Enseignement de l'informatique

POLYCOPIE

Méthodes d'optimisation

LEBBAH Yahia
Professeur, Université d'Oran 1

Département Informatique, Faculté des Sciences Exactes et Appliquées, Université Oran1
B.P. 1524, El-M'Naouar Oran, Algérie

Version du 1er octobre 2025

Table des matières

Chapitre 1

Introduction

Chapitre 2

Préliminaires sur l'optimisation

2.1 optimisation combinatoire : problèmes faciles et problèmes difficiles[Prins, 1994]

2.1.1 Problèmes faciles

Exploration d'un graphe Donnée : Un graphe orienté $G = (X, U)$, deux sommets s et t de X. Question : Existe-t-il un chemin de s à t ? Algorithme en $O(M)$.

Chemin de coût minimal Données : $G = (X, U, C)$ un graphe orienté valué, et deux sommets s et t de X. Question : Trouver un chemin de coût minimal de s à t. Algorithme de Bellman en $O(NM)$. Algorithme de Dijkstra en $O(N^2)$.

Flot maximal Données : Un réseau de transport $G = (X, U, C, s, t)$. Question : Maximiser le débit du flot qui peut s'écouler dans le réseau entre s et t. Algorithme de Ford-Fulkerson en $O(NM^2)$.

Arbre recouvrant de poids minimal Données : $G = (X, E, W)$ un graphe simple valué. Question : Trouver un arbre recouvrant de poids minimal. Algorithme de Prim en $O(N^2)$. Algorithme de Kruskal en $O(M \log N)$. Si G est orienté, on pourrait calculer une arborescence recouvrante en $O(MN)$ avec l'algorithme de Edmonds.

Couplage Données : Soit $G = (X, E)$ un graphe simple. Un couplage de G (matching) est un sous-ensemble d'arêtes tel que deux quelconques d'entre elles n'aient aucun sommet commun. Question : Trouver un couplage de cardinalité maximale.

Parcours eulériens et chinois Données : Un parcours eulérien passe une fois par chaque arc ou arête. Le problème d'existence est solvable en $O(M)$. Un parcours chinois visite au moins une fois chaque arête.

Test de planarité Données : Un graphe simple $G = (X, E)$. Question : G est-il planaire, c'est-à-dire dessinable dans le plan sans croisement d'arêtes ? Un algorithme en $O(M)$ dû à Hopcroft et Tarjan.

Test de bipartisme On peut démontrer qu'un graphe est biparti ssi ne contient pas de cycle impair. Algorithme en $O(M)$.

Recherche d'une information parmi N Il s'agit de la recherche d'un élément dans un tableau de N éléments.

Tri de N **nombre** Solvable avec l'algorithme du tri par tas en $O(n \log n)$.

Programmation linéaire Il s'agit de résoudre le problème d'optimisation :

$$\begin{cases} min\ c.x \\ A.x \leq b \\ x \in I\!R^n, x \geq 0 \end{cases}$$

L'algorithme du simplexe est performant en moyenne, mais exponentiel dans le pire des cas. Karmarkar a proposé en 1984 un algorithme polynomial en $O(n^{3.5}L)$ où L est le nombre de bits pour coder A, b et c.

2.1.2 Problèmes difficles

Stable maximal Données : Un graphe simple $G = (X, E)$. Un sous-ensemble de sommets S est un ensemble stable s'il n'y a pas d'arête entre deux sommets quelconques de S.

Transversal minimal Données : Un graphe simple $G = (X, E)$. Un sous-ensemble de sommets T est un Transversal si toute arête de G a au moins une extrémité dans T.

Clique maximale Données : Un graphe simple $G = (X, E)$. Un sous-ensemble de sommets Q est une clique si toute paire de sommets de Q est reliée par une arête. Q engendre donc un graphe complet.

Coloration minimale Données : Un graphe simple $G = (X, E)$. G est k-colorable si on peut colorer ses sommets avec k couleurs distinctes, sans que deux sommets voisins aient la même couleur. Le plus petit k pour lequel G est k-colorable est le nombre chromatique de G.

Problèmes hamiltoniens Données : $G = (X, U, C)$ un graphe orienté valué. Le problème d'existence d'un parcours hamiltonien dans un graphe G est difficile. Le problème du voyageur de commerce consiste à trouver un circuit ou un cycle hamiltonien, de coût minimal, dans un graphe valué complet.

Problème SAT Données : une formule clausale. Question : Peut-on affecter à chaque variable propositionnelle de façon à rendre toutes les clauses vraies.

Sac à dos en variables entières Il s'agit de résoudre le problème :

$$\begin{cases} min\ c.x \\ a.x \leq b \\ x \text{ entier} \end{cases}$$

Bin packing On donne n objets de poids a_i et un nombre non limité de boites de capacité b. Le but est de répartir les objets en un nombre minimal de boites.

2.1.3 NP-complétude et les classes P et NP [Prins, 1994]

Certains problèmes d'optimisation combinatoire disposent d'algorithmes polynomiaux, tandis que d'autres n'en ont toujours pas. Existe-t-il réellement une classe de problèmes combinatoires pour lesquels on ne trouvera jamais d'algorithmes polynomiaux, ou est ce que les problèmes difficiles ont en fait de tels algorithmes, mais non encore découverts ? On conjecture depuis longtemps l'existence d'une classe de problèmes intrinsèquement difficiles, car plusieurs problèmes difficiles (comme le problème