sont les inconnues. On le mettra sous la forme

$$\left\{ \begin{array}{l} \min J(X) = \dfrac{1}{2}\,(M\,X, X) + (b, X) \\ g(X) \leq 0 \, , \end{array} \right. \tag{2}$$

   où $X = \left[\begin{array}{c} \alpha \\ \beta \end{array} \right]$. On précisera la matrice (2×2) M, le vecteur b et la fonction g.

   (b) **Deuxième approche**
   Sans résoudre le système (1), formuler le problème comme un problème de contrôle optimal. On précisera l'état, le contrôle et les contraintes.
   Ecrire le système d'optimalité en faisant intervenir l'état adjoint.

   (c) **Résolution**
   Résoudre le problème (2) pour une vitesse initiale $v_0 = 60$ km/h et une distance $x^* = 800$ m. La voiture est-elle complètement stabilisée ?

---

2. On considère dans $\mathbb{R}$, l'équation scalaire sur $[0, \frac{\pi}{2}]$, $\frac{dx}{dt} = x + u,\ x(0) = 0$ et la fonction $f(t) = \cos(t)$; on veut asservir x à f en minimisant l'expression

$$\int_0^{\frac{\pi}{2}} [(x_u(s) - \cos(s))^2 + u(s)^2]\, ds + x(\frac{\pi}{2})^2 .$$

Trouver le contrôle et le coût correspondant.

---

3. On considère dans $\mathbb{R}$ l'équation $\frac{dx}{dt} = u$ qui relie le moment angulaire d'un axe à un couple de contrôle u. On suppose x_0 donné pour $t = 0$, et on veut réduire $x(1)$ en minimisant l'expression $x(1)^2 + \int_0^1 u^2(s)\, ds$.
Trouver le contrôle et le coût correspondant.

---

4. On considère dans $\mathbb{R}^2$ le système suivant contrôlé par $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$:

$$(\mathcal{S}) \qquad \begin{cases} \frac{dx}{dt} = ax + by + u_1(t) \\ \frac{dy}{dt} = cx + dy + u_2(t)\ , \\ x(0) = y(0) = 0\ . \end{cases}$$

Soient $\alpha \geq 0,\ \beta \geq 0,\ \lambda > 0$ et $\mu > 0$; on considère le coût

$$\frac{1}{2} \int_0^T [\alpha\, x_u^2(t) + \beta\, y_u^2(t) + \lambda u_1^2(t) + \mu u_2^2(t)]\, dt\ .$$

Ecrire les équations permettant de trouver u_1 et u_2. Le système $(\mathcal{S})$ est un système "Prédateurs et proies" que l'on essaie de contrôler.
Application numérique : $a = b = c = 1,\ d = -1;\ \alpha = 1,\ \beta = 2,\ \lambda = \mu = 1.$

---

5. Soit A une matrice (n,n) inversible et soit B une matrice de (n,m). Pour f donnée dans $\mathbb{R}^n$, on considère pour chaque $v \in \mathbb{R}^m$ la solution $z = z(v) \in \mathbb{R}^n$ du problème

$$Az(v) = f + Bv \tag{3}$$

On définit ensuite la fonction J de $\mathbb{R}^m$ dans $\mathbb{R}$ par

$$J(v) = \frac{1}{2}\|z(v) - z_0\|_n^2 + \frac{N}{2}||v||_m^2 \tag{4}$$

où z_0 est donné dans $\mathbb{R}^n$, $N > 0$; $\|\cdot\|_n$ désigne la norme euclidienne et $(\cdot,\cdot)_n$ le produit scalaire (canonique) de $\mathbb{R}^n$.

(a) Montrer que l'application $v \to z(v)$ est affine de $\mathbb{R}^m$ dans $\mathbb{R}^n$ (c'est à dire que $z(v) = y_0 + Cv$ où $C \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$). En déduire que la fonction $v \to J(v)$ est convexe puis strictement convexe.

(b) Soit K un sous ensemble convexe fermé non vide de $\mathbb{R}^m$. Montrer qu'il existe une et une seule solution u au problème

$$\left\{ \begin{array}{l} u \in K \\ J(u) = \min\limits_{v \in K} J(v) \end{array} \right. \tag{5}$$

(c) Pour $w \in \mathbb{R}^m$, calculer $(J'(u), w)_m = \dfrac{dJ}{dt}(u + tw)_{|t=0}$. Montrer que u est solution de (5) si et seulement si

$$\left\{ \begin{array}{l} (J'(u), v - u)_m \geq 0, \forall v \in K \\ u \in K \end{array} \right. \tag{6}$$

Que devient (6) si $K = \mathbb{R}^m$ tout entier ?

(d) En utilisant (6) et le calcul de $(J'(u), w)$, montrer que u est solution de (5) si et seulement si il existe un triplet (y, p, u) unique vérifiant (en notant M^t la transposée de M)

$$\left\{ \begin{array}{l} Ay = f + Bu \\ A^t p = y - z_0 \\ (B^t p + Nu, v - u)_m \geq 0 \ , \ \ \forall v \in K \\ u \in K \end{array} \right. \tag{7}$$

Dans le cas où $K = \mathbb{R}^m$, simplifier le système (7) et montrer qu'il équivaut à

$$\left\{ \begin{array}{l} Ay + BB^t \dfrac{p}{N} = f \\ -y + A^t p = -z_0 \end{array} \right. \tag{8}$$

(e) On suppose que l'ensemble K est tel que l'opérateur de projection P_K sur K dans $\mathbb{R}^m$ soit aisément calculable. Ecrire l'algorithme du gradient projeté pour approcher la solution u de (5). Traduire cet algorithme en termes des inconnues (y, p, u). Comparer avec le système (7).

---

6. Soit N un entier supérieur à 2 ; Y désigne le vecteur de $\mathbb{R}^N$ de composantes $(y_i, i = 1, \ldots, N)$ et V le vecteur de $\mathbb{R}^N$ de composantes $(v_i, i = 0, \ldots, N-1)$.$y_0$ est fixé.
Soient **a** = $(a_i, i = 0, \ldots, N-1)$ et **b**= $(b_i, i = 0, \ldots, N-1)$ des vecteurs de $\mathbb{R}^N$; on considère le système des N équations suivantes :

$$\left\{ \begin{array}{ll} y_{i+1} = a_i y_i + b_i v_i & i = 0, \ldots, N-1 \\ y_0 & \text{donné} \end{array} \right. \tag{9}$$

(a) Le système d'équations (9) admet-il une solution unique ?
Dans la suite du problème on notera $Y(V)$ la solution de (9) quand elle existe (on rappelle que y_0 est fixé une fois pour toutes ...).
Montrer que l'application (encore notée Y) de $\mathbb{R}^N$ vers $\mathbb{R}^{N+1}$ qui à V associe $Y(V)$ est linéaire. En déduire sa Gâteaux-différentielle : on explicitera $DY(V)\cdot W$ où V et W sont des vecteurs quelconques de $\mathbb{R}^N$.

(b) On se donne à présent la fonctionnelle J définie de $\mathbb{R}^N$ vers $\mathbb{R}$ par :

$$\forall V \in \mathbb{R}^N \qquad J(V) = \frac{1}{2}\left[\sum_{i=0}^{N} \alpha_i y_i(V)^2 + \sum_{i=0}^{N-1} \beta_i v_i^2\right] , \tag{10}$$

et on considère le problème de contrôle optimal suivant :

$$\min \{ J(V) \mid V \in \mathcal{U} \}, \tag{11}$$

où $\mathcal{U}$ est un convexe fermé non vide (pas nécessairement borné) de $\mathbb{R}^N$. Quelles sont les conditions que doivent satisfaire les réels $(\alpha_i, i = 0, \ldots, N)$ et $(\beta_i, i = 0, \ldots, N-1)$ pour que le problème (11) admette une solution unique ?

(c) Quelle est la solution de (11) quand $y_0 = 0$?

(d) Montrer que J est Gâteaux-différentiable et calculer $DJ(V)\,(W - V)$ pour V et W dans $\mathbb{R}^N$.

(e) En utilisant un résultat général du cours montrer que la solution $\overline{V}$ du problème (11) vérifie la relation suivante :

$$\forall V \in \mathcal{U} \qquad \sum_{i=0}^{N} \alpha_i \bar{y}_i (y_i - \bar{y}_i) + \sum_{i=0}^{N-1} \beta_i \bar{v}_i (v_i - \bar{v}_i) \geq 0 , \tag{12}$$

où $y_i = Y(V)_i, i = 1, \ldots, N$ et $\overline{Y} = Y(\overline{V})$.

(f) On définit maintenant l'état adjoint $\overline{P} = (\bar{p}_0, \ldots, \bar{p}_N)$ de la manière suivante :

$$\begin{cases} \bar{p}_i = a_i \bar{p}_{i+1} + \alpha_i \bar{y}_i \quad , \; i = 0, \ldots, N-1 \\ \bar{p}_N = \alpha_N \bar{y}_N \end{cases} \tag{13}$$

Montrer que l'inéquation (12) peut s' écrire alors :

$$\forall V \in \mathcal{U} \qquad \sum_{i=0}^{N-1} (b_i \bar{p}_{i+1} + \beta_i \bar{v}_i)(v_i - \bar{v}_i) \geq 0 , \tag{14}$$

On suppose à présent que $\mathcal{U} = \mathbb{R}^N$.
Donner l'expression du contrôle optimal $\overline{V}$ en fonction de l'état adjoint $\overline{P}$.
Ecrire alors, dans ce cas, le système d'optimalité complet $(\mathcal{S})$ vérifié par la solution de (11). Montrer que $\overline{Y}$ vérifie un système linéaire de la forme

$$\mathbb{T}\,\overline{Y} = Y_0 ,$$

où $\mathbb{T}$ est une matrice triangulaire que l'on précisera, et Y_0 est le vecteur de $\mathbb{R}^{N+1}$ suivant : $(y_0, \frac{b_1^2 a_0}{\beta_0}, 0, \cdots, 0)$.

(Pour alléger les calculs on pourra poser : $\gamma_i = \frac{b_i^2}{\beta_i}$.)

(g) **On ne suppose plus maintenant que** $\mathcal{U} = \mathbb{R}^N$.
Ecrire le Lagrangien du problème : $\mathcal{L}(y, v, \lambda)$ où $(y, v, \lambda) \in \mathbb{R}^{N+1} \times \mathbb{R}^N \times \mathbb{R}^N$.
Décrire précisément les différentes étapes de la méthode d'Uzawa pour résoudre le problème (11).

(h) Connaissez-vous d'autres méthodes pour résoudre le problème (11) ? Si oui, présentez-les brièvement et indiquez (succinctement) comment procéder.

---

7. On considère un système dynamique dont l'état x est donné par l'équation différentielle du second ordre suivante :

$$\frac{d^2x}{dt^2}(t) = u(t) \text{ sur }]\,0, T\,[, x(0) = x_0,\ \frac{dx}{dt}(0) = x_0' \ ; \tag{15}$$

T est un réel strictement positif, u désigne une fonction de contrôle de $L^2(0, T; \mathbb{R})$, à valeurs dans un intervalle $[a, b]$ de $\mathbb{R}$. On notera $\mathcal{U}_{[a,b]}$ l'ensemble de telles fonctions.

(a) Reconnaissez-vous ce modèle ? Quelle en est l'interprétation physique ? Ecrire l'équation différentielle sous la forme d'un système différentiel en posant $y = (x, \frac{dx}{dt}) = (y_1, y_2)$.
Ce système admet une solution unique que l'on notera $y(u)$.

(b) On se donne maintenant la fonction coût suivante :

$$J(y, u) = \frac{1}{2}\left[(y_1(T) - z_1)^2 + (y_2(T) - z_2)^2\right] + \frac{\alpha}{2}\int_0^T u(t)^2\, dt \ , \tag{16}$$

où $\alpha,\ z_1,\ z_2 \in \mathbb{R}$.
A quelle condition (suffisante) sur α le problème de contrôle optimal suivant

$$\min\ \left\{\ J(y(u), u) \mid u \in \mathcal{U}_{[a,b]}\ \right\}\ , \tag{17}$$

admet-il une solution unique ?
On notera désormais $\bar{u}$ le contrôle optimal et $\bar{y}$ l'état correspondant.

(c) Ecrire le système d'optimalité pour ce problème. On notera $\bar{p}$ l'état adjoint .

(d) Résoudre entièrement le problème quand $a = -\infty$ et $b = +\infty$; on prendra $y_0 = 0$ pour simplifier les calculs. Que peut-on dire du contrôle optimal ? Que devient ce contrôle lorsque T tend vers $+\infty$? Que signifie ce résultat selon vous ?
Que devient ce contrôle lorsque α tend vers $+\infty$? Comment interpréter ce résultat ?

Chapitre 6

Contrôle à horizon infini : contrôlabilité et stabilité

Si on veut généraliser le chapitre précédent et étudier un problème de contrôle optimal à horizon infini, en gardant une fonctionnelle quadratique, la première question qu'on peut se poser est de savoir si on peut borner $x(t)$ quand t tend vers $+\infty$. On peut aussi exiger que l'état $x(t)$ soit nul à l'infini ou pour un t assez grand. Nous n'étudierons pas de problèmes de contrôle optimal à horizon infini mais nous allons nous focaliser sur la contrôlabilité et la satbilité des systèmes étudiés.

6.1 Généralités

Rappelons ce qui a été présenté dans l'introduction (chapitre 4) et précisons un peu les notations et les problèmes qui se posent.
On considère un système dynamique dont l'équation d'état est l'EDO suivante :

$$\left\{ \begin{array}{l} \dfrac{dx}{dt}(t) = f(t, x(t), u(t)) \ , \ x(0) = x_0 \ , \ t \in \mathbb{R}^+ \\ x(t) \in \mathbb{R}^n \ , u \in \mathcal{U} \end{array} \right. \tag{6.1.1}$$

- f est choisie suffisamment régulière (au moins $\mathcal{C}_1$ et vérifiant, par exemple, les conditions du Théorème de Cauchy-Lipschitz A.3.1- Annexe A) pour que l'équation (6.1.1) ait une solution unique : $x[u, x_0]$. De plus on suppose que $f(\cdot, 0, 0) = 0$.
- Le contrôle u est une fonction de $\mathbb{R}^+$ dans $\mathbb{R}^p$. L'ensemble des contrôles $\mathcal{U}$ est supposé **convexe et symétrique** (i.e. $u \in \mathcal{U} \Longleftrightarrow -u \in \mathcal{U}$). On peut choisir, par exemple
  ▷ $\mathcal{U} = \bigcup_{t>0} \mathcal{U}(t)$ où

$$\mathcal{U}(t) = \{ u \text{ mesurable sur } [0,t] \ \mid \ |u_i(t)| \leq 1 \ , 1 \leq i \leq q \ \} \ .$$

  ▷ Plus généralement

$$\mathcal{U}(t) = \{ u \text{ mesurable sur } [0,t] \text{ pour tout } t \text{ et } \ u(t) \in U \ \} \ ,$$

où U est un **compact, convexe** de $\mathbb{R}^p$.
- Pour simplifier, on choisit l'ensemble cible égal à 0 : $\mathcal{T}(t) = \{0\}$.

Remarque 6.1.1 *On pourrait choisir l'instant initial égal à $t_0 \geq 0$ en toute généralité. On choisit $t_0 = 0$ pour alléger l'exposé. On peut toujours se ramener à ce cas par changement de variable : $t \mapsto t - t_0$.*

On se pose alors le problème de **contrôlabilité** suivant :

Décrire $\mathcal{C}$ l'ensemble contrôlable (ou commandable) défini par

$$\mathcal{C} = \{\ x_0 \in \mathbb{R}^n \ \mid\ \exists \tau > 0,\ \exists u \in \mathcal{U}\ , x[u, x_0](\tau) = 0\ \}.$$

Définition 6.1.1 (Etat atteignable)
L'ensemble des états atteignables depuis x_0 en un temps τ est

$$\mathcal{A}(x_0, \tau) = \{\ x[u, x_0](\tau)\ \mid\ u \in \mathcal{U}\ \} \subset \mathbb{R}^n\ .$$

L'ensemble des états atteignables depuis x_0 est : $\mathcal{A}(x_0) = \bigcup_{\tau \geq 0} \mathcal{A}(x_0, \tau)$.

On note aussi $\mathcal{A} = \mathcal{A}(0) = \bigcup_{\tau \geq 0} \mathcal{A}(0, \tau)$ *, l'ensemble des états atteignables depuis* 0.

Le problème de contrôlabilité revient à chercher tous les points $x_0 \in \mathbb{R}^n$ tels que $0 \in \mathcal{A}(x_0)$.

Remarque 6.1.2 *Un autre problème que l'on peut se poser est de décrire le comportement de $\mathcal{C}$ lorsque $\mathcal{U}$ varie.*

On va en fait s'intéresser à deux propriétés importantes :

1. **Quand a-t'on $\mathcal{C} = \mathbb{R}^n$?**
   On dit alors que le système est **complètement contrôlable** .
2. **Quand a-t'on $0 \in$ int($\mathcal{C}$) ?** (où int($\mathcal{C}$) désigne l'intérieur de l'ensemble $\mathcal{C}$)
   En effet $0 \in \mathcal{C}$ car on a supposé que $f(\cdot, 0, 0) = 0$. Donc la solution nulle convient. Si on perturbe un peu la condition initiale on obtient l'EDO :

   $$\frac{dx}{dt}(t) = f(t, x(t), u(t))\ ,\ \ x(0) = \varepsilon\ ,\ t \in \mathbb{R}^+$$

   et on note x_ε sa solution. Peut-on affirmer que l'on va rester dans $\mathcal{C}$? Ceci est très important du point de vue numérique. On souhaite donc que $\mathcal{C}$ contienne au moins une boule centrée en 0, c'est-à dire que $0 \in$ int($\mathcal{C}$).
   Plus généralement, on souhaite que cette propriété soit vraie pour tout élément x_0 de $\mathcal{C}$. On souhaite donc que $\mathcal{C}$ **soit ouvert**.

Théorème 6.1.1 *Pour le système décrit par (6.1.1), l'ensemble $\mathcal{C}$ est connexe par arcs. De plus $\mathcal{C}$ est ouvert si et seulement si* $0 \in$ int *($\mathcal{C}$).*

Démonstration - Si x_0 est dans $\mathcal{C}$, il existe un chemin continu reliant x_0 à 0 qui est la trajectoire $x[u, x_0](\cdot)$ où u est un contrôle ad-hoc. De plus, tous les points de cette trajectoire sont dans $\mathcal{C}$. Donc si x^* est dans $\mathcal{C}$ on peut le relier à x_0 via 0 et nous obtenons la connexité par arcs.

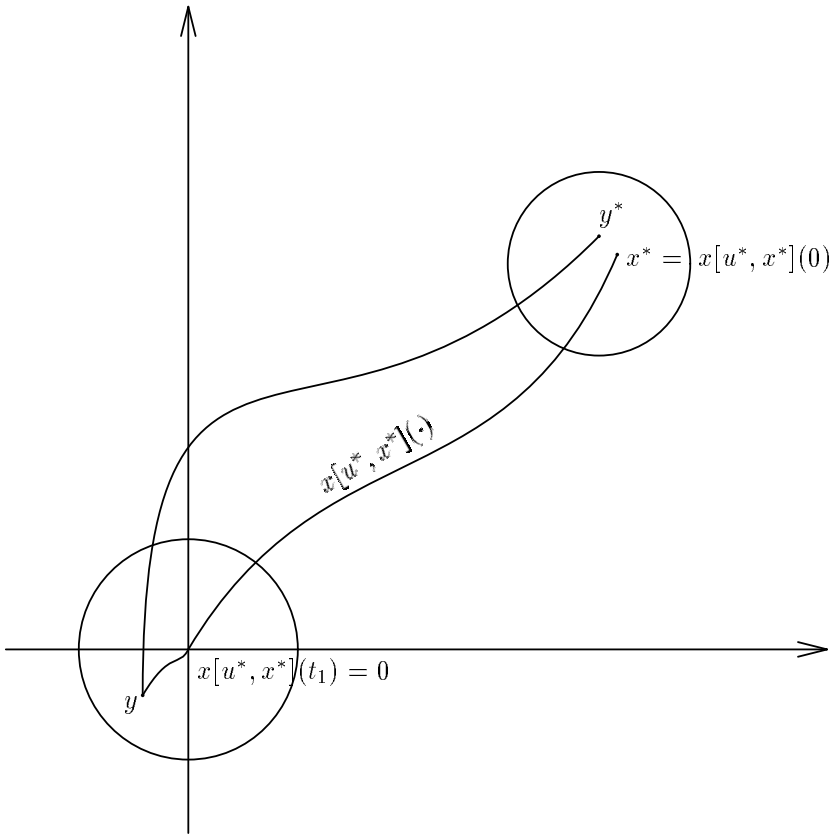

Figure 6.1. : $\mathcal{C}$ est ouvert

Nous savons que 0 appartient à $\mathcal{C}$; donc si $\mathcal{C}$ est ouvert, $0 \in \text{int}(\mathcal{C})$. Il faut maintenant démontrer l'affirmation réciproque. On suppose que $0 \in \text{int}(\mathcal{C})$: il existe une boule $B_0 = B(0, \delta_0) \subset \mathcal{C}$. Soit x^* dans $\mathcal{C}$: on veut montrer qu'il existe une boule $B(x^*, \delta)$ incluse dans $\mathcal{C}$.

$$x^* \in \mathcal{C} \Longrightarrow \exists u^* \in \mathcal{U},\ \exists t_1 > 0 \qquad x[u^*, x^*](t_1) = 0\ .$$

Comme f est différentiable, les solutions de $\dfrac{dx}{dt} = f(\cdot, x, u^*)$ dépendent continuement de la donnée initiale ; autrement dit la fonction $\varphi : x_0 \mapsto x[u^*, x_0,](t_1)$ est continue . Donc pour tout voisinage B_0 de $0 = \varphi(x^*)$, il existe un voisinage $B^* = B(x^*, \delta^*)$ de x^* tel que $\varphi(B^*) \subset B_0$.
Comme $\varphi(x_0) = x[u^*, x_0,](t_1) = \tilde{x}$ avec $x[u^*, x_0,](0) = x_0$, si $x_0 \in B^*$, alors $\tilde{x} \in B_0 \subset \mathcal{C}$. Donc il existe un contrôle $\tilde{u}$ et $t_2 > 0$ tels que $x[\tilde{u}, \tilde{x}](t_2) = 0$ (en d'autres termes, on relie la position finale de la première trajectoire $\tilde{x}$, à 0 par une autre trajectoire.)

Soit alors le contrôle u défini par :

$$u(t) = \begin{cases} u^*(t) & \text{si } 0 \leq t \leq t_1 \\ \tilde{u}(t-t_1) & \text{si } t_1 < t \leq t_1 + t_2 \ . \end{cases}$$

On a alors $x[u, x_0](t_2) = 0$ (on a raccordé les deux trajectoires).
Par conséquent $x^* \in \mathcal{C}$ et $B^* \subset \mathcal{C}$. □

Remarque 6.1.3
– *Dire que $\mathcal{C}$ est connexe par arcs revient à dire qu'on peut toujours joindre deux points quelconques de $\mathcal{C}$ par un arc. Si $x^* \in \mathcal{C}$ on peut toujours le relier à 0 : on suit une trajectoire provenant de 0 dans le sens rétrograde ; puis on suit une autre trajectoire reliant 0 à x_0.*
– *Le caractère ouvert provient de la régularité de f qui entraîne la continuité de la trajectoire par rapport à la donnée initiale. On peut donc envoyer une boule centrée en un point x^* de $\mathcal{C}$ sur une boule centrée en 0 et on repart ensuite (tout point de la boule centrée en 0 étant contrôlable).*

6.2 Cas d'une EDO linéaire

Nous allons étudier l'ensemble contrôlable dans le cas où le système dynamique est décrit par un système d'équations différentielles linéaires de la forme (voir Annexe A).

$$(EL) \qquad \begin{cases} \dfrac{dx}{dt}(t) &= A(t)\,x(t) + B(t)\,u(t)\ ,\ t \in I =]0, T[\\ x(0) &= x_0\ . \end{cases}$$

x est la fonction inconnue de I dans $\mathbb{R}^n$, $u : I \to \mathbb{R}^p$, $A(t)$ est une matrice carrée $n \times n$ pour tout t dans I (T peut être infini) et $B(t)$ est une matrice $n \times p$ pour tout t dans I.

On note X la résolvante de ce système (cf Annexe A.)

Théorème 6.2.1 *Si l'ensemble $\mathcal{U}$ des contrôles est convexe alors l'ensemble contrôlable $\mathcal{C}$ est convexe.*
Si $\mathcal{U}$ est symétrique par rapport à 0, $\mathcal{C}$ l'est aussi.

Démonstration -

$$x_0 \in \mathcal{C} \Longleftrightarrow \exists u \in \mathcal{U},\ \exists t_1 > 0 \qquad x[u, x_0](t_1) = 0\ ,$$

c'est-à-dire

$$X(t_1)X^{-1}(0)\,x_0 + X(t_1)\int_0^{t_1} X^{-1}(s)\,B(s)\,u(s)\,ds = 0\ ,$$

où X ne dépend pas de u, c'est-à-dire

$$x_0 = -X(0)\int_0^{t_1} X^{-1}(s)\,B(s)\,u(s)\,ds\ .$$

Si $\mathcal{U}$ est symétrique, $-u \in \mathcal{U}$ et donc $-x_0 \in \mathcal{C}$. Par conséquent $\mathcal{C}$ est aussi symétrique.
Supposons maintenant que $\mathcal{U}$ est convexe.

$$x_1 \in \mathcal{C} \Longleftrightarrow \exists u_1 \in \mathcal{U},\ \exists t_1 > 0\ x_1 = -X(0) \int_0^{t_1} X^{-1}(s)\, B(s)\, u_1(s)\, ds\ .$$

$$x_2 \in \mathcal{C} \Longleftrightarrow \exists u_2 \in \mathcal{U},\ \exists t_2 > 0\ x_2 = -X(0) \int_0^{t_2} X^{-1}(s)\, B(s)\, u_2(s)\, ds\ .$$

On peut supposer $t_1 \leq t_2$; soit $\alpha \in [0,1]$. Soit

$$u = \begin{cases} \alpha\, u_1 + (1-\alpha)\, u_2 & \text{sur } [0, t_1] \\ (1-\alpha)\, u_2 & \text{sur } [t_1, t_2] \end{cases} \in \mathcal{U}.$$

On voit que $\alpha\, x_1 + (1-\alpha)\, x_2$ est associé à u et donc est dans $\mathcal{C}$. □

6.2.1 Cas des équations différentielles linéaires à coefficients constants

On suppose maintenant que pour tout t, $A(t) \equiv A$ et $B(t) \equiv B$ et que U est un sous-ensemble non vide de $\mathbb{R}^p$ et

$$\mathcal{U} = \{\ u \text{ mesurable sur } [0,t],\ \forall t;\ u(s) \in U \quad \text{p.p. } \}\ .$$

On note $\mathcal{C}_U$ l'ensemble contrôlable (qui dépend de U).

$$x_0 \in \mathcal{C}_U \Longleftrightarrow x_0 = -\int_0^t e^{As} B u(s)\, ds\ . \tag{6.2.1}$$

Définition 6.2.1 (Matrice de contrôlabilité)
On appelle $M(A,B)$ la matrice de contrôlabilité :

$$M(A,B) = \left[B, AB, A^2B, \ldots, A^{n-1}B\right]\ .$$

C'est une matrice $n \times nq$.

Théorème 6.2.2 *Si U est convexe et $0 \in$ int U (dans $\mathbb{R}^p$) ; alors*

$$\textit{rang}(M) = n \Longleftrightarrow 0 \in \text{int}\, \mathcal{C}\ .$$

Démonstration - **Montrons que rang $(M) < n \Longrightarrow 0 \notin$ int $\mathcal{C}$.**
Si rang$(M) < n$, il existe un vecteur $p \neq 0$ tel que $M^t\, p = 0$, c'est-à-dire

$$p^t M = 0\ ,$$

$$p^t\ \left[B, AB, A^2B, \ldots, A^{n-1}B\right] = 0\ ,$$

$$\forall k \in \{\, 0, \ldots n-1\,\} \qquad p^t A^k B = 0\ .$$

Soit $P(\lambda) = \det(\lambda I - A)$ le polynôme caractéristique de A. D'après le théorème de Cayley-Hamilton, on a $P(A) = 0$. Par conséquent A^n est combinaison linéaire de $A^o, A, \ldots, A^{n-1}$. Donc

$$\begin{aligned} A^n &= \beta_1 A^{n-1} + \ldots + \beta_n I \ , \\ A^n B &= \beta_1 A^{n-1} B + \ldots + \beta_n B \ , \\ p^t A^n B &= \beta_1 p^t A^{n-1} B + \ldots + \beta_n p^t B \ ; \end{aligned} \tag{6.2.2}$$

donc $p^t A^n B = 0$. On réitère le procédé en multipliant (6.2.2) par A. Par récurrence on obtient

$$\forall k \geq 0 \qquad p^t A^k B = 0 \ .$$

Comme $e^{-As} = \sum_0^{+\infty} \frac{A^k}{k!} s^k$ on en déduit que $p^t e^{-As} B$ et avec (6.2.1) on a finalement

$$\forall x_0 \in \mathcal{C} \qquad p^t x_0 = 0 \ .$$

Donc $\mathcal{C}$ est inclus dans l'hyperplan orthogonal à p qui est fermé, int $\mathcal{C} = \emptyset$ et $0 \notin$int $\mathcal{C}$.
Remarque : on n'a pas utilisé la convexité de U (et donc de $\mathcal{C}$).
Réciproquement
Supposons que $\mathcal{C}$ est convexe et que $0 \notin$ int $\mathcal{C}$. D'après le corollaire A.4.2 du Théorème de Hahn-Banach cité en Annexe A- Section A.4, il existe un hyperplan fermé séparant 0 et $\mathcal{C}$: autrement dit,

$$\exists p \in \mathbb{R}^n, \ p \neq 0 \ \text{ tel que } \forall x_0 \in \mathcal{C} \ \ p^t x_0 \geq 0 \ .$$

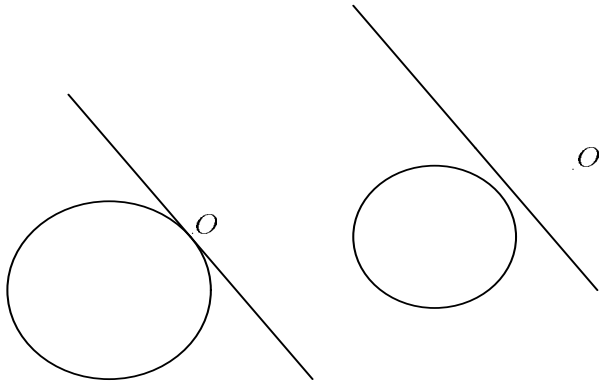

Figure 6.2. : séparation d'un convexe fermé et d'un point 0

Comme x_0 est de la forme $x_0 = -\int_0^t e^{As} B u(s)\, ds$ on a donc

$$\forall u \in \mathcal{U}, \ \forall t \geq 0 \qquad \int_0^t p^t e^{As} B u(s)\, ds \ \leq 0 \ .$$

Comme $0 \in \text{int}\,\mathcal{U}$ on peut prendre des accroissements dans une boule centrée en 0 et on obtient

$$\forall u \in \mathcal{U},\ \forall t \geq 0 \qquad \int_0^t p^t e^{As} B u(s)\, ds\ = 0\ .$$

Par conséquent $p^t e^{-As} = 0$ pour tout s et

$$\forall k \quad p^t A^k B = 0\ ,$$

ce qui implique que le rang de M est strictement inférieur à n.

□

Remarque 6.2.1 *Le théorème précédent reste vrai si on remplace $\mathcal{C}$ par $\mathcal{C}(\tau)$ (l'ensemble des états contrôlables en un temps donné $\tau > 0$.)*

En réalité l'hypothèse que U est convexe est superflue. Elle servait à assurer que $\mathcal{C}_U(t)$ est convexe pour pouvoir séparer. En fait on a le résultat plus précis suivant :

Théorème 6.2.3 *Si U est compact, alors $\mathcal{C}_U(t)$ est convexe et compact.*

Démonstration - Nous renvoyons à [12] pour une démonstration complète et nous ne donnons ici que les grandes lignes de la preuve.
• Montrons d'abord que $\mathcal{C}_U(t)$ est convexe.
Pour cela, nous admettrons le Lemme suivant [12]

Lemme 6.2.1 (Lyapounov)
Si $v \in \left(L^1(0,t)\right)^n$ alors le vecteur $v_E = \displaystyle\int_E v(s)\, ds$ décrit un ensemble convexe de $\mathbb{R}^n$: $\mathcal{E}$, lorsque E décrit l'ensemble des sous-ensembles mesurables de $[\,0,t\,]$.

Soient x_1 et x_2 dans $\mathcal{C}_U(t)$: on a donc l'existence de $u_i, i = 1,2$ dans U tels que

$$x_i = -\int_0^t e^{-A\,s} B\, u_i(s)\, ds\ ,\ i = 1,2\ .$$

Soit E un sous-ensemble mesurable de $[\,0,t\,]$ et

$$y_E = \begin{bmatrix} -\int_E e^{-A\,s} B\, u_1(s)\, ds \\ -\int_E e^{-A\,s} B\, u_2(s)\, ds \end{bmatrix}\ ,\quad y_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}\ ,\quad y_{[0,t]} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\ .$$

Soit $\lambda \in [0,1]$. Par convexité de $\mathcal{E}$ (due au lemme 6.2.1), nous avons $(1-\lambda)\, y_0 + \lambda\, y_{[0,t]} \in \mathcal{E}$. Donc, il existe D_λ un sous-ensemble mesurable de $[0,t]$ tel que

$$y_{D_\lambda} = (1-\lambda)\, y_0 + \lambda\, y_{[0,t]} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\ .$$

Par complémentarité :

$$y_{D_\lambda^c} = (1-\lambda) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\ .$$

Définissons le contrôle $u_\lambda \in \mathcal{U}$ de la manière suivante :

$$u_\lambda(s) = \begin{cases} u_1(s) & \text{si } s \in D_\lambda \\ u_2(s) & \text{si } s \notin D_\lambda \end{cases}$$

et calculons l'état "initial" correspondant à ce contrôle :

$$\begin{aligned} -\int_0^t e^{-A\,s} B\, u_\lambda(s)\, ds &= -\int_{D_\lambda} e^{-A\,s} B\, u_1(s)\, ds - \int_{D_\lambda^c} e^{-A\,s} B\, u_2(s)\, ds \\ &= (y_{D_\lambda})_1 + (y_{D_\lambda^c})_2 \\ &= \lambda\, x_1 + (1-\lambda)\, x_2\ . \end{aligned}$$

Par conséquent $\lambda\, x_1 + (1-\lambda)\, x_2$ est contrôlable et appartient bien à $\mathcal{C}_U(t)$.
• Montrons que $\mathcal{C}_U(t)$ est borné.
Tout élément de $\mathcal{C}_U(t)$ s'écrit

$$x = -\int_0^t e^{-A\,s} B\, u(s)\, ds\ .$$

Donc

$$\|x\| \leq \int_0^t \|e^{-A\,s}\|\, \|B\|\, \|u(s)\|\, ds\ .$$

Or $u(s) \in U$ et U est compact donc borné. Par conséquent, $\|x\| \leq C_t$ et $\mathcal{C}_U(t)$ est borné.
• Il reste à montrer que $\mathcal{C}_U(t)$ est fermé. C'est le point le plus technique de la démonstration et nous renvoyons à [12], pour une preuve détaillée. □

Dès lors, on peut en déduire des résultats pratiques sur la contrôlabilité ou l'atteignabilité du système dynamique suivant

$$\begin{cases} \dfrac{dx}{dt}(t) = A\, x(t) + B\, u(t)\ ,\ x(0) = x_0 \\ u(s) \in U \text{ p.p.} \end{cases}$$

S'il y a des contraintes sur le contrôle telles que $0 \in$ int U, alors il existe un voisinage de 0 contrôlable (avec rang(M) = n). S'il est possible d'atteindre ce voisinage à partir de tout état initial x_0, le système sera (complètement) contrôlable. C'est le cas si les valeurs propres de A sont telles que leur partie réelle Re$\lambda(A) < 0$, car alors la trajectoire correspondant au contrôle nul : $x(t) = e^{At}x_0$ converge vers 0 quand $t \to +\infty$ et donc atteint le voisinage en temps fini.
Par conséquent

Théorème 6.2.4 *Si* $0 \in$ int U, rang $(M) = n$ *et* Re $\lambda(A) < 0$, *alors* $\mathcal{C} = \mathbb{R}^n$.

On a même mieux :

Théorème 6.2.5 *Supposons U compact et* $0 \in$ int U.

$$\text{rang } M = n \text{ et Re } \lambda(A) \leq 0 \Longleftrightarrow \mathcal{C} = \mathbb{R}^n\ .$$

Démonstration - Voir [12]. □

6.2.2 Cas des EDO linéaires à coefficients constants sans contraintes sur le contrôle

On se place dans le cas où $U = \mathbb{R}^p$ (pas de contraintes sur le contrôle). Nous pouvons alors préciser les résultats précédents :

Théorème 6.2.6 *Les propriétés suivantes sont équivalentes :*
*(i) Le système est contrôlable (ou de manière équivalente la paire **(A,B)** est contrôlable (i.e. $\mathcal{C} = \mathbb{R}^n$)*
(ii) La matrice de contrôlabilité $\mathbb{M} = [B, AB, \dots, A^{n-1}B]$ est de rang maximal (égal à n).
De plus si $\mathcal{C} = \mathbb{R}^n$, la matrice $[A - \lambda I, B]$ est de rang maximal pour tout $\lambda \in \mathbb{C}$.

Démonstration - La démonstration de l'équivalence $(i) \Longleftrightarrow (ii)$ est une conséquence du théorème 6.2.2 avec $U = \mathbb{R}^p$.
Supposons que nous avons *(ii)* et que la matrice $[A - \lambda I, B]$ n'est pas de rang maximal . Alors, il existe un vecteur $x^* \in \mathbb{C}^n$ tel que $(x^*)^t A = \lambda (x^*)^t$ et $(x^*)^t B = 0$.

$$(x^*)^t \mathbb{M} = [(x^*)^t B, (x^*)^t A\, B, \dots, (x^*)^t A^{n-1} B] = [(x^*)^t B, \lambda (x^*)^t\, B, \dots, \lambda^{n-1} (x^*)^t B] = 0\ ;$$

cela contredit le fait que $\mathbb{M}$ est de rang maximal n. □

6.3 Stabilité et observabilité

6.3.1 Stabilité

Nous avons défini la notion de stabilité d'un système dans le chapitre 4. Nous allons étudier plus particulièrement le cas des systèmes décrits par des EDO linéaires à coefficients constants. Nous allons donner une caractérisation simple de la stabilité (asymptotique) dans ce cas. On considère donc le système (autonome) décrit par

$$(EH) \qquad \begin{cases} \dfrac{dx}{dt}(t) = A\, x(t) \quad \text{pour } t > 0 \\ x(0) = x_0\ . \end{cases}$$

Théorème 6.3.1 *Le système (EH) est asymptotiquement stable si et seulement si toutes les valeurs propres de A sont à partie réelle strictement négative.*

Démonstration - La définition de la stabilité asymptotique a été donnée au Chapitre 4 - Définition 4.2.4.
La solution de (EH) est de la forme $x(t) = e^{A\,t}\, x_0$. Ici le système a un point d'équilibre $\bar{x}$ vérifiant $e^{A\,t}\bar{x} = \bar{x}$, pour tout t. Soit $t > t_0$:

$$\|x(t) - \bar{x}\| = \|e^{A\,t}\,(x_0 - \bar{x})\| = \|e^{A\,(t-t_0)}\, e^{A\,t_0}\,(x_0 - \bar{x})\| = \|e^{A\,(t-t_0)}(x(t_0) - \bar{x})\|\ .$$

Donc

$$\|x(t) - \bar{x}\| \leq \|x(t_0) - \bar{x}\|\ \|e^{A\,(t-t_0)}\|\ .$$

On ne peut donc avoir stabilité (et du même coup stabilité asymptotique) que si toutes les valeurs propres de A ont une partie réelle (qui seule intervient dans le calcul de la norme de $e^{A\,t}$) strictement négative. □

Remarque 6.3.1 *On constate que les notions de contrôlabilité et de stabilité sont liées via les valeurs propres de A.*

Ceci nous amène à la notion de **stabilisation**. Considérons maintenant le système contrôlé

$$(EC) \qquad \begin{cases} \dfrac{dx}{dt}(t) = A\,x(t) + B\,u(t) \quad \text{pour } t > 0 \\ x(0) = x_0 \end{cases}$$

Définition 6.3.1 (Stabilisation) *Le système (EC) est stabilisable si on peut trouver une loi de commande F donnant un contrôle feedback $u = Fx$, telle que le système associé (i.e. (EH) avec $A + BF$ au lieu de A) est stable.*

Nous avons le résultat suivant qui est une conséquence du théorème 6.2.6.

Théorème 6.3.2 *Les propriétés suivantes sont équivalentes :*
(i) Le système (EC) est stabilisable
(ii) Il existe une matrice F telle que $A+BF$ a toutes ses valeurs propres à partie réelle strictement négative.

6.3.2 Observabilité

Nous envisageons enfin, pour terminer, un système dynamique linéaire "complet" :
$u : \mathbb{R} \to \mathbb{R}^p$ est la commande ou variable d'entrée (**input**) , $x : \mathbb{R} \to \mathbb{R}^n$ est la variable d'état et $y : \mathbb{R} \to \mathbb{R}^q$ est la variable de sortie (observée) (**output**). Le système est décrit par

$$\begin{array}{l} \dfrac{dx}{dt} = A\,x + B\,u,\ x(0) = x_0\ , \\ y = C\,x + D\,u\ . \end{array} \tag{6.3.1}$$

$A,\ B,\ C$ et D sont des matrices réelles constantes de taille appropriée.

Définition 6.3.2 (Observabilité)
*Le système décrit par (6.3.1) est dit observable , ou la paire **(C,A)** est dite observable, si pour tout instant t_1, l'état initial $x(0) = x_0$ peut être déterminé à partir de l'entrée (commande) u et de la sortie (observation) y dans l'intervalle $[\,0, t_1\,]$.*

La notion d'observabilité est la notion **duale** de la contrôlabilité. Elle peut être étudiée sur le système non contrôlé comme le montre le résultat suivant.

Théorème 6.3.3 *Les propriétés suivantes sont équivalentes :*
(i) Le système (6.3.1) (ou paire (C, A)) est observable
(ii) la matrice d'observabilité

$$\mathbb{O} = \begin{bmatrix} C \\ C\,A \\ C\,A^2 \\ \vdots \\ C\,A^{n-1} \end{bmatrix}$$

est de rang maximal (ce qui est équivalent à $\bigcap_{i=1}^{n} \ker (C\,A^{i-1}) = \{\,0\,\}$*).*

(iii) La matrice $\begin{bmatrix} A-\lambda I \\ C \end{bmatrix}$ *est de rang maximal pour tout* $\lambda \in \mathbb{C}$*.*

(iv) le système décrit par (A^*, C^*) *est contrôlable (où* A^* *est la matrice conjuguée de* A*).*

Démonstration - Il suffit de montrer l'équivalence de (i) et (ii). Le reste se déduit par dualité, du théorème 6.2.6.

• Montrons d'abord $(ii) \Longrightarrow (i)$

Si on se donne u et la condition initiale $x(0)$, la sortie y sur l'intervalle $[\,0, t_1\,]$ est donnée par

$$y(t) = C\,e^{A\,t}x(0) + \int_0^t C\,e^{A\,(t-s)}\,B\,u(s)\,ds + D\,u(t)\;.$$

Comme $y(t)$ et $u(t)$ sont connus, on peut supposer que $u(t) \equiv 0$. Nous obtenons alors

$$y(t) = C\,e^{A\,t}x(0)\;, t \in [\,0, t_1\,]\;.$$

Ceci entraîne

$$\begin{bmatrix} y(0) \\ \dfrac{dy}{dt}(0) \\ \vdots \\ y^{(n-1)}(0) \end{bmatrix} = \begin{bmatrix} C \\ C\,A \\ \vdots \\ C\,A^{n-1} \end{bmatrix} x(0)\;,$$

où $y^{(i)}$ est la dérivée i-ème de y. Comme $\mathbb{O}$ est de rang maximal, le système ci-dessus a une solution unique.

• Réciproquement : supposons que (C, A) est observable mais que $\mathbb{O}$ n'est pas de rang maximal, c'est-à-dire qu'il existe un vecteur x_0 tel que $\mathbb{O}\,x_0 = 0$, ou de manière équivalente

$$C\,A^i x_0 = 0, \text{ pour tout } i$$

par le théorème de Cayley-Hamilton. Supposons maintenant que l'état du système x est déterminé par $x(0) = x_0$. Alors $y(t) = C\,e^{A\,t}x(0) = 0$. Le système n'est donc pas observable puisque $x(0)$ ne peut pas être déterminé à partir de $y \equiv 0$. □

[Exercices]

1. On considère un système dynamique décrit par

$$\begin{aligned} &\frac{dx}{dt} = A\,x + B\,u,\; x(0) = x_0\;, \\ &y = C\,x + D\,u\;. \end{aligned} \tag{1}$$

et une loi de commande déterminant un contrôle feedback

$$u = F\,x + v\ .$$

Ce système en boucle fermée est schématisé par la figure suivante

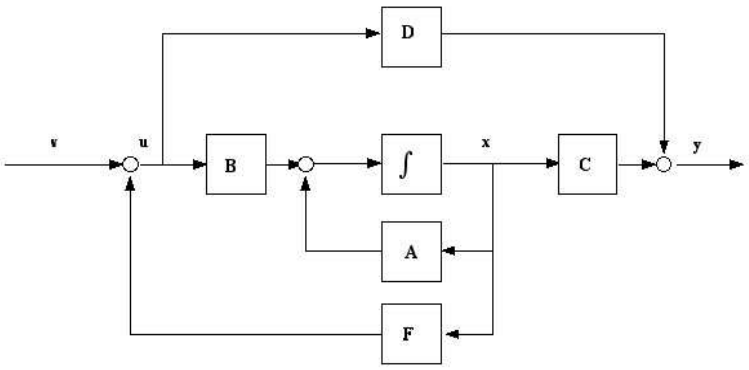

Figure 6.3. : système en boucle fermée

(a) Montrer que (A, B) est contrôlable (respectivement stabilisable) si et seulement si $(A + BF, B)$ est contrôlable (respectivement stabilisable) .

(b) On se donne

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\ ,\ B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}\ ,\ C = \begin{bmatrix} 1 & 0 \end{bmatrix}\ \text{ et }\ D = 0.$$

Montrer que ce système est contrôlable et observable.

(c) On se donne la loi de commande : $u = F\,x$ avec $F = [-1 \quad -1]$. Montrer que le système n'est pas complètement observable.

---

2. On considère le système suivant (n=2, p=1) :

$$x = \begin{bmatrix} p \\ q \end{bmatrix},\ \frac{dx}{dt} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t), \quad \mathcal{T}(t) \equiv 0,$$

qui est une petite variante du problème du véhicule à réaction. Utiliser les théorèmes du cours pour montrer que $\mathcal{C} \neq \mathbb{R}^2$. Trouver $\mathcal{C}$ par calcul direct.

---

3. *On va montrer que même si $x_0 \in \mathcal{C}$, les états décrits par la trajectoire issue de x_0 peuvent ne pas être dans $\mathcal{C}$.*

On considère le système dont les équations d'état sont :

$$\left.\begin{array}{rcl} \dfrac{dp}{dt} & = & -(1-t)q \\ \dfrac{dq}{dt} & = & (1-t)p \end{array}\right\} \quad \text{pour } 0 \leq t \leq 1$$

$$\left.\begin{array}{rcl} \dfrac{dp}{dt} & = & 0 \\ \dfrac{dq}{dt} & = & [u(t)-2](t-1) \end{array}\right\} \quad \text{pour } 1 \leq t < +\infty \quad .$$

L'ensemble U des contrôles est inclus dans $]-\infty, 2]$.
Montrer que pour $0 \leq t \leq 1$ les trajectoires sont supportées par des cercles parcourus dans le sens trigonométrique, dans le plan (p, q), tandis que pour $t > 1$ elles sont supportées par des verticales parcourues de haut en bas.
Décrire $\mathcal{C}$ et trouver une trajectoire issue de $\mathcal{C}$. Montrer qu'aucun état de cette trajectoire n'est contrôlable sauf l'état initial.

---

4. Décrire l'ensemble commandable $\mathcal{C}$ pour le système suivant (avec $\mathcal{T}(t) = 0$).

$$\frac{dx}{dt} = \begin{bmatrix} 0 & -1 & 1 \\ 2 & -3 & 1 \\ 1 & -1 & -1 \end{bmatrix} x + \begin{bmatrix} -1 & 1 \\ 0 & 2 \\ 1 & 3 \end{bmatrix} u \quad (n = 3, p = 2) \; .$$

---

5. On considère le système suivant (n=3, p=1) :

$$\frac{dx}{dt} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix} x + u(t)\, b \; .$$

Pour quels vecteurs constants $b \in \mathbb{R}^3$ a-t'on $\mathcal{C} = \mathbb{R}^n$?

Chapitre 7

Commande en temps minimum de systèmes linéaires à coefficients constants

Dans ce chapitre on ne fixe plus le temps final T mais la valeur finale. On cherche à atteindre cette valeur en un temps minimal. On ne considère que des systèmes dynamiques décrits par des EDO linéaires à coefficients constants.

7.1 Existence d'un temps optimal

Comme dans le chapitre précédent, on considère un système dynamique décrit par l'EDO

$$\left\{ \begin{array}{rcl} \dfrac{dx}{dt} & = & A\,x(t) + B\,u(t)\ ,\ \text{sur } [0,+\infty[\ , \\ x(0) & = & x_0\ . \end{array} \right. \tag{7.1.1}$$

Le contrôle u est dans l'ensemble

$$\mathcal{U} = \{\ u \text{ mesurable sur } \mathbb{R}\ \mid u(s) \in U \text{ presque partout }\}$$

où U est un **compact** de $\mathbb{R}^p$ contenant 0.
L'état associé est une fonction de la variable t et dépend du contrôle u et du point de départ x_0 : on le note $x[u,x_0](\cdot)$.

On cherche à atteindre l'état final 0 en temps **minimum**. Il faut donc qu'il existe des états **contrôlables** , c'est-à-dire tels que 0 soit **atteignable**. On cherche ensuite un contrôle u qui permet de minimiser

$$J(u) = \int_0^t 1\ ds\ .$$

On désigne par u^* et t^* (s'ils existent) un contrôle optimal et la durée optimale.
Rappelons que dire que 0 est atteignable revient à dire que x_0 est contrôlable, c'est-à-dire

$$\exists \tau > 0,\ \exists u \in \mathcal{U} \text{ tels que } x[u,x_0](\tau) = 0\ .$$

Pour un couple (τ, x_0) de $]\,0, +\infty[\times\mathbb{R}^n$ donné, on définit l'ensemble des points atteignables à partir de x_0 en un temps τ par

$$\mathcal{A}(x_0, \tau) = \{\ x[u, x_0](\tau)\ \mid\ u \in \mathcal{U}\ \} \subset \mathbb{R}^n\ . \tag{7.1.2}$$

Comme nous connaissons l'expression explicite de la solution de (7.1.1), on a également une caractérisation des éléments de $\mathcal{A}(x_0, \tau)$:

$$y \in \mathcal{A}(x_0, \tau) \Longleftrightarrow \exists u \in \mathcal{U}\ \ y = e^{A\tau}x_0 + e^{A\tau}\int_0^\tau e^{-As}Bu(s)\,ds\ .$$

Comme U est compact, il est facile de voir que $\mathcal{A}(x_0, \tau)$ est **compact**. D'autre part, comme dans le chapitre 6 (Théorème 6.2.3) on peut montrer que $\mathcal{A}(x_0, \tau)$ est **convexe** (grâce à la compacité de U).
Nous avons le résultat d'existence suivant :

Théorème 7.1.1 *Supposons que l'état (final)* 0 *est atteignable. Alors il existe un contrôle optimal qui permet de l'atteindre en temps minimum.*

Démonstration - Comme 0 est atteignable, il existe $\tilde{t} > 0$ tel que $0 \in \mathcal{A}(x_0, \tilde{t})$. Soit

$$\mathcal{E} = \{\ t \geq 0\ \mid\ 0 \in \mathcal{A}(x_0, t)\ \}\ .$$

Comme $\tilde{t} \in \mathcal{E}$, $\mathcal{E} \neq \emptyset$; $\mathcal{E}$ est minoré par 0 et donc la borne inférieure existe. Soit

$$t^* = \inf\ \{\ t \geq 0\ \mid\ 0 \in \mathcal{A}(x_0, t)\ \}\ .$$

Montrons que t^* est atteint, c'est-à-dire $t^* \in \mathcal{E}$.
Soit t_n une suite minimisante i.e.

$$t_n \in \mathcal{E} \text{ et } t_n \to t^*\ .$$

Comme $t_n \in \mathcal{E}$, $0 \in \mathcal{A}(x_0, t_n)$ et il existe $u_n \in \mathcal{U}$ tel que $x[u_n, x_0](t_n) = 0$.
Or pour tout $s \in [0, \tilde{t}]$, $u_n(s)$ appartient à U qui est compact dans $\mathbb{R}^p$. Donc (quitte à extraire une sous-suite) $u_n(s) \to u^*(s)$ pour presque tout s ; de plus u^* est aussi à valeurs dans U. D'autre part, comme U est borné

$$\forall s \qquad |u_n(s)| \leq M\ ,$$

et

$$\int_0^{\tilde{t}} |u_n(s)|\,ds \leq M\,\tilde{t}\ .$$

Donc u_n est bornée dans L^1 et par le théorème de Lebesgue u_n converge vers u^* dans L^1. Nous admettrons que la solution de l'EDO est continue par rapport au paramètre pour la norme L^1. Par conséquent,

$$x[u_n, x_0](t_n) \to x[u^*, x_0](t^*) = 0\ .$$

Donc $t^* \in \mathcal{E}$ et l'inf est atteint. □
Il n'y a en général pas unicité.

7.2 Principe du minimum de Pontryagin

Nous allons essayer de caractériser le temps minimum t^*. Pour cela nous commençons par montrer qu'on peut séparer (au sens large) le point "cible" 0 de l'ensemble atteignable $\mathcal{A}(x_0, t^*)$.

Théorème 7.2.1 *Il existe un hyperplan séparant* 0 *et* $\mathcal{A}(x_0, t^*)$. *Plus précisément, on peut trouver* $p \in \mathbb{R}^n$, $p \neq 0$ *tel que :*

$$\forall y \in \mathcal{A}(x_0, t^*) \qquad \langle p, y \rangle_n \geq 0 \, . \tag{7.2.1}$$

Démonstration - Soit t_k une suite de réels positifs, strictement croissante vers t^*. Comme $\mathcal{A}(x_0, t_k)$ est convexe et fermé, le projeté de 0 sur $\mathcal{A}(x_0, t_k)$ existe et on le note x_k. (C'est donc l'élément de $\mathcal{A}(x_0, t_k)$ de norme minimale). D'après les résultats sur la projection du chapitre 3- Section 3.4.1, x_k vérifie en particulier

$$\forall x \in \mathcal{A}(x_0, t_k) \qquad \langle x_k, x - x_k \rangle_n \geq 0 \, . \tag{7.2.2}$$

Comme t^* est minimal, $0 \notin \mathcal{A}(x_0, t_k)$; par conséquent aucun x_k n'est nul et on peut définir $p_k = \dfrac{x_k}{\|x_k\|}$. Nous avons, bien sûr

$$\langle p_k, x_k \rangle_n = \|x_k\| \geq 0 \, ,$$

et grâce à (7.2.2)

$$\forall x \in \mathcal{A}(x_0, t_k) \qquad \langle p_k, x - x_k \rangle_n = \frac{\langle x_k, x - x_k \rangle_n}{\|x_k\|} \geq 0 \, .$$

Donc finalement

$$\forall k \, , \forall x \in \mathcal{A}(x_0, t_k) \qquad 0 \leq \langle p_k, x_k \rangle_n \leq \langle p_k, x \rangle_n \, . \tag{7.2.3}$$

Or x_k est bornée ainsi que p_k. Quitte à extraire des sous-suites que l'on note de la même façon, on peut donc trouver $x^* \in \mathbb{R}^n$ et $p \in \mathbb{R}^n$ tels que $x_k \to x^*$ et $p_k \to p^*$.
D'autre part, tout point y de $\mathcal{A}(x_0, t^*)$ est limite d'une suite y_k de points de $\mathcal{A}(x_0, t_k)$: en effet, on peut trouver $u \in \mathcal{U}$ tel que

$$y = e^{A\,t^*} x_0 + e^{A\,t^*} \int_0^{t^*} e^{-As} B u(s) \, ds \, .$$

On pose alors

$$y_k = e^{A\,t_k} x_0 + e^{A\,t_k} \int_0^{t_k} e^{-As} B u(s) \, ds \in \mathcal{A}(x_0, t_k) \, .$$

On peut donc passer à la limite dans (7.2.3) , ce qui donne

$$\forall x \in \mathcal{A}(x_0, t^*) \qquad 0 \leq \langle p, x^* \rangle_n \leq \langle p, x \rangle_n \, .$$

□

Remarque 7.2.1 *Nous avons choisi de restreindre l'ensemble cible (à atteindre) au seul singleton* $\{0\}$*, pour simplifier l'exposé. Il est facile de passer au cas où l'ensemble cible est réduit à un point quelconque (pas nécessairement* $\{0\}$*), par translation (le système est linéaire).*
On peut même envisager un ensemble cible plus "grand" mais toujours convexe, fermé : les résultats énoncés ici sont toujours vrais et les démonstrations sont à peine plus compliquées.

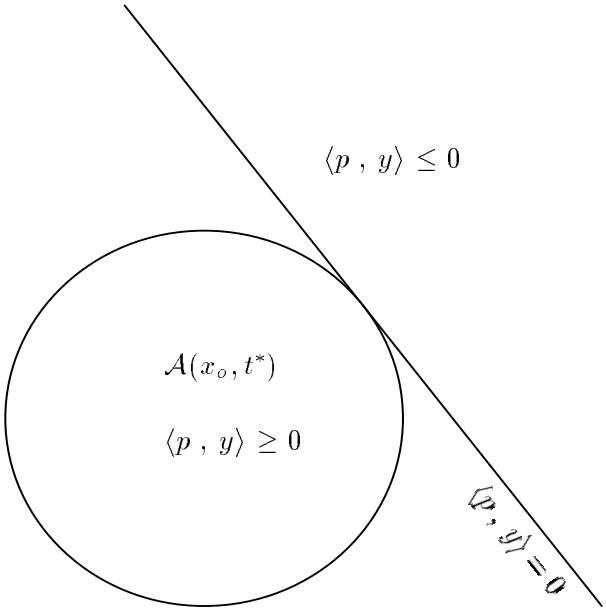

Figure 7.1

Le théorème 7.2.1 donne de manière indirecte un renseignement sur la position de 0 par rapport à $\mathcal{A}(x_0, t^*)$. On sait déjà que $0 \in \mathcal{A}(x_0, t^*)$; le fait qu'on puisse séparer 0 du reste de l'ensemble $\mathcal{A}(x_0, t^*)$ prouve que 0 est nécessairement sur la frontière de $\mathcal{A}(x_0, t^*)$:

$$0 \in \partial\mathcal{A}(x_0, t^*)\ .$$

En effet, si 0 était dans l'intérieur on ne pourrait pas séparer en raison du théorème A.4.3 de l'Annexe A - Section A.4.

Nous allons maintenant détailler le résultat du théorème 7.2.1. Dire que $y \in \mathcal{A}(x_0, t^*)$ revient à dire que l'on peut trouver u dans $\mathcal{U}$ tel que $y = x[u, x_0](t^*)$ où x est solution de (7.1.1), et réciproquement.
Ecrivons explicitement cette solution :

$$y = x[u, x_0](t^*) = e^{A\,t^*}x_0 + e^{A\,t^*}\int_0^{t^*} e^{-A\,s}Bu(s)\,ds\ .$$

D'autre part $0 \in \mathcal{A}(x_0, t^*)$ donne l'existence de u^* dans $\mathcal{U}$ tel que $x[u^*, x_0](t^*) = 0$. On obtient

$$\left\{\begin{array}{ll} y = & e^{A\,t^*}x_0 + e^{A\,t^*}\displaystyle\int_0^{t^*} e^{-A\,s}Bu(s)\,ds\ , \\ 0 = & e^{A\,t^*}x_0 + e^{A\,t^*}\displaystyle\int_0^{t^*} e^{-A\,s}Bu^*(s)\,ds\ , \end{array}\right.$$

La relation (7.2.1) s'écrit alors

$$\forall u \in \mathcal{U} \qquad \left\langle p\,,\, \int_0^{t^*} e^{A\,(t^*-s)} B(u(s)-u^*(s))\,ds \right\rangle_n \geq 0\,,$$

c'est-à-dire

$$\forall u \in \mathcal{U} \qquad \int_0^{t^*} \left\langle p\,,\, e^{A\,(t^*-s)} B(u(s)-u^*(s)) \right\rangle_n ds \geq 0\,,$$

ce qui est finalement équivalent à

$$\text{Pour presque tout } s,\ \forall v \in U\,, \qquad \left\langle p\,,\, e^{A\,(t^*-s)} B(v-u^*(s)) \right\rangle_n \geq 0\,. \tag{7.2.4}$$

C'est une condition nécessaire appelée **Principe du minimum de Pontryagin**.
On va transformer un peu la relation précédente. Soit la fonction vectorielle p^* définie par

$$p^*(s) = e^{A^t(t^*-s)} p\,.$$

D'une part, cette fonction vérifie, pour presque tout s

$$\forall v \in U\,, \qquad \langle\, p^*(s)\,,\, B(v-u^*(s))\,\rangle_n \geq 0\,; \tag{7.2.5}$$

d'autre part elle est solution de l'équation différentielle (homogène)

$$\left\{ \begin{array}{rcl} \dfrac{dp^*}{dt}(s) & = & -A^t p^*(s) \\ p^*(t^*) & = & p\,. \end{array} \right. \tag{7.2.6}$$

La condition (7.2.5) donne alors

$$\text{Pour presque tout } s,\ \forall v \in U \qquad \langle\, p^*(s)\,,\, Bv\,\rangle_n \geq \langle\, p^*(s)\,,\, Bu^*(s)\,\rangle_n\,.$$

Autrement dit, pour presque tout s de $\mathbb{R}^+$, $u^*(s)$ est solution du problème $(\mathcal{P}_s)$:

$$(\mathcal{P}_s) \qquad \left\{ \begin{array}{l} \inf\ H_s(v) \stackrel{\text{def}}{=} \langle\, p^*(s)\,,\, Bv\,\rangle_n \\ v \in U\,. \end{array} \right.$$

La fonction H ainsi minimisée à chaque instant par $u^*(s)$ est le **Hamiltonien** du système.
Nous pouvons résumer la situation sous la forme du théorème suivant :

Théorème 7.2.2 (Principe de Pontryagin, cas linéaire à temps minimum)
Soit u^* une commande admissible transférant le système de x_0 en $x[u^*,x_0](t^*) = 0$, dans le temps t^*. Si t^* est minimum, il existe une solution p^* non identiquement nulle aux équations adjointes telle que pour presque tout s, $u^*(s)$ réalise le minimum de l'Hamiltonien $v \mapsto H(v) = \langle\, p^*(s)\,,\, Bv\,\rangle_n$ sur U.

Une façon synthétique d'exprimer cette condition nécessaire est de dire que toute trajectoire $\bar{x}$ et tout contrôle optimal u^* vérifient le système d'(in)équations :

$$\left\{ \begin{array}{l} \dfrac{dx^*}{dt} = A\,x^* + B\,u^*\ ,\ \dfrac{dp^*}{dt} = -A^t\,p^*\ , \\ x^*(0) = x_0\ ,\ \ x^*(t^*) = 0\ , \\ \langle\, p^*(s)\,,\, Bu^*(s)\,\rangle_n \leq \langle\, p^*(s)\,,\, Bv\,\rangle_n\ ,\ \forall v \in U,\ \text{pp. } s\ . \end{array} \right. \tag{7.2.7}$$

Cet ensemble de conditions est typique en contrôle optimal (on pourra le comparer aux conditions obtenues dans le chapitre précédent) : x^* est l'état optimal, p^* est **l'état adjoint**.

7.3 Unicité

On va regarder de plus près les contrôles susceptibles de vérifier (7.2.7). On suppose de plus dans cette section que U est **convexe** et **compact** (en pratique U est souvent un polyèdre borné).

Nous avons vu que si u^* est un contrôle optimal alors pour presque tout s, $u^*(s)$ réalise le minimum d'une fonctionnelle linéaire ($v \mapsto \left\langle\, B^t p^*(s)\,,\, v\,\right\rangle_p$) sur U : c'est le principe de Pontryagin. D'après des résultats classiques de programmation linéaire, $u^*(s)$ est nécessairement un point extrêmal de U, sauf bien sûr si la fonctionnelle est nulle sur une arête, c'est-à-dire si $B^t p^*(s)$ est orthogonal à une arête. On va faire une hypothèse qui assure que ça ne peut arriver qu'à des instants isolés et que la commande est par conséquent nécessairement **bang-bang**, c'est-à-dire pour presque tout s, $\ u^*(s)$ est un point extrêmal de U.

On rappelle qu'une **arête** est un segment joignant deux points extrêmaux (ou **sommets**) de U. On fait donc l'hypothèse suivante

$$(\mathcal{H}) \qquad \left\{ \begin{array}{c} \forall p \in \mathbb{R}^n\ ,\ p \neq 0\ ,\ \forall w \in \mathbb{R}^p\ \text{ arête de } U \\ \varphi(s) = p^t e^{As} Bw \text{ n'a qu'un nombre fini de zéros sur tout segment compact.} \end{array} \right.$$

Comme φ est analytique $(\mathcal{H})$ signifie que φ n'est pas identiquement nulle. En réalité cette hypothèse "technique" n'est pas utilisable en l'état. On admettra qu'une forme équivalente est donnée par le critère suivant :

$$(\mathcal{H}) \qquad \left\{ \begin{array}{c} \forall p \in \mathbb{R}^n\ ,\ p \neq 0\ ,\ \forall w \in \mathbb{R}^p\ \text{ arête de } U \\ p^t \left[B, AB, \ldots, A^{n-1}B\right] w \neq 0\ , \end{array} \right.$$

ce qui revient encore à dire que :
pour toute arête w de U le rang de la matrice $\left[Bw, ABw, \ldots, A^{n-1}Bw\right]$ est maximal, égal à n. Cette condition justifie la définition suivante :

Définition 7.3.1 (Système dynamique linéaire normal)
Le système décrite par l'EDO : $\dfrac{dx}{dt} = A\,x + B\,u,\ u(s) \in U$, ***convexe***, ***compact***, *presque partout, est* ***normal*** *si*

$$\forall w \textit{ arête de } U\ ,\ Rg\left(\left[Bw, ABw, \ldots, A^{n-1}Bw\right]\right) = n\ .$$

Exemples
- Si U est strictement convexe (par exemple si $U = \{\ v\ \mid\ \|v\|_2 \leq 1\ \}$), il n'y a pas d'arête, et donc le système est toujours *normal*.
- Si U est le cube $\{\ v\ \mid\ |v_i| \leq 1\ , i = 1, \ldots, p\ \}$, il y a p directions d'arêtes qui sont les p vecteurs de base de $\mathbb{R}^p$; donc le système est *normal* si et seulement si le rang de la matrice $\left[b_i, Ab_i, \ldots, A^{n-1}b_i\right]$ est maximal (égal à n) pour les p vecteurs colonnes b_i de B.

Si un système dynamique linéaire est normal, le contrôle en temps optimal, s'il existe, est nécessairement bang-bang, donc **unique**. En effet si u_1^* et u_2^* sont optimaux, alors $v^* = \dfrac{u_1^* + u_2^*}{2}$ est aussi optimal et donc $v^* = u_1^* = u_2^*$ car composés de points extrêmaux.
En résumé :

Théorème 7.3.1 *Si le système est **normal** et si 0 est **atteignable**, il existe une unique commande en temps minimum qui est une commande bang-bang.*

Remarque 7.3.1 *Si U a un nombre fini de points extrêmaux (si U est un polyèdre compact et convexe par exemple), la commande optimale est constante par morceaux : il n'y a qu'un nombre fini de commutations.*

7.4 Réciproque du principe du minimum

Le principe du minimum est une condition nécessaire. On va montrer que sous certaines hypothèses c'est aussi une condition suffisante.

Théorème 7.4.1 *Si $0 \in$ int U et si le rang de $C = \left[B, AB, \ldots, A^{n-1}B\right]$ est maximal (égal à n), alors toute trajectoire vérifiant les conditions (7.2.7) du principe du minimum est optimale, c'est-à-dire correspond à t^* minimal.*

Démonstration - Soient $x^*,\ t^*,\ u^*$ et p^* vérifiant (7.2.7) et supposons qu'il existe $\tilde{u}$ dans $\mathcal{U}$ transférant le système en 0 en un temps $\tilde{t} < t^*$. On a donc

$$x[\tilde{u}, x_0](\tilde{t}) = e^{A\,\tilde{t}}x_0 + \int_0^{\tilde{t}} e^{A\,(\tilde{t}-s)}B\,\tilde{u}(s)\,ds = 0\ .$$

De plus (7.2.7) donne

$$e^{A\,t^*}x_0 + \int_0^{t^*} e^{A\,(t^*-s)}B\,u^*(s)\,ds = 0\ .$$

En multipliant la première relation par $p^{*\,t}e^{A\,(t^*-\tilde{t})}$ et la deuxième par $p^{*\,t}$, puis en faisant la différence on obtient

$$\int_0^{\tilde{t}} p^{*\,t}e^{A\,(t^*-s)}B\,\tilde{u}(s)\,ds - \int_0^{t^*} p^{*\,t}e^{A\,(t^*-s)}B\,u^*(s)\,ds = 0\ ,$$

$$\int_0^{\tilde{t}} p^{*\,t}e^{A\,(t^*-s)}B\,(\tilde{u}(s) - u^*(s))\,ds - \int_{\tilde{t}}^{t^*} p^{*\,t}e^{A\,(t^*-s)}B\,u^*(s)\,ds = 0\ .$$

La relation (7.2.7) entraîne que

$$\forall v \in U \qquad {p^*}^t e^{A\,(t^*-s)} B\,(v - u^*(s)) \geq 0\ ,$$

et $u^*(s)$ est un élément de U ; donc

$$\int_0^{\tilde{t}} {p^*}^t e^{A\,(t^*-s)} B\,(\tilde{u}(s) - u^*(s))\ ds \geq 0\ ,$$

ce qui implique

$$\int_{\tilde{t}}^{t^*} {p^*}^t e^{A\,(t^*-s)} B\,u^*(s)\ ds \geq 0\ .$$

Par conséquent sur un sous-ensemble I de $[\tilde{t}, t^*]$ de mesure strictement positive, on a

$$\forall v \in U \qquad {p^*}^t e^{A\,(t^*-s)} B\,v \geq {p^*}^t e^{A\,(t^*-s)} B\,u^*(s) \geq 0\ .$$

Comme 0 est dans l'intérieur de U, on peut prendre localement v et $-v$ (de normes assez petites) autour de 0 en restant dans U et on obtient

$$\forall v \in U, \forall s \in I \qquad {p^*}^t e^{A\,(t^*-s)} B\,v = 0\ ,$$

c'est-à-dire

$$\forall s \in I \qquad {p^*}^t e^{A\,(t^*-s)} B = 0\ ,$$

et donc ${p^*}^t \left[B, AB, \ldots, A^{n-1}B\right] = 0$. Ceci contredit le fait que la matrice $\left[B, AB, \ldots, A^{n-1}B\right]$ est de rang n. □

[Exercices]

1. **On va montrer que le rang de la matrice de contrôlabilité peut être n, sans que le système soit normal.**

   Soient n=p=2 et le système : $\dfrac{dx}{dt} = u$, où

   $$u \in U := \{\ u \in \mathbb{R}^p\ \mid\ |u_i| \leq 1\ , i = 1, \ldots, p\ \}\ .$$

   Quel est le rang de la matrice de contrôlabilité M ?
   Montrer que le système n'est pas normal.

   ---

2. On considère le système dont les équations d'état sont :

   $$\frac{dx}{dt} = u\ \ ,\ \ x_0 = -1\ ,\ \ n = p = 1\ ,$$

   où

   $$u \in \mathcal{U} = \{\ u \text{ mesurable sur } \mathbb{R}\ \mid u(s) \in U \text{ presque partout }\}$$

et

$$U := \{\ u \in \mathbb{R}^p\ \mid\ |u_i| \leq 1\ , 1 \cdots p\ \}\ .$$

Montrer qu'il y a un contrôle extrêmal, qui n'est pas optimal.

---

3. Avec $n = p = 2$, on considère : $\dfrac{dx}{dt} = u\ \ ,\ \ x_0 = (-1, 0)$. où

$$u \in \mathcal{U} = \{\ u \text{ mesurable sur } \mathbb{R}\ \mid u(s) \in U \text{ presque partout }\}$$

et

$$U := \{\ u \in \mathbb{R}^p\ \mid\ |u_i| \leq 1\ , 1 \cdots p\ \}\ .$$

   (a) Montrer que $\mathcal{A}(x_0, t)$ est toujours un carré.

   (b) Montrer que le contrôle : $u_1(t) \equiv 1\ ,\ u_2(t) = \varphi(t)$, où φ est une fonction telle que $|\varphi(t)| \leq 1$ pour tout t et $\displaystyle\int_0^1 \varphi(s)\ ds = 0$, est optimal.

Il y a donc une infinité de contrôles bang-bang. Comment expliquer cela en fonction des théorèmes connus ?

---

4. **Traversée d'un bateau**

Un bateau quitte le point $P_0(x_0, y_0)$ pour aller au point $P_1(0, 0)$ (on ajuste les repères pour cela). La rivière à traverser a un courant de vitesse $\vec{c} = c\vec{\imath}\,(c \geq 0)$. Le bateau traverse à la vitesse $\vec{v} = v(\cos\gamma\vec{\imath} + \sin\gamma\vec{\jmath})$ avec $v = \|\vec{v}\| \leq 1$.

Le contrôle du bateau s'exerce sur sa vitesse, par le choix de la direction. Soit donc u le contrôle, fonction mesurable sur tout intervalle $[a, b]$ de $\mathbb{R}$ définie par :

$$u(t) = v(t) \left[\begin{array}{c} \cos\gamma(t) \\ \sin\gamma(t) \end{array}\right] \quad \text{où } 0 \leq v(t) \leq 1.$$

La position $(x(t), y(t))$ du bateau à l'instant t est donnée par le système différentiel suivant :

$$\left\{\begin{array}{rcl} \dfrac{dx}{dt}(t) & = & c + v(t)\sin\gamma(t) \\ \dfrac{dy}{dt}(t) & = & v(t)\cos\gamma(t) \end{array}\right. \qquad \gamma \in [0, 2\pi]\ .$$

   (a) Ecrire le système décrivant l'état sous la forme

$$\frac{dX}{dt} = AX(t) + Bu(t) + F\ . \tag{1}$$

   Préciser les vecteurs X et F et les matrices A et B. Décrire l'ensemble des contrôles $\mathcal{U}$.

(b) Peut-on toujours atteindre le point P_1, en partant d'un point P_0 quelconque de la rive opposée ?
Décrire l'ensemble $\mathcal{C}$ des points P_0 contrôlables dans le cas où la vitesse du courant est inférieure à 1.
Dans ce qui suit on suppose que $c = 0$.

(c) Pour un point P_0 fixé dans $\mathcal{C}$, existe-t'il un contrôle permettant d'atteindre P_1 en un temps minimum ? Si oui, ce contrôle est-il unique ?

(d) Ecrire le principe du minimum vérifié par un couple optimal $(\bar{X}, \bar{u})$. Montrer que l'état adjoint est constant.

(e) Calculer explicitement le contrôle optimal u^* et l'état correspondant dans le cas où la vitesse v est supposée constante et $x_0 = -1$ et $y_0 = -1$.

---

5. **Problème du pendule linéarisé**

Le déplacement angulaire (compté à partir de la verticale) $\theta(t)$ d'un pendule libre vérifie l'équation différentielle ordinaire suivante : $\dfrac{d^2\theta}{dt^2}(t) + \sin\theta(t) = 0$. Quand θ est petit on linéarise autour de 0 (avec $\sin\theta \simeq \theta$) et on a :

$$\frac{d^2\theta}{dt^2}(t) + \theta(t) = 0 \, .$$

On veut contrôler le pendule pour l'amener au repos, à la verticale en un temps fini, en appliquant un contrôle u qui sera une force de "freinage". Le système est alors décrit par

$$\left\{ \begin{array}{rcl} \dfrac{d^2\theta}{dt^2}(t) + \theta(t) & = & u(t) \, , \\ \theta(0) & = & \theta_0 \, , \\ \omega(0) & = & \omega_0 \, , \text{ avec } \omega = \dfrac{d\theta}{dt} \, , \end{array} \right. \tag{2}$$

où $u \in \mathcal{U} = \{ \, u \text{ mesurable } | \, |u(t)| \leq 1 \text{ p.p. } \}$.

(a) Ecrire le système décrivant l'état sous la forme

$$\frac{dx}{dt} = Ax(t) + Bu(t) \, , x(0) = x_0 \, , \quad \text{où } x = [\theta, \omega]^t \quad . \tag{3}$$

Préciser les matrices A et B. Le système est-il contrôlable ?
Décrire l'ensemble $\mathcal{C}$ des points x_0 contrôlables.

(b) Pour x_0 fixé dans $\mathcal{C}$, peut-on trouver un contrôle permettant d'amener le pendule au repos à la verticale en un temps minimum ? Si oui, ce contrôle est-il unique ?

(c) Ecrire le principe du minimum vérifié par un couple optimal $(\bar{x}, \bar{u})$. Est-ce une condition suffisante ?

(d) Montrer que l'état adjoint associé à l'état optimal $\bar{x}$ et au contrôle optimal $\bar{u}$ est de la forme :

$$\bar{p}(t) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \begin{bmatrix} p_0^1 \\ p_0^2 \end{bmatrix} .$$

On pose $p_0^1 = \rho \cos\delta$ et $p_0^2 = \rho \sin\delta$ où $\rho > 0$. Montrer que :

$$\bar{u}(t) = \text{signe}[\sin(t-\delta)] \, , p.p.$$

(e) Montrer que les trajectoires correspondant à $u \equiv 1$ (resp. $u \equiv -1$) sont des cercles de centre (1,0) (resp. (-1,0)).

Donner un exemple de point de départ x_0 d'où on peut atteindre 0 sans changer de trajectoire.

---

6. **Exemple du Rendez-vous spatial.-** On considère l'exemple de la sous-section 4.3.2 du chapitre 4 et on prend en compte la limite de puissance des moteurs en introduisant la contrainte :

$$\forall s \in \mathbb{R} \qquad u(s) \in U \subset \mathbb{R}^p \ , \ U \text{ compact} \ .$$

(Par exemple, s'il y a trois moteurs indépendants, $|u_i(s)| \leq 1$).

On cherche à atteindre $y(T)$ en temps minimum T.

Etudier l'existence et l'unicité du problème. Ecrire les (in)équations du principe du minimum de Pontryagin.

---

7. On considère le système suivant :

$$\frac{dx}{dt} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} u(t) \, , x(0) = x_0 \, ,$$

où $u \in \mathcal{U} = \{u \text{ mesurable} \, , u(t) \in \mathbb{R}^2 \text{ et } |u_i| \leq 1, \, i = 1, 2 \text{ p.p.}\}$ et l'ensemble cible est $\{0\}$.

(a) Montrer que le problème n'est pas normal.

(b) Etudier la contrôlabilité du système.

(c) On prend $x_0 = (-1, -1)$. Décrire $\mathcal{A}(t; x_0)$.

(d) Montrer qu'il y a unicité du contrôle optimal amenant l'état de x_0 en 0 en temps optimal.

On peut donc avoir unicité du contrôle même si le problème n'est pas normal.

---

8. Etudier complètement le problème de contrôle en temps optimal pour :

$$\frac{dx}{dt} = b \, x(t) + u(t) \quad , \quad b < 0 \ , \ n = p = 1 \, ,$$

où

$$u \in \mathcal{U} := \{u \text{ mesurable et } u(t) \in [-1,1] \text{ p.p.}t \} ,$$

et l'ensemble cible est $\{0\}$.

---

9. On considère le système suivant (n=p=2) :

$$\frac{dx}{dt} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} u(t) \ , \ x_0 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} ,$$

où

$$u \in \mathcal{U} := \{u \text{ mesurable }, u(t) \in \mathbb{R}^2 \text{ et } |u_i| \leq 1 \, , i = 1,2 \text{ p.p.}t \} ,$$

et l'ensemble cible est $\{0\}$. Montrer que $T = 1$ est optimal. Montrer qu'il y a une infinité de contrôles optimaux mais que l'état est unique.

---

10. On se propose de faire l'étude complète du contrôle en temps optimal du système dynamique décrit par l'équation d'état suivante :

$$p''(t) + 3p'(t) + 2p(t) = u(t) \text{ sur }]0, +\infty[, p(0) = p_{0,1} \, , p'(0) = p_{0,2} \, , \tag{4}$$

où p est une fonction deux fois dérivable de $\mathbb{R}^+$ vers $\mathbb{R}$, p' désigne la dérivée première de p et p'' sa dérivée seconde ; u est une fonction de contrôle continue par morceaux de $\mathbb{R}^+$ vers $\mathbb{R}$ vérifiant

$$\forall t \geq 0 \qquad |u(t)| \leq 1;$$

p_0 et q_0 sont des réels.
On veut atteindre la position $p(t) = p'(t) = 0$, en un temps minimal t^*.

(a) Ecrire l'équation différentielle du second ordre (4) sous la forme d'un système différentiel

$$z'(t) = Az(t) + Bu(t) \ , \ z(0) = z_0 \ , \tag{5}$$

où $z(t) = [p(t), p'(t)]^t$. On précisera les matrices A et B. Le système est-il contrôlable ?

(b) On pose

$$Q = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix} .$$

Montrer que la matrice $D = Q^{-1}AQ$ est diagonale. On pose alors $y = Q^{-1}z$. Ecrire le système différentiel vérifié par y.

(c) Peut-on atteindre le point $O = (0,0)$ en un temps minimal ? Si oui, que peut-on dire du contrôle optimal que l'on notera u^* ?

(d) On se place désormais en variable y. Ecrire le principe du minimum de Pontryagin. Quelle est la relation entre le contrôle optimal et le signe de l'état adjoint q ? Calculer l'état adjoint en fonction de sa valeur en $t = 0$ notée $(q_{0,1}, q_{0,2})$. Montrer que u^* change de signe au plus une fois.

(e) Résoudre l'équation d'état pour $u(t) \equiv 1$ et $u(t) \equiv -1$; montrer que les trajectoires sont des paraboles. Dans chaque cas, donner la relation que doivent vérifier $y_{0,1}$ et $y_{0,2}$ pour que les trajectoires passent par O. Cette relation donne l'équation cartésienne de deux courbes $\mathcal{P}^+$ (pour $u(t) \equiv 1$) et $\mathcal{P}^-$ (pour $u(t) \equiv -1$) ; ce sont les courbes de commutation. Dessiner ces deux courbes.

(f) Que se passe-t'il si le point de départ y_0 ne se situe pas sur $\mathcal{P}^+ \cup \mathcal{P}^-$?

Chapitre 8

Programmation dynamique

8.1 Cas discret

8.1.1 Motivation

Jusqu'ici nous avons considéré des systèmes "continus" : l'inconnue est une fonction (continue) d'un intervalle de $\mathbb{R}$ dans $\mathbb{R}^n$. Dans le cas du contrôle optimal à coût quadratique on a exhibé un système d'optimalité "continu" que l'on peut résoudre ensuite par différentes méthodes. On peut faire appel à des logiciel de calcul formel (type MAPLE©) permettant de résoudre analytiquement un système différentiel linéaire, ou bien discrétiser le système et le résoudre en utilisant des logiciels de calcul scientifique (type MATLAB© ou SCILAB©).

On peut aussi discrétiser le problème de contrôle optimal dès le début pour en faire un problème de contrôle discret, l'équation différentielle devenant un système aux différences. Précisons ce point de vue. Lorsque l'équation différentielle ordinaire est très générale (non nécessairement linéaire) le problème de contrôle optimal s'écrit

$$\min\ \tilde{J}(v) = \tilde{J}(x_v, v)\ ,\ v \in \mathcal{U}_{ad}\ , \tag{8.1.1}$$

où $x_v : [t_0, T] \to \mathbb{R}^n$ est (la) solution de

$$\frac{dx}{dt}(t) = \varphi(x(t), v(t), t)\ ,\ t \in]\, t_0, T[\ ,\ x(t_0) = x_0\ ; \tag{8.1.2}$$

- $x_0 \in \mathbb{R}^n$ est donné,
- $\varphi : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^n$ est une fonction suffisamment régulière (voir les conditions d'application du théorème de CAUCHY-LIPSCHITZ dans l'annexe A. par exemple) pour que l'équation (8.1.2) admette une solution unique,
- $\mathcal{U}_{ad}$ est un sous-ensemble fermé , convexe de $L^2(0, T; \mathbb{R}^p)$ de la forme

$$\mathcal{U}_{ad} = \{v \in L^2(0, T; \mathbb{R}^p)\ \mid\ v(t) \in U\ \text{ p.p. } t \in]\, t_0, T[\ \}, \tag{8.1.3}$$

  où U est un sous-ensemble fermé convexe de $\mathbb{R}^p$.

On discrétise alors (8.1.2) par une méthode classique (par exemple la méthode d'Euler explicite [7, 8]) ; on obtient un **système aux différences** de la forme :

$$\left\{ \begin{array}{lcl} x(t_{k+1}) & = & f(x(t_k), v(t_k), t_k)\ ,\ k = 0, \ldots, N-1 \\ x(t_0) & = & x_0 \text{ donné,} \end{array} \right. \tag{8.1.4}$$

où on a posé pour $N \in \mathbb{N}^*,\ t_k = t_0 + k\dfrac{T - t_0}{N}$.
Supposons que la fonction coût $\tilde{J}$ soit donnée par

$$\tilde{J}(v) = \tilde{F}(x_v(T), T) + \int_{t_0}^{T} \tilde{L}(x_v(t), v(t), t)\ dt\ ,$$

avec $\tilde{F},\ \tilde{L} \geq 0$. Là encore, des formules de quadrature classiques permettent de donner une approximation de $\tilde{J}$ qui peut s'écrire par exemple :

$$\tilde{J}(v) \simeq J(v) \stackrel{\text{def}}{=} F(x_v(T), T) + \sum_{k=0}^{N-1} L(x_v(t_k), v(t_k), t_k)\ .$$

En définitive le problème de contrôle discret s'écrit maintenant

$$(\mathcal{P}) \qquad \left\{ \begin{array}{l} \min\ F(x_v(T), T) + \displaystyle\sum_{k=0}^{N-1} L(x_v(t_k), v(t_k), t_k) \\ v \in U_{ad} \\ \\ x_v(t_{k+1}) = f(x_v(t_k), v(t_k), t_k)\ ,\ k = 0, \ldots, N-1 \\ x_v(t_0) = x_0\ . \end{array} \right.$$

Remarque 8.1.1 *Les suites (t_k), $x_v(t_k)$ et $v(t_k)$ sont des suites finies ($k = 0, \ldots, N$); nous préférons garder la notation $x_v(t_k)$ au lieu de $x_{v,k}$ (par exemple) qui semblerait plus standard, mais qui ne reflète pas très bien la dépendance de x_v par rapport à t_k.*

Avant d'étudier le problème $(\mathcal{P})$, nous allons préciser sur un exemple simple le processus de discrétisation et d'approximation du problème continu.

Exemple

Considérons le problème de contrôle à coût quadratique déjà étudié dans le chapitre 5 :

$$\left\{ \begin{array}{l} \min\ \tilde{J}(v) = \dfrac{1}{2}\displaystyle\int_0^T [x_v(t) - z_d(t)]^2\ dt + \frac{\alpha}{2}\int_0^T v(t)^2\ dt + \frac{1}{2}[x_v(T) - z_d(T)]^2 \\ v \in \mathcal{U}_{ad}\ , \\ \\ \dfrac{dx_v}{dt}(t) = \varphi(x_v(t), v(t), t)\ ,\ t \in]0, T[\\ x_v(0) = x_0\ , \end{array} \right.$$

où x_v, v, z_d sont a priori des fonctions de carré intégrable de $[\,0,T]$ dans $\mathbb{R}$ qu'on choisira pour cet exemple continues et $\mathcal{U}_{ad}$ est un sous-ensemble fermé , convexe de $L^2(0,T)$ de la forme

$$\mathcal{U}_{ad} = \{v \in L^2(0,T) \ \mid \ v(t) \in U \ \text{ p.p. } t \in]\,0,T[\,\},$$

et U est un sous-ensemble fermé convexe de $\mathbb{R}$.
On considère une subdivision de $[\,0,T]:\ \{\,t_k = \dfrac{kT}{N}\ , k = 0,\ldots,N\}$ de pas $\Delta t = \dfrac{T}{N}$.

– Commençons par discrétiser la fonction coût. On peut par exemple utiliser la formule des trapèzes pour calculer l'intégrale d'une fonction ψ sur $[\,t_k, t_{k+1}]$; plus précisément

$$\int_0^T \ \psi(t)\ dt = \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \ \psi(t)\ dt$$

$$\simeq \frac{\Delta t}{2} \sum_{k=0}^{N-1} [\psi(t_{k+1} + \psi(t_k)] = \frac{T}{N}\left[\frac{\psi(0)}{2} + \frac{\psi(T)}{2} + \sum_{k=1}^{N-1} \psi(t_k)\right]\ .$$

La fonction coût discrétisée est alors

$$\begin{aligned} J_N(x,v) &= \frac{1}{2}\left[\,[x(T) - z_d(T)]^2 + \frac{T}{2N}\,[x(0) - z_d(0)]^2 + \frac{T}{2N}\,[x(T) - z_d(T)]^2\right] \\ &+ \frac{T}{2N}\left[\sum_{k=1}^{N-1}[x(t_k) - z_d(t_k)]^2\right] \\ &+ \frac{\alpha}{2}\left[\frac{T}{2N}\,[v(0)^2 + v(T)^2] + \frac{T}{N}\ \sum_{k=1}^{N-1} v(t_k)^2\right]. \end{aligned}$$

Finalement nous obtenons

$$J_N(x,v) = F(x(T),T) + \sum_{k=1}^{N-1} L(x(t_k), v(t_k), t_k) + C\ ,$$

avec

$$F(x(T),T) = \frac{1}{2}\,(1 + \frac{T}{N})[x(T) - z_d(T)]^2 + \frac{\alpha\, T}{4N}\, v(T)^2\ ,$$

$$L(x(t_k), v(t_k), t_k) = \frac{T}{2N}\ [x(t_k) - z_d(t_k)]^2 + \frac{\alpha\, T}{2N}\, v(t_k)^2\ , k = 1,\ldots,N-1$$

$$L(x(0), v(0), 0) = \frac{T}{4N}\ \left[[x(0) - z_d(0)]^2 + \alpha\, v(0)^2\right]\ .$$

– L'équation différentielle se discrétise grâce à la méthode d'Euler explicite par exemple. Cela donne

$$\begin{cases} x_v(0) = x_0 \\ x_v(t_{k+1}) = x_v(t_k) + \dfrac{T}{N}\,\varphi(x_v(t_k), v(t_k), t_k)\ ,\ k = 0,\ldots,N-1\ . \end{cases}$$

Dans ce cas $f(x_v(t_k), v(t_k), t_k) = x_v(t_k) + \dfrac{T}{N}\varphi(x_v(t_k), v(t_k), t_k)$.

8.1.2 Principe d'optimalité de Bellman

On considère donc désormais le problème discret

$$(\mathcal{P}) \qquad \begin{cases} \min\ F(x_v(T),T) + \displaystyle\sum_{k=0}^{N-1} L(x_v(t_k),v(t_k),t_k) \\ v \in U_{ad} \\ \\ x_v(t_{k+1}) = f(x_v(t_k),v(t_k),t_k)\ ,\ k=0,\dots,N-1 \\ x_v(t_0) = x_0 \end{cases}$$

Dans tout ce qui suit on adopte la notation suivante

$$J_K(\mathbf{x},\mathbf{v}) = \sum_{k=K}^{N-1} L(\mathbf{x}(t_k),\mathbf{v}(t_k),t_k) + F(\mathbf{x}(T),T)\ , \tag{8.1.5}$$

où $0 \leq K \leq N-1$ et $\mathbf{x}$ est solution de

$$\begin{cases} \mathbf{x}(t_{k+1}) = f(\mathbf{x}(t_k),\mathbf{v}(t_k),t_k)\ ,\ k=K,\dots,N-1 \\ \mathbf{x}(t_K) = x\ . \end{cases}$$

En d'autres termes $\mathbf{x}$ est l'unique trajectoire associée à $\mathbf{v}$ et passant par x à l'instant t_K.
Résoudre le problème $(\mathcal{P})$ revient à minimiser $J_0(x_0,\mathbf{v})$ pour $\mathbf{v} \in U_{ad}$. Rappelons que l'ensemble U_{ad} des contrôles admissibles est de la forme

$$U_{ad} = \{\ \mathbf{v} : \{\ t_k\ \}_{0\leq k\leq N-1} \to (\mathbb{R}^p)^N\ \mid\ \mathbf{v} \text{ est à valeurs dans } U \subset \mathbb{R}^p\ \}\ .$$

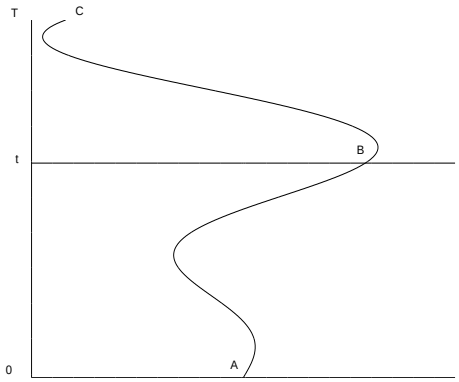

Figure 8.1

Le principe d'optimalité que nous allons établir est fondé sur la remarque suivante :
Le point A correspond à l'instant $t=0$ (coordonnées $(0,x_0)$), le point C à l'instant $t=T$ (coordonnées $(T,x(T))$) et le point B à l'instant $t=t_K$ (coordonnées $(t_K,x(t_K))$. Le coût pour aller de A à C est la somme des coûts pour aller de A à B puis de B à C. En effet

$$J_0(x,v) = J_{AB} + J_{BC} = \sum_{k=0}^{K-1} L(\mathbf{x}(t_k),\mathbf{v}(t_k),t_k) + \sum_{k=K}^{N-1} L(\mathbf{x}(t_k),\mathbf{v}(t_k),t_k) + F(\mathbf{x}(T),T)$$

$$= \sum_{k=0}^{K-1} L(\mathbf{x}(t_k), \mathbf{v}(t_k), t_k) + J_K(\mathbf{x}(t_K), \mathbf{v}) \ .$$

Le minimum pour aller de A à C est égal au minimum pour aller de A à B plus le minimum pour aller de B à C. En effet, s'il y avait une loi moins coûteuse sur BC on améliorerait (sans bouger J_{AB}) le coût global.
La politique optimale à une phase donnée B (c'est-à-dire commençant à l'instant t_K) ne dépend que de cette phase et non pas de la trajectoire antérieure : c'est le **principe d'optimalité de Bellman.**

Pour résoudre notre problème, on ne va pas chercher le contrôle optimal $\mathbf{v}(x_0, t_k)$ directement mais scinder le problème en plusieurs étapes. Plus généralement, on va chercher $\mathbf{v}(x, t_k)$ où x sera la valeur de l'état ($\mathbf{x}_\mathbf{v}(t_k) = x$) aux différentes étapes intermédiaires.

8.1.3 L'algorithme de programmation dynamique

L'équation de Bellman

Soit $(x, t_K) \in \mathbb{R}^n \times [t_0, T]$. On va considérer le problème aux différences qui possède (x, t_K) comme phase initiale :

$$(\mathcal{P}_K(x, \mathbf{v})) \qquad \left\{ \begin{array}{rcl} \mathbf{x}(t_{k+1}) & = & f(\mathbf{x}(t_k), \mathbf{v}(t_k), t_k) \quad K \leq k \leq N-1 \ , \\ \mathbf{x}(t_K) & = & x \ . \end{array} \right.$$

On suppose que ce problème a une solution et on note $\mathcal{V}_K(x)$ la valeur optimale du coût $J_K(\mathbf{x}, \mathbf{v})$ pour le problème suivant :

$$\begin{array}{rcl} \mathcal{V}_K(x) & = & \min J_K(\mathbf{x}, \mathbf{v}) \\ & = & \min\limits_{\mathbf{v}_K \in U_{ad,K}} \sum\limits_{k=K}^{N-1} L(\mathbf{x}(t_k), \mathbf{v}(t_k), t_k) + F(\mathbf{x}(T), T) \end{array}$$

où $\mathbf{x}(.)$ est solution de $(\mathcal{P}_K(x, \mathbf{v}))$, $\mathbf{v}_K = (\mathbf{v}(t_k),\ K \leq k \leq N-1)$ et

$$U_{ad,K} = \{\ \mathbf{v}_K \ \mid \ \mathbf{v} \in U_{ad}\ \} = \{\ \mathbf{v}_K(t_k),\ K \leq k \leq N-1 \ \mid \ \mathbf{v} \in U_{ad}\ \} \ .$$

Supposons $\mathcal{V}_{K+1}(y)$ connue pour tout y de $\mathbb{R}^n$: on peut alors calculer la valeur de $\mathcal{V}_K(y)$.
On se donne x dans $\mathbb{R}^n$ et on part de (x, t_K) pour obtenir la solution $\mathbf{x}(.)$ de $(\mathcal{P}_K(x, \mathbf{v}))$. A l'instant suivant, t_{K+1} l'état du système vaut $\mathbf{x}(t_{K+1})$: on sait alors trouver la trajectoire optimale entre t_{K+1} et T puisqu'on a supposé qu'on connaissait $\mathcal{V}_{K+1}(.)$. Plus précisément le coût optimal entre t_{K+1} et T est

$$\mathcal{V}_{K+1}(\mathbf{x}(t_{K+1})) = \mathcal{V}_{K+1}(f(\mathbf{x}(t_K), \mathbf{v}(t_K), t_K)) = \mathcal{V}_{K+1}(f(x, \mathbf{v}(t_K), t_K)) \ .$$

De plus

$$J_K(\mathbf{x}, \mathbf{v}) = L(\mathbf{x}(t_K), \mathbf{v}(t_K), t_K) + J_{K+1}(\mathbf{x}, \mathbf{v}) \ .$$

On sait optimiser le terme $J_{K+1}(\mathbf{x}, \mathbf{v})$: il vaut $\mathcal{V}_{K+1}(\mathbf{x}(t_{K+1}))$. Par conséquent

$$J_K(\mathbf{x}, \mathbf{v}) = \underbrace{L(x, \mathbf{v}(t_K), t_K)}_{\text{point de départ}} + \mathcal{V}_{K+1}(f(x, \mathbf{v}(t_K), t_K)) \ .$$

On minimise sur la valeur $\mathbf{v}(t_K) = u \in U \subset \mathbb{R}^p$. Finalement

$$\mathcal{V}_K(x) = \min_{u \in U} \left[L(x, u, t_K) + \mathcal{V}_{K+1}(f(x, u, t_K)) \right] \ . \tag{8.1.6}$$

C'est l'équation de **Bellman**.

L'algorithme de programmation dynamique

Cette équation permet de calculer $\mathcal{V}_k$ et $\mathbf{v}$ de proche en proche en partant **de l'ensemble d'arrivée** $\mathcal{C}$ et en "remontant le temps". En effet, à l'arrivée $K = N$ et J_K vaut :

$$J_N(\mathbf{x}, \mathbf{v}) = \underbrace{F(\mathbf{x}(T), T)}_{\text{arrivée = départ}} \ .$$

Par conséquent, pour tout $x \in \mathcal{C}$:

$$\mathcal{V}_N(x) = F(x, t_N) = F(x, T) \ . \tag{8.1.7}$$

On obtient ainsi l'algorithme de **Programmation Dynamique** :

Algorithme de Programmation Dynamique

1. **Initialisation**
   $k = N$: Résolution de l'équation (8.1.7) pour tous les points de l'ensemble cible $\mathcal{C}$.
2. **Itération** k

$$\mathcal{V}_{k-1}(x) = \min_{u \in U} \left[L(x, u, t_{k-1}) + \mathcal{V}_k(f(x, u, t_{k-1})) \right] \ .$$

3. $k = k - 1$ et on retourne à 2.

L'équation (8.1.6) avec la condition "initiale" (8.1.7) est une condition suffisante pour que $\mathcal{V}$ soit le coût optimal mais la démonstration exige que $\mathcal{V}$ existe pour tout $x \in \mathbb{R}^n$. En fait $\mathcal{V}$ doit exister pour toutes les phases accessibles (atteignables) depuis l'ensemble d'arrivée.

8.1.4 Exemples

Course au trésor

Etant donné 5 villes A, B, C, D et E possédant chacune un trésor (10, 5, 7, 4,3) et compte tenu du réseau de transport et des coûts, on pose la question du chemin optimal (i.e. permettant de gagner le plus d'argent).

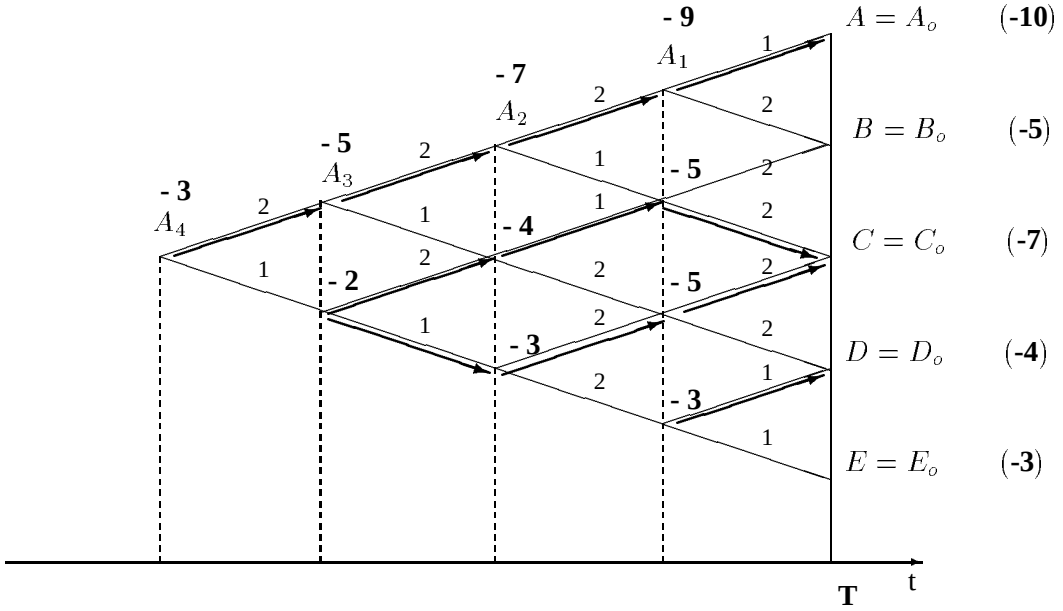

Figure 8.2 . Course au trésor

On rappelle que maximiser le gain revient à minimiser la perte ou le coût. L'étape d'initialisation revient à calculer $\mathcal{V}(T)$. Ici

$$\mathcal{V}(x,T) \in \begin{cases} -10 \\ -5 \\ -7 \\ -4 \\ -3 \end{cases}$$

A l'étape $T-1$, il y a 4 villes antécédentes possibles : $A_1,\ B_1,\ C_1$ et D_1 ;

$$\mathcal{V}_{T-1}(A_1) = \min_u [\ \underbrace{L(A_1,u,T-1)}_{\text{coût}} + \underbrace{\mathcal{V}_T(f(A_1,u))}_{\text{ville atteinte en partant de } A_1 \text{ avec le chemin } u}\ ].$$

Chemin pour aller de A_1 à A ou à B

Par exemple :

$$A_1 : \begin{cases} -10+1 & \rightarrow & -9\ (A) \\ -9+2 & \rightarrow & -7\ (B) \end{cases} \ : \text{ chemin du haut } A_1 \text{ vers } A \text{ correspondant à } \mathcal{V} = -9.$$

De même

$$B_1 : \begin{cases} -5+2 & \rightarrow & -3\ (B) \\ -7+2 & \rightarrow & -5\ (C) \end{cases} \ : \text{ chemin de } B_1 \text{ vers } C \text{ correspondant à } \mathcal{V} = -5.$$

Les chemins optimaux sont indiqués par des flèches sur la figure 8.2.

Affectation optimale de ressources

On considère le problème d'affectation de ressources suivant : à chaque instant $t_i = i\Delta t,\ i = 1, 2, \ldots$ on dispose des ressources $y_i \in \mathbb{R}^+$; on peut affecter ces ressources dans la période (t_i, t_{i+1}) entre deux usages :
(a) rapporte le gain 2 par unité de ressource affectée mais il y a disparition de 20 % de ces ressources (amortissement).
(b) rapporte plus : le gain est de 3 par unité mais l'amortissement est de 50 % .

La décision à prendre à l'instant t_i est donc la quantité de ressources $u_i \in [0, y_i]$ à affecter à l'usage (a) ($y_i - u_i$ étant affecté à l'usage (b)). L'équation d'état est :

$$y_{i+1} = 0.8\, u_i + 0.5\, (y_i - u_i) = 0.5\, y_i + 0.3\, u_i = f(y_i, u_i, t_i)\ .$$

La quantité de ressources à l'instant initial est ξ. Le critère à minimiser pour T étapes est

$$J(u, y) = J_1(u_1, y_1) + \ldots + J_N(u_N, y_N)$$

où

$$y = (y_i)_{1\le i\le N},\ u = (u_i)_{1\le i\le N},\ J_i(u_i, y_i) = -[2\, u_i + 3(y_i - u_i)] = u_i - 3\, y_i\ .$$

Ecrivons les équations de la programmation dynamique pour les différentes étapes :

1. **Instant** $T = t_N = N\Delta t$ **:**

$$\mathcal{V}_N(y) = \min_{0\le u\le y} u - 3\, y = -3\, y \ \text{ et } u_N(y) = 0\ .$$

2. **Instant** t_{N-1} **:**

$$\begin{aligned}
\mathcal{V}_{N-1}(y) &= \min_{0\le u\le y} J_{N-1}(u, y) + \mathcal{V}_N(f(y, u, t_{N-1})) \\
&= \min_{0\le u\le y} u - 3\, y - 3(0.5\, y + 0.3\, u) \\
&= \min_{0\le u\le y} 0.1\, u - 4.5\, y \\
&= -4.5\, y \\
u_{N-1}(y) &= 0\ .
\end{aligned}$$

3. **Instant** t_{N-2} **:**

$$\begin{aligned}
\mathcal{V}_{N-2}(y) &= \min_{0\le u\le y} (u - 3\, y) - 4.5\, (0.5y + 0.3\, u) \\
&= \min_{0\le u\le y} -0.35\, u - 5.25\, y \\
&= -\max_{0\le u\le y} 0.35\, u + 5.25\, y \\
&= -5.6\, y \\
u_{N-2}(y) &= y\ .
\end{aligned}$$

4. **Instant t_{N-3} :**

$$\begin{aligned}\mathcal{V}_{N-3}(y) &= \min_{0\le u\le y}(u-3\,y)-5.6\,(0.5y+0.3\,u)\\ &= \min_{0\le u\le y} -0.68\,u-5.8\,y\\ &= -6.48\,y\\ u_{N-3}(y) &= y\ .\end{aligned}$$

On continue ainsi jusqu'au nombre d'étapes désiré. Par exemple, si on se donne un problème à 4 étapes, la solution sera :

$$u_4=u_3=0\ ,\ u_2=y_2,\ u_1=y_1,\ y_0=\xi\ .$$

La valeur minimale du critère est $-6.48\,\xi$. Le gain maximum en 4 étapes est donc $6.48\,\xi$.

8.2 Programmation dynamique en dimension infinie

L'examen du cas "discret" en dimension finie nous a permis de présenter les idées de base de la programmation dynamique. On peut bien sûr étendre le principe de la programmation dynamique à des problèmes non discrets, c'est-à-dire posés dans des espaces de dimension infinie. Nous présentons dans cette section quelques résultats dans ce cadre (issus de [16]). La théorie est vaste et compliquée. Pour une étude approfondie on peut se référer par exemple à [3, 17].

8.2.1 Présentation du problème

On considère à présent un système dont l'état $\mathbf{x}$ est décrit par l'équation différentielle (a priori non linéaire) (8.1.2) que nous rappelons

$$\frac{d\mathbf{x}}{dt}(t)=\varphi(\mathbf{x}(t),v(t),t)\ ,\ t\in]\,t_0,T[\ ,\ \mathbf{x}(t_0)=x_0\ .$$

On a fait des hypothèses générales au début de la section 8.1.1 pour que cette équation admette une solution unique $\mathbf{x}[v,t_0,x_0](\cdot)$. On considère également la fonction coût

$$\tilde{J}(\mathbf{x},v)=F(\mathbf{x}[v,t_0,x_0](T))+\int_{t_0}^{T}L(\mathbf{x}[v,t_0,x_0](t),v(t),t)\ dt\ ;$$

On s'intéresse au problème de contrôle optimal (8.1.1) :

$$\min\{J_{x_0,t_0}(v)\stackrel{\text{def}}{=}\tilde{J}(\mathbf{x}[v,t_0,x_0],v),\ v\in\mathcal{U}_{ad}\ \}.$$

L'ensemble des contrôles admissibles est donné par (8.1.3). Pour assurer l'existence et l'unicité d'une solution de l'équation d'état et du problème de contrôle nous faisons désormais les hypothèses suivantes :

– φ et L sont uniformément continues sur $K\times U\times I$ pour toutes parties bornées K de $\mathbb{R}^n$ et I de $\mathbb{R}$.

- φ est dérivable par rapport à x sur $\mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}$ et sa dérivée partielle par rapport à x est continue et bornée sur $\mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}$.
- $\exists \kappa \geq 0,\ \|\varphi(x,u,t)\|_n \leq \kappa(1 + \|u\|_p + \|x\|_n) \qquad \forall (x,u,t) \in \mathbb{R}^n \times U \times \mathbb{R}$.
- $\exists \nu > 0,\ \exists \kappa \in \mathbb{R},\ L(x,u,t) \geq \nu \|u\|_p^2 + \kappa \qquad \forall (x,u,t) \in \mathbb{R}^n \times U \times \mathbb{R}$ (coercivité de L).
- F est continue et minorée sur $\mathbb{R}^n$

Ces hypothèses assurent l'existence de la fonction valeur pour le problème (8.1.1)

$$\mathcal{V}(x_0,t_0) \stackrel{\text{def}}{=} \inf\{J_{x_0,t_0}(v),\ v \in \mathcal{U}_{ad}\}. \tag{8.2.1}$$

Lorsque $t_0 = T$, $\mathcal{V}$ est donnée par

$$\mathcal{V}(x,T) = F(x) \qquad \forall x \in \mathbb{R}^n\ . \tag{8.2.2}$$

Principe d'optimalité de Bellman

Tout comme dans le cas où la fonction valeur vérifiait l'équation de Bellman, nous allons montrer que $\mathcal{V}$ vérifie aussi une équation. C'est une équation aux dérivées partielles non linéaire du premier ordre, appelée équation de Hamilton-Jacobi-Bellman (HJB). Pour cela, nous commençons par montrer le principe d'optimalité de Bellman, analogue au principe discret énoncé dans la section précédente.

Théorème 8.2.1 *On suppose les hypothèses précédentes vérifiées.*
Soit $(x,t) \in \mathbb{R}^n \times]-\infty, T[$ *et* $\tau \in [t,T]$. *Alors, nous avons*

$$\mathcal{V}(x,t) = \inf_{u \in \mathcal{U}_{ad}} \left(\int_t^\tau L(\mathbf{x}[x,t,u](s), u(s), s)\, ds + \mathcal{V}(\mathbf{x}[x,t,u](\tau), \tau) \right)\ . \tag{8.2.3}$$

Démonstration - On note $\mathcal{U}_{ad,1} = L^2(]\,t,\tau\,[,U)$ et $\mathcal{U}_{ad,2} = L^2(]\,\tau,T\,[,U)$. Pour tout contrôle $u \in \mathcal{U}_{ad}$ on pose $u_1 = u_{|]\,t,\tau\,[} \in \mathcal{U}_{ad,1}$ et $u_2 = u_{|]\,\tau,T\,[} \in \mathcal{U}_{ad,2}$. L'état correspondant à u_1 partant de x au temps t est la solution de (8.1.2) sur $]t,\tau[$ avec $t_0 = t$ et $x_0 = x$. On le note $\mathbf{x}_1 = \mathbf{x}[x,t,u_1]$. Cet état coïncide bien sûr avec $\mathbf{x} = \mathbf{x}[x,t,u]$ sur l'intervalle $[t,\tau]$. De la même façon, l'état correspondant à u_2 et partant de $\mathbf{x}(\tau) = \mathbf{x_1}(\tau)$ au temps τ est donné par $\mathbf{x}_2$ et coïncide avec $\mathbf{x} = \mathbf{x}[x,t,u]$ sur l'intervalle $[\tau,T]$. Par conséquent, nous avons

$$\begin{aligned}
\mathcal{V}(x,t) &= \inf_{u \in \mathcal{U}_{ad}} \left(\int_t^T L(\mathbf{x}[x,t,u](s), u(s), s)\, ds + F(\mathbf{x}[x,t,u](T)) \right) \\
&= \inf_{(u_1,u_2) \in \mathcal{U}_{ad,1} \times \mathcal{U}_{ad,2}} \left(\int_t^\tau L(\mathbf{x_1}(s), u_1(s), s)\, ds + \int_\tau^T L(\mathbf{x_2}(s), u_2(s), s)\, ds + F(\mathbf{x_2}(T)) \right) \\
&= \inf_{u_1 \in \mathcal{U}_{ad,1}} \left(\int_t^\tau L(\mathbf{x_1}(s), u_1(s), s)\, ds + \inf_{u_2 \in \mathcal{U}_{ad,2}} \left(\int_\tau^T L(\mathbf{x_2}(s), u_2(s), s)\, ds + F(\mathbf{x_2}(T)) \right) \right) \\
&= \inf_{u_1 \in \mathcal{U}_{ad,1}} \left(\int_t^\tau L(\mathbf{x_1}(s), u_1(s), s)\, ds + \mathcal{V}(\mathbf{x_1}(\tau), \tau) \right) \\
&= \inf_{u \in \mathcal{U}_{ad}} \left(\int_t^\tau L(\mathbf{x}[x,t,u](s), u(s), s)\, ds + \mathcal{V}(\mathbf{x}[x,t,u](\tau), \tau) \right)\ .
\end{aligned}$$

□

Equation de Hamilton-Jacobi-Bellman

On peut déduire du principe d'optimalité l'équation de Hamilton-Jacobi-Bellman, aux points où la fonction valeur $\mathcal{V}$ est différentiable.

Théorème 8.2.2 *On suppose les hypothèses précédentes vérifiées. Soit* $(x_0, t_0) \in \mathbb{R}^n \times]-\infty, T[$. *Si* $\mathcal{V}$ *est différentiable en* (x_0, t_0) *alors l'équation de Hamilton-Jacobi-Bellman a lieu en* (x_0, t_0) *:*

$$\frac{\partial \mathcal{V}}{\partial t}(x_0, t_0) + \mathcal{H}(x_0, t_0, \nabla_x \mathcal{V}(x_0, t_0)) = 0 \,, \tag{8.2.4}$$

où $\mathcal{H}$ *est* ***l'Hamiltonien*** *minimisé donné par*

$$\mathcal{H}(x, t, p) = \inf_{w \in U} \left(L(x, t, w) + p \cdot \varphi(x, t, w) \right) \,, \tag{8.2.5}$$

Démonstration - On ne présente la démonstration que dans le cas (plus simple) où U est borné. Pour le cas général, on pourra se référer à [16]. Nous partons du principe d'optimalité en prenant $x_0 = x,\ t = t_0$ et $\tau = t_0 + h$ (avec $h \in]\, 0, T - t_0\, [$). On obtient

$$\mathcal{V}(x_0, t_0) = \inf_{u \in \mathcal{U}_{ad}} \left(\int_{t_0}^{t_0+h} L(\mathbf{x}(s), u(s), s)\, ds + \mathcal{V}(\mathbf{x}(t_0 + h), t_0 + h) \right) \,, \tag{8.2.6}$$

où $\mathbf{x}$ est la trajectoire (associée à u) vérifiant $\mathbf{x}(t_0) = x_0$. Comme U est borné et que nous avons supposé que

$$\exists \kappa \geq 0, \forall (x, u, t) \in \mathbb{R}^n \times U \times \mathbb{R} \qquad \|\varphi(x, u, t)\|_n \leq \kappa(1 + \|u\|_p + \|x\|_n),$$

on déduit que

$$\|\mathbf{x}'(t)\|_n \leq \kappa \left(1 + \|\mathbf{x}(t)\|_n\right)$$

et donc

$$\forall s \in [t_0, t_0 + h] \qquad \|\mathbf{x}(s) - x_0\|_n \leq \kappa \left| s - t_0 \right| , \tag{8.2.7}$$

pour toute trajectoire $\mathbf{x}$, κ désignant une constante générique indépendante de h et u. On obtient avec l'équation différentielle

$$\mathbf{x}(s) = x_0 + \int_{t_0}^{s} \varphi(\mathbf{x}(\sigma), u(\sigma), \sigma)\, d\sigma = x_0 + \int_{t_0}^{s} \varphi(\mathbf{x}(x_0), u(\sigma), t_0)\, d\sigma + o(h) \,,$$

pour tout $s \in [t_0, t_0 + h]$; ici et dans tout ce qui suit $o(h)$ est uniforme par rapport à $u \in \mathcal{U}_{ad}$. On a donc en particulier

$$\mathbf{x}(t_0 + h) = x_0 + \int_{t_0}^{t_0+h} \varphi(x_0, u(s), t_0)\, ds + o(h) \,.$$

En outre, $\mathcal{V}$ est supposée différentiable en (x_0, t_0) : en appliquant le théorème de dérivation des fonctions composées, on obtient

$$\mathcal{V}(\mathbf{x}(t_0+h), t_0+h) = \mathcal{V}(x_0, t_0) + h \frac{\partial \mathcal{V}}{\partial t}(x_0, t_0) + \nabla_x \mathcal{V}(x_0, t_0) \cdot \left(\int_{t_0}^{t_0+h} \varphi(x_0, u(s), t_0)\, ds \right) + o(h) \,.$$

En utilisant (8.2.7), la relation précédente et les hypothèses sur L, on déduit finalement de (8.2.6) que :

$$\mathcal{V}(x_0,t_0)+o(h)= \inf_{u\in\mathcal{U}_{ad}}\left(\int_{t_0}^{t_0+h} L(x_0,u(s),t_0)\,ds+\mathcal{V}(x_0,t_0)+h\,\frac{\partial\mathcal{V}}{\partial t}(x_0,t_0) \right.$$
$$\left. +\nabla_x\mathcal{V}(x_0,t_0)\cdot\left(\int_{t_0}^{t_0+h}\varphi(x_0,u(s),t_0)\,ds\right)\right)\ .$$

Par conséquent, la quantité

$$\inf_{u\in\mathcal{U}_{ad}}\left(\frac{1}{h}\int_{t_0}^{t_0+h} L(x_0,u(s),t_0)\,ds+\frac{\partial\mathcal{V}}{\partial t}(x_0,t_0)+\nabla_x\mathcal{V}(x_0,t_0)\cdot\left(\frac{1}{h}\int_{t_0}^{t_0+h}\varphi(x_0,u(s),t_0)\,ds\right)\right)$$

est en $O(h)$. On fait tendre h vers 0 en remarquant que pour tout $p\in\mathbb{R}^n$, nous avons

$$\lim_{h\to 0}\inf_{u\in\mathcal{U}_{ad}}\left(\frac{1}{h}\int_{t_0}^{t_0+h} L(x_0,u(s),t_0)\,ds+p\cdot\left(\frac{1}{h}\int_{t_0}^{t_0+h}\varphi(x_0,u(s),t_0)\,ds\right)\right)$$
$$=\inf_{u\in U} L(x_0,u,t_0)+p\cdot\varphi(x_0,u,t_0)\ .$$

□

8.2.2 Cas quadratique

Considérons maintenant le cas d'un coût quadratique avec une EDO linéaire comme dans le chapitre 5. Rappelons brièvement les données : l'équation d'état est la suivante

$$\left\{\begin{array}{rcl}\dfrac{d\mathbf{x}}{dt} &=& A\,\mathbf{x}(t)+B\,v(t)\ ,\ \text{sur }]t,T[\ ,\\ \mathbf{x}(t) &=& x\ ,\end{array}\right. \tag{8.2.8}$$

(on a choisi $f=0$ pour simplifier). La fonction coût est :

$$\mathcal{J}(\mathbf{x},v)=\frac{1}{2}\int_t^T\langle\mathbf{x}(s),Q\,\mathbf{x}(s)\rangle_n\,ds+\langle\mathbf{x}(T),D\,\mathbf{x}(T)\rangle_n+\frac{1}{2}\int_t^T\langle v(s),R\,v(s)\rangle_p\,ds\ . \tag{8.2.9}$$

Les propriétés des matrices $A,\ B,\ Q,\ D$ et R ont été précisées dans le chapitre 5. On suppose aussi qu'il n'y a pas de contraintes sur le contrôle.
Posons

$$\mathcal{V}(x,t)=\inf_{L^2(]t,T[,\mathbb{R}^p)}\mathcal{J}(\mathbf{x},v)\ .$$

L'équation de Hamilton-Jacobi-Bellman s'écrit :

$$\frac{\partial\mathcal{V}}{\partial t}+\inf_{u\in\mathbb{R}^p}\left(\frac{1}{2}\langle x,Q\,x\rangle_n+\frac{1}{2}\langle u,R\,u\rangle_p+\langle A\,x+B\,u,\nabla_x\mathcal{V}\rangle_n\right)=0\ . \tag{8.2.10}$$

Après avoir calculé le minimum, on obtient

$$\frac{\partial \mathcal{V}}{\partial t} + \left(\frac{1}{2} \langle x, Q\, x \rangle_n - \frac{1}{2} \left\langle R^{-1} B^t \nabla_x \mathcal{V}, B^t \nabla_x \mathcal{V} \right\rangle_p + \langle A\, x, \nabla_x \mathcal{V} \rangle_n \right) = 0 \,. \tag{8.2.11}$$

Cette équation a lieu en tout point de $\mathbb{R}^n \times \mathbb{R}$ où la fonction valeur est différentiable. De plus $\mathcal{V}$ vérifie la condition finale :

$$\forall x \in \mathbb{R}^n \qquad \mathcal{V}(x,T) = \frac{1}{2} \langle D\, x, x \rangle_n \,.$$

Les résultats obtenus dans le chapitre 5 (Théorème 5.3.1) nous donnent l'expression de $\mathcal{V}$. Soit P la matrice solution de l'équation de Ricatti (5.3.3)

$$\frac{dP}{dt} = -A^t P - P\, A + P\, B\, R^{-1} B^t P - Q \text{ pour } t < T, \;\; P(T) = D \,.$$

On vérifie que la fonction valeur $\mathcal{V}$ est définie par

$$\mathcal{V}(x,t) = \frac{1}{2} \langle P(t)\, x, x \rangle_n \,.$$

On retrouve également le fait que le contrôle optimal est donné par :

$$u(x,t) = -R^{-1} B^t P(x).$$

[Exercices]

[Cas discret]

1. On se pose le problème d'allocation suivant : on a un certain nombre de magasins qui doivent être approvisionnés dans une denrée déterminée.
   Pour éviter qu'ils n'entrent en rupture de stock ces magasins ont besoin d'une quantité suffisante de marchandises. Le besoin de chaque magasin est exprimé par une fonction d'utilité : utilité de n unités.
   On veut connaître l'allocation optimale des marchandises en fonction de la quantité totale disponible, inférieure ou égale à 8. Les fonctions d'utilité sont données ci-dessous.

Magasin	→	1	2	3	4
Unités	8	100	80	80	80
↓	7	80	71	80	70
	6	60	64	72	60
	5	53	55	60	50
	4	48	39	44	40
	3	40	32	11	30
	2	20	20	10	20
	1	8	13	9	10
	0	0	0	0	0

   Suggestion : traiter le nombre de magasins comme le "temps" et le nombre total d'unités allouées aux p premiers magasins comme $x(p)$ ($x(p) - x(p-1)$ = allocation au magasin n$^{\text{o}}$ p), et progresser de gauche à droite.

2. Reprendre l'exemple d'allocation de ressources de la section 8.1.4. On dispose d'une quantité initiale de ressources $\xi > 0$. Montrer que $\mathcal{V}(x,n) = \left(5.5\,(0.8)^{N-n-2} - 10\right) x$ pour $n \leq N-2$. En déduire le gain maximal possible en N étapes (pour $N \geq 2$). Quelles est la stratégie optimale ? Combien faut-il d'étapes pour avoir un gain maximum au moins égal à $10\,\xi$?

3. On considère le problème de minimisation où la fonction coût est donnée par

$$J(x,u) = \sum_{i=0}^{N-1} \left[a_i x_i^2 + b_i x_i u_i + c_i u_i^2 + d_i x_i + e_i u_i + f_i\right] + l\, x_N^2 + w\, x_N + z\ ,$$

et la dynamique du système est donnée par l'équation aux différences

$$x_{i+1} = \varphi(x_i, u_i) = q_i x_i + h_i u_i + k_i,\ i = 0, \ldots, N-1\ .$$

On pose $J_i(x,u) = a_i x_i^2 + b_i x_i u_i + c_i u_i^2 + d_i x_i + e_i u_i + f_i$. Montrer par récurrence que la fonction valeur optimale $\mathcal{V}_i$ est de la forme

$$\mathcal{V}_i(x) = p_i x^2 + q_i x + r_i,\ i = 0, \ldots, N\ ,$$

et déterminer les formules de récurrence pour $p_i,\ q_i$ et r_i.

4. On reprend le problème de l'exercice précédent et on suppose que $x(0)$ n'est pas spécifié, et qu'on doit le choisir (comme toute la suite) aussi pour minimiser J. Comment utiliser la solution de l'exercice précédent pour résoudre ce problème ?

5. On considère le problème suivant : trouver $u(0),\ u(1)$ et $u(2)$ minimisant J donnée par :

$$J(x,y,u) = \underbrace{u^2(0)}_{J_0} + \underbrace{x^2(1) + y^2(1) + u^2(1)}_{J_1} + \underbrace{x^2(2) + y^2(2) + u^2(2)}_{J_2}\ ,$$

avec les contraintes

$$x(0) = 7\ , y(0) = 1\ , x(3) = 2\ ,\ y(3) = 1\ ,$$

où

$$x(i+1) = x(i) + y(i) + 2u(i)\ ,\ \text{et}\ y(i+1) = x(i) - y(i) + u(i)\ .$$

6. Résoudre de deux façons le problème de contrôle quadratique suivant :

$$\min\left\{J(v) \mid v \in L^2(0,T)\right\}\ ,$$

où

$$J(v) = \frac{1}{2}\int_0^T (y(t) - z_d(t))^2\, dt + \frac{N}{2}\int_0^T v^2(t)\, dt\ ,$$

avec y et v fonctions de $[\,0,T\,]$ dans $\mathbb{R}$, et y vérifie l'équation d'état suivante :

$$\frac{dy}{dt}(t) = ay(t) + bv(t)\ , y(0) = 0\ ,\ a, b \in \mathbb{R}^2\ .$$

(a) En écrivant le principe du maximum et en calculant explicitement le contrôle.

(b) En discrétisant par la méthode d'Euler et en appliquant l'algorithme de programmation dynamique.

[Cas continu]

7. **La fonction valeur n'est pas toujours dérivable en tout point.**
Soit le système suivant dans $\mathbb{R}$:

$$\frac{d\mathbf{x}}{dt} = u\ , \mathbf{x}(t) = x\ , \qquad U = [-1, +1], \qquad L \equiv 0\ .$$

Soit F une fonction paire, régulière vérifiant $F'(z) < 0$ pour tout $z > 0$.

(a) Montrer que la fonction valeur définie par

$$\mathcal{V}(x,t) = \inf_{u \in \mathcal{U}_{ad}} F(\mathbf{x}(T))\ ,$$

est paire et donnée par

$$\mathcal{V}(x,t) = \begin{cases} F(x+T-t) & \text{si } x \geq 0\ , \\ F(x-(T-t)) & \text{si } x \leq 0\ . \end{cases}$$

Elle correspond aux contrôles optimaux $u \equiv 1$ si $x \geq 0$ et $u \equiv -1$ si $x \leq 0$.

(b) Montrer que $\mathcal{V}$ est régulière sur $\mathbb{R}^* \times]-\infty, T\,[$ mais qu'elle n'est pas dérivable (par rapport à x) en $(0,t)$ pour tout $t < T$.

---

8. Soient F une fonction de classe $\mathcal{C}^1$ sur $\mathbb{R}$ et $\mathcal{U} = L^2(]\,t, T\,[, [-1,1])$. On considère le problème de contrôle optimal gouverné par l'équation d'état :

$$\frac{d\mathbf{x}}{dt}(s) = u(s) \text{ sur }]\,t, T\,[, \qquad \mathbf{x}(t) = x\ ,$$

et le coût : $J(u) = F(\mathbf{x}(T))$, avec $u \in \mathcal{U}$. On note $\mathcal{V}$ la fonction valeur.

(a) Montrer que

$$\forall x \in \mathbb{R},\ \forall t < T \qquad \mathcal{V}(x,t) = \inf_{y \in \mathbb{R}, |y-x| \leq T-t} F(y)\ .$$

(b) Montrer qu'il existe toujours au moins un contrôle optimal $\bar{u}$. Est-il unique ?

(c) Ecrire la condition nécessaire d'optimalité en $\bar{u}$. On définit l'état adjoint $\bar{p}$ en généralisant la définition du chapitre 5 par

$$\frac{d\bar{p}}{dt}(s) = 0 \text{ sur }]\,t, T[\ ,\ \bar{p}(T) = \frac{dF}{dx}(\mathbf{x}(T))\ .$$

Montrer que

$$\bar{p}(s)\bar{u}(s) = \min_{v \in [-1,1]} \bar{p}(s)\, v\ ,\ p.p.s \in]\,t, T[\ . \tag{1}$$

En déduire l'expression du contrôle optimal.

(d) Ecrire l'équation HJB vérifiée par $\mathcal{V}$ en tout point où elle est différentiable et déterminer un contrôle feedback optimal.

Annexe A

Rappels de quelques notions

A.1 Rappels d'algèbre linéaire

A.1.1 Normes sur $\mathbb{R}^n$

On se place dans $\mathbb{R}^n$ (ensemble des vecteurs réels d'ordre n). On notera $\{e_1, \cdots, e_n\}$ la base canonique de $\mathbb{R}^n$ où le vecteur e_i a toutes ses composantes nulles sauf la i-ème qui vaut 1. On rappelle que $\mathbb{R}^n$ est un espace euclidien de dimension finie quand on le munit du produit scalaire

$$x \cdot y \stackrel{\text{def}}{=} (x, y) = \sum_{i=1}^{n} x_i \, y_i \, ,$$

où $x = (x_1, \cdots, x_n)$ et $y = (y_1, \cdots, y_n)$. On peut aussi écrire sous forme matricielle

$$(x, y) = X^t Y$$

où X et Y désignent les vecteurs colonnes correspondants à x et à y :

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \, , \, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

et $X^t = [x_1, \cdots, x_n]$ est le vecteur transposé de X. Ce produit scalaire induit une norme euclidienne

$$\|x\| = (x, x)^{\frac{1}{2}} = \left[\sum_{i=1}^{n} x_i^2 \right]^{\frac{1}{2}} \, .$$

La distance entre deux vecteurs de $\mathbb{R}^n$ est donnée par

$$d(x, y) = \|x - y\| = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{\frac{1}{2}} \, .$$

On rappelle que sur $\mathbb{R}^n$ toutes les normes sont équivalentes. On peut donc munir $\mathbb{R}^n$ de différentes normes ; les plus courantes sont les normes "ℓ_p" où p est un entier non nul. Elles sont définies par :

$$\|x\|_p = \left[\sum_{i=1}^{n} x_i^p\right]^{\frac{1}{p}} . \tag{A.1.1}$$

La norme euclidienne correspond à $p = 2$: c'est la seule des normes ℓ_p qui soit issue d'un produit scalaire. La norme uniforme ou norme ℓ_∞ est définie de la manière suivante

$$\|x\|_\infty = \max_{i=1,..,n} |x_i| . \tag{A.1.2}$$

A.1.2 Généralités sur les matrices

Dans tout ce qui suit on considère l'ensemble $\mathcal{M}_n(K)$ des matrices carrées $n \times n$ à coefficients dans un corps K (en pratique $\mathbb{R}$ ou $\mathbb{C}$).

Définition A.1.1 (Matrice symétrique) *La matrice transposée d'une matrice carrée de $\mathcal{M}_n(K)$ $A = [a_{ij}]_{1\leq i,j\leq n}$ est $A^t = [a_{ji}]_{1\leq i,j\leq n}$. On dit que A est symétrique si $A = A^t$.*

Définition A.1.2 (Matrice (semi-définie) positive)
Soit A une matrice carrée d'ordre n à coefficients réels. On dit que A est semi-définie positive si

$$\forall x \in \mathbb{R}^n \quad (Ax, x) = x^t\, Ax \geq 0$$

Attention **cela ne signifie pas** automatiquement que tous les coefficients de A sont positifs.

Définition A.1.3 (Matrice définie positive)
Soit A une matrice carrée d'ordre n. On dit que A est définie positive si A est semi-définie positive et

$$(Ax, x) = 0 \Leftrightarrow x = 0 .$$

A.1.3 Propriétés spectrales

Définition A.1.4 *Soit A une matrice carrée d'ordre n. On dit que $\lambda \neq 0$ est **valeur propre** de A si on peut trouver $x_\lambda \in \mathbb{R}^n$, $x_\lambda \neq 0$ tel que $Ax_\lambda = \lambda x_\lambda$. Le vecteur x_λ est appelé **vecteur propre** associé à la valeur propre λ. L'ensemble des valeurs propres de A est le **spectre** de A.*

Théorème A.1.1 *Soit $A \in \mathcal{M}_n(K)$ et $\lambda \in K$. Les propriétés suivantes sont équivalentes :*

1. *λ est valeur propre de A.*
2. *$A - \lambda I_n$ n'est pas inversible*
3. *$\det(A - \lambda I_n) = 0$.*

On appelle **polynôme caractéristique** de A :

$$P_A(X) = \det(A - XI_n) .$$

Théorème A.1.2 *Soit $A \in \mathcal{M}_n(\mathbb{C})$. A possède n valeurs propres distinctes ou confondues.*

Théorème A.1.3 (Théorème de Cayley-Hamilton)
Soit $A \in \mathcal{M}_n(K)$ telle que toutes les racines du polynôme caractéristique de A soient dans K. Alors $P_A(A) = 0$.

Pour des matrices symétriques réelles nous avons un résultat important.

Théorème A.1.4 *Les valeurs propres d'une matrice symétrique réelle sont toutes réelles. De plus toute matrice symétrique réelle est diagonalisable dans une base orthogonale de vecteurs propres.*

Rappelons qu'une matrice de $\mathcal{M}_n(\mathbb{R})$ est diagonalisable (dans une base de vecteurs propres) si on peut trouver une base (de vecteurs propres) de $\mathbb{R}^n$ dans laquelle la matrice est diagonale.
Pour des matrices définies positives le résultat est encore plus précis.

Théorème A.1.5 *Les valeurs propres d'une matrice symétrique réelle semi- définie positive sont toutes réelles et positives. Si de plus, la matrice est définie positive, les valeurs propres sont strictement positives.*

A.1.4 Conditionnement des systèmes linéaires

La notion de conditionnement d'une matrice permet de "mesurer" la stabilité d'un système linéaire. Plus précisément, soit A une matrice réelle inversible. On se donne un vecteur quelconque (non nul) b de $\mathbb{R}^n$ et on s'intéresse au système linéaire : $A\,x = b$. On note x sa solution. Souvent A et b sont le résultat de calculs ou de mesures et sont entachés d'erreur. On ne résoud donc pas vraiment le système $A\,x = b$ mais plutôt : $(A + \Delta A)(x + \Delta x) = b$, où ΔA est une "perturbation" de A ou encore $A(x + \delta x) = b + \delta b$, δb étant une perturbation de b. On peut alors montrer que les erreurs relatives commises sur la solution x sont contrôlées par un nombre noté cond (A) appelé **nombre de conditionnement** de A. Plus précisément si $x + \Delta x$ est la solution du système $(A + \Delta A)(x + \Delta x) = b$, alors

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \text{cond}\,(A)\frac{\|\Delta A\|}{\|A\|}\ .$$

De même si $x + \delta x$ est la solution du système $A(x + \delta x) = b + \delta b$, nous obtenons

$$\frac{\|\delta x\|}{\|x\|} \leq \text{cond}\,(A)\frac{\|\delta b\|}{\|b\|}\ .$$

Ici $\|\cdot\|$ désigne indifféremment une norme vectorielle sur $\mathbb{R}^n$ et la norme matricielle induite définie par : $\|A\| = \sup_{\|x\|=1} \|A\,x\|$.

Le nombre cond (A) qui est égal à $\|A\|\,\|A^{-1}\|$, mesure donc la stabilité du système. On dit que la matrice A est **bien conditionnée** si cond (A) est voisin de 1. Dans le cas contraire, la matrice est mal conditionnée.

Le nombre cond (A) dépend de la norme matricielle choisie pour la matrice A. On choisit habituellement la norme matricielle induite par la norme euclidienne $\|\cdot\|_2$ et on note alors $\text{cond}_2(A)$. Pour plus de détails, on peut se référer au livre de P. CIARLET [7].

A.2 Calcul différentiel dans $\mathbb{R}^n$

Nous rappelons dans cette section les notions de base du calcul différentiel dans $\mathbb{R}^n$. Pour plus de détails on pourra consulter le *Cours de calcul différentiel* de H. CARTAN [6].

A.2.1 Dérivées, différentielles

Dérivée, différentielle d'une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$

Soit f une fonction de plusieurs variables $x = (x_1, \cdots, x_n)$ à valeurs réelles : f est donc définie de $\mathbb{R}^n$ dans $\mathbb{R}$.

Définition A.2.1 *On dit que f est **différentiable** (ou dérivable au sens de* FRÉCHET *) en un point x_0 de $\mathbb{R}^n$ si on peut trouver une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$ notée $Df(x_0)$ ou $f'(x_0)$ telle que*

$$\lim_{\|h\|\to 0} \frac{|f(x_0+h) - f(x_0) - Df(x_0)(h)|}{\|h\|} = 0 \ .$$

$Df(x_0)$ est l'application dérivée (ou différentielle) de f au point x_0.

L'application dérivée Df associe à $x_0 \in \mathbb{R}^n$ un **vecteur** de $\mathbb{R}^n$ appelé aussi gradient de f en x_0 et noté $\nabla f(x_0)$. Nous allons préciser les coefficients de ce vecteur gradient ce qui permettra de calculer "facilement" $Df(x_0)(h)$.

Définition A.2.2 *On dit que f est **dérivable dans la direction** $d \in \mathbb{R}^n$ au point x_0 de $\mathbb{R}^n$ si*

$$\lim_{t\to 0^+} \frac{f(x_0+td) - f(x_0)}{t} \ ,$$

*existe et est linéaire par rapport à d. Quand d est le i ème vecteur de base e_i de $\mathbb{R}^n$, on dit que f admet une **dérivée partielle** par rapport à x_i et on la note $\dfrac{\partial f}{\partial x_i}$. On a donc*

$$\lim_{t\to 0} \frac{f(x_0+te_i) - f(x_0)}{t} = \frac{\partial f}{\partial x_i}(x_0) \ ,$$

On a le résultat important suivant :

Théorème A.2.1 *Si f est différentiable en un point x_0 de $\mathbb{R}^n$ alors toutes les dérivées partielles existent et le gradient s'écrit :*

$$\nabla f(x_0) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1}(x_0) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(x_0) \end{bmatrix} . \tag{A.2.1}$$

On a alors

$$Df(x_0)(h) = (\nabla f(x_0), h) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) \cdot h_i \ \ (\in \mathbb{R})$$

où $h = (h_1, \cdots, h_n)$.

Dérivée, différentielle d'une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$

Soit u une fonction de plusieurs variables $x = (x_1, \cdots, x_n)$ à valeurs dans $\mathbb{R}^n$:

$$\begin{array}{rccc} u : & \mathbb{R}^n & \to & \mathbb{R}^n \\ & x = (x_1, \cdots, x_n) & \mapsto & u(x) = (u_1(x), \cdots, u_n(x)) \end{array}$$

La définition de la dérivabilité pour u est la même mais on fait intervenir la norme de $\mathbb{R}^n$ "à l'arrivée" au lieu de celle de $\mathbb{R}$. Cette fois la dérivée (ou différentielle) en x n'est plus identifiable à un vecteur mais à une matrice carrée d'ordre n appelée matrice **Jacobienne** que l'on calcule de manière analogue au calcul du gradient :

$$Du(x) = \begin{bmatrix} \dfrac{\partial u_1}{\partial x_1}(x) & \cdots & \dfrac{\partial u_1}{\partial x_j}(x) & \cdots & \dfrac{\partial u_1}{\partial x_n}(x) \\ \vdots & \ddots & \dfrac{\partial u_i}{\partial x_j}(x) & \cdots & \vdots \\ \dfrac{\partial u_n}{\partial x_1}(x) & \cdots & \dfrac{\partial u_n}{\partial x_j}(x) & \ddots & \dfrac{\partial u_n}{\partial x_n}(x) \end{bmatrix} \tag{A.2.2}$$

de sorte que $Du(x)(h)$ est un vecteur de $\mathbb{R}^n$.

Dérivée seconde d'une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$

On peut alors définir la dérivée seconde de f (de $\mathbb{R}^n$ dans $\mathbb{R}$), comme étant la dérivée de Df. Plus précisément

Définition A.2.3 *On dit que $f : \mathbb{R}^n \to \mathbb{R}$ est dérivable sur un sous-ensemble (ouvert) S de $\mathbb{R}^n$ si f est dérivable en tout point x de S (S peut éventuellement être $\mathbb{R}^n$ tout entier).*

L'application dérivée $Df : S \to \mathbb{R}^n$ associe à un point x de S le vecteur $Df(x) = \nabla f(x)$. Cette application est définie dans $\mathbb{R}^n$ et prend ses valeurs dans $\mathbb{R}^n$. Si elle est elle-même dérivable en un point x_0 de S on dit que f est deux fois dérivable en x_0. Dans ce cas, la dérivée seconde de f en x_0 est définie comme la dérivée première de Df. D'après le paragraphe précédent elle est identifiable à une matrice carrée appelée matrice **Hessienne** et définie par

$$D^2 f(x_0) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(x_0) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(x_0) & \cdots & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(x_0) \\ \vdots & \cdots & \dfrac{\partial^2 f}{\partial x_i \partial x_j}(x_0) & \cdots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1}(x_0) & \cdots & \cdots & \dfrac{\partial^2 f}{\partial x_n \partial x_{n-1}}(x_0) & \dfrac{\partial^2 f}{\partial x_n^2}(x_0) \end{bmatrix}$$

où

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x_0) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x_0) = \frac{\partial}{\partial x_i}\left[\frac{\partial f}{\partial x_j}\right](x_0)\,.$$

Cette matrice est symétrique.
Donnons pour terminer quelques définitions

Définition A.2.4 *On dit que f de $\mathbb{R}^n$ dans $\mathbb{R}$ est de **classe** $\mathcal{C}^1$ (ou tout simplement $\mathcal{C}^1$) sur $S \subset \mathbb{R}^n$ si elle est dérivable sur S et si sa dérivée est continue sur S.*

Définition A.2.5 *On dit que f de $\mathbb{R}^n$ dans $\mathbb{R}$ est $\mathcal{C}^2$ sur $S \subset \mathbb{R}^n$ si elle est deux fois dérivable sur S et si sa dérivée seconde est continue sur S.*

A.2.2 Le théorème des fonctions implicites

On rappelle ci-dessous le théorème des fonctions implicites dans sa forme générale ([6] p.61) :

Théorème A.2.2 *Soient $E,\ F$ et G trois espaces de Banach, U un ouvert de $E \times F$ et $f : U \to G$ une application de classe $\mathcal{C}^1$. Soit (a, b) un point de U tel que*

$$f(a, b) = 0 \ .$$

On suppose que la dérivée partielle $f'_y(a, b) \in \mathcal{L}(F; G)$ est un isomorphisme de F sur G. Alors, il existe un voisinage ouvert $\mathcal{V}$ de (a, b) dans $E \times F$, un voisinage ouvert $\mathcal{W}$ de a dans E et une application de classe $\mathcal{C}^1$

$$g : \mathcal{W} \to F \ ,$$

tels que la relation

$$(x, y) \in \mathcal{V} \text{ et } f(x, y) = 0 \ ,$$

est équivalente à

$$x \in \mathcal{W} \text{ et } y = g(x) \ .$$

Rappelons qu'un espace de **Banach** est un espace vectoriel normé tel que toute suite de Cauchy est convergente. En particulier tout espace vectoriel réel de dimension finie (isomorphe à $\mathbb{R}^n$) est un espace de Banach.
Nous pouvons détailler le résultat en dimension finie : soient $y = (y_1, \ldots, y_m, y_{m+1}, \ldots, y_q) \in \mathbb{R}^q$ et m fonctions $g_i,\ i = 1, \ldots, m$ de $\mathbb{R}^q$ dans $\mathbb{R}^m$ $(q \geq m)$ vérifiant les propriétés suivantes :

- Les fonctions g_i sont $\mathcal{C}^1$ dans un voisinage de y.
- $g_i(y) = 0,\ i = 1, \ldots, m$.
- La matrice Jacobienne $J = \begin{bmatrix} \dfrac{\partial g_1(y)}{\partial y_1} & \cdots & \dfrac{\partial g_1(y)}{\partial y_m} \\ \vdots & & \vdots \\ \dfrac{\partial g_m(y)}{\partial y_1} & \cdots & \dfrac{\partial g_m(y)}{\partial y_m} \end{bmatrix}$, est inversible.

Alors il existe un voisinage $\mathcal{V}(\hat{y})$ de $\hat{y} = (y_{m+1}, y_{m+2} \ldots, y_q)$ dans $\mathbb{R}^{q-m}$ et une fonction $\Phi = (\Phi_1, \ldots, \Phi_m) : \mathcal{V} \to \mathbb{R}^m,\ \mathcal{C}^1$ tels que

- $y_i = \Phi_i(\hat{y}),\ i = 1, \ldots, m$
- $g_i(\Phi_1(\hat{x}), \Phi_2(\hat{x}), \ldots, \Phi_m(\hat{x}), \hat{x}) = 0,\ i = 1, \ldots, m$, pour tout $\hat{x} \in \mathcal{V}(\hat{y})$.

A.2.3 La formule de Taylor

La formule de Taylor est un outil important en convexité. Nous la rappelons dans le cas général ([6] p. 77).

Théorème A.2.3 (Reste intégral) *Soit $f : U \to F$ une application de classe $\mathcal{C}^{n+1}$ (E et F sont deux espaces de banach et U est un ouvert de E). Si le segment $[\, a, a+h \,]$ est contenu dans U, on a :*

$$f(a+h) = f(a) + f'(a)\cdot h + \ldots + \frac{1}{n!} f^{(n)}(a)\cdot (h)^n + \int_0^1 \frac{(1-t)^n}{n!} f^{(n+1)}(a+th)\cdot (h)^{n+1}\, dt\,.$$

Théorème A.2.4 (Reste de Lagrange) *Soit $f : U \to F$ une application $n+1$ fois différentiable ; supposons*

$$\|f^{n+1}(x)\| \leq M \text{ pour tout } x \in U\ .$$

Alors

$$\|f(a+h) - f(a) - f'(a)\cdot h - \ldots - \frac{1}{n!} f^{(n)}(a)\cdot (h)^n\| \leq M \frac{\|h\|^{n+1}}{(n+1)!}\ .$$

A.3 Equations différentielles ordinaires (EDO)

A.3.1 Le théorème de Cauchy-Lipschitz

La plupart des lois d'état en physique sont données par des **équations différentielles ordinaires** ou EDO. On cherche une fonction y d'une **seule variable** réelle t à valeurs dans $\mathbb{R}^n$ (c'est-à-dire à n composantes) dérivable sur I et solution de

$$\frac{dy}{dt}(t) = F(t, y(t)) \text{ sur } I\ . \tag{A.3.1}$$

I est un intervalle ouvert ou une réunion d'intervalles ouverts de $\mathbb{R}$; F est une fonction de $\mathbb{R}\times\mathbb{R}^n$ à valeurs dans $\mathbb{R}^n$. En réalité l'équation (A.3.1) est un **système différentiel** dès que $n \geq 2$. Pour assurer l'unicité d'une éventuelle solution, il faut ajouter des conditions supplémentaires. Les primitives d'une fonction différant toutes d'une constante on peut fixer la valeur de la fonction en un point donné. Il faut donc ajouter à la relation différentielle (A.3.1) des conditions "ponctuelles" ; dans le cas où on impose des conditions ponctuelles à la fonction (et à ses dérivées) **en un point donné** on dit qu'on a un problème de CAUCHY . On dispose alors d'un théorème général qui permet d'assurer l'existence et l'unicité des solutions des problèmes de Cauchy.

Théorème A.3.1 (Théorème de Cauchy-Lipschitz)
Soit le problème de CAUCHY *suivant : trouver y continue et dérivable sur I intervalle (ou réunion d'intervalles) de $\mathbb{R}$ telle que*

$$\frac{dy}{dt}(t)(t) = F(t, y(t)) \text{ sur } I,\ \ y(t_0) = y_0\ , \tag{A.3.2}$$

où $t_0 \in I$ et $y_0 \in \mathbb{R}^n$ sont donnés.

On suppose que la fonction F est continue sur $I \times \mathbb{R}^n$ et que F est lipschitzienne par rapport à la deuxième variable y uniformément par rapport à la première t sur I, c'est-à-dire

$$\exists M > 0\ , \forall t \in I\ , \forall (y,z) \in \mathbb{R}^n \quad \|F(t,y) - F(t,z)\| \leq M\|y - z\|\ .$$

Alors le problème de CAUCHY *(A.3.2) admet une solution unique définie sur I.*

Démonstration - On trouvera la démonstration de ce théorème dans le livre de M.CROUZEIX et A.MIGNOT [8]. On y trouvera aussi un exposé détaillé de la théorie des équations différentielles. □

La condition $y(t_0) = y_0$ dans le problème de CAUCHY est une **condition initiale** .

A.3.2 Equations différentielles linéaires

On donne dans cette section quelques résultats concernant les équations différentielles ordinaires **linéaires** de la forme

$$(EL) \qquad \left\{ \begin{array}{rcl} \dfrac{dx}{dt}(t) & = & A(t)\,x(t) + B(t)\,u(t)\ ,\ t \in I =]\,0,T[\\ x(0) & = & x_0\ . \end{array} \right.$$

x est la fonction inconnue de I dans $\mathbb{R}^n$, $u : I \to \mathbb{R}^p$, $A(t)$ est une matrice carrée $n \times n$ pour tout t dans I et B est une matrice $n \times p$ pour tout t dans I.
On considère l'équation différentielle homogène associée

$$(EH) \qquad \left\{ \begin{array}{rcl} \dfrac{dx}{dt}(t) & = & A(t)\,x(t)\ ,\ t \in I =]\,0,T[\\ x(0) & = & x_0\ , \end{array} \right.$$

et on note x^i la solution de (EH) correspondant à une donnée initiale $x_0^i \in \mathbb{R}^n$, c'est-à-dire la solution de

$$\frac{dx^i}{dt}(t) = A(t)\,x^i(t)\ , t \in I =]\,0,T[\ , x^i(0) = x_0^i\ .$$

Nous avons un résultat donnant la structure de l'ensemble des solutions de (EH) :

Proposition A.3.1 *L'ensemble des solutions de (EH) est un espace vectoriel de dimension n. De plus, les deux conditions suivantes sont équivalentes :*
i) $(x_0^1, \ldots, x_0^n)$ sont linéairement indépendants dans $\mathbb{R}^n$
ii) Pour tout t dans I, $(x^1(t), \ldots, x^n(t))$ sont linéairement indépendants dans $\mathbb{R}^n$

Dans ce cas, $(x^1(\cdot), \ldots, x^n(\cdot))$ est un ensemble de solutions fondamentales de (EH) (en fait une base de l'espace des solutions). La matrice

$$X(t) = \left[x^1(t), \ldots, x^n(t)\right]\ ,$$

est une matrice **fondamentale** du système. Elle est régulière pour tout t.

Lemme A.3.1 *Si $X(\cdot)$ et $Y(\cdot)$ sont deux matrices fondamentales du système, alors il existe C matrice $n \times n$, régulière et* ***constante*** *telle que*

$$X(t) = Y(t)\, C \text{ pour tout } t\ .$$

On peut maintenant donner la forme générale des solutions de (EL).

Théorème A.3.2 *Pour x_0 donné dans $\mathbb{R}^n$, l'unique solution de (EL) est donnée (grâce à la méthode de la variation de la constante) par*

$$x[u, x_0](t) = X(t)X^{-1}(0)\, x_0 + X(t) \int_0^t X^{-1}(s)B(s)u(s)\ ds\ .$$

Ceci est un résultat général pour des données A et B dépendant de t. Dans le cas où A ne dépend pas du temps le résultat est plus précis :

Théorème A.3.3 *Si A est constante, une matrice fondamentale pour (EH) est*

$$X(t) = e^{At} \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} \frac{A^k t^k}{k!}\ ,$$

et l'unique solution de (EL) est

$$x[u, x_0](t) = e^{At}x_0 + e^{At} \int_0^t e^{-As}B(s)u(s)\ ds\ .$$

A.4 Le Théorème de Hahn-Banach

Le Théorème de Hahn-Banach, sous sa forme géométrique, permet de séparer des ensembles convexes. Il est très important en analyse convexe et sert en particulier à exhiber des multiplicateurs en optimisation. Nous rappelons ici la forme géométrique de ce théorème qui est la seule utile dans notre cas ainsi que des corollaires importants. Pour les démonstrations et plus de précisions nous renvoyons au livre de H. BRÉZIS [5].
Dans tout ce qui suit E désigne un espace de Banach (sur $\mathbb{R}$) et E' est son espace dual (c'est-à-dire l'ensemble des formes linéaires $f : E \to \mathbb{R}$, continues).

Définition A.4.1 (Hyperplan affine) *Un* ***hyperplan*** *affine* ***fermé*** *est un ensemble de la forme*

$$H = \{\ x \in E\ \mid\ \alpha(x) + \beta = 0\ \}\ ,$$

où $\alpha \in E'$ est une forme linéaire continue non nulle sur E et $\beta \in \mathbb{R}$.

Dans le cas où E est un espace de Hilbert $\mathbb{H}$ (en particulier si $E = \mathbb{R}^n$), on peut identifier $\mathbb{H}$ à son dual et tout hyperplan affine fermé est de la forme

$$H = \{\ x \in \mathbb{H}\ \mid (\alpha, x)_{\mathbb{H}} + \beta = 0\ \}\ ,$$

où $(\cdot, \cdot)_{\mathbb{H}}$ désigne le produit scalaire de $\mathbb{H}$, $\alpha \in \mathbb{H},\ \alpha \neq 0$ et $\beta \in \mathbb{R}$.

Définition A.4.2 (Séparation) *Soient A et B deux sous-ensembles non vides de E. On dit que l'hyperplan affine H d'équation : $\alpha(x) + \beta = 0$, sépare A et B* ***au sens large*** *si*

$$\forall x \in A \quad \alpha(x) + \beta \leq 0 \quad \textit{et} \quad \forall y \in B \quad \alpha(y) + \beta \geq 0 \, .$$

On dit que H sépare A et B ***strictement*** *s'il existe $\varepsilon > 0$ tel que*

$$\forall x \in A \quad \alpha(x) + \beta \leq -\varepsilon \quad \textit{et} \quad \forall y \in B \quad \alpha(y) + \beta \geq \varepsilon \, .$$

Donnons à présent la première forme géométrique du théorème de Hahn-Banach :

Théorème A.4.1 *Soient A et B deux sous-ensembles de E convexes, non vides et disjoints. On suppose que A est ouvert. Alors, il existe un hyperplan affine fermé qui sépare A et B au sens large .*

Ce sont les corollaires suivants que l'on utilise essentiellement dans ce livre :

Corollaire A.4.1 *Soit C convexe ouvert et $x^* \notin C$. Alors, il existe un hyperplan affine fermé qui sépare C et x^* au sens large, c'est-à-dire*

$$\exists \alpha \in E', \alpha \neq 0, \exists \beta \in \mathbb{R} \quad \textit{tels que} \quad \alpha(x^*) + \beta \leq 0 \quad \textit{et} \quad \forall x \in C \quad \alpha(x) + \beta \geq 0 \, .$$

Supposons connu le fait que l'intérieur d'un ensemble convexe est convexe (s'il est non vide). On a alors

Corollaire A.4.2 *Soit C un convexe fermé et $x^* \notin$ Int(C). Alors, il existe un hyperplan affine fermé qui sépare C et x^* au sens large.*

Corollaire A.4.3 *Soit C un convexe non vide de $\mathbb{R}^n$ et fermé et $x^* \in C$. Alors :*
$x^* \in$ Int (C) si et seulement si il n'existe aucune forme linéaire séparant x^* et C.

Citons enfin pour mémoire la deuxième forme géométrique du théorème de Hahn-Banach :

Théorème A.4.2 *Soient A et B deux sous-ensembles de E convexes, non vides et disjoints. On suppose que A est fermé et que B est compact. Alors, il existe un hyperplan affine fermé qui sépare A et B strictement.*

Annexe B

Correction des exercices

Chapitre 1

- **Exercice 1**
  Soit N une norme d'un espace vectoriel E. Soient $x,\ y \in E$ et $t \in [0,1]$.

  $$N(tx+(1-t)y) \leq N(tx) + N((1-t)y) = tN(x) + (1-t)N(y)\ .$$

  N est donc convexe.
- **Exercice 2**
  (a) Supposons K convexe et soient $x,\ y \in \mathbb{R}^n$ et $t \in [0,1]$.
  - Si x et y sont dans K alors $tx+(1-t)y$ est dans K et

    $$I_K(tx+(1-t)y) = 0 = t \underbrace{I_K(x)}_{=0 \text{ car } x\in K} + (1-t) \underbrace{I_K(y)}_{=0 \text{ car } y\in K}\ .$$

  - Si x (ou y) n'est pas dans K, alors $I_K(x)$ (ou $I_K(y)$) $= +\infty$, et l'inégalité de convexité est trivialement vérifiée.

  (b) *Réciproquement.* Soient $x,\ y \in K$ et $t \in [0,1]$; par convexité de I_K

  $$I_K(tx+(1-t)y) \leq t \underbrace{I_K(x)}_{=0 \text{ car } x\in K} + (1-t) \underbrace{I_K(y)}_{=0 \text{ car } y\in K} = 0\ .$$

  Comme I_K ne prend que les valeurs 0 ou $+\infty$, $I_K(tx+(1-t)y) = 0$ et $tx+(1-t)y) \in K$.
- **Exercice 3**
  (a) Supposons f convexe ; soient (u,α) et (v,β) dans epi(f) et $t \in [0,1]$. Comme U est convexe, $tu+(1-t)v \in U$ et

  $$f(tu+(1-t)v) \leq tf(u) + (1-t)f(v) \leq t\alpha + (1-t)\beta\ ;$$

  donc $t(u,\alpha) + (1-t)(v,\beta) \in$ epi (f).
  (b) *Réciproquement.*
  Comme $(u,f(u))$ et $(v,f(v))$ sont dans epi(f), $t(u,f(u)) + (1-t)(v,f(v))$ aussi. La convexité de f en découle.

– **Exercice 4**
Soient $x,\ y \in U$ et $t \in [0,1]$.

$\forall i \in I \qquad f_i(tx+(1-t)y) \leq tf_i(x)+(1-t)f_i(y) \leq t \sup_{i\in I} f_i(x) + (1-t)\sup_{i\in I} f_i(y)\ .$

Donc $\sup_{i\in I} f_i(tx+(1-t)y) \leq t \sup_{i\in I} f_i(x) + (1-t)\sup_{i\in I} f_i(y)$.

– **Exercice 5**
On utilise la convexité de la fonction exponentielle avec $\frac{1}{p}+\frac{1}{q}=1$. On obtient

$$\exp\left(\frac{1}{p}\log a^p + \frac{1}{q}\log b^q\right) \leq \frac{1}{p}\exp\left(\log a^p\right) + \frac{1}{q}\exp\left(\log b^q\right)$$

pour $a,\ b>0$, c'est-à-dire l'inégalité voulue.

– **Exercice 6**
On raisonne par récurrence : c'est vrai pour $p=2$. Supposons que c'est vrai pour $p-1$. Soit $(\lambda_i)_{1\leq i\leq p} \in (\mathbb{R}^+)^p$ tel que $\sum_{i=1}^{p}\lambda_i = 1$. Il existe donc i_0 tel que $\lambda_{i_0} \neq 0$. Posons $\mu = \sum_{i=1, i\neq i_0}^{p} \lambda_i$. Il est clair que $\lambda_{i_0}, \mu \in]0,1[$ et $\lambda_{i_0}+\mu = 1$. Soit $(x_i)_{1\leq i\leq p} \in \mathbb{R}^p$. On appelle x le barycentre des points $(\lambda_i, x_i)_{i\neq i_0}$ de sorte que $\sum_{i=1, i\neq i_0}^{p}\lambda_i x_i = \mu x$. La convexité de f donne

$$f(\sum_{i=1}^{p}\lambda_i x_i) = f(\mu x + \lambda_{i_0}x_{i_0}) \leq \mu f(x) + \lambda_{i_0} f(x_{i_0})\ .$$

comme $x = \sum_{i=1, i\neq i_0}^{p} \frac{\lambda_i}{\mu} x_i$, on utilise l'hypothèse de récurrence pour conclure.

– **Exercice 7**
La condition suffisante est que les fonctions g_i soient convexes et continues et que les fonctions h_j soient affines et continues.

– **Exercice 8**
Soient $\lambda_2 \geq \lambda_1 > 0$. Posons $t = \frac{\lambda_1}{\lambda_2} \in]\,0,1\,]$.

$F(u+\lambda_2\, v) = F(u+t\,\lambda_1 v) = F\left((1-t)u + t(u+\lambda_1 v)\right) \leq (1-t)\,F(u) + t\,F(u+\lambda_1 v)\ .$

Donc $\qquad F(u+\lambda_2\, v) - F(u) \leq t\ (F(u+\lambda_1 v) - F(u)) = \frac{\lambda_1}{\lambda_2}\left(F(u+\lambda_1 v) - F(u)\right)\ ,$

c'est-à-dire $\Phi(\lambda_1) \leq \Phi(\lambda_2)$.

– **Exercice 9**
(a) $x \in \text{dom}\ (f)\ \Rightarrow f(x) < +\infty$.
(b) f^* est convexe :

$$f^*(ty+(1-t)z) = \sup_{x\in \text{dom}\ f} \langle x, t\,y + (1-t)z\rangle - f(x)$$

$$= \sup_{x \in \text{dom } f} \left(t(\langle x, y\rangle - f(x)) + (1-t)\left(\langle x, z\rangle - f(x)\right)\right).$$

Comme $\sup(a+b) \leq \sup(a) + \sup(b)$ et $\sup(ta) \leq t\ \sup(a)$ pour $t > 0$, on a

$$f^*(ty + (1-t)z) \leq t\, f^*(y) + (1-t) f^*(z)\ .$$

(c) Evident avec les propriétés du sup.
(d)

- Pour $f(x) = \langle x, b\rangle$, $f^*(y) = 1_{\{b\}}(y) \stackrel{\text{def}}{\equiv} \begin{cases} 0 & \text{si } y = b\ , \\ +\infty & \text{sinon.} \end{cases}$
- Pour $f(x) = k$, $f^*(0) = -k$ et $f^*(y) = +\infty$ si $y \neq 0$.
- Pour $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$, $f^*(y) = \frac{1}{2}\left\langle A^{-1}(b+y), b+y\right\rangle$.

- **Exercice 10**

  On montre que si f est lipschitzienne sur $[\,0,1\,]$ et dérivable, alors f' est bornée sur $]\,0,1]$: Soit $x \in]\,0,1]$ et h assez petit pour que $x - h \in [\,0,1\,]$. Comme f est lipschitzienne sur $[\,0,1\,]$

  $$\|f(x-h) - f(x)\| \leq k\,|h|$$

  où k est indépendant de x. En divisant et avec $h \to 0$, on obtient que la dérivée à gauche est bornée. On procède de même avec la dérivée à droite (sauf en 1).
- **Exercice 11**
  - $f :\ x \mapsto \|x\|_1$ est différentiable sur $\mathbb{R}^n - \{0\}$. Elle ne l'est pas en 0 (elle n'est pas Gâteaux - différentiable non plus). En $x^* \neq 0$,

    $$f'(x^*) \cdot x = \sum_{x_i^* > 0} x_i - \sum_{x_i^* < 0} x_i\ .$$

  - $f :\ x \mapsto \|x\|_2$ n' est différentiable que sur $\mathbb{R}^n - \{0\}$. En $x^* \neq 0$, $f'(x^*) = \dfrac{x^*}{\|x^*\|_2}$.
- **Exercice 12**
  (a) On applique le théorème de représentation de Riesz à l'application linéaire $a(u,\cdot)$: $v \mapsto a(u,v)$, pour u fixé. On définit ainsi une forme linéaire A_u ; on voit facilement que $u \mapsto A_u$ est linéaire :$A_u = A\,u$.
  (b) $\nabla J(v) = A\,v - b$ et $D^2 J(v) = A$.
- **Exercice 13**
  Résultat classique de calcul différentiel (avec les dérivées partielles).
- **Exercice 14**

  $$J(y+tz) - J(y) = \int_a^b \left(t^2 z^2(x) + t\, y(x)\, z(x)\right)\ dx$$

  pour tout $t > 0$. Donc $\nabla J(y) \cdot z = 2 \int_a^b y(x)\ z(x)\ dx$.

Chapitre 2

- **Exercice 1**

1. Non car $\lim_{x\to-\infty} J(x) = -\infty$.
2. Si $a = 0$ alors J est constante et ne peut pas être coercive. Si $a \neq 0$, il existe $i_0, 1 \leq i_0 \leq n$ tel que $a_{i_0} \neq 0$. On prend la suite $x_k = -k a_{i_0} e_{i_0}$ (où e_i est le i ème vecteur de base). Lorsque $k \to +\infty$, on a $\|x_k\| \to +\infty$ et $J(x_k) \to -\infty$. J n'est donc jamais coercive.
3. Non : prendre la suite $x^n = (0, -n)$.
4. Non : prendre la suite $x^n = (0, -n)$.
5. Oui car $J(x_1, x_2) = (x_1 - 500)^2 + x_2^2 - 255\,000$.

– **Exercice 2**
A est symétrique, donc il existe une base de vecteurs propres orthonormés $(u_i)_{i=1,\dots,n}$. les valeurs propres associées $(\lambda_i)_{i=1,\dots,n}$ sont strictement positives puisque A est définie positive. Soit $x = \sum_{i=1}^{n} x_i u_i$ dans $\mathbb{R}^n$. Nous avons

$$(Ax, x) = \sum_{ij=1}^{n} \lambda_i x_i x_j \,(u_i, u_j) = \sum_{i=1}^{n} \lambda_i x_i^2 \geq \lambda_{\min} \|x\|^2 \,.$$

La constante α peut être prise égale à la plus petite valeur propre $\lambda_{\min} > 0$.

– **Exercice 3**
Il suffit de considérer la fonction de $\mathbb{R}$ vers $\mathbb{R}$ définie par $f(x) = x^3$.

– **Exercice 4**
(a) Il y a deux points critiques : $(0, 0)$ et $(\frac{1}{2}, 1)$. La matrice hessienne vaut $\begin{bmatrix} 2 & -1 \\ -1 & x_2 \end{bmatrix}$. Pour $x_2 = 0$, la matrice a deux valeurs propres de signes différents. D'après le théorème 2.2.3 le point $(0, 0)$ n'est ni un minimum, ni un maximum. Pour $x_2 = 1$, la matrice est définie positive. D'après le théorème 2.2.4 , le point $(\frac{1}{2}, 1)$ est un minimum strict.
(b) Le point $(0, 0)$ est un point critique mais ce n'est ni un minimum, ni un maximum.
(c) Les deux points critiques sont $(0, 0)$ et $(3, 3)$. Le point $(0, 0)$ n'est ni un maximum ni un minimum car la matrice hessienne n'est ni semi-positive ni semi-négative. $(3, 3)$ est un minimum strict.

– **Exercice 5**
Un rapide calcul donne pour tous $u,\ v \in \mathbb{R}^n$ et $t \in [0, 1]$:

$$J(t\,u + (1-t)\,v) - t\,J(u) - (1-t)J(v) = \frac{t\,(t-1)}{2}\,(A(u-v), u-v) \,.$$

D'où (a) et (b).
La question (c) est une application directe du cours.
(d) Soient $u,\ v \in \mathbb{R}^n$ et $t > 0$:

$$J(u+tv) - J(u) = t\,(A\,u - b, v) + \frac{t^2}{2}\,(A\,u, u) \,. \tag{B.1}$$

S'il existe $u \in \mathbb{R}^n$ tel que : $\forall v \in \mathbb{R}^n,\ \ J(u) \leq J(v)$, (B.1) donne après division par t

$$\forall v \in \mathbb{R}^n \quad (A\,u - b, v) + \frac{t}{2}\,(A\,u, u) \geq 0 \,.$$

En faisant tendre t vers 0 on voit que $(A\,u - b, v) \geq 0$ pour tout v et donc $Au - b = 0$ (l'ensemble $\{w \in \mathbb{R}^n \mid Aw = b\}$ n'est donc pas vide) ; par conséquent

$$\forall v \in \mathbb{R}^n, \forall t > 0 \qquad \frac{t}{2}\,(A\,u, u) \geq 0\,,$$

ce qui signifie que A est semi-définie positive.
Réciproquement, on choisit u dans l'ensemble $\{w \in \mathbb{R}^n \mid Aw = b\}$ qui n'est pas vide. Si de plus A est semi-définie positive, la relation (B.1) montre que $J(u) \leq J(u + tv)$ pour tout $v \in \mathbb{R}^n$.
(e) C'est en partie la contraposée de (d). Elle s'en déduit immédiatement en supposant par exemple que $\inf_{v \in \mathbb{R}^n} J(v) > -\infty$.

– **Exercice 6**
Soient x la largeur, y la longueur et z la hauteur du wagon. $V = xyz$ et la somme des aires et du plancher vaut $A = xy + 2yz + 2xz$. Les côtés sont de longueur non nulle donc $xy > 0$ (par exemple) et $z = \dfrac{V}{xy}$. On doit donc minimiser la fonction

$$A(x, y) = xy + 2V\frac{(x+y)}{xy} = xy + \frac{2V}{y} + \frac{2V}{x}\,.$$

Le système d'optimalité est : $\left\{ \begin{array}{rcl} y^3 & = & 2V \\ x^3 & = & 2V \end{array} \right.$ et on obtient $x = y = \sqrt[3]{2V}$.

– **Exercice 7**
(a) Le problème s'écrit

$$\min J(a, b)\,,\ (a, b, c) \in \mathbb{R}^3\,,$$

où $J(a, b, c) = \displaystyle\sum_{i=1}^{n} (x_i - at_i^2 - bt_i - c)^2$. Les inconnues sont (a, b, c) et il n'y a pas de contraintes.
(b) Il y a solution unique si la matrice $A = \begin{bmatrix} S_4 & S_3 & S_2 \\ S_3 & S_2 & S_1 \\ S_2 & S_1 & N \end{bmatrix}$ associée à la forme quadratique est définie positive ($S_k = \displaystyle\sum_{i=1}^{N} t_i^k$) .
(c) Le système d'optimalité s'écrit : $A \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \displaystyle\sum_{i=1}^{N} x_i t_i^2 \\ \displaystyle\sum_{i=1}^{N} x_i t_i \\ \displaystyle\sum_{i=1}^{N} x_i \end{bmatrix}$.

– **Exercice 8**
(a) $q_1 = 20$ et $q_2 = 12$.
(b) $q_1 = 30$ et $q_2 = 16$.

– **Exercice 9**

1. Comme $e_k = x_k - \bar{x}$, on obtient

$$e_{k+1} = e_k + \alpha\, r_k = e_k + \alpha\,(b - A(e_k + \bar{x}))$$
$$= e_k + \alpha\,(\underbrace{b - A\bar{x}}_{=0} - A\, e_k) = (I - \alpha\, A)e_k\ .$$

On conclut par récurrence.

2. Dire que e_k converge vers 0 est équivalent à dire que le rayon spectral de $I - \alpha\, A$ est strictement inférieur à 1. Les valeurs propres de $I - \alpha\, A$ sont $(1 - \alpha\,\lambda_i)_{i=1,\ldots,n}$ où $(\lambda_i)_{i=1,\ldots,n}$ sont les valeurs propres de A rangées de manière croissante. On doit donc avoir : $\max\limits_{i=1,\ldots,n} |1 - \alpha\,\lambda_i| < 1$ c'est-à-dire :

$$-1 < 1 - \alpha\,\lambda_1 \leq \ldots 1 - \alpha\,\lambda_{n-1} \leq 1 - \alpha\,\lambda_n < 1\ .$$

On doit donc avoir d'une part $\alpha > 0$ (car les valeurs propres de A sont toutes strictement positives), et d'autre part

$$-\frac{2}{\alpha} < -\lambda_1 \leq \ldots - \lambda_{n-1} \leq -\lambda_n\ ,$$

c'est-à-dire $\dfrac{2}{\alpha} > \lambda_1 \geq \ldots \lambda_{n-1} \geq \lambda_n$; il suffit donc que $\alpha < \dfrac{2}{\lambda_1}$.

3. Le meilleur choix de α correspond au cas où le rayon spectral $\rho(I - \alpha\, A)$ est minimal. Or $\rho(I - \alpha\, A) = \max\{|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|\}$. Une résolution graphique montre que $\min\ \max\{|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|\}$ est atteint lorsque $1 - \alpha\lambda_n = \alpha\lambda_1 - 1$, c'est-à-dire lorsque $\alpha = \dfrac{2}{\lambda_1 + \lambda_n}$.

– **Exercice 10**

1. $A = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{c}{2} \end{pmatrix}$ et les valeurs propres sont $\dfrac{1}{2}$ et $\dfrac{c}{2}$.

2. $r = \begin{pmatrix} -\frac{x}{2} \\ -\frac{cy}{2} \end{pmatrix}$ et la fonctionnelle J associée est :

$$J(x, y) = \frac{1}{2} \begin{pmatrix} \frac{x}{2} \\ \frac{cy}{2} \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{4}(x^2 + cy^2)\ .$$

Cherchons le minimum de $\varphi(\alpha) = J(x_k + \alpha r_k)$.

$$\varphi(\alpha) = \frac{1}{4}\left[x_k^2(1 - \frac{\alpha}{2})^2 + c\, y_k^2(1 - \frac{\alpha\, c}{2})^2\right]\ .$$

On vérifie que le minimum de φ est atteint pour $\alpha = 2\,\dfrac{x_k^2 + c^2 y_k^2}{x_k^2 + c^3 y_k^2}$.

3. Un calcul élémentaire donne

$$x_{k+1} = \frac{c^2(c-1)x_k y_k^2}{x_k^2 + c^3 y_k^2} \qquad \text{et} \qquad y_{k+1} = -\frac{(c-1)x_k^2 y_k}{x_k^2 + c^3 y_k^2} \ .$$

4. $t_{k+1} = -\dfrac{1}{c^2 t_k}$ pour tout k. Donc $t_{k+2} = t_k$ ce qui signifie que les droites OP_{k+2} et OP_k sont parallèles. Par conséquent, pour tout k, les points $O,\ P_k$ et P_{k+2} sont alignés.

5. Calcul de τ^2 :

$$\frac{y_{k+1}}{y_k} = \frac{(1-c)x_k^2}{x_k^2 + c^3 y_k^2} = \frac{(1-c)}{1 + c^3 t_k^2} \ .$$

Donc

$$\tau^2 = \frac{(1-c)^2}{(1 + c^3 t_k^2)(1 + c^3 t_{k+1}^2)} = \frac{(1-c)^2}{(1 + c^3 t_k^2)(1 - \dfrac{1}{ct_k^2})} \ .$$

$$\tau^2 = \frac{(1-c)^2}{(1 + c^3 t^2)(1 - \dfrac{1}{ct^2})} \ .$$

On retrouve après calcul la formule annoncée. La quantité τ est maximale quand $(c\,t - \dfrac{1}{ct})^2$ est minimum c'est-à-dire nul. La valeur de t correspondante est $\dfrac{1}{c}$.

– **Exercice 11**

1. Evident
2. Comme A est symétrique (réelle) on peut trouver une base de vecteurs propres orthonormée ; ainsi $(u_1, \ldots, u_k)$ est une base orthonormée de W_k et $(u_{k+1}, \ldots, u_n)$ est une base orthonormée de $W_k^{\perp}$.

   Soit $x \in W_k$. On peut écrire $x = \sum_{i=1}^{k} x_i u_i$. Donc

$$(Ax, x) = \sum_{i=1}^{k} \lambda_i x_i^2 \geq \lambda_k \sum_{i=1}^{k} x_i^2 = \lambda_k \|x\|^2,$$

   puisque les valeurs propres sont rangées en ordre décroissant. Par conséquent :

$$\min_{x \in W_k} \frac{(Ax, x)}{\|x\|^2} \geq \lambda_k,$$

   et le min est atteint pour $x = u_k$.
3. Cette question se traite comme précédemment en décomposant $x \in W_k^{\perp}$ sur la base ad-hoc.

4. Soit $\eta = Ax - \lambda x$.

$$\|\eta\|_2^2 = \|Ax - \lambda x\|_2^2 = \sum_{i=1}^{n} (\lambda_i - \lambda)^2 x_i^2 \geq (\min_{1 \leq i \leq n} |\lambda_i - \lambda|)^2 \|x\|_2^2 \, .$$

On a donc

$$\min_{1 \leq i \leq n} |\lambda_i - \lambda| \leq \frac{\|\eta\|_2}{\|x\|_2}.$$

D'autre part $\|\eta\|_2^2 = \|Ax - \lambda x\|_2^2 = \|Ax\|_2^2 + \lambda^2 \|x\|_2^2 - 2\lambda \, (Ax, x)$. C'est un trinôme du second degré en λ dont on voit qu'il est minimal pour $\lambda = R_A(x)$.

– **Exercice 12**

1. Montrons que les $(v_i)_{0 \leq i \leq N-1}$ forment une famille libre (donc une base de $\mathbb{R}^N$).

$$\sum_{i=0}^{N-1} \alpha_i v_i = 0 \Longrightarrow \sum_{i=0}^{N-1} \alpha_i A \, v_i = 0 \Longrightarrow \forall j \quad \left(\sum_{i=0}^{N-1} \alpha_i A \, v_i, v_j \right) = 0$$

$$\Longrightarrow \forall j \quad \sum_{i=0}^{N-1} \alpha_i \underbrace{(A \, v_i, v_j)}_{=0 \text{ si } i \neq j} = 0 \Longrightarrow \forall j \;\; \alpha_j \underbrace{(A \, v_j, v_j)}_{\neq 0 \text{ car } v_j \neq 0} = 0 \Longrightarrow \forall j \;\; \alpha_j = 0 \, .$$

2. Soit $j \in \{0, \ldots, k-1\}$;

$$C_k A v_j = \sum_{i=0}^{k-1} \frac{v_i v_i^t A v_j}{(A \, v_i, v_i)} = \sum_{i=0}^{k-1} \frac{v_i \, (A \, v_j, v_i)}{(A \, v_i, v_i)} = v_j \, .$$

$$D_k v_j = v_j - C_k A v_j = 0 \, .$$

$$D_k^t A v_j = A v_j - A^t C_k^t A v_j \, ;$$

or C_k et A sont symétriques. Donc $A^t C_k^t A v_j = A C_k A v_j = A v_j$ et $D_k^t A v_j = 0$.
Quand $k = N$, D_N s'annule sur la base $(v_i)_{0 \leq i \leq N-1}$, donc $D_N = 0$ et $C_N = A^{-1}$.

3. Si $D_k = 0$ alors $C_k = A^{-1}$. Sinon on choisit v tel que $D_k(v) \neq 0$ (par exemple un des vecteurs de la base canonique de $\mathbb{R}^N$) et on pose : $v_k = D_k v$. Soit $j \in \{0, \ldots, k-1\}$;

$$(A \, v_k, v_j) = (A D_k v, v_j) = (D_k v, A v_j)$$

$$= (v, A v_j) - (C_k A v, A v_j) = (A v, v_j) - (A v, C_k A v_j) = (A v, v_j) - (A v, v_j) = 0 \, .$$

4. Algorithme
1. Initialisation : $v_0 \in \mathbb{R}^N - \{0\}$
2. Itération k : $v_0, \ldots, v_{k-1}$ connus.
Calcul de C_k et D_k .
– Si $D_k = 0$ alors $C_k = A^{-1}$
– Sinon, on choisit v comme dans la question précédente (par exemple le premier vecteur colonne non nul de D_k) et on continue.

– **Exercice 13**
Si on cherche les zéros de la fonction $f : x \mapsto \alpha - \dfrac{1}{x}$, la méthode de Newton donne précisément l'itération indiquée.
En généralisant aux matrices, l'algorithme devient :

$$\begin{cases} B_0 \text{ donnée,} \\ B_{k+1} = B_k(2\,I - AB_k)\ . \end{cases}$$

Si on pose $C_k = AB_k$, l'itération k s'écrit $C_{k+1} = 2\,C_k - C_k^2$, c'est-à-dire $I - C_{k+1} = (I - C_k)^2$ et par récurrence : $I - C_k = (I - C_0)^{2^k} = (I - AB_0)^{2^k}$. Donc la méthode converge si et seulement si $\rho(I - AB_0) < 1$.
Lorsque A est symétrique, définie positive elle admet N valeurs propres réelles strictement positives λ_i. Si on choisit pour $B_0 = \dfrac{1}{\rho(A)} I$, un calcul rapide montre que $\rho((I - AB_0) < 1$.

– **Exercice 14**
Dans le cas où les fonctions f_i sont affines on obtient

$$\partial_i f_j(x_1, x_2, \ldots, x_n) = a_{ij}\ .$$

La méthode de Newton relaxée s'écrit alors pour $1 \leq i \leq n$

$$x_i^{k+1} = x_i^k - \frac{\displaystyle\sum_{j=1}^{i-1} a_{ij} x_j^{k+1} + \sum_{j=i}^{n} a_{ij} x_j^k - b_i}{a_{ii}}\ ,$$

c'est-à-dire

$$\sum_{j=1}^{i} a_{ij} x_j^{k+1} = -\sum_{j=i+1}^{n} a_{ij} x_j^k + b_i\ .$$

On reconnait la forme matricielle : $(D - L)x^{k+1} = U\,x^k + b$, caractéristique de la méthode de Gauss-Seidel (voir [7]) où D est la diagonale de A, L (respectivement U) l'opposée de la partie triangulaire inférieure (respectivement supérieure) de A.

Chapitre 3

– **Exercice 1**
Il suffit de démontrer qu'on peut trouver trouver $d \neq 0 \in \mathbb{R}^n$ tel que

$$(\nabla h_i(x^*), d) = 0,\ i = 1, \ldots, p \ \text{ et } \ (\nabla g_j(x^*), d) < 0,\ \ j \in I(x^*)\ .$$

Cherchons d sous la forme $d = \displaystyle\sum_{i=1}^{p} d_i \nabla h_i(x^*) + \sum_{j \in I(x^*)} \delta_j \nabla g_j(x^*)$. On va imposer

$$(\nabla h_i(x^*), d) = 0,\ i = 1, \ldots, p \ \text{ et } \ (\nabla g_j(x^*), d) = -1,\ \ j \in I(x^*)\ ,$$

ce qui conduit au système linéaire suivant

$$\forall k = 1, \dots, p \qquad \sum_{i=1}^{p} d_i \left(\nabla h_i(x^*), \nabla h_k(x^*)\right) + \sum_{j \in I(x^*)} \delta_j \left(\nabla g_j(x^*), \nabla h_k(x^*)\right) = 0 \,,$$

$$\forall k \in I(x^*) \qquad \sum_{i=1}^{p} d_i \left(\nabla h_i(x^*), \nabla g_k(x^*)\right) + \sum_{j \in I(x^*)} \delta_j \left(\nabla g_j(x^*), \nabla g_k(x^*)\right) = -1.$$

La matrice associée à ce système est une matrice de Gram, de la forme $[(e_i, e_k)]_{j,k}$ où la famille (e_i) est une base. Elle est donc inversible. Par conséquent le système considéré admet une solution (unique) ce qui permet de trouver d.

– **Exercice 2**

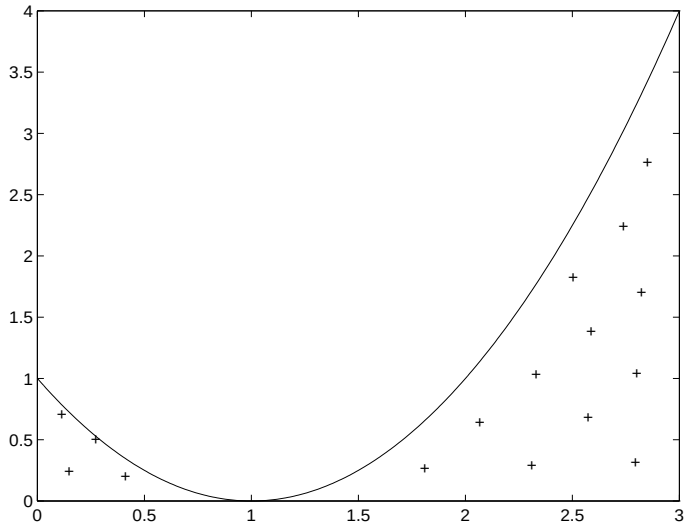

Figure B.1 : Ensemble des contraintes

Le point $x^* = (1, 0)$ est réalisable car les contraintes sont vérifiées. Calculons les gradients des contraintes actives en x^*.

$$g_1(x^*) = -x_1 \ , \ g_2(x^*) = -x_2 \ , \ g_3(x^*) = x_2 - (x_1 - 1)^2 \ .$$

$$\nabla g_2(x^*) = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \ , \ \nabla g_3(x^*) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \ .$$

Ces vecteurs ne sont pas indépendants car $\nabla g_3(x^*) = -\nabla g_2(x^*)$, donc x^* n'est pas régulier.

– **Exercice 3**

Soient $f \in V$ et g son projeté sur H (sous-espace vectoriel du Hilbert V). Une caractérisation de g est donnée par le théorème 3.4.1 :

$$\forall h \in H \qquad (f - g, h - g) \leq 0 \ .$$

Il suffit de choisir $h = g \pm \tilde{h}$ avec $\tilde{h}$ parcourant H pour obtenir l'équivalence annoncée.

– **Exercice 4**

1. Cf les relations (3.4.1) et (3.4.2).
2. La relation (3.4.1) avec $P(y) \in C$ donne $0 \leq (x - P(x), P(x) - P(y))$. De plus

$$(x-P(x), P(x)-P(y)) = (x-y, P(x)-P(y)) + \underbrace{(y-P(y), P(x)-P(y))}_{\leq 0 \text{ avec } (3.4.1)} - \|P(y)-P(x)\|^2.$$

3. Montrons l'inégalité de gauche :

$$\|y - P(y)\|^2 - \|x - P(x)\|^2 - 2 \ (x - P(x), y - x)$$

$$= \|y - x + x - P(x) + P(x) - P(y)\|^2 - \|x - P(x)\|^2 - 2\ (x - P(x), y - x)$$

$$= \|y - x + P(x) - P(y)\|^2 + 2\ \underbrace{(P(x) - P(y), x - P(x))}_{\geq 0}.$$

Montrons l'inégalité de droite :

$$\|y - P(y)\|^2 - \|x - P(x)\|^2 - 2\ (x - P(x), y - x)$$

$$= \|y - P(y)\|^2 - \left(\|y - P(x)\|^2 - \|y - x\|^2\right)$$

$$= \|y - P(y)\|^2 - \|y - P(y) + P(y) - P(x)\|^2 + \|y - x\|^2$$

$$= \|y-P(y)\|^2-\|y-P(y)\|^2-\|P(y)-P(x)\|^2+\|y-x\|^2-2\underbrace{(y-P(y), P(y)-P(x))}_{\geq 0 \text{ avec } (3.4.1)}$$

$$\leq \|y - x\|^2 - \|P(y) - P(x)\|^2 .$$

4. L'inégalité précédente avec $y = x + h$ et $h \in \mathbb{R}^n$ donne

$$0 \leq f(x + h) - f(x) - 2\ (x - P(x), h) \leq \|h\|^2 - \|P(x + h) - P(x)\|^2 \leq \|h\|^2 .$$

Donc f est différentiable et sa différentielle est $\nabla f(x) = 2\,(x - P(x))$.

5. La fonction g a pour expression

$$g(x_1, x_2) \begin{cases} 0 & \text{si } x_1 \geq 0 \text{ et } x_2 \geq 0 \\ |x_1| & \text{si } x_1 \leq 0 \text{ et } x_2 \geq 0 \\ |x_2| & \text{si } x_1 \geq 0 \text{ et } x_2 \leq 0 \\ \sqrt{x_1^2 + x_2^2} & \text{si } x_1 \leq 0 \text{ et } x_2 \leq 0\,. \end{cases}$$

Cette fonction n'est pas différentiable aux points de la frontière de C.

– **Exercice 5**
L'équation de la droite de régression D est : $x = at + b$ avec $a \simeq 0.512$ et $b \simeq -0.6$. Le point le plus éloigné de D est $M = (8, -2)$. La droite D' est : $x = at + b$ avec $a \simeq 0.715$ et $b \simeq -1.045$.
La droite Δ se calcule en posant $b = 0$ et $a = \dfrac{S_{xt}}{S_{t^2}} \simeq 0.427$ puis $b = 1$ et $a \simeq 0.28416$ en utilisant les relations de Karush-Kuhn-Tucker.

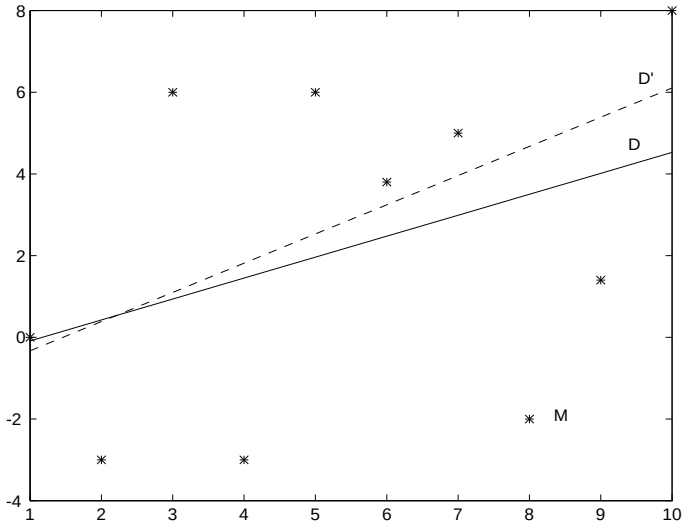

Figure B.2 : Droite de régression

– **Exercice 6**

$$\nabla J(y) = Ay + b = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2y_1 - y_2 + 3 \\ -y_1 + 2y_2 - y_3 - 1 \\ -y_2 + 2y_3 + 2 \end{bmatrix}.$$

Le problème s'écrit

$$(\mathcal{P}) \begin{cases} \min\ J(y) \\ g(y) = -y_1 \leq 0 \\ h(y) = y_2 + y_3 = 0 \\ y \in \mathbb{R}^3 \end{cases}$$

$$\nabla g(y) = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}\ ,\ \nabla h(y) \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}\ .$$

1. Le problème a une solution unique car la fonction J est quadratique avec A définie positive.
2. Les relations de KKT donnent $\lambda \in \mathbb{R}\ ,\ \mu \in \mathbb{R}$ et

$$\lambda \geq 0 \tag{B.2a}$$

$$y_1 \geq 0 \text{ et } y_2 + y_3 = 0\ . \tag{B.2b}$$

$$\lambda y_1 = 0\ . \tag{B.2c}$$

$$\nabla J(y) + \lambda \nabla g(y) + \mu \nabla h(y) = 0. \tag{B.2d}$$

c'est à dire

$$y_1 \geq 0,\ \lambda \geq 0\ ,\ \lambda y_1 = 0\ . \tag{B.3a}$$

$$y_2 + y_3 = 0 \tag{B.3b}$$

$$\begin{cases} 2y_1 - y_2 + 3 - \lambda & = 0 \\ -y_1 + 2y_2 - y_3 - 1 + \mu & = 0 \\ -y_2 + 2y_3 + 2 + \mu & = 0 \end{cases} . \tag{B.3c}$$

Si on reporte (B.3b) dans (B.3c) et on obtient

$$\begin{cases} 2y_1 - y_2 & = -3 + \lambda \\ -y_1 + 6y_2 & = 3 \end{cases} . \tag{B.4}$$

- Si $y_1 > 0$, avec (B.3a) on obtient $\lambda = 0$. On résout alors (B.4) ce qui donne $y_2 = 3/11$ et $y_1 = -15/11$. On obtient une contradiction : ce cas est impossible.
- On a donc $y_1 = 0$ et (B.4) donne $y_2 = 1/2$ puis $y_3 = -1/2$. La solution est donc $(0, 0.5, -0.5)$.

– **Exercice 7**
Maximiser la fonction J revient à minimiser la fonction $-J(x, y)$ c'est-à-dire (comme la constante ne sert à rien dans une minimisation) à minimiser $F(x, y) = x^2 + y^2 - 14x - 6y$. Cette fonction est quadratique, associée à la matrice $A = 2I$ qui est définie positive. Par conséquent, le problème a une solution unique.

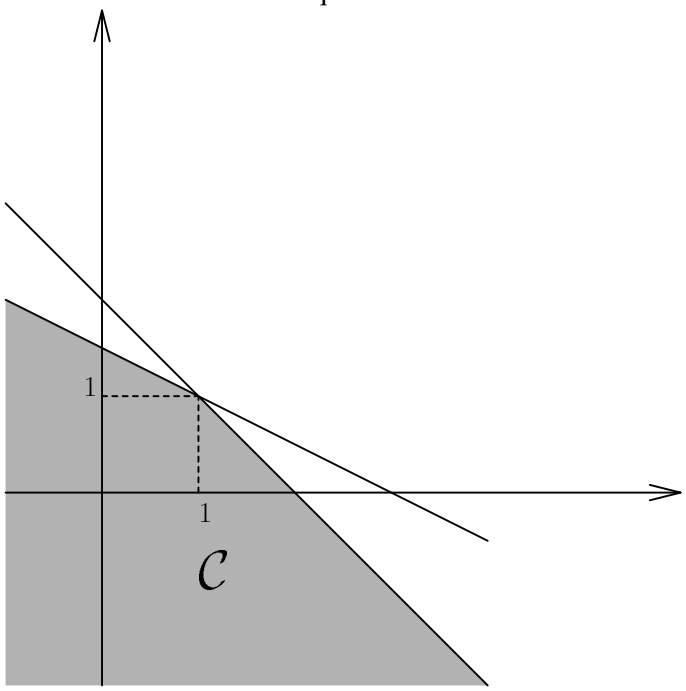

Figure B.3 : Ensemble des contraintes

Les relations de KKT appliquées à ce problème donnent

$$\begin{cases} \lambda_1 \geq 0, & \lambda_2 \geq 0 \\ x + y - 2 \leq 0, & x + 2y - 3 \leq 0 \\ \lambda_1(x + y - 2) = 0, & \lambda_2(x + 2y - 3) = 0 \\ 2x - 14 + \lambda_1 + \lambda_2 = 0, & 2y - 6 + \lambda_1 + 2\lambda_2 = 0 \, . \end{cases} \tag{B.5}$$

On obtient alors successivement

$$x = -\frac{\lambda_1 + \lambda_2 - 14}{2} \ , \ y = -\frac{\lambda_1 + 2\lambda_2 - 6}{2}$$

$$\lambda_1(-\lambda_1 - \frac{3\lambda_2}{2} + 8) = 0 \ , \ \lambda_2(-\lambda_1 - 5\lambda_2 + 20) = 0 \ .$$

Si $\lambda_1 \neq 0$, la relation précédente donne $\lambda_1 = -\frac{3\lambda_2}{2} + 8$ puis $\lambda_2(-\lambda_2 - 8) = 0$. Comme λ_2 est positif on a nécessairement $\lambda_2 = 0$. On obtient alors $\lambda_1 = 8, \ x = 3 \ , y = -1$. Toutes les relations de KKT sont satisfaites. Comme la fonction est strictement convexe elles sont aussi suffisantes. Par conséquent le point trouvé est bien la solution

– **Exercice 8**
Pour trouver le minimum sans contraintes on résout le système $Ax = b$. On obtient $x_0 = (-3/2, -1/2, -1/2)$ qui n'est pas réalisable. Le minimum avec contraintes (s'il existe) sera donc différent.
Une fois de plus la résolution des relations de KKT fournit une solution : $x^* = (1, 1, 0)$. On peut même vérifier que $J(x_0) = -3/4 < J(x^*) = 1/2$.

– **Exercice 9**

1. Maximiser f revient à minimiser $-f$; on s'intéresse donc à $g = -f$.
Les lignes de niveau de g sont des paraboles :plus précisément

$$g(x_1, x_2) = \alpha \Longleftrightarrow x_2 = (x_1 - 1)^2 + 1 - \alpha \ .$$

On veut trouver α le plus petit possible, donc x_2 le plus grand possible tout en restant dans l'ensembles de contraintes K. On translate donc la parabole $(\mathcal{P})_\alpha$ vers "le haut" jusqu'au moment où on "sort" de K, c'est-à-dire quand elle est tangente à la droite Δ d'équation $x_2 = 1 - x_1$. Le point de tangence est la solution cherchée. Il faut donc écrire que ce point est $(\mathcal{P})_\alpha \cap \Delta$ et que les dérivées sont égales en x_1 ; cela donne d'une part

$$x_2 = (x_1 - 1)^2 + 1 - \alpha \ \text{ et } x_2 = 1 - x_1 \ ,$$

c'est-à-dire

$$(x_1 - 1)^2 + x_1 - \alpha = 0 \ ;$$

d'autre part

$$-1 = 2(x_1 - 1) \text{ c'est-à-dire } x_1 = 0.5 \ .$$

On en conclut que la solution est $(0.5, 0.5)$.

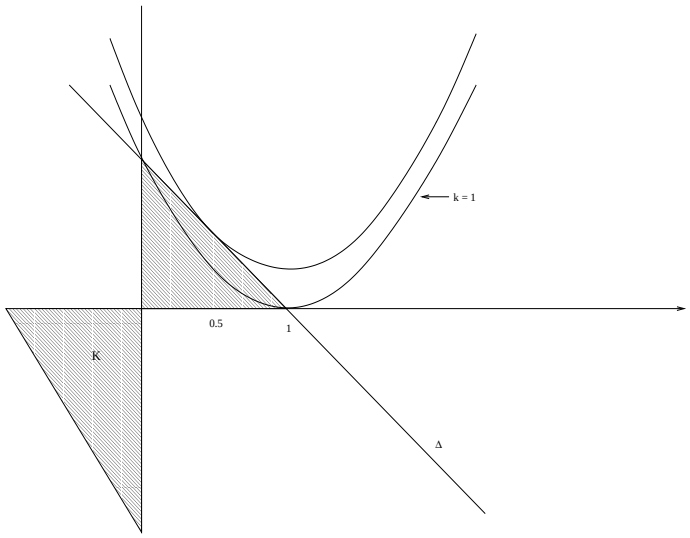

Figure B.4 : Résolution graphique

2. Ecrivons a priori les équations de KKT (on pourra vérifier a posteriori que la solution trouvée est bien un point régulier) :

$$x_1 + x_2 - 1 \leq 0 \text{ et } -x_1 x_2 \leq 0 \,, \tag{B.6a}$$

$$\nabla g(x_1, x_2) + \lambda_1 \nabla(x_1 + x_2 - 1) - \lambda_2 \nabla(x_1 x_2) = 0 \,, \tag{B.6b}$$

$$\lambda_i \geq 0 \,, i = 1, 2 \,, \tag{B.6c}$$

$$\lambda_1(x_1 + x_2 - 1) = 0 \quad \text{et} \quad \lambda_2 x_1 x_2 = 0 \,. \tag{B.6d}$$

L'équation (B.6b) est équivalente à

$$2x_1 - 2 + \lambda_1 - \lambda_2 \, x_2 = 0 \quad \text{et} \quad -1 + \lambda_1 - \lambda_2 x_1 = 0 \,.$$

- Si $x_1 = 0$, on montre successivement que $\lambda_1 = 1,\ x_2 = 1$ et $\lambda_2 = -1$ ce qui est impossible.
- Si $x_2 = 0$, on montre successivement que $x_1 = 1,\ \lambda_1 = 0$ et $\lambda_2 = -1$ ce qui est impossible.
- Finalement $x_1 x_2 \neq 0$, donc $\lambda_2 = 0$; on en déduit que $x_1 = 0.5$ puis que $x_2 = 0.5$.

3. On ne peut pas appliquer la méthode d'Uzawa (manque de convexité).

4. Se fait comme précédemment.

– **Exercice 10**
La droite de régression D est donnée par $x = at + b$ avec $a \simeq 7.6836$ et $b \simeq -13.161$.
En réalité la concentration initiale est positive. La droite de régression D' est alors donnée par $x = at + b$ avec $a \simeq 5.96$ et $b = 0$.

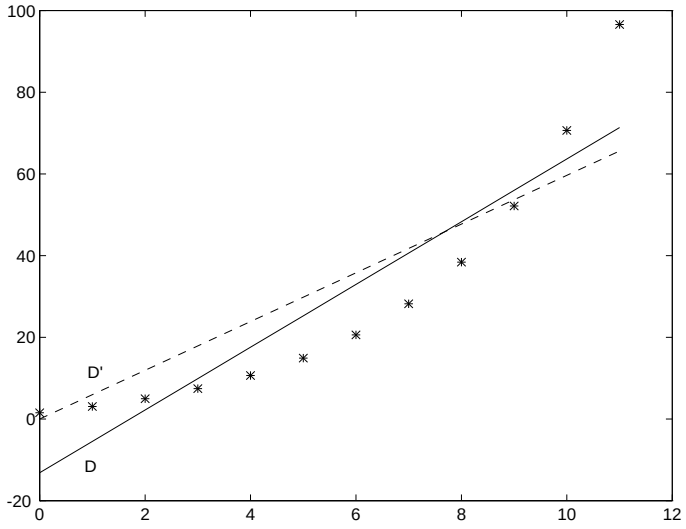

Figure B.5 : Droites de régression

On voit sur le dessin que la courbe des données est très loin de la droite de régression qui est la "meilleure possible". La loi n'est donc pas linéaire. En réalité, elle est de la forme $c(t) = ae^t + be^{-t}$, sachant que la concentration initiale est positive. Il faut donc résoudre le problème de moindres carrés suivant :

$$\min\{J(a,b) = \sum_{i=0}^{11}[ae^{t_i} + be^{-t_i} - x_i]^2 \ \mid b \geq 0\}$$

où les couples (t_i, x_i) sont donnés par le tableau. On écrit ensuite les relations de KKT que l'on résout.

– **Exercice 11**

On se place dans $\mathbb{R}^2$, et on note $x = (x_1, x_2)$.

$$f(x) = \frac{1}{2}(Bx, x) + (b, x) = \frac{1}{2}(x_1^2 + \alpha x_2^2) + x_1 \ .$$

1. $B = \begin{bmatrix} 1 & 0 \\ 0 & \alpha \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ et $\nabla f(x) = Bx + b = \begin{bmatrix} x_1 + 1 \\ \alpha x_2 \end{bmatrix}$.
2. Tout minimum est un point stationnaire et doit donc vérifier $Bx = -b$, c'est-à-dire :

$$x_1 = -1 \text{ et } \alpha x_2 = 0 \ .$$

(a) Si $\alpha = 0$, la matrice B n'est pas définie. On n'a donc pas unicité. On peut vérifier que tous les éléments de la forme $(-1, x_2)$ minimisent f. En effet

$$\forall (x_1, x_2) \in \mathbb{R}^2 \ \ f(x_1, x_2) = \frac{1}{2}(x_1 + 1)^2 - \frac{1}{2} \geq -\frac{1}{2} = f(-1, x_2) \ .$$

Il y a donc une infinité de minima. Si $\alpha \neq 0$ le seul élément susceptible de réaliser le minimum est $x^* = (-1, 0)$.

(b) Si $\alpha > 0$, la matrice B est définie positive et donc x^* est le minimum (unique).

(c) Si $\alpha < 0$, la matrice B n'est pas positive et d'après la condition nécessaire du second ordre, il ne peut y avoir de minimum.

3. $\alpha = 2$. Alors $\nabla f(x) = \begin{bmatrix} x_1 + 1 \\ 2x_2 \end{bmatrix}$.

(a) Le problème avec contraintes admet une solution unique car B est définie positive (et donc f est strictement convexe, coercive) et l'ensemble des contraintes

$$\mathcal{C} = \{ (x_1, x_2) \in \mathbb{R}^2 \mid \sqrt{x_1^2 + x_2^2} \leq \frac{1}{2} \}$$

est non vide, fermé (et même borné).

(b) Pour écrire les conditions de KKT nous allons écrire la contrainte sous la forme équivalente

$$(x_1, x_2) \in \mathcal{C} \Longleftrightarrow x_1^2 + x_2^2 \leq \frac{1}{4} .$$

Soit $\lambda \geq 0$ le multiplicateur associé à cette contrainte ; on obtient

$$x_1^2 + x_2^2 \leq \frac{1}{4} \ , \ \lambda(x_1^2 + x_2^2 - \frac{1}{4}) = 0 ,$$

$$x_1 + 1 + 2\lambda x_1 = 0 \ , \ 2x_2 + 2\lambda x_2 = 0 .$$

Comme λ est positif on a donc $2 + 2\lambda \neq 0$ et $x_2 = 0$. Donc $x_1 \in [-\frac{1}{2}, \frac{1}{2}]$.

Si $\lambda = 0$ alors $x_1 = -1$ ce qui est impossible. Donc $\lambda \neq 0$ et $x_1 = \frac{1}{2}$ ou $x_1 = -\frac{1}{2}$. Si $x_1 = \frac{1}{2}$ alors $\lambda = -\frac{3}{2}$ ce qui est impossible.

Donc $x_1 = -\frac{1}{2}$ (et dans ce cas $\lambda = \frac{1}{2}$). La solution est donc $x^* = (-0.5, 0)$.

– **Exercice 12**

(a) g^+ est le max de deux fonctions convexes, donc convexe. $(g^+)^2$ est la composée deux fonctions convexes dont la première est croissante : elle est donc convexe.
La dérivée de $s \mapsto (s^+)^2$ est $2s^+$ et celle de $(g^+)^2$ est $2g'.g^+$.
(b) Classique, car J est continue et coercive.
(c) J_ε est coercive comme J car supérieure à J. Elle est aussi continue et strictement convexe. Donc $(\mathcal{P}_\varepsilon)$ a une solution unique x_ε.
(d) J_ε est différentiable, donc $\nabla J(x_\varepsilon) = 0$. On obtient la relation $(\mathcal{E}_\varepsilon)$ désirée.
(e) Soit α la plus petite valeur propre de A. On a, pour tout $x \in K$

$$\frac{\alpha}{2}\|x_\varepsilon\|^2 - (b, x_\varepsilon) + \sum_{i=1}^{m} \left(\frac{g_i^+(x_\varepsilon)}{\sqrt{\varepsilon}} \right)^2 \leq J_\varepsilon(x_\varepsilon) \leq J_\varepsilon(x) = J(x) . \tag{B.7}$$

En particulier, pour $x = \bar{x}$: $\frac{\alpha}{2}\|x_\varepsilon\|^2 - (b, x_\varepsilon) \leq J(\bar{x}) = \bar{\jmath} < +\infty$.

Donc $\alpha\|x_\varepsilon\|^2 - \|b\| \, \|x_\varepsilon\| - \bar{\jmath} \leq 0$. Cela entraîne que x_ε est borné indépendamment de ε.

De même, avec (B.7), on voit que $\dfrac{g_i^+(x_\varepsilon)}{\sqrt{\varepsilon}}$ reste borné indépendamment de ε, pour tout i.

(f) On peut extraire de (x_ε) une sous-suite $(x_n \stackrel{\text{def}}{=} x_{\varepsilon_n})$ qui converge vers $\tilde{x}$. D'après (e), $g_i^+(x_n)$ converge vers $g_i^+(\tilde{x})$ d'une part et vers 0 d'autre part. Donc $\tilde{x} \in K$. De plus,

$$J(x_n) \leq J_{\varepsilon_n}(x_n) \leq J_{\varepsilon_n}(\bar{x}) = J(\bar{x}) \ ,$$

implique $J(\tilde{x}) \leq J(\bar{x})$. Par conséquent, $\tilde{x}$ est une solution de $(\mathcal{P})$. Par unicité de cette solution $\tilde{x} = \bar{x}$. Le raisonnement étant valable pour toute valeur d'adhérence de la famille (x_ε), il n'y a donc qu'une seule valeur d'adhérence et la famille entière converge vers $\bar{x}$ quand $\varepsilon \to 0$.

(g) Il suffit de chercher $d_i = \sum_{i=1}^{m} \delta_j c_j$ en résolvant le système $\sum_{i=1}^{m} \delta_j (c_j, c_k) = 0$ pour $k \neq i$ et $\sum_{i=1}^{m} \delta_j (c_j, c_i) = 1$. La matrice associée est une matrice de Gram inversible car les (c_i) sont indépendants, d'où l'existence de d_i qu'on peut ensuite normer.

En multipliant $(\mathcal{E}_\varepsilon)$ par d_i on obtient $h_\varepsilon^i = \dfrac{(b, d_i) - (Ax_\varepsilon, d_i)}{(c_i, d_i)}$ qui est borné puisque x_ε l'est.

(h) On peut donc extraire (par un procédé diagonal) une sous-suite de la famille $(h_\varepsilon^i)_{\varepsilon,i}$ telle que chaque $h_{\varepsilon_n}^i$ converge vers un réel h^i. On passe ensuite à la limite dans $(\mathcal{E}_\varepsilon)$ pour obtenir $(\mathcal{E})$. Remarquons aussi que chaque h_ε^i est positif ; donc $h^i \geq 0$ pour tout i. Enfin, si $g_i(\bar{x}) < 0$, pour tout ε_n assez petit on a $g_i^+(x_{\varepsilon_n}) = 0$ et donc $h_{\varepsilon_n}^i = 0$. A la limite $h^i = 0$. Ceci permet de retrouver la condition de complémentarité.

– **Exercice 13**

1. Dire que x^* est régulier revient à dire que $h'(x^*) \neq 0$ (et $h(x^*) = 0$). Le système d'optimalité permettant de calculer x^* est

$$(\mathcal{S}_0) \qquad \begin{cases} f'(x^*) + \lambda^* h'(x^*) = 0 \\ h(x^*) = 0 \ . \end{cases}$$

2. Le système d'optimalité permettant de calculer une éventuelle solution de $(\mathcal{P}_t)$ est

$$(\mathcal{S}_t) \qquad \begin{cases} f'(x) + \lambda h'(x) = 0 \\ h(x) = t \ . \end{cases}$$

3. Soit la fonction G définie de $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$ dans $\mathbb{R} \times \mathbb{R}$ par

$$G(x, \lambda, t) = (f'(x) + \lambda h'(x), h(x) - t) \ .$$

$G(x^*, \lambda^*, 0) = (0, 0)$ et la différentielle de G en $(x^*, \lambda^*, 0)$ par rapport à (x^*, λ^*) est $H(x^*, \lambda^*)$ qui est inversible car x^* est régulier. On peut donc appliquer le théorème des fonctions implicites.

4. $F(t) = f(x(t))$. Le théorème des fonctions implicites et la composition des fonctions dérivables donne
$$F'(0) = f'(x(0)).x'(0) = f'(x^*).x'(0) \ .$$
De plus : $1 = (h\,o\,x)'(0) = h'(x(0)).x'(0) = h'(x^*).x'(0)$. Donc $x'(0) = \dfrac{1}{h'(x^*)}$ et $F'(0) = \dfrac{f'(x^*)}{h'(x^*)} = -\lambda^*$ d'après $(\mathcal{S}_0)$.

– **Exercice 14**

1. Le raisonnement est le même pour les sous-questions (i) et (ii) (on le fait sur $-F$ pour (ii)).
L'application $u \in U \mapsto \mathcal{L}(u, v)$ est strictement convexe et U est un convexe, fermé et borné. Donc l'inf est atteint en un point unique noté $\varphi(v)$.
(iii) Soit v_n une suite maximisante, c'est-à-dire $v_n \in V$ et $G(v_n) \to m = \sup_{v\in V} G(v)$. Comme V est compact, on peut extraire de v_n une sous-suite (encore notée v_n) convergente vers $v^* \in V$. (Remarquons qu'on pourrait conclure si on savait que G est continue, ce qu'on n'a pas montré.) Soit u fixé dans U ; comme $v \in V \mapsto \mathcal{L}(u, v)$ est continue on a : $\lim_{n\to+\infty} \mathcal{L}(u, v_n) = \mathcal{L}(u, v^*)$.
Or $G(v_n) \leq \mathcal{L}(u, v_n)$ pour tout u ; donc, en passant à la limite on obtient
$$m = \sup_{v\in V} G(v) \leq \mathcal{L}(u, v^*) \text{ pour tout } u \in U \ .$$
Donc $m \leq \inf_{u\in U} \mathcal{L}(u, v^*) = G(v^*)$ lequel est inférieur à m . D'où la conclusion.
(iv) Posons $v_n = (1 - \dfrac{1}{n})v^* + \dfrac{1}{n}v$. La suite $u_n = \varphi(v_n)$ est dans U qui est compact. On peut donc en extraire une sous-suite (notée de la même façon) convergente vers $u^* \in U$. Comme $G(v^*) = \mathcal{L}(\varphi(v^*), v^*) \leq \mathcal{L}(u, v^*)$ pour tout $u \in U$, on a en particulier $G(v^*) \leq \mathcal{L}(u^*, v^*)$. D'autre part, $\mathcal{L}(u, \cdot)$ est concave donc
$$G(v^*) \geq G(v_n) = \mathcal{L}(u_n, v_n) \geq (1 - \frac{1}{n})\mathcal{L}(u_n, v^*) + \frac{1}{n}\mathcal{L}(u_n, v) \ .$$
Par continuité de $\mathcal{L}(\cdot, v)$ et par passage à la limite, nous obtenons $G(v^*) \geq \mathcal{L}(u^*, v^*)$. Par conséquent
$$G(v^*) = \mathcal{L}(u^*, v^*) = \mathcal{L}(\varphi(v^*), v^*) \ ;$$
par unicité $u^* = \varphi(v^*)$. Cela prouve aussi que la suite n'a qu'une seule valeur d'adhérence et qu'elle converge entièrement. Montrons maintenant que pour tous $u \in U$ et $v \in V$
$$\mathcal{L}(u^*, v) \leq \mathcal{L}(u^*, v^*) \leq \mathcal{L}(u, v^*) \ .$$
On a déjà $\mathcal{L}(u^*, v^*) = G(v^*) \leq \mathcal{L}(u, v^*)$ pour tout $u \in U$. D'autre part,
$$\begin{aligned} \mathcal{L}(u^*, v^*) = G(v^*) &\geq \mathcal{L}(u_n, v_n) \\ &\geq (1 - \frac{1}{n})\mathcal{L}(u_n, v^*) + \frac{1}{n}\mathcal{L}(u_n, v) \\ &\geq (1 - \frac{1}{n})G(v^*) + \frac{1}{n}\mathcal{L}(u_n, v) \ . \end{aligned}$$
Donc $G(v^*) \geq \mathcal{L}(u_n, v)$ pour tout n et par passage à la limite $G(v^*) \geq \mathcal{L}(u^*, v)$.

2. On applique ce qui précède à $\mathcal{L}_n$ qui est strictement convexe par rapport à u. Donc il existe un point selle (u_n^*, v_n^*) de $\mathcal{L}_n$ sur $U \times V$, c'est-à-dire

$$\mathcal{L}(u_n^*, v) + \frac{1}{n}\|u_n^*\|^2 \leq \mathcal{L}(u_n^*, v_n^*) + \frac{1}{n}\|u_n^*\|^2 \leq \mathcal{L}(u, v_n^*) + \frac{1}{n}\|u\|^2 \, ;$$

comme U et V sont compacts on peut extraire des sous-suites convergentes (notées de la même façon) vers u^* et v^* respectivement. Comme en particulier

$$\mathcal{L}(u_n^*, v) \leq \mathcal{L}(u_n^*, v) + \frac{1}{n}\|u_n^*\|^2 \leq \mathcal{L}(u, v_n^*) + \frac{1}{n}\|u\|^2 \, ,$$

et avec la continuité des différentes applications, on peut passer à la limite et obtenir

$$\forall u \in U, \; \forall v \in V \qquad \mathcal{L}(u^*, v) \leq \mathcal{L}(u, v^*) \, ,$$

ce qui est exactement dire que (u^*, v^*) est un point selle de $\mathcal{L}$ sur $U \times V$.

– **Exercice 15**
Le système d'optimalité permettant de calculer une solution u du problème est

$$(\mathcal{S}) \qquad \left\{ \begin{array}{l} A\,u + C^t\lambda = 0 \\ C\,u = 0 \, . \end{array} \right.$$

(a) L'algorithme d'Uzawa s'écrit ici :

1. **Initialisation**
   $k = 0$: choix de $\lambda^o \in \mathbb{R}^m$
2. **Itération** k :
   $\lambda^k \in \mathbb{R}^m$ est connu ; calcul de $u^k = -A^{-1}C^t\lambda^k$.
3. Calcul de $\lambda^{k+1} = \lambda^k + \rho C u^k$.
4. Critère d'arrêt

(b) Voir l'exercice 9. du chapitre 2.
(c) Grâce à $(\mathcal{S})$, il existe au moins un vecteur $\lambda \in \mathbb{R}^m$ vérifiant $Au + C^t\lambda = b$.
D'une part $\lambda^{k+1} = \lambda^k + \rho_1\rho_2 C u^{k+1}$ par l'algorithme, et $\lambda = \lambda + \rho_1\rho_2 C u$, d'autre part. Donc

$$\|\lambda^{k+1} - \lambda\|_m^2 = \|\lambda^k - \lambda\|_m^2 + \rho_1^2\rho_2^2\|C(u^{k+1} - u)\|_m^2 + 2\,\rho_1\rho_2\left(\lambda^k - \lambda, C(u^{k+1} - u)\right)_m$$

$$\begin{aligned} \|\lambda^{k+1} - \lambda\|_m^2 \leq \; & \|\lambda^k - \lambda\|_m^2 + (\rho_1\rho_2)^2\,\|CC^t\|\,\|u^{k+1} - u\|_n^2 \\ & +2\rho_1\rho_2\left(C^t(\lambda^k - \lambda), u^{k+1} - u\right)_n \, . \end{aligned}$$

Comme

$$u^{k+1} = u^k - \rho_1(Au^k - b + C^t\lambda^k) \text{ et } u = u - \rho_1(Au - b + C^t\lambda)$$

nous avons

$$\rho_1 C^\top(\lambda^k - \lambda) = (I - \rho_1\,A)(u^k - u) - (u^{k+1} - u)$$

et donc

$$\begin{aligned}\|\lambda^{k+1}-\lambda\|_m^2 \leq \quad & \|\lambda^k-\lambda\|_m^2+(\rho_1\rho_2)^2\|CC^t\|-2\rho_2)\,\|u^{k+1}-u\|_n^2\\ & +2\rho_2\beta\|u^{k+1}-u\|_n\,\|u^k-u\|_n\,,\end{aligned}$$

$$\begin{aligned}\|\lambda^{k+1}-\lambda\|_m^2 \leq \quad & \|\lambda^k-\lambda\|_m^2+\rho_2(\rho_1^2\rho_2)\|CC^t\|-2)\,\|u^{k+1}-u\|_n^2\\ & +\rho_2\beta(\|u^{k+1}-u\|_n^2+\|u^k-u\|_n^2)\,;\end{aligned}$$

D'où

$$\begin{aligned}0\leq \quad & \frac{\|\lambda^k-\lambda\|_m^2}{\rho_2}-\frac{\|\lambda^{k+1}-\lambda\|_m^2}{\rho_2}+\left(\rho_1^2\rho_2\|CC^t\|-2+2\beta\right)\,\|u^{k+1}-u\|_n^2\\ & -\beta(\|u^{k+1}-u\|_n^2+\|u^k-u\|_n^2)\,,\end{aligned}$$

c'est-à-dire en posant $\gamma=-(\rho_1^2\rho_2\|CC^t\|-2+2\beta)$

$$\gamma\|u^{k+1}-u\|_n^2\leq\left(\frac{\|\lambda^k-\lambda\|_m^2}{\rho_2}+\beta\|u^k-u\|_n^2\right)-\left(\frac{\|\lambda^{k+1}-\lambda\|_m^2}{\rho_2}+\beta\|u^{k+1}-u\|_n^2\right)\,.$$

Si on choisit $\rho_2<\dfrac{2(1-\beta)}{\rho_1^2\|CC^t\|}$ alors $\gamma>0$.

(d) Posons alors $\alpha_k=\dfrac{\|\lambda^k-\lambda\|_m^2}{\rho_2}+\beta\|u^k-u\|_n^2$. Ce qui précède montre que (α_k) est une suite positive et décroissante. Donc elle converge et $\alpha_{k+1}-\alpha_k\to 0$. L'inégalité de la question (c) montre alors que $u^k\to u$ car $\|u^k-u\|_n^2\leq\dfrac{\alpha_k-\alpha_{k+1}}{\gamma}$.

(e) A priori on ne peut rien dire de la suite (λ_k) sauf si le rang de C est m. Dans ce cas $\lambda^k\to\lambda$ car $C^t\lambda^k=-Au^k\to -Au=C^t\lambda$.

– **Exercice 16**

(a) J est continue, coercive et strictement convexe et l'ensemble des contraintes est fermé : le problème admet donc une solution unique.

(b) Comme les vecteurs lignes de B sont indépendants, la condition de régularité est satisfaite pour tout point. Donc toute solution vérifie les conditions de KKT qui sont les relations (2). C'est une condition nécessaire et suffisante car J est convexe.

(c) C'est le théorème 3.5.5.

(d) Si (u,p) est point selle de $\mathcal{L}$ alors

$$\forall v\in\mathbb{R}^N,\forall q\in\mathbb{R}^M\qquad \mathcal{L}(u,q)\leq\mathcal{L}(u,p)\leq\mathcal{L}(v,p)\,.$$

Cela entraîne en particulier que $Bu=0$ et donc $\mathcal{L}_r(u,p)=\mathcal{L}(u,p)$ et $\mathcal{L}_r(u,q)=\mathcal{L}(u,q)$. Comme $\mathcal{L}(v,p)\leq\mathcal{L}_r(v,p)$ on obtient que (u,p) est point selle de $\mathcal{L}_r$ également.
Réciproquement, si (u,p) est point selle de $\mathcal{L}_r$, par convexité et grâce au théorème 3.5.5. on déduit que u réalise le minimum de $J_r=J+\dfrac{r}{2}\|B\cdot\|_M^2$ sous la contrainte $Bu=0$. Donc, de nouveau $\mathcal{L}_r(u,p)=\mathcal{L}(u,p)$ et $\mathcal{L}_r(u,q)=\mathcal{L}(u,q)$. La troisième inégalité s'en déduit.

(e) On reconnait l'algorithme d'Uzawa. La relation (3) correspond aux relations de KKT

avec J_r au lieu de J.
(f) La relation (i) s'obtient en prenant la norme au carré de $\bar{p}^{n+1} = \bar{p}^n + \rho_n B\bar{u}^n$. La relation (ii) se déduit de (6)- (3) qui donne $A\bar{u}^n + rB^tB\bar{u}^n + B^t\bar{p}^n = 0$; il suffit ensuite de prendre le produit scalaire avec $\bar{u}^n$. (iii) est immédiate.
(g) Immédiat
(h) La suite $\|\bar{p}^n\|_M^2$ est décroissante, minorée donc convergente. La différence $\|\bar{p}^n\|_M^2 - \|\bar{p}^{n+1}\|_M^2$ tend vers 0. Donc $\lim_{n\to+\infty} \langle A\bar{u}^n, \bar{u}^n\rangle_N = 0$. et $u_n \to u$.
(i) Similaire à ce qui précède

– **Exercice 17**
(a) Le Lagrangien est

$$\mathcal{L}(y,\mu) = \frac{1}{2}\,(\,Ay, y\,)_n - (b, y\,)_n + \sum_{i=1}^{p} \mu_i\ [(\,c_i, y\,)_n - \gamma_i]\ .$$

Cette fonction est strictement convexe et coercive par rapport à la variable y et donc le problème

$$h(\mu) = \min_{y\in\mathbb{R}^n} \mathcal{L}(y,\mu)\ ,$$

a une solution unique y_μ caractérisée par $\nabla_y \mathcal{L}(y_\mu,\mu) = 0$.
Donc $y_\mu = A^{-1}\left(b - \sum_{i=1}^{p} \mu_i\, c_i\right)$ et en reportant dans $\mathcal{L}(y_\mu,\mu)$, on obtient

$$h(\mu) = \mathcal{L}(y_\mu,\mu) = -\frac{1}{2}\,(B\mu,\mu)_p + (d,\mu)_p + \alpha\ ,$$

avec

$$B = \left(\left(\,c_i, A^{-1}c_j\,\right)_n\right)_{1\le i,j\le p},\ \ d = \left(\left(\,c_i, A^{-1}b\,\right)_n - \gamma_i\right)_{1\le i\le p} \text{ et } \alpha = \frac{1}{2}\left(b, A^{-1}b\,\right)_n\ .$$

(b) $\lambda \in \mathbb{R}^p$ vérifie les conditions de KKT pour le problème $(\mathcal{P})$ (dont la solution est y^*), si et seulement si

$$Ay^* - b + \sum_{i=1}^{p} \lambda_i\, c_i = 0 \text{ et } (\,c_j, y^*\,)_n = \gamma_j,\ j = 1,\dots,p\ ,$$

c'est-à-dire si et seulement si $y^* = A^{-1}\left(b - \sum_{i=1}^{p} \lambda_i\, c_i\right)$ et

$$\left(\,c_j, A^{-1}b\right)_n - \sum_{i=1}^{p} \lambda_i\ \left(\,c_j, A^{-1}c_i\,\right)_n = \gamma_j,\ j = 1,\dots,p\ ,$$

ce qui revient à $-B\,\lambda = d$, c'est-à-dire $\nabla h(\lambda) = 0$ qui est la condition nécessaire et suffisante pour que λ soit solution du problème dual

$$h(\lambda) = \max_{\mu\in\mathbb{R}^p} h(\mu) \qquad (\mathcal{D})$$

(c) L'algorithme d'Uzawa pour le problème $(\mathcal{P})$ s'écrit :

1. **Initialisation**
   $k = 0$: choix de $\lambda^o \in \mathbb{R}^p$
2. **Itération** k :
   $\lambda^k \in \mathbb{R}^p$ est connu ; calcul de $y^k \in \mathbb{R}^n$ solution de

$$\min_{y \in \mathbb{R}^n} \mathcal{L}(y, \lambda^k) \ ,$$

   c'est-à-dire $y^k = A^{-1} \left(b - \sum_{i=1}^{p} \lambda_i^k \, c_i \right)$.
3. Calcul de λ^{k+1} avec : $\lambda_i^{k+1} = \lambda_i^k + \rho \ \left(c_i, y^k, \right)_n - \gamma_i, \ i = 1, \ldots, p$,
   où $\rho > 0$ est un réel fixé (choisi par l'utilisateur), c'est-à-dire après calcul

$$\lambda^{k+1} = \lambda^k - \rho \left(B \, \lambda^k - d \right) \ .$$

4. Critère d'arrêt

(d) L'algorithme du gradient pour le problème $(\mathcal{D})$ s'écrit :

1. **Initialisation**
   $k = 0$: choix de $\lambda^o \in \mathbb{R}^p$
2. **Itération** k :
   $\lambda^k \in \mathbb{R}^p$ est connu ; calcul de λ^{k+1} avec :$\lambda^{k+1} = \lambda^k + \rho \, \nabla h(\lambda^k)$,
   où $\rho > 0$ est un réel fixé (choisi par l'utilisateur), c'est-à-dire

$$\lambda^{k+1} = \lambda^k - \rho \left(B \, \lambda^k - d \right) \ .$$

3. Critère d'arrêt

(e) On remarque d'après ce qui précède que l'algorithme d'Uzawa appliqué au problème primal $(\mathcal{P})$ revient à appliquer l'algorithme du gradient sur le problème dual $(\mathcal{D})$.

– **Exercice 18**

(a) C'est le théorème 3.2.1. La convexité de $\mathcal{P}$ provient directement de celle de U et J.
(b) Voir l'exercice 2.6.1.2. $\mathcal{P}$ est non vide et est réduit à un seul élément d'après le théorème 3.1.2.
(c) i. J est continue sur un compact, donc il existe au moins une solution et $\mathcal{P}$ est non vide. Il n'y a en général pas unicité.
ii. La relation (9) se montre comme dans l'exercice 2.6.1.5 (relation (B.1)) avec $t = 1, \ u = u_m$ et $v = u_{m+1} - u_m$.
La caractérisation (3.4.1) du projeté de $u_m - \rho(Au_m - f)$ donne

$$\forall u \in U \qquad \langle u_m - \rho(Au_m - f) - u_{m+1}, u - u_{m+1} \rangle \leq 0 \ .$$

On l'applique avec $u = u_m$ pour obtenir la relation voulue.
Avec (9) on obtient alors

$$J(u_{m+1}) - J(u_m) \leq -\frac{1}{\rho} \|u_{m+1} - u_m\|^2 + \|A\| \|u_{m+1} - u_m\|^2 = (\|A\| - \frac{1}{\rho}) \, \|u_{m+1} - u_m\|^2 \ .$$

Donc si $0 \in]0, \rho_0[$ avec $\rho_0 \leq \dfrac{1}{\|A\|}$ la suite $(J(u_m))$ est décroissante. Comme elle est minorée, elle converge et la différence $J(u_{m+1}) - J(u_m)$ tend vers 0. La relation précédente montre alors que $\|u_{m+1} - u_m\| \to 0$.
La suite (u_m) est bornée, donc on peut en extraire une sous-suite (u_{m_k}) convergente vers $\tilde{u}$. D'après ce qui précède $u_{m_k+1} \to \tilde{u}$. De proche en proche on voit que la suite entière converge vers $\tilde{u}$. A la limite $\tilde{u}$ vérifie : $\tilde{u} = \pi(\tilde{u} - \rho(A\tilde{u} - f)$ c'est-à-dire

$$\forall u \in U \qquad \langle \rho(A\tilde{u} - f), u - \tilde{u} \rangle \leq 0 \, .$$

Ceci signifie que $\tilde{u} \in \mathcal{P}$. Si on change ρ on change la suite (u_m) et on peut ainsi trouver un autre élément de $\mathcal{P}$.

– **Exercice 19**
(a) L'inégalité de gauche est démontrée dans la preuve du théorème 2.4.1 grâce à la formule de Taylor avec reste intégral. On y a aussi prouvé la coercivité de J. L'inégalité de droite vient de la relation (11) et de la convexité de J due à l'ellipticité (en effet le gradient est alors fortement monotone) (Théorèmes 1.3.2 et 1.3.2) .

$$J(v) - J(u) \leq (\nabla J(v), v - u) = (\nabla J(u), v - u)_n + (\nabla J(v) - \nabla J(u), v - u)_n$$

$$\leq (\nabla J(u), v - u)_n + \|\nabla J(v) - \nabla J(u)\|_n \, \|v - u\|_n \leq (\nabla J(u), v - u)_n + M \, \|v - u\|_n^2 \, .$$

(b) U est un convexe fermé non vide (car les φ_i sont convexes continues). J est continue et coercive, strictement convexe car elliptique. Donc le problème $(\mathcal{P})$ a une solution unique notée u.
(c) Voir le chapitre 3.
(d) Soit $\rho > 0$ fixé . Utilisons la caractérisation du projeté de $\lambda + \rho\Phi(u)$ sur $\mathbb{R}^m_+$ (Théorème 3.4.1) :

$$\lambda = P_+(\lambda + \rho\Phi(u)) \, , \ \rho > 0 \text{ fixé } \Leftrightarrow \left\{ \begin{array}{l} \lambda \in \mathbb{R}^m_+, \\ (\lambda + \rho\Phi(u) - \lambda, \mu - \lambda)_m \leq 0 \, , \forall \mu \in \mathbb{R}^m_+ \end{array} \right.$$

$$\Leftrightarrow \lambda \in \mathbb{R}^m_+ \text{ et } (\Phi(u), \mu - \lambda)_m \leq 0 \, , \text{pour tout } \mu \in \mathbb{R}^m_+ \, .$$

En choisissant $\mu = \lambda + \varepsilon$, avec $\varepsilon > 0$ puis $\mu = \lambda - \varepsilon$ si $\lambda > 0$ on obtient le résultat voulu.
(e) Un couple (u, λ) est un point-selle du Lagrangien $\mathcal{L}$ si et seulement si :

$$\mathcal{L}(u, \mu) \leq \mathcal{L}(u, \lambda) \leq \mathcal{L}(v, \lambda) \, .$$

L'inégalité de gauche est équivalente à

$$\lambda \in \mathbb{R}^m_+ \, , \ \Phi(u) \leq 0, \text{ et } (\lambda, \Phi(u))_m = 0$$

c'est-à-dire $\lambda = P_+(\lambda + \rho\Phi(u)) \, , \ \rho > 0$ fixé Considérons maintenant l'inégalité de droite. Soient $t \in]\, 0, \, 1[, \ w \in \mathbb{R}^n$ et $v = u + t\,(w - u)$. On obtient

$$\mathcal{L}(u, \lambda) = J(u) \leq J(v) + (\lambda, \Phi(v)\,)_m \ ;$$

$$J(u) \leq J(u + t\,(w - u)) + (\lambda, \Phi(u + t\,(w - u))\,)_m$$

$$\leq J(u + t\,(w-u)) + (1-t)\,(\lambda, \Phi(u)\,)_m \; + t\;(\lambda, \Phi(w)\,)_m \;,$$

par convexité de Φ et $\lambda \geq 0$. Comme $(\lambda, \Phi(u)\,)_m = 0$ nous obtenons

$$J(u + t\,(w-u)) - J(u) + t\;(\lambda, \Phi(w)\,)_m \;\geq 0\;,$$

puis par division par t et en passant à la limite ($t \to 0$) :

$$(\nabla J(u), w-u)_n + (\lambda, \Phi(w))_m \geq 0\;.$$

La réciproque se démontre avec la convexité de J.
(f) Soit

$$J^k_\varepsilon(v) = \frac{1}{2}\|v\|^2_n + (\varepsilon\nabla J(u^k) - u^k, v)_n + \varepsilon(\lambda^k, \Phi(v))_m\;;$$

il est clair que J^k_ε est coercive et continue, donc le problème définissant u^{k+1} est bien défini et a une solution unique vérifiant

$$\forall v \in \mathbb{R}^n\; \frac{1}{2}(\|v\|^2_n - \|u^{k+1}\|^2_n)\left(\varepsilon\nabla J(u^k), v-u^{k+1}\right)_n + \varepsilon\left(\lambda^k, \Phi(v) - \Phi(u^{k+1})\right)_m \geq 0.$$

En utilisant la même technique que dans la question précédente (puisque Φ n'est pas supposée dérivable) on obtient le résultat désiré.
(g)
– Montrons (12) : on sait que

$$\lambda = P_+(\lambda + \rho\Phi(u)) \text{ et } \lambda^{k+1} = P_+(\lambda^k + \rho\Phi(u^{k+1}))\;.$$

En utilisant que P_+ est une contraction, nous avons

$$\begin{aligned}\|\lambda^{k+1} - \lambda\|^2_m \leq\; & \|\lambda^k + \rho\Phi(u^{k+1}) - \lambda - \rho\Phi(u)\|^2_m \\ \leq\; & \|\lambda^k - \lambda\|^2_m + \rho^2\|\Phi(u^{k+1}) - \Phi(u)\|^2_m + 2\rho\left(\lambda - \lambda^k, \Phi(u) - \Phi(u^{k+1}\right)_m.\end{aligned}$$

On obtient

$$2(\lambda^k - \lambda, \Phi(u) - \Phi(u^{k+1})_m \leq \frac{1}{\rho}(\|\lambda^k - \lambda\|^2_m - \|\lambda^{k+1} - \lambda\|^2_m) + \rho C^2\|u^{k+1} - u\|^2_n\;.$$

– Montrons (13) : nous avons d'après la question (e), avec $v = u^{k+1}$:

$$\varepsilon\left((\nabla J(u), u^{k+1} - u)_n + (\lambda, \Phi(u^{k+1}))_m\right) \geq 0$$

et d'après la question (f) avec $v = u$

$$(u^{k+1} - u^k + \varepsilon\nabla J(u^k), u - u^{k+1})_n + \varepsilon(\lambda^k, \Phi(u) - \Phi(u^{k+1}))_m \geq 0\;.$$

Après sommation on obtient (13).

– Montrons (14) :

$$(\nabla J(u^k) - \nabla J(u), u - u^{k+1})_n$$

$$= (\nabla J(u^k), u - u^k)_n + (\nabla J(u^k), u^k - u^{k+1})_n + (\nabla J(u), u^{k+1} - u)_n$$

$$\leq J(u) - J(u^k) - \frac{\alpha}{2}\|u^k - u\|_n^2 + (\nabla J(u^k), u^k - u^{k+1})_n + J(u^{k+1}) - J(u) - \frac{\alpha}{2}\|u^{k+1} - u\|_n^2 ,$$

d'après la question (a). Donc

$$(\nabla J(u^k) - \nabla J(u), u - u^{k+1})_n$$

$$\leq J(u^{k+1}) - J(u^k) + (\nabla J(u^k), u^k - u^{k+1})_n - \frac{\alpha}{2}\left(\|u^k - u\|_n^2 + \|u^{k+1} - u\|_n^2\right)$$

$$\leq \frac{M}{2}\|u^k - u^{k+1}\|_n^2 - \frac{\alpha}{2}\left(\|u^k - u\|_n^2 + \|u^{k+1} - u\|_n^2\right)$$

toujours d'après la question (a).

– Montrons (15) : on reporte (12) et (14) dans (13) ce qui donne

$$\begin{aligned} 0 \leq \quad & (u^{k+1} - u^k, u - u^{k+1})_n \\ & + \varepsilon\left[\frac{M}{2}\|u^k - u^{k+1}\|_n^2 - \frac{\alpha}{2}\left(\|u^k - u\|_n^2 + \|u^{k+1} - u\|_n^2\right)\right] \\ & + \frac{1}{2\rho}\left(\|\lambda^k - \lambda\|_m^2 - \|\lambda^{k+1} - \lambda\|_m^2\right) + \rho C^2\left(\|u^{k+1} - u\|_n^2\right) . \end{aligned}$$

En utilisant $(u^{k+1} - u^k, u - u^{k+1})_n = (u^{k+1} - u^k, u - u^k)_n - \|u^{k+1} - u\|_n^2)$ et

$$(u^{k+1} - u^k, u - u^k)_n = (u^{k+1} - u, u - u^k)_n + \|u^k - u\|_n^2$$

$$= \frac{1}{2}\|u^k - u^{k+1}\|_n^2 - \frac{1}{2}\left(\|u^{k+1} - u\|_n^2 - \|u^k - u\|_n^2\right)$$

on obtient la relation cherchée.

(h) Posons

$$\beta_k = -\frac{\varepsilon\alpha - 1}{2}\|u^k - u\|_n^2 + \frac{\varepsilon}{2\rho}\|\lambda^k - \lambda\|_m^2;$$

la dernière inégalité de la question précédente devient

$$\beta_{k+1} - \beta_k \leq \frac{\varepsilon M - 1}{2}\|u^k - u^{k+1}\|_n^2 + \varepsilon(\rho\frac{C^2}{2} - \alpha)\|u^{k+1} - u\|_n^2 . \qquad \text{(B.8)}$$

Si $0 < \varepsilon < \frac{1}{M}$ et $0 < \rho < \frac{2\alpha}{C^2}$ alors $\beta_{k+1} - \beta_k \geq 0$ et la suite (β_k) est alors décroissante. Comme $\alpha \leq M$ la suite (β_k) est positive. Donc elle converge et la différence $\beta_{k+1} - \beta_k$ tend vers 0. En reportant dans (B.8) on voit que $\|u^{k+1} - u\|_n^2 \to 0$. Comme β_k converge, elle est bornée, ce qui permet de conclure.

Chapitre 4

– **Exercice 1**
La solution de l'équation différentielle proposée est : $x(t) = \dfrac{x_1}{1 - \left(\frac{x_0 - x_1}{x_0}\right) e^{(r-E)t}}$.
Si $x_0 = x_1$ la solution est constante, égale à x_1. On est et on reste à l'état d'équilibre.
De même, si $x_0 \neq x_1$ et $t \to +\infty$ on constate que la solution tend vers l'état d'équilibre x_1.

Chapitre 5

– **Exercice 1**
(a) La solution du système (1) est donnée par $\begin{cases} v(t) = \dfrac{\alpha}{2}t^2 + \beta t + v_0 \ , \\ x(t) = \dfrac{\alpha}{6}t^3 + \dfrac{\beta}{2}t^2 + v_0 t + x_0 \ . \end{cases}$
On suppose que $x_0 = 0$. Il faut résoudre $x(3) = x^*$ et $v(3) = 0$. S'il n'y a aucune contrainte sur α et β on obtient $\alpha = \dfrac{2}{9}(3v_0 - 2x^*)$ et $\beta = \dfrac{2(x^* - 2v_0)}{3}$. Dans le cas de contraintes on met le problème sous la forme

$$\begin{cases} \min J(X) = \dfrac{1}{2}\,(M\,X, X) + (b, X) \\ g(X) \leq 0 \ , \end{cases}$$

où $X = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$. La fonction coût est

$$J(\alpha, \beta) = (x(3) - x^*)^2 + v(3)^2 = \left(\frac{9}{2}\alpha + \frac{9}{2}\beta + 3v_0 - x^*\right)^2 + \left(\frac{9}{2}\alpha + 3\beta + v_0\right)^2 \ .$$

Après simplifications, la matrice M et le vecteur b obtenus sont

$$M = \begin{bmatrix} 9 & \dfrac{15}{2} \\ \dfrac{15}{2} & \dfrac{13}{2} \end{bmatrix} \text{ et } b = \begin{bmatrix} 4v_0 - x^* \\ \dfrac{11}{3}v_0 - x^* \end{bmatrix} \ .$$

Enfin $g(X) = -\beta \leq 0$.
(b) Le problème de contrôle optimal s'écrit

$$\begin{cases} \min \tilde{J}(u) = \dfrac{1}{2}(y(3) - y_d)^2 + \rho \displaystyle\int_0^T u(t)^2 \, dt \\ \dfrac{dy}{dt} = Ay(t) + Bu(t) \ , \ y(0) = \begin{bmatrix} 0 \\ v_0 \end{bmatrix} \\ u(t) = \alpha t + \beta \ , \ \beta \geq 0 \ , \end{cases}$$

avec

$$y(t) = \begin{bmatrix} x(t) \\ v(t) \end{bmatrix} \ , \ y_d = \begin{bmatrix} x^* \\ 0 \end{bmatrix} \ , \ A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \ , \ B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \text{ et } \rho > 0 \ .$$

Ici $Q = 0$, $R = \rho I$ et $D = I$ (avec les notations du cours) et $\mathcal{U} = \{ u$ affine, avec $\beta = u(0) \geq 0 \}$.
Le système d'optimalité pour la solution $(\bar{y}, \bar{u} = \bar{\alpha}t + \bar{\beta})$ est :

$$\begin{cases} \dfrac{d\bar{y}}{dt}(t) = A\,\bar{y}(t) + B\,\bar{u}(t) \text{ sur }]\,0,3[,\ \bar{y}(0) = y_0 \\ \dfrac{d\bar{p}}{dt}(t) = -A^t\bar{p}(t) \text{ sur }]\,0,3[,\ \bar{p}(T) = \bar{y}(3) - y_d \\ \forall(\alpha,\beta) \in \mathbb{R}\times\mathbb{R}^+ \qquad \displaystyle\int_0^3 \left\langle B^t\,\bar{p}(t) + \bar{\alpha}t + \bar{\beta}, (\alpha-\bar{\alpha}) + (\beta-\bar{\beta})t \right\rangle dt\ \geq 0 \end{cases}$$

Posons $\bar{p} = (\bar{p}_1, \bar{p}_2)^t$. La dernière relation devient

$$\forall(\alpha,\beta) \in \mathbb{R}\times\mathbb{R}^+ \qquad \int_0^3 \left[\bar{p}_2(t) + \bar{\alpha}t + \bar{\beta}\right]\left[(\alpha-\bar{\alpha}) + (\beta-\bar{\beta})t\right] dt\ \geq 0\,.$$

(c) **Résolution** : on le fait avec la première approche en écrivant les équations de KKT qui donnent ici : $(\bar{X}, \bar{\lambda})$ est solution de

$$\begin{aligned} M\bar{X} + b - \begin{bmatrix} 0 \\ \bar{\lambda} \end{bmatrix} &= 0 \\ \bar{\beta} \geq 0\,, \bar{\lambda} \geq 0\,, \bar{\lambda}\bar{\beta} &= 0\,. \end{aligned}$$

Dans ce cas $b = \begin{bmatrix} 3200 \\ \dfrac{8600}{3} \end{bmatrix}$ car v_0= 60 km/h = 1000 m/mn.
• Si $\bar{\beta} \neq 0$ alors $\bar{\lambda} = 0$ et on obtient $\beta \simeq -800$ ce qui impossible.
• Par conséquent $\bar{\beta} = 0$ et le système nous donne :

$$\bar{\alpha} = -\frac{3200}{9} \simeq -355 \text{ et } \bar{\lambda} = \frac{8600}{3} + \frac{15}{2}\bar{\alpha} = 200\,.$$

$x(3) = 600$m et $v(3) = -600$m/mn. La voiture est loin d'être stabilisée. C'est même peu réaliste d'avoir une vitesse négative ! !
• Remarquons que le cas sans contraintes fournit la stabilisation complète mais avec $\beta = -800 < 0$. C'est donc la contrainte $\beta > 0$ qui n'est pas réaliste.

– **Exercice 2**
Nous avons (avec les notations du cours)

$$n = 1\ ,\ Q = 1\ ,\ R = 1\ ,\ D = 1\ ,\ B = 1\ ,\ T = \frac{\pi}{2} \text{ et } z_d(s) = \cos(s)\ .$$

Le système d'optimalité s'écrit :

$$\begin{cases} \dfrac{dx}{dt} = x + u\ ,\ x(0) = 0\ , \\ \dfrac{dp}{dt} = -p - x + z_d\ ,\ p(\dfrac{\pi}{2}) = x(\dfrac{\pi}{2}) \\ u = -p\ . \end{cases}$$

Posons $X = (x, p)^t$. Nous avons $X' = AX + f$ avec $f = (0, z_d)^t$. La solution de cette EDO est

$$X(t) = e^{At}X(0) + e^{At}\int_0^t e^{-As}f(s)\,ds\ ,\ \text{avec } A = \begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix}\ .$$

La solution (obtenue par exemple avec MAPLE©) est :

$$x(t) = -\left(\frac{2+\sqrt{2}+3\sqrt{2}C}{12}\right)e^{\sqrt{2}\,t} + \left(\frac{3C\sqrt{2}-2+\sqrt{2}}{12}\right)e^{-\sqrt{2}\,t} + \frac{1}{3}\cos(t)\ ,$$

et

$$p(t) = \left(\frac{6C+\sqrt{2}-3\sqrt{2}C}{12}\right)e^{\sqrt{2}\,t} + \left(\frac{3\sqrt{2}C+6C-\sqrt{2}}{12}\right)e^{-\sqrt{2}\,t} + \frac{1}{3}\sin(t) + \frac{1}{3}\cos(t),$$

la constante C étant ajustée avec la relation $p(\frac{\pi}{2}) = x(\frac{\pi}{2})$.

– **Exercice 3**
On procède comme précédemment :

$$n = 1\ ,\ Q = 0\ ,\ R = 1\ ,\ D = 1\ ,\ B = 1\ ,\ A = 0\ ,\ T = 1 \text{ et } z_d \equiv 0.\ .$$

Le système d'optimalité est : $\begin{cases} \dfrac{dx}{dt} = u\ ,\ x(0) = x_0\ , \\ \dfrac{dp}{dt} = 0\ ,\ p(1) = x(1) \\ u = -p\ . \end{cases}$

Cela donne $p(t) = x(1)\ , x(t) = -x(1)t + x_0$. Par conséquent, $x(1) = \dfrac{x_0}{2}$. En définitive

$$p(t) \equiv \frac{x_0}{2}\ ,\ u(t) \equiv -\frac{x_0}{2}\ ,\ x(t) = x_0(1 - \frac{t}{2}) \text{ et } J = \frac{x_0^2}{2}.$$

– **Exercice 4**
Le problème de contrôle optimal est associé aux matrices

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad , \quad B = I_2 \quad , \quad Q = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \quad , \quad R = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix} \quad \text{et } D = 0\ .$$

De plus on note $X = \begin{bmatrix} x \\ y \end{bmatrix}$, $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ et $f \equiv z_d \equiv 0$. On peut appliquer les résultats sur l'équation de Ricatti. Avec les données du problème, on obtient

$$\bar{u}(t) = -R^{-1}B^t\bar{p}(t) = -\begin{bmatrix} \dfrac{\bar{p}_1}{\lambda} \\ \dfrac{\bar{p}_2}{\mu} \end{bmatrix}\ ,\ \bar{p}(t) = \bar{P}(t)\,\bar{X}(t)\ ,$$

avec

$$\frac{d\bar{P}}{dt} = -A\bar{P} - \bar{P}A^t + \bar{P}^2 - Q\ ,\ \bar{P}(T) = 0 \ \text{ et } \frac{d\bar{X}}{dt} = \left(A - R^{-1}\bar{P}(t)\right)\bar{X}(t)\ ,\ \bar{X}(0) = 0\ .$$

La résolution se fait ensuite soit par un logiciel de calcul formel comme MAPLE© soit par approximation numérique (avec MATLAB© par exemple).

– **Exercice 5**

(a) Application directe du cours.

(b) Utiliser le théorème 3.1.1.

(c) L'application $v \to z(v)$ est affine de $\mathbb{R}^m$ dans $\mathbb{R}^n$ donc $z'(u)\cdot w = A^{-1}B\,w$. Par composition

$$\left(J'(u), w\right)_m = \left(z(u) - z_0, z'(u)\cdot w\right)_n + N\,(u,w)_m = \left(z(u) - z_0, A^{-1}B\,w\right)_n + N\,(u,w)_m\ .$$

Par application du théorème 3.2.1, on obtient la relation (6). Si $K = \mathbb{R}^m$ cette relation est équivalente à $J'(u) = 0$.

(d) On pose $y = z(u)$ et $A^t p = y - z_0$. La relation (6) implique que u est solution de (5) si et seulement si

$$\begin{aligned}
&\forall v \in K \qquad \left(y - z_0, A^{-1}B\,(v-u)\right)_n + N\,(u, v-u)_m \geq 0\\
&\iff \left(A^t p, A^{-1}B\,(v-u)\right)_n + N\,(u, v-u)_m \geq 0\\
&\iff \left(B^t p, v-u\right)_m + N\,(u, v-u)_m \geq 0\\
&\iff \forall v \in K \qquad \left(B^t p + N\,u, v-u\right)_m \geq 0\ .
\end{aligned}$$

Si $K = \mathbb{R}^m$, c'est équivalent à $B^t p + N\,u = 0$, c'est-à-dire $u = -\dfrac{1}{N}B^t p$. Par conséquent, u est solution si et seulement si le couple (y,p) est solution de (8).

(e) L'itération courante de l'algorithme du gradient projeté donne

$$u_{n+1} = P_K(u_n - \rho J'(u_n))\ ,\ \text{où } \rho > 0\ .$$

Comme $J'(u_n) = (A^{-1}B)^t(z(u_n) - z_0) + Nu_n = B^t p_n + Nu_n$ où on a posé

$$A^t p_n = z(u_n) - z_0 \ \text{ et } \ Az(u_n) = f + Bu_n\ ,$$

on obtient $u_{n+1} = P_K(u_n - \rho B^t p_n + Nu_n)$. Ceci est équivalent au système (7) avec (y_n, u_n, p_n) à la place de (y,u,p).

– **Exercice 6**

(a) C'est une suite récurrente donnée de manière explicite : elle est définie de manière unique. On peut toutefois écrire le système sous la forme $AY = BV + F$ où

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix},\ V = \begin{bmatrix} v_0 \\ \vdots \\ v_{N-1} \end{bmatrix},\ A = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -a_1 & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & -a_{N-1} & 1 \end{bmatrix},\ F = \begin{bmatrix} -a_0 y_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

et B = diag$(b_i)_{i=0,\dots,N-1}$. Les calculs sont comme ceux de l'exercice 5.6.5. : $DY(V)\cdot W = A^{-1}BW$.
(b) La fonctionnelle J peut s'écrire de la manière suivante :

$$J(Y) = \frac{1}{2}Y^t\mathbb{A}Y + \frac{1}{2}\alpha_0 y_0^2 + \frac{1}{2}V^t\mathbb{B}V \ ,$$

où $\mathbb{A}$ = diag$(\alpha_i)_{i=1,\dots,N}$ et $\mathbb{B}$ = diag$(\beta_i)_{i=0,\dots,N-1}$. Une condition suffisante pour que le problème admette une solution (unique) est que $\mathbb{B}$ soit définie positive, c'est-à-dire $\beta_i > 0$, pour tout $i = 0, \dots, N-1$.
(c) Lorsque $y_0 = 0$, la solution est $V = 0$.
(d)-(e) Voir l'exercice 5.6.5.
(f) Exprimons le premier terme de la relation (12) :

$$\begin{aligned}
\sum_{i=0}^{N} \alpha_i\bar{y}_i(y_i - \bar{y}_i) &= \sum_{i=0}^{N-1} (\bar{p}_i - a_i\bar{p}_{i+1})(y_i - \bar{y}_i) + \bar{p}_N(y_N - \bar{y}_N) \\
&= \sum_{i=0}^{N} \bar{p}_i(y_i - \bar{y}_i) - \sum_{i=1}^{N} a_{i-1}\bar{p}_i(y_{i-1} - \bar{y}_{i-1}) \\
&= \sum_{i=0}^{N} \bar{p}_i(y_i - \bar{y}_i) - \sum_{i=1}^{N} \bar{p}_i\left[y_i - \bar{y}_i + b_{i-1}(v_{i-1} - \bar{v}_{i-1})\right] \ \text{(avec (9) ,} \\
&= \bar{p}_0 \underbrace{(y_0 - \bar{y}_0)}_{=0} - \sum_{i=0}^{N-1} b_i(v_i - \bar{v}_i) \ .
\end{aligned}$$

La relation (12) est donc équivalente à (14).
Si $\mathcal{U} = \mathbb{R}^m$, on a donc $\bar{v}_i = -\dfrac{b_i}{\beta_i}\bar{p}_{i+1}$ pour tout $i = 0, \dots, N-1$. Si on pose

$$\Gamma = \begin{bmatrix} 0 & \gamma_0 & 0 & \dots & 0 \\ 0 & 0 & \gamma_1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 0 & -\gamma_{N-1} \\ 0 & \dots & \dots & 0 & 0 \end{bmatrix} ,$$

avec $\gamma_i = \dfrac{b_i^2}{\beta_i}$, on obtient $B\overline{V} = -\Gamma\overline{P}$. Comme $A^t\overline{P} = \mathbb{A}\overline{Y}$ et $A\overline{Y} = BV + F$, on obtient $A\overline{Y} = -\Gamma(A^t)^{-1}\mathbb{A}\overline{Y} + F$, c'est-à-dire $\mathbb{T}\,\overline{Y} = F$ avec $\mathbb{T} = (A + \Gamma(A^t)^{-1}\mathbb{A})$.
(g) -(h) Question de cours.

– **Exercice 7**
(a) L'équation correspond à l'équation de la chaleur stationnaire en dimension 1. Elle décrit (par exemple) la distribution de la température x dans un milieu sous l'effet d'une source

distribuée de chaleur u. On pose $y = (x, \frac{dx}{dt})^t$ de sorte que l'EDO est équivalente au système différentiel suivant :

$$\frac{dy}{dt} = A\,y + B\,u\ ,\ y(0) = (x_0, x_0')^t\ ,$$

où

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \text{ et } B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\ .$$

(b) Le problème de contrôle optimal est associé aux données

$$Q = 0\ ,\ D = I\ ,\ R = \alpha \text{ et } z_d = (z_1, z_2)^t\ .$$

Une condition suffisante pour que le problème admette une solution unique est $\alpha > 0$.
(c) Le système d'optimalité pour la solution $(\bar{y}, \bar{u})$ est :

$$\begin{cases} \dfrac{d\bar{y}}{dt}(t) = A\,\bar{y}(t) + B\,\bar{u}(t) \text{ sur }]\,0, T[,\ \bar{y}(0) = y_0 \\ \dfrac{d\bar{p}}{dt}(t) = -A^t\bar{p}(t) \text{ sur }]\,0, T[,\ \bar{p}(T) = \bar{y}(T) - z_d \\ \forall u \in \mathcal{U}_{[a,b]} \qquad \displaystyle\int_0^T \left(B^t\,\bar{p}(t) + \alpha\,\bar{u}(t)\right)\ (u(t) - \bar{u}(t))\ dt\ \geq 0 \end{cases}$$

(d) Lorsque $\mathcal{U}_{[a,b]} = L^2(0, T, \mathbb{R})$ la dernière équation devient : $\bar{p}_2(t) = -\alpha\,\bar{u}(t)$. Notons $y_1(T) = y_1$ et $y_2(T) = y_2$. L'état adjoint est donné par

$$\bar{p}_1(t) \equiv y_1 - z_1 \text{ puis } \bar{p}_2(t) = -(y_1 - z_1)t + (y_2 - z_2) + (y_1 - z_1)\,T\ .$$

On en déduit l'état

$$y_2(t) = \frac{1}{\alpha}\left[(y_1 - z_1)\frac{t^2}{2} - ((y_2 - z_2) + (y_1 - z_1)\,T)\ t\ \right],$$

$$y_1(t) = \frac{1}{\alpha}\left[(y_1 - z_1)\frac{t^3}{6} - ((y_2 - z_2) + (y_1 - z_1)\,T)\,\frac{t^2}{2}\right]\ .$$

On en tire $y_1(T)$ et $y_2(T)$ et on conclut.

Chapitre 6

– **Exercice 1**
(a) Avec la loi de commande $u = F\,x + v$ le système est en boucle fermée et est équivalent à

$$\begin{cases} \dfrac{dx}{dt} = (A + BF)\,x + B\,v,\ x(0) = x_0\ , \\ \quad y = (C + DF)\,x + D\,v\ , \end{cases}$$

où v joue le rôle d'un contrôle en boucle ouverte. La contrôlabilité et la stabilisation des deux systèmes (A, B) et (A+ B F, B) sont donc équivalentes.
(b) La matrice de contrôlabilité de (A, B) est :

$$\mathbb{M} = [B\ ,\ AB] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ avec } n = 2\ .$$

Le rang de $\mathbb{M}$ est maximal donc le système est contrôlable d'après le théorème 6.2.6. De même, la matrice d'observabilité de (A, B) est :

$$\mathbb{O} = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \text{ avec } n = 2\ .$$

Le rang de $\mathbb{O}$ est maximal donc le système est observable d'après le théorème 6.3.3.
(c) Avec la loi de commande associée à F le système s'écrit

$$\begin{cases} \dfrac{dx}{dt} = (A + BF)\, x + Bv,\ x(0) = x_0\ , \text{ (équation autonome)} \\ \quad y = (C + DF)\, x + Dv\ , \text{ (équation de sortie)} \end{cases}$$

Le système est observable si la matrice $\tilde{\mathbb{O}} = \begin{bmatrix} C + DF \\ (C + DF)(A + BF) \end{bmatrix}$ est de rang maximal 2. Or $\tilde{\mathbb{O}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ est de rang 1. Le système n'est donc pas complètement observable.

– **Exercice 2**

Ici $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ et la matrice de contrôlabilité est $\mathbb{M} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. Elle est de rang 1. Comme $U = \mathbb{R}$, le théorème 6.2.6 montre que $\mathcal{C} \neq \mathbb{R}^2$.
Retrouvons ce résultat directement : la solution du système est

$$p(t) = q_0 t + \int_0^t u(s)\, ds + p_0, \quad q(t) = q_0\ .$$

Donc $x_0 = (p_0, q_0)^t \in \mathcal{C}$ si et seulement si $q_0 = 0$ et $p_0 = -\displaystyle\int_0^t u(s)\, ds$. Par conséquent $\mathcal{C} = \{(p_0, 0)\ \mid\ p_0 \in \mathbb{R}\ \}$.

– **Exercice 3**

Pour $t \in [0, 1]$ le système différentiel donne $p\dfrac{dp}{dt} = -(1-t)q\, p$ et $q\dfrac{dq}{dt} = (1-t)p\, q$. Donc $p\ \dfrac{dp}{dt} + q\ \dfrac{dq}{dt} = 0$ et l'intégration donne $p^2(t) + q^2(t) = \rho^2$. Par conséquent les trajectoires sont supportées par des cercles de centre $(0, 0)$ et passant par $x_0 = (p_0, q_0)$. Si $p_0 > 0$ alors $\dfrac{dq}{dt}$ est positif et q est croissant. De même, si $p_0 < 0$, q est décroissant. Par conséquent les cercles sont parcourus dans le sens trigonométrique. On peut alors préciser la valeur de la

solution : $p(t) = \rho \sin(\varphi(t) + \alpha)$ et $q(t) = \rho \cos(\varphi(t) + \alpha)$. Un rapide calcul de $\dot{p}$ et $\dot{q}$ montre que $\varphi(t) = \dfrac{(t-1)^2}{2}$ et on obtient

$$p(t) = \rho \sin\left(\frac{(t-1)^2}{2} + \alpha\right),\ q(t) = \rho \cos\left(\frac{(t-1)^2}{2} + \alpha\right)$$

avec $\rho = \sqrt{p_0^2 + q_0^2}$ et $\alpha = -\dfrac{1}{2} + \arctan\left(\dfrac{p_0}{q_0}\right)$.

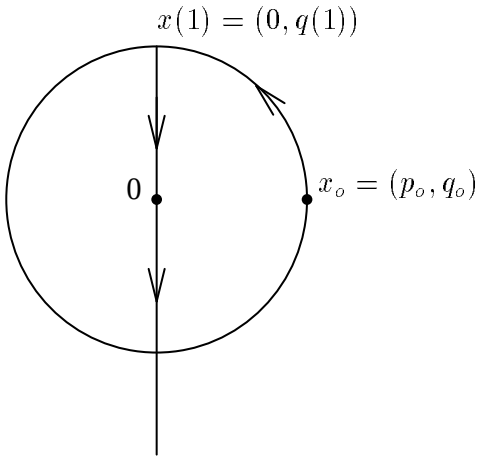

Figure B.6. Allure des trajectoires

Si $t \geq 1$, $p(t) \equiv p(1)$, donc les trajectoires sont supportées par des droites verticales. Comme $u(t) \leq 2$ et $t \geq 1$, $\dot{q} \leq 0$ et q est décroissant. Les droites sont parcourues de bas en haut .

Supposons que $x_0 = (p_0, q_0) \in \mathcal{C}$. La trajectoire doit passer par 0. Tant que $t \in [0, 1]$, le système n'est pas contrôlé et la trajectoire décrit un cercle de centre 0. A l'instant $t = 1$ le point se trouve en $(\rho \sin \alpha, \rho \cos \alpha)$. La seule possibilité est donc que $p_0 = \rho \sin \alpha = 0$; comme $\alpha = -\dfrac{1}{2}$, cela implique que $\rho = 0$, c'est-à-dire $x_0 = (0, 0)$. L'ensemble contrôlable est donc réduit à $(0, 0)$. Il est clair alors qu'aucun autre point de la trajectoire n'est contrôlable.

– **Exercice 4**

$A = \begin{bmatrix} 0 & -1 & 1 \\ 2 & -3 & 1 \\ 1 & -1 & -1 \end{bmatrix}$, $B = \begin{bmatrix} -1 & 1 \\ 0 & 2 \\ 1 & 3 \end{bmatrix}$ et la matrice de contrôlabilité est

$$\mathbb{M} = \begin{bmatrix} -1 & 1 & 1 & 1 & -1 & -3 \\ 0 & 2 & -1 & -1 & 3 & 1 \\ 1 & 3 & -2 & -4 & 4 & 6 \end{bmatrix}.$$

Elle est de rang 3, donc $\mathcal{C} = \mathbb{R}^3$.

– **Exercice 5**

$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$, $B = b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$. La matrice de contrôlabilité est

$$\mathbb{M} = [b, Ab, A^2b] = \begin{bmatrix} b_1 & b_2 & -b_2 + b_3 \\ b_2 & -b_2 + b_3 & b_2 - 2b_3 \\ b_3 & -b_3 & b_3 \end{bmatrix} .$$

Le déterminant de cette matrice est $-b_1\, b_3^2 - b_2\, b_3^2 - b_3^3 = -b_3^2(b_1 + b_2 + b_3)$.
Par conséquent, si on veut que $\mathcal{C} = \mathbb{R}^3$, il faut choisir b en dehors des plans d'équations $b_1 + b_2 + b_3 = 0$ et $b_3 = 0$.

Chapitre 7

– **Exercice 1**
Ici $n = p = 2$, $A = O_2$ (matrice nulle) et $B = I_2$ (matrice identité). La matrice de contrôlabilité de (A, B) est :

$$\mathbb{M} = [B\ ,\ AB] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} .$$

Elle est de rang 2. D'autre part, le système est normal si et seulement si le rang de toutes les matrices $\mathbb{P}_i = \left[b_i, Ab_i, \ldots, A^{n-1}b_i\right]$ est égal à 2 pour les p vecteurs colonnes b_i de B (voir l'exemple 7.3). Ici

$$\mathbb{P}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \text{ et } \mathbb{P}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} .$$

Elles sont toutes les deux de rang 1 : le système n'est donc pas normal.

– **Exercice 2**
Les points extrêmaux de $U = [-1, 1]$ sont -1 et 1. Les contrôles extrêmaux sont les fonctions qui ne prennent que les valeurs -1 ou 1. La matrice de contrôlabilité est réduite à 1, donc le système est contrôlable et il existe un temps minimal que l'on note t^*. D'après le principe de Pontryagin, si u^* est un contrôle optimal, nous avons

$$p^* \text{ constant et } \forall v \in [-1, 1] \qquad p^* u^*(s) \leq p^* v\ .$$

Si $p^* < 0$ alors $u^*(s) = 1$ en prenant $v = -1$. On raisonne de même si $p^* > 0$. Par conséquent $u^* = -\text{signe}(p^*)$ est toujours égal soit à 1 soit à -1. Tout autre contrôle extrêmal (par exemple $u = \begin{cases} -1 & \text{sur } [0\ , \dfrac{t^*}{2}[\\ 1 & \text{sur } [\dfrac{t^*}{2}, t^*] \end{cases}$) n'est pas optimal.

– **Exercice 3**
(a) $\mathcal{A}(x_0, t) = \{\, x[u, x_0](t) \ \mid\ u \in \mathcal{U} \,\}$. Ici

$$x_1(t) = -1 + \int_0^t u_1(s)\, ds \ \text{ et } \ x_2(t) = \int_0^t u_2(s)\, ds\ .$$

Il est clair que pour tout t, $|x_1(t)+1| \leq t$ et $|x_2(t)| \leq t$ d'après le choix de $\mathcal{U}$. Donc $\mathcal{A}(x_0,t) \subset [-t-1,t-1]\times[-t,t]$ qui est un carré de côté t. Réciproquement, si on choisit (ξ_1,ξ_2) dans $[-t-1,t-1]\times[-t,t]$, il suffit de prendre comme contrôles, les fonctions constantes

$$u_1 \equiv \frac{1+\xi_1}{t} \text{ et } u_2 \equiv \frac{\xi_2}{t}\ ,$$

pour vérifier que (ξ_1,ξ_2) est atteignable.
(b) Soit $u_1(t)\equiv 1$ et $u_2(t)=\varphi(t)$ où φ est une fonction telle que $|\varphi(t)|\leq 1$ pour tout t et $\displaystyle\int_0^1 \varphi(s)\,ds = 0$. L'état associé à ce contrôle vérifie $x_1(1)=0$ et $x_2(1)=0$. Donc $(0,0)$ est atteignable par une infinité de contrôles en un temps $t^*=1$.
Ici $n=p=2$, $A=O_2$ (matrice nulle) et $B=I_2$ (matrice identité). Vérifions que t^* est optimal grâce au théorème 7.4.1. On sait que $0\in$ Int (U) et la matrice de contrôlabilité $\mathbb{M}=[I_2\ ,\ 0_2]$ est bien de rang maximal (égal à 2). Il reste à montrer que le principe de Pontryagin est vérifié. L'état adjoint p^* doit être constant : $p^*=[p_1,p_2]^t$ et tel que

$$p_1u_1(s)+p_2u_2(s) \leq p_1v_1+p_2v_2 \qquad \forall (v_1,v_2)\in U\ ,\ \text{p.p. } s,$$

c'est-à-dire $p_1+p_2\varphi(s)\leq p_1v_1+p_2v_2$. Si on choisit $p_1=-1$ et $p_2=0$, il est clair que la relation précédente est toujours vérifiée. Par conséquent, il existe une infinité de contrôles optimaux (associés au temps optimal $t^*=1$).Le fait qu'il n'y ait pas unicité du contrôle s'explique par le fait que le système n'est pas normal. En effet la matrice $[b_1, A\,b_1] = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ est de rang 1.

– **Exercice 4**
(a)

$$X=\begin{bmatrix} x \\ y \end{bmatrix}\ ,\ F=\begin{bmatrix} c \\ 0 \end{bmatrix}\ ,\ A=0_2 \text{ et } B=I_2\ .$$

(b) On part de $P_0=(x_0,y_0)$. Si $P_1=(0,0)$ est atteignable, alors il existe $t>0$ et u un contrôle tels que

$$x_0=-ct-\int_0^t v(s)\sin(\gamma(s))\,ds \text{ et } y_0=-\int_0^t v(s)\cos(\gamma(s))\,ds\ .$$

Comme $0\leq v(s)\leq 1$, il faut que $-(1+c)t\leq x_0\leq (1-c)t$ et $-t\leq y_0\leq t$. On voit que si la vitesse du courant c est supérieure à 1, on ne pourra pas toujours atteindre P_1 car $\displaystyle\bigcup_{t\geq 0}[-(1+c)t,(1-c)t]=\mathbb{R}^-$.
En pratique, si P_0 est en amont de P_1 on pourra atteindre P_1. Si P_0 est en aval ce sera impossible (voir la figure B.7). C'est tout-à-fait cohérent, puisque la vitesse du bateau est inférieure à la vitesse du courant : on ne peut pas remonter la rivière ...

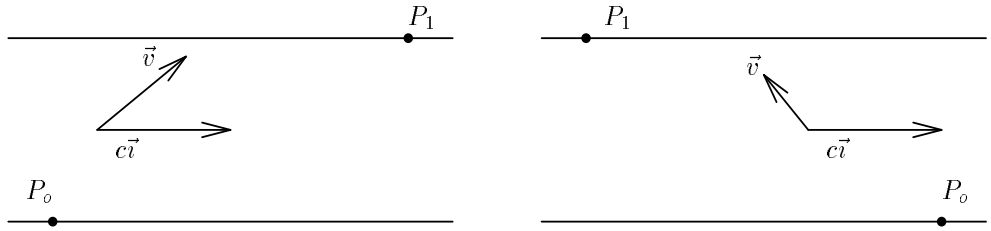

Figure B.7.

Précisons $\mathcal{C}$ dans le cas où $c < 1$. Prenons par exemple $v \equiv 1$ et $\gamma = \text{cste}$. On doit trouver t et γ tels que

$$x_0 = -ct - t\sin\,\gamma \text{ et } y_0 = -t\cos\,\gamma\ .$$

Comme $(x_0 + ct)^2 + y_0^2 = v^2,\ t$ vérifie l'équation $t^2(c^2-1) + 2x_0ct + x_0^2 + y_0^2 = 0$. Le discriminant Δ est égal à $(1-c^2)y_0^2 + x_0^2 \geq 0$. On choisit donc (par exemple)

$$t^* = \frac{x_0 c + \sqrt{\Delta}}{1-c^2} \text{ et } \gamma = \arctan\left(\frac{y_0}{x_0 + ct^*}\right)\ ;$$

ceci prouve que tout point P_0 est contrôlable.
(c) Dans le cas où $c = 0$ on voit qu'il existe un contrôle permettant d'atteindre P_1 en temps minimal (Théorème 7.1.1). On ne peut pas conclure à l'unicité car le système n'est pas normal.
(d) Le principe du minimum donne

$$\frac{dX^*}{dt} = Bu^*\ ,\ X^*(0) = P_0,\ X^*(t^*) = P_1 = 0\ , p^* = (p_1^*, p_2^*)^t = \text{cste, car } \frac{dp^*}{dt} = 0\ ,$$

et

$$\forall\alpha \in \mathbb{R}\ ,\ \forall v \in [0,1] \qquad v^*(s)\left(p_1^*\sin\gamma^*(s) + p_2^*\cos\gamma^*(s)\right) \leq v\left(p_1^*\sin\alpha + p_2^*\cos\alpha\right)\ .$$

(e) Posons $\rho = \sqrt{(p_1^*)^2 + (p_2^*)^2}$ et $p_1^* = \rho\cos\delta$, $p_2^* = \rho\sin\delta$. On obtient

$$v^*(s)\sin(\gamma^*(s) + \delta) \leq v\sin(\alpha + \delta)\ .$$

$v^*(s)$ est supposée constante , donc $\sin(\gamma^*(s)+\delta) \leq \sin(\alpha+\delta)$ pour tout α. Par conséquent $\sin(\gamma^*(s) + \delta) = -1$ et $\gamma^*(s) = -\dfrac{\pi}{2} - \delta$.
En définitive le contrôle $u^*(t) = v \cdot (\sin\gamma, \cos\gamma)^t$ est une fonction constante ; d'autre part

$$x^*(t) = -1 + vt\sin\gamma \quad \text{et} \quad y^*(t) = -1 + vt\cos\gamma\ ,$$

avec $x^*(t^*) = 0$ et $y^*(t^*) = 0$. D'où $\tan\gamma = 1$, $\delta = \frac{\pi}{4}$ et $t^* = \frac{\sqrt{2}}{v}$. On retrouve ainsi le résultat évident : sans courant la trajectoire la plus rapide à vitesse constante correspond à la ligne droite.

– **Exercice 5**

(a) Si on pose $x = [\theta, \omega]^t$, le système est équivalent à $\dfrac{dx}{dt} = Ax(t) + Bu(t)\ , x(0) = x_0\ ,$ avec

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\ ,\ B = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \text{ et } x_0 = \begin{bmatrix} \theta_0 \\ \omega_0 \end{bmatrix}\ .$$

La matrice de contrôlabilité est $\mathbb{M} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Elle est de rang 2, donc le système est contrôlable et $\mathcal{C} = \mathbb{R}^2$.

(b) Il existe donc un contrôle permettant d'amener le système au repos en un temps minimal (on applique le théorème 7.1.1). Ce contrôle est unique car le système est normal (il est facile de vérifier que la matrice $[B\ AB]$ est de rang 2).

(c) Le principe du minimum vérifié par un couple optimal $(\bar{x}, \bar{u})$ donne

$$\frac{d\bar{x}}{dt}(t) = A\bar{x}(t) + B\bar{u}(t)\ , \bar{x}(0) = x_0\ , \bar{x}(t^*) = 0\ ,$$

$$\frac{d\bar{p}}{dt}(t) = -A^t\bar{p}(t) \quad \text{et} \quad \forall v \in [-1, 1] \qquad \bar{p}_2(t)\bar{u}(t) \leq \bar{p}_2(t)v\ ,$$

où $\bar{p} = [\bar{p}_1, \bar{p}_2]^t$. Ces conditions sont suffisantes car le système est normal (avec le théorème 7.4.1).

(d) Le calcul de l'état adjoint se fait avec les équations précédentes.

On obtient $\dfrac{d^2\bar{p}_1}{dt^2}(t) = -\bar{p}_1(t)$, c'est-à-dire $\bar{p}_1(t) = p_1^o \cos t + p_2^o \sin t$. Donc $\bar{p}_2(t) = -p_1^o \sin t + p_2^o \cos t$ ce qui donne avec les notations de l'énoncé $\bar{p}_2(t) = -\rho \sin(t - \delta)$.

D'autre part, un raisonnement classique (voir les exercices précédents) montre que $\bar{u}(t) = -$ signe $(\bar{p}_2(t))$. On obtient donc $\bar{u}(t) = \text{signe}[\sin(t - \delta)]$ (p.p.).

(e) Supposons que u soit constant. L'état associé est de la forme

$$\theta(t) = r\sin(t + \alpha) \text{ et } \omega(t) = r\cos(t + \alpha) + u\ .$$

Les trajectoires sont donc des cercles d'équation $\theta^2 + (\omega - u)^2 = r^2$. Quand $u \equiv 1$, c'est un cercle de centre $(0, 1)$ et passant par $x_0 = (\theta_0, \omega_0)$. Si $u \equiv -1$, c'est un cercle de centre $(0, -1)$ et passant par. Le contrôle optimal est bang-bang et vaut 1 ou -1. Si x_0 est tel que le cercle de centre $(0, 1)$ (ou $(0, -1)$) passant par x_0, passe aussi par 0, on peut aller à 0 sans changer de trajectoire. C'est le cas si $1 = \theta_0^2 + (\omega_0 \pm 1)^2$. Sinon, on suit une trajectoire circulaire issue de x_0 jusqu'au moment où on rencontre une trajectoire associée au contrôle de signe opposé qui nous amène à 0 (voir la figure B8).

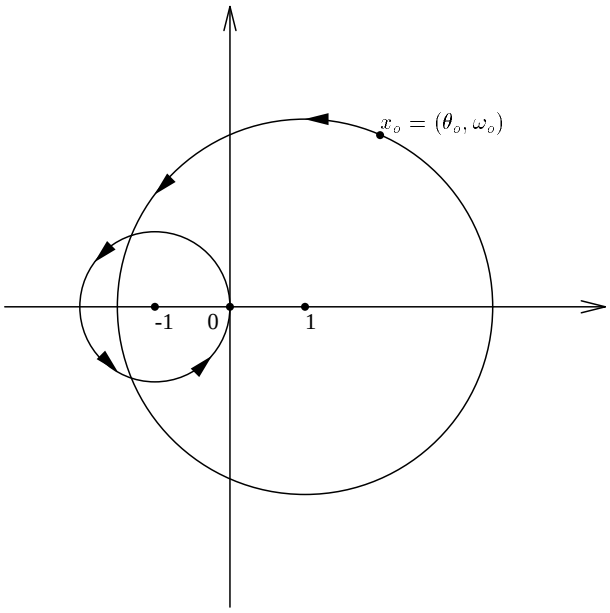

Figure B.8.

– **Exercice 6**

Les équations générales du système s'écrivent $\frac{dX}{dt}(t) = AX(t) + BU(t)\ , X(0) = X_0$, avec

$$X = \begin{bmatrix} x_1 \\ x_1' \\ x_2 \\ x_2' \\ x_3 \\ x_3' \end{bmatrix} \quad U = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \quad A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 3\omega^2 & 0 & 0 & 2\omega & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -2\omega & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -\omega^2 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Un logiciel de calcul formel type MAPLE© permet de calculer la matrice de contrôlabilité et de vérifier qu'elle est bien de rang 6. Donc le problème de contrôle en temps optimal a une solution. On vérifie de la même façon que le système est normal. La solution est donc unique. Ecrire les inéquations du principe de Pontryagin est désormais classique.

– **Exercice 7**

(a) La matrice A est la matrice nulle 0_2 et $B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Pour $i = 1, 2$ le vecteur colonne b_i de B vaut $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. La matrice $[b_i\ Ab_i] = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ est de rang 1, pour tout i : le système n'est donc pas normal.

(b) La matrice de contrôlabilité est

$$\mathbb{M} = [B\ AB] = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix};$$

elle est de rang 1 : donc le système n'est pas entièrement contrôlable. Précisons l'ensemble

$\mathcal{C}$ des points contrôlables. La solution de l'EDO est

$$x_1(t) = x_{0,1} + \int_0^t (u_1(s) + u_2(s))\, ds\ ,\ x_2(t) = x_{0,2} + \int_0^t (u_1(s) + u_2(s))\, ds\ ;$$

pour avoir $x_1(t) = x_2(t) = 0$, on doit avoir $x_{0,1} = x_{0,2}$. Par conséquent

$$\mathcal{C} = \{\ x_0 \in \mathbb{R}^2\ \mid\ x_{0,1} = x_{0,2}\ \}\ .$$

(c) On choisit $x_0 = (-1, 1)$; l'ensemble atteignable est alors

$$\mathcal{A}(t; x_0) = \{\ -1 + \int_0^t (u_1(s) + u_2(s))\, ds\ \mid |u_i(s)| \leq 1, \text{p.p.}, i = 1, 2\ \}\ \subset [-1-2t, -1+2t].$$

En fait $\mathcal{A}(t; x_0) = [-1 - 2t, -1 + 2t]$: on montre l'inclusion inverse en choisissant des contrôles u_i constants.
(d) Calculons un contrôle optimal amenant l'état de x_0 à 0 en temps minimal. Le principe de Pontryagin donne $\dfrac{dx}{dt} = Bu\ , x(0) = x_0,\ x(t^*) = 0$ et $\dfrac{dp}{dt} = 0$. Donc $p = (p_1, p_2)$ est constant. De plus

$$\forall (v_1, v_2) \in [-1, 1] \times [-1, 1] \qquad p_1 u_1(s) + p_2 u_2(s) \leq p_1 v_1 + p_2 v_2\ .$$

En prenant successivement $(v_2 = 0, v_1 = -\text{signe}\, p_1)$ et $(v_1 = 0, v_2 = -\text{signe}\, p_2)$, on obtient $u_i \equiv$-signe p_i pour $i = 1, 2$. De plus

$$x_i(t^*) = -1 + \int_0^{t^*} (u_1 + u_2)\, ds = -1 + t^*(u_1 + u_2) = 0.$$

Donc $u_1 + u_2 > 0$. La seule possibilité est $u_1 = 1, u_2 = 1$. Le contrôle est donc unique (avec $t^* = \dfrac{1}{2}$). Le problème admet une solution unique bien qu'il ne soit pas normal.

– **Exercice 8**
Ici $A = b < 0$, $B = 1$ et la matrice de contrôlabilité est 1. Le système est donc contrôlable et normal. Par conséquent il existe un temps optimal unique t^* (et un contrôle optimal unique $\bar{u}$) et les conditions du prncipe de Pontryagin sont nécessaires et suffisantes. Elles donnent : $p(t) = p_0 e^{-bt}$ et $p(t)\bar{u}(t) \leq p(t)v$ pour tout $v \in [-1, 1]$. Un raisonnement désormais classique (avec $v = -$ signe $p(t)$) donne $\bar{u}(t) = -$ signe $p(t) = -$ signe p_0. Par conséquent $\bar{u}$ est constant et vaut soit 1, soit -1. L'état est donné par : $\bar{x}(t) = \left(x_0 + \dfrac{u}{b}\right) e^{bt} - \dfrac{u}{b}$. Comme $\bar{x}(t^*) = 0$, nous obtenons $t^* = \dfrac{1}{b} \log\left(\dfrac{u}{u + bx_0}\right)$ avec bien sûr : $0 < \dfrac{u}{u + bx_0} \leq 1$.
Si $x_0 = 0$, le problème est sans intérêt : le système est déjà à 0.
Si $x_0 > 0$, alors $bx_0 < 0$ et

$$\frac{u}{u + bx_0} = \begin{cases} \dfrac{1}{1 + bx_0} & \text{si } u \equiv 1\ , \\ \dfrac{1}{1 - bx_0} & \text{si } u \equiv -1\ . \end{cases}$$

La seule solution possible est donc dans ce cas $u \equiv -1$.
On montre de la même façon que si $x_0 < 0$ la solution est $u \equiv 1$.

– **Exercice 9**

Une résolution partielle de l'EDO donne $\frac{dx_1}{dt} = \frac{dx_2}{dt}$; donc $x_1(t) = x_2(t)+C$ où C est une constante. Comme $x_{0,1} = x_{0,2}$, $x_1(t) = x_2(t)$ pour tout t. Ecrivons a priori le principe de Pontryagin : $\frac{dp}{dt} = 0$ donc $p = (p_1, p_2)$ est constant et $p_1u_1(t) \leq p_1v$ pour tout $v \in [-1, 1]$. Par un raisonnement classique, il vient que $u_1 = -$ signe $p_1 = \pm 1$, u_2 étant quelconque. L'état associé est alors $x_1(t) = -1+\int_0^t u_1(s)\,ds = tu_1 - 1 = x_2(t)$. On peut donc amener le système à 0 en prenant $tu_1 - 1 = 0$. Comme $t > 0$, on voit que la seule valeur possible de u_1 est 1. Dans ce cas $t = 1$ est le temps optimal. Toutefois, il y a une infinité de contrôles optimaux puisque tous les contrôles de la forme $(1, u_2)$ où u_2 est mesurable à valeurs dans $[-1, 1]$ conviennent.

– **Exercice 10**

(a) L'équation différentielle du second ordre (4) s'écrit sous la forme (5) :

$$\frac{dz}{dt}(t) = Az(t) + Bu(t)\ ,\ z(0) = z_0\ ,\quad \text{où}$$

$$z(t) = \begin{bmatrix} p(t) \\ p'(t) \end{bmatrix}\ , A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}\ ,\ B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\ .$$

La matrice de contrôlabilité est $\mathbb{M} = \begin{bmatrix} 0 & 1 \\ 1 & -3 \end{bmatrix}$; elle est de rang 2 : le système est contrôlable.

(b)

$$Q = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix}\ ,\ Q^{-1} = \begin{bmatrix} 2 & 1 \\ -1 & -2 \end{bmatrix}\ ,\ D = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}\ .$$

On pose $y = Q^{-1}z$: le système différentiel vérifié par y est alors

$$\frac{dy}{dt} = Dy + Q^{-1}Bu\ ,\ y(0) = Q^{-1}z_0 = y_0 \quad \text{avec}\ Q^{-1}B = \begin{bmatrix} 1 \\ -1 \end{bmatrix}\ .$$

(c) Le système est contrôlable et il est facile de vérifier qu'il est normal. Donc il existe un temps minimal unique et le principe de Pontryagin est nécessaire et suffisant. On note le contrôle optimal (unique) u^*.

(d) L'état adjoint donné par le principe de Pontryagin vérifie $\frac{dq}{dt} = -D^tq$. Cela donne $q_1(t) = q_{0,1}e^t$ et $q_2(t) = q_{0,2}e^{2t}$. De plus

$$\forall v \in [-1, 1] \qquad q_1(s)u(s) - q_2(s)u(s) \leq q_1(s)v - q_2(s)v\ ;$$

donc

$$u(s) = -\text{ signe } (q_{0,1}e^t - q_{0,2}\,e^{2t}) = -\text{ signe } (q_{0,1} - q_{0,2}\,e^t)\ .$$

Comme la fonction $t \mapsto q_{0,1} - q_{0,2}\,e^t$ est monotone, u change de signe au plus une fois.

(e) On prend u constant ($= \pm 1$). L'état est donné par

$$y_1(t) = (y_{0,1} - u)e^{-t} + u\ ,\ y_2(t) = (y_{0,2} - \frac{u}{2})e^{-2t} + \frac{u}{2}\ .$$

La trajectoire a pour équation :

$$y_2 = \frac{y_{0,2} - \dfrac{u}{2}}{(y_{0,1} - u)^2}(y_1 - u)^2 + \frac{u}{2} \, .$$

C'est donc une parabole. On note $\mathbb{P}^+$ la parabole correspondant à $u \equiv 1$ et $\mathbb{P}^-$ la parabole correspondant à $u \equiv -1$.

Pour que $0 \in \mathbb{P}^+$ il faut que $\dfrac{y_{0,2} - \dfrac{1}{2}}{(y_{0,1} - 1)^2} + \dfrac{1}{2} = 0$, c'est-à-dire

$$y_{0,2} = -\frac{1}{2}(y_{0,1} - 1)^2 + \frac{1}{2} \, .$$

C'est une parabole notée $\mathcal{P}^+$. De même, dans le cas où $u \equiv -1$ on obtient la parabole $\mathcal{P}^-$, d'équation

$$y_{0,2} = \frac{1}{2}(y_{0,1} + 1)^2 - \frac{1}{2} \, .$$

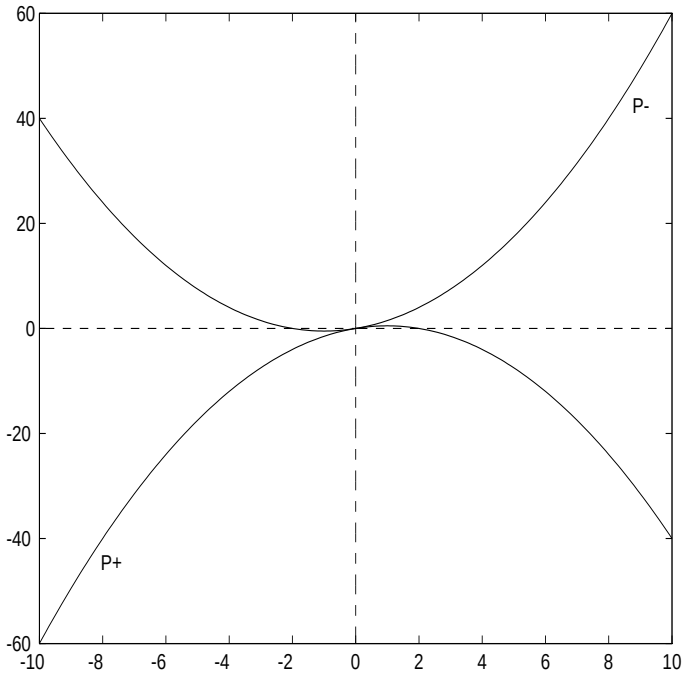

Figure B.9.

(f) Si $y_0 \in \mathcal{P}^+ \cup \mathcal{P}^-$, on va directement à 0 en suivant la trajectoire concernée. Sinon, on suit la trajectoire $\mathbb{P}^+$ (ou $\mathbb{P}^-$) jusqu'au moment où on coupe $\mathcal{P}^-$ (ou $\mathcal{P}^+$). A ce moment là, le contrôle change de signe et on suit la courbe $\mathcal{P}^-$ (ou $\mathcal{P}^+$) jusqu'à 0.

Chapitre 8

– **Exercice 1**
Les solutions minimales obtenues par l'algorithme de programmation dynamique sont données par :

Magasin	1	2	3	4
Solution 1	8	0	0	0
Solution 2	3	0	5	0

Le coût vaut alors 100.

– **Exercice 2**
Fixons $N \geq 2$. On a vu dans la section 8.1.4 que

$$\mathcal{V}_N(x) = -3x, \;\; \mathcal{V}_{N-1}(x) = -4.5x \text{ et } \; \mathcal{V}_{N-2}(x) = -5.6x \; ,$$

avec $u_N(x) = u_{N-1}(x) = 0$ et $u_{N-2}(x) = x$. Nous allons montrer par une récurrence descendante que

$$\forall n \leq N-2 \qquad \mathcal{V}_n(x) = \left(5.5\,(0.8)^{N-2-n} - 10\right)\, x \text{ et } u_n(x) = x \; .$$

D'après ce qui précède, c'est vrai pour $n = N-2$. Le principe d'optimalité de Bellman donne

$$\begin{aligned}
\mathcal{V}_{n-1}(x) &= \min_{0\leq u\leq x} J_{n-1}(u,x) + \mathcal{V}_n(f(x,u,t_{n-1})) \\
&= \min_{0\leq u\leq x} (u-3x) + \left(5.5\,(0.8)^{N-2-n} - 10\right)(0.5x + 0.3u) \qquad (*) \\
&= \min_{0\leq u\leq x} \left[1 + 0.3\left(5.5(0.8)^{N-2-n} - 10\right)\right] u + \left[0.5\left(5.5(0.8)^{N-2-n} - 10\right) - 3\right] x \\
&= -\max_{0\leq u\leq x} \underbrace{\left(2 - 1.65\,(0.8)^{N-2-n}\right)}_{\geq 0} u + \underbrace{\left(8 - 2.75\,(0.8)^{N-2-n}\right)}_{\geq 0} x
\end{aligned}$$

Le maximum est atteint pour $u = u_{n-1} = x$ et on obtient avec la ligne (*)

$$\mathcal{V}_{n-1}(x) = -2x + \left(5.5\,(0.8)^{N-2-n} - 10\right)(0.8x) = \left(5.5\,(0.8)^{N-2-(n-1)} - 10\right)\, x \; .$$

Le gain maximum en N étapes est donc $G_N = (10 - 5.5\,(0.8)^{N-2})\,\xi$.
Comme $\lim_{N\to+\infty} G_N = 10\xi$, on ne peut pas espérer un gain supérieur à 10ξ celui-ci étant "obtenu" en un nombre infini d'étapes.

– **Exercice 3**
Grâce au principe d'optimalité de Bellman, la fonction valeur est donnée par

$$\mathcal{V}_i(x) = \min_{u\in\mathbb{R}} J_i(x,u) + \mathcal{V}_{i+1}(\varphi(x,u)),$$

c'est-à-dire

$$\begin{aligned}
\mathcal{V}_i(x) = \min_{u\in\mathbb{R}} \; & \left\{ \left[a_i x^2 + b_i x\, u + c_i u^2 + d_i x + e_i u + f_i\right] \right. \\
& \left. + p_{i+1}\left(q_i x + h_i u + k_i\right)^2 + q_{i+1}\left(q_i x + h_i u + k_i\right) + r_{i+1} \right\} \; .
\end{aligned}$$

Donc

$$u = -\frac{[b_i + 2p_{i+1}g_ih_i]\,x + e_i + 2p_{i+1}h_ik_i + q_{i+1}h_i}{2(c_i + p_{i+1}h_i^2)}\ .$$

En reportant dans $\mathcal{V}_i(x)$ on obtient

$$p_i = a_i + p_{i+1}g_i^2 - \frac{b_i + 2p_{i+1}g_ih_i^2}{4(c_i + p_{i+1}h_i^2)}$$

$$q_i = d_i + 2p_{i+1}k_ig_i + q_{i+1}g_i - \frac{(b_i + 2p_{i+1}g_ih_i)(e_i + 2p_{i+1}h_ik_i + q_{i+1}h_i)}{2(c_i + p_{i+1}h_i^2)}\ ,$$

$$r_i = f_i + p_{i+1}k_i^2 + q_{i+1}k_i + r_{i+1} - \frac{(e_i + 2p_{i+1}h_ik_i + q_{i+1}h_i)^2}{4(c_i + p_{i+1}h_i^2)}\ ,$$

$$p_N = l\ ,\ q_N = w\ ,\ r_N = z\ .$$

– **Exercice 4**
Comme $\mathcal{V}_0(x)$ est le coût minimal pour un système dynamique commençant à x et finissant à 0, le choix optimal du x_0 minimise $\mathcal{V}_0$. Donc, si on suppose que p_0 est positif, nous obtenons $x_0 = -\dfrac{q_0}{2p_0}$.

– **Exercice 5**
Dans cet exercice, il faut prendre en compte le fait que les conditions finales sont imposées et modifier en conséquence l'équation d'état (ou son équivalent discret) pour l'avant dernière valeur de l'état (à $N-1$). Nous avons

$2 = x(3) = x(2)+y(2)+2u(2) = 2x(1)+3u(1)+2u(2)$ et $y(3) = 1 = 2y(1)+u(1)+u(2)$.

Donc $u(1) = -2x(1)+4y(1)$ et $u(2) = 1+2x(1)-6y(1)$. En outre $x(2) = -3x(1)+9y(1)$ et $y(2) = -x(1)+3y(1)$: ceci nous donne une nouvelle "équation d'état". Estimons la fonction valeur à l'aide du principe d'optimalité : $\mathcal{V}_3(x,y) = 0$ et $\mathcal{V}_2(x,y) = J_2(x,y,u(2))$.

$$\mathcal{V}_1(x,y) = J_1(x,y,u(1)) + J_2(\underbrace{-3+9y}_{x(2)}, \underbrace{-x+3y}_{y(2)}, \underbrace{1+2x-6y}_{u(2)})$$

$$= x^2 + y^2 + (-2x+4y)^2 + (-3x+9y)^2 + (-x+3y)^2 + (1+2x-6y)^2$$

$$= 19x^2 - 100x\,y + 143y^2 + 4x - 12y + 1\ .$$

$$\mathcal{V}_0(x,y) = \min_u \left(J_0(x,y,u) + \mathcal{V}_1(x+y+2u, x-y+u)\right)\ .$$

On obtient, avec $x = x(0) = 7$ et $y = y(0) = 1$: $\mathcal{V}_0(7,1) = \min\limits_u u^2 + \mathcal{V}_1(8+2u, 6+u)$.

Ceci donne

$$u = -8 = \qquad \mathcal{V}_0(7,1) = 64 + \mathcal{V}_1(-8,-2) = 245\ .$$

Nous obtenons alors successivement $x(0) = 7,\ y(0) = 1,\ u(0) = -8,$

$$x(1) = x(0) + y(0) + 2u(0) = -8\ ,\ y(1) = x(0) - y(0) + u(0) = -2,$$

$$u(1) = -2x(1) + 4y(1) = 8\ ,\ u(2) = 1 + 2x(1) - 6y(1) = -3\ ,$$

$$x(2) = -3x(1) + 9y(1) = 6 \quad \text{et} \quad y(2) = -x(1) + 3y(1) = 2\ .$$

On vérifie que le coût correspondant est $J = 245 = \mathcal{V}_0(7,1)$.

– **Exercice 6**
Cet exercice est une synthèse de tout ce qui a été fait précédemment. Nous renvoyons donc aux autres exercices pour la correction.
– **Exercice 7**
(a) L'état du système est donné par

$$\mathbf{x}(\tau) = x + \int_t^\tau u(s)\, ds \ ,$$

de sorte que $F(\mathbf{x}(T)) = F(x + \displaystyle\int_t^T u(s)\, ds)$. Posons $v = \displaystyle\int_t^T u(s)\, ds$. Il est facile de voir que

$$\{v \ \mid \ u \in \mathcal{U}_{ad}\ \} = [\,-(T-t), T-t\,] \ ;$$

par conséquent la fonction valeur vaut

$$\mathcal{V}(x,t) = \inf_{u\in\mathcal{U}_{ad}} F(\mathbf{x}(T)) = \inf_{v\in[\,-(T-t),T-t]} F(x+v) \ . \tag{B.9}$$

Montrons que $\mathcal{V}$ est paire par rapport à x :

$$\begin{aligned}
\mathcal{V}(-x,t) &= \inf_{v\in[\,-(T-t),T-t]} F(-x+v) \\
&= \inf_{v\in[\,-(T-t),T-t]} F(x-v) \text{ car } F \text{ est paire} \\
&= \inf_{-v\in[\,-(T-t),T-t]} F(x-v) \\
&= \inf_{v\in[\,-(T-t),T-t]} F(x+v) = \mathcal{V}(x,t) \ .
\end{aligned}$$

Grâce à la relation (B.9), la parité de F et la décroissance de F sur $\mathbb{R}^+$, on constate que

$$\mathcal{V}(x,t) = \min\left\{F(x+T-t), F(x-T+t)\right\} \ .$$

Supposons $x \geq 0$:
– Si $T-t \geq x$ alors $0 \leq T-t-x \leq T-t+x$ et la décroissance de F sur $\mathbb{R}^+$ donne : $F(-x+T-t) \leq F(x+T-t)$; c'est-à-dire avec la parité de F :

$$F(x-T+t) \leq F(x+T-t) \ .$$

Donc $\mathcal{V}(x,t) = F(x+T-t)$.
– Si $T-t \leq x$ alors $0 \leq x-T+t \leq x+T-t$ car $t \leq T$. On conclut de la même façon que $\mathcal{V}(x,t) = F(x+T-t)$.
Le cas $x \leq 0$ se traite de la même façon et on obtient :

$$\mathcal{V}(x,t) = \begin{cases} F(x+T-t) & \text{si } x \geq 0 \ , \\ F(x-(T-t)) & \text{si } x \leq 0 \ . \end{cases}$$

(b) Il est alors clair que $\mathcal{V}$ est régulière sur $\mathbb{R}^* \times]-\infty, T\,[$. Si on calcule les dérivées partielles en $x = 0$, à droite et à gauche, on obtient

$$\mathcal{V}'_x(0^+,t) = F'(T-t) \neq \mathcal{V}'_x(0^-,t) = F'(t-T) \ .$$

$\mathcal{V}$ n'est donc pas dérivable en $(0,t)$ pour tout $t < T$.

– **Exercice 8**

L'état étant donné par $\mathbf{x}(\tau) = x + \displaystyle\int_t^\tau u(s)\,ds$,, la fonction valeur vaut :

$$\mathcal{V}(x,t) = \inf_{u\in\mathcal{U}} F(x + \int_t^T u(s)\,ds)\ .$$

Posons $y = x + \displaystyle\int_t^T u(s)\,ds$; $u \in \mathcal{U}$ est équivalent à $|y-x| \le T-t$. Donc

$$\mathcal{V}(x,t) = \inf_{|y-x|\le T-t} F(y)\ .$$

(a) La fonction F est continue, donc elle admet un minimum sur le compact $|y-x| \le T-t$. Il existe donc toujours au moins un contrôle optimal. Ce contrôle n'est a priori pas unique puisqu'on a aucune hypothèse de stricte convexité (par exemple) sur F.
(b) Le problème de contrôle optimal s'écrit de la manière suivante :

$$\min F(\mathbf{x}(T))\ ,\ \mathbf{x}' = u \text{ sur }]\,t,T[\ ,\ \mathbf{x}(t) = x\ , u \in \mathcal{U}\ .$$

Comme au chapitre 5, on peut écrire une condition d'optimalité en $(\bar{\mathbf{x}}, \bar{u})$:

$$\forall v \in \mathcal{U}\ ,\ \forall \mathbf{z} \text{ vérifiant } \mathbf{z}' = v,\ \mathbf{z}(t) = x \quad F'(\bar{\mathbf{x}}(T))(\mathbf{z}(T) - \bar{\mathbf{x}}(T)) \ge 0\ . \tag{B.10}$$

Posons $\bar{p}'(s) = 0$ sur $]\,t,T[\ ,\ \bar{p}(T) = F'(\mathbf{x}(T))$; nous obtenons alors pour tous $v \in \mathcal{U}$ et $\mathbf{z}$ vérifiant $\mathbf{z}' = v,\ \mathbf{z}(t) = x$:

$$F'(\bar{\mathbf{x}}(T))(\mathbf{z}(T) - \bar{\mathbf{x}}(T)) = \bar{p}(T)(\mathbf{z}(T) - \bar{\mathbf{x}}(T))$$

$$= \int_t^T \underbrace{\bar{p}'(s)}_{=0}(\mathbf{z}(s) - \bar{\mathbf{x}}(s))\,ds + \int_t^T \bar{p}(s)(\mathbf{z}'(s) - \bar{\mathbf{x}}'(s))\,ds + \bar{p}(t)\underbrace{(\mathbf{z}(t) - \bar{\mathbf{x}}(t))}_{=x-x=0}$$

$$= \int_t^T \bar{p}(s)(v(s) - \bar{u}(s))\,ds\ .$$

La relation (B.10) devient

$$\forall v \in \mathcal{U} \qquad \int_t^T \bar{p}(s)(v(s) - \bar{u}(s))\,ds\ \ge 0\ ,$$

ce qui entraîne (1) :

$$\forall w \in [-1,1] \qquad \bar{p}(s)\bar{u}(s)\,ds\ \le \bar{p}(s)w\ ,\ \text{p.p. } s \in]\,t,T[\ .$$

Comme $\bar{p}' = 0$, $\bar{p}$ est constant égal à $F'(\bar{\mathbf{x}}(T))$. La relation (1) donne alors

$$\bar{u}(s) \equiv -\ \text{signe}\ F'(\bar{\mathbf{x}}(T) = \pm 1\ .$$

Par conséquent, $\bar{\mathbf{x}}(s) = x + \bar{u} \cdot (s-t)$ et $\bar{\mathbf{x}}(T) = x + \bar{u} \cdot (T-t) = x \pm (T-t)$. La fonction valeur vaut alors $\mathcal{V}(x,t) = F(x \pm (T-t))$.
(c) Ecrivons l'équation de Hamilton-Jacobi-Bellman vérifiée par $\mathcal{V}$ en tout point où elle est différentiable ; ici $L \equiv 0$ et $\varphi(x,u,t) = u$. Le Hamiltonien est donc

$$\mathcal{H}(x,t,p) = \inf_{w \in [-1,1]} p \cdot w = \begin{cases} -p & \text{si } p \geq 0 \\ p & \text{si} p \leq 0 \ . \end{cases}$$

L'équation HJB est

$$\frac{\partial \mathcal{V}}{\partial t} + \mathcal{H}(x,t,\frac{\partial \mathcal{V}}{\partial x}) = 0 \ , \ \mathcal{V}(x,T) = F(x).$$

Si $\dfrac{\partial \mathcal{V}}{\partial x} \geq 0$, on obtient

$$\frac{\partial \mathcal{V}}{\partial t} - \frac{\partial \mathcal{V}}{\partial x} = 0 \ , \ \mathcal{V}(x,T) = F(x);$$

c'est une équation de transport dont la solution est $\mathcal{V}(x,t) = F(x+t-T)$.
Si $\dfrac{\partial \mathcal{V}}{\partial x} \leq 0$, on obtient $\mathcal{V}(x,t) = F(x-t+T)$.
On retrouve ainsi le résultat de la question précédente. On conclut dès qu'on connaît explicitement F et surtout le signe de F'.

Errata

- p 9 : Exemple 1.3.2. : $\|x\|^2$ au lieu de $\|x\|$
- p 12 : Définition 1.3.7. : Rajouter linéaire **continue**
- p 13 : Ligne 6 : $f(x,y) =$ **y** si $x = y^2$
  Théorème 1.3.2 : $\mathcal{C}$ au lieu de $\mathbb{H}$
  Démonstration du Théorème 1.3.2 : première ligne : soient u et v dans $\mathcal{C}$.
  Ligne 7 : $J(v) \geq J(u + t(v-o))+(1-t)\nabla J(\cdots$
  Théorème 1.3.3 : Enlever : opérateur monotone **[de $\mathbb{H}$ dans $\mathbb{H}$]**
  Démonstration du Théorème 1.3.3 : première ligne : soient u et v dans $\mathcal{C}$.
- p14 : Ligne 1 : soient (u, v) dans $\mathcal{C} \times \mathcal{C}$.
- p 36 : Ligne suivant 2.4.3 : $r = \min(r_o, \dfrac{2}{mM}, 1)$.
- p 49. 2.2 : rajouter « **positive** »
- p 51 : c) Rajouter :« où $\rho(M)$ désigne le rayon spectral de M »
- p 62. Théorème 3.3.1 : ($\mathcal{CQ}$**2**) de la définition 3.2.**4**
- p 74 : Formule 3.5.2 : $\nabla_x \mathcal{L}(x_k, \lambda_k)$ au lieu de $\nabla_x \mathcal{L}(x_k, \lambda_k)^t$
  et dernière ligne $\nabla J(x_k)$ au lieu de $\nabla J(x_k)^t$
- p 92 : Exercice 3.8 : matrice A : le coefficient $a_{2,3}$ est -1 et non 0
- p 108 : Définition 4.2.1 : x_o au lieu de y_o.
- p 113 : $\tau_1 = -q_o + \sqrt{p^* + \frac{q_o^2}{2}}$ et $\tau_2 = -\left(q_o + \sqrt{p^* + \frac{q_o^2}{2}}\right)$. Puis :

  pour que τ_1 soit solution il faut que $q_o \leq \sqrt{p^* + \frac{q_o^2}{2}}$ c'est-à-dire $|q_o| \leq \sqrt{2p^*}$.
  Supprimer les 7 lignes après *ne démarre pas ...*
- p 120 . dernière ligne $\int_o^t$ au lieu de $\int_0^T$.
- p 125. $L^2(0, T; \mathbb{R}^p)$ au lieu de $\mathcal{U}$
- p 136 : $\frac{1}{2r}$ au lieu de $\frac{r}{2}$
- p 171. b) ma**t**rice
- p 217 : Correction de l'exercice 3.8 : $x_o = (-3/2, -1/2, -1/2)$, $x^* = (1, 1, 0)$, $\lambda^* = 0$ et $\mu^* = 1$. $J(x_o) = -3/4 < J(x^*) = 1/2$.
- p 233 : ligne -4 en partant du bas : $3C\sqrt{2}$ au lieu de $3C$ dans l'expression de $x(t)$
- p 253. dernière ligne : connaît

Index

Bibliographie

[1] **M. Amouroux - A. El Jai**, *Automatique des systèmes distribués*, Traité des nouvelles technologies, série Automatique, Hermès, Paris 1990.

[2] **D.P. Bertsekas**,*Projected Newton Method for Optimization Problems with Simple Constraints*, SIAM Journal on Control and Optimization, Vol. 20, n° 2, Mars 82, pp. 221-246.

[3] **G. Barles**, *Solutions de viscosité des équations d'Hamilton-Jacobi*, Collection Mathématiques et Applications, Vol. 17, Springer-verlag, 1994.

[4] **J.F. Bonnans- J.C. Gilbert - C. Lemaréchal - C. Sagastizabal**, *Optimisation numérique. Aspects théoriques et pratiques* , Collection Mathématiques et Applications, Vol. 27, Springer-verlag, 1998.

[5] **H. Brezis**, *Analyse fonctionnelle*, Collection Mathématiques Appliquées pour la Maîtrise, Masson,1987.

[6] **H. Cartan**, *Cours de calcul différentiel* , Hermann, Paris, 1977.

[7] **P.G. Ciarlet**, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Collection Mathématiques appliquées pour la maîtrise, Masson, Paris, 1988.

[8] **M.Crouzeix - A.Mignot**, *Analyse Numérique des équations Différentielles*, Collection Mathématiques appliquées pour la maîtrise, Masson, Paris, 1989.

[9] **M. Duflo**, *Algorithmes stochastiques*, Mathématiques et Applications 23, Springer-Verlag, 1996.

[10] **F. Ecoto**, *Initiation à la Recherche Opérationnelle*, Ellipses, Paris, 1986.

[11] **I.Ekeland - R.Temam**, *Analyse Convexe et Problèmes Variationnels*, Dunod, Gauthier-Villars, Paris, 1974.

[12] **P. Faurre**, *Analyse Numérique- Notes d'optimisation*, Collection X, Ellipses, Paris, 1988.

[13] **M. Fortin - R. Glowinski**, *Méthodes de Lagrangien Augmenté. Application à la résolution de problèmes aux limites*, Dunod-Bordas, Paris, 1982.

[14] **N. Yamashita - M. Fukushima**, *On the rate of Convergence of the Levenberg-Marquardt Method* , preprint Octobre 2000, www-optima.amp.i.kyoto-u.ac.jp/ fuku/index-e.html

[15] **J.B. Hiriart-Urruty**, *L'optimisation*, Que sais-je ?, PUF, Paris, 1996.

[16] **B. Larrouturou - P.L. Lions**, *Méthodes mathématiques pour les sciences de l'ingénieur : optimisation et analyse numérique*, Cours de l'Ecole Polytechnique, 1994.

[17] **P.L. Lions**, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Londres, 1982.

[18] **D.G. Luenberger**, *Linear and Nonlinear Programming*, deuxième édition, Addison-Wesley, Reading, Massachusets,1984.

[19] **D.G. Luenberger**, *Optimization by Vector Space Methods*, Wiley, New York,1969.

[20] **G.L. Nemhauser - A.H.G. Rinnooy Kan - M.J. Todd (eds)**, *Optimization*, Handbook in operations research and management science, Vol.1, North-Holland , 1989.

[21] **P.A. Raviart-J.M. Thomas**, *Introduction à l'analyse numérique des EDP*, Collection Mathématiques appliquées pour la maîtrise, Masson, Paris,1988.

[22] **H. Reinhard**, *Equations différentielles, fondements et applications*, Dunod Université, Bordas, Paris,1989.

[23] **K. Zhou - J.C.Doyle - K. Glover**, *Robust and Optimal control*, Prentice-Hall, New-Jersey,1996.

Table des matières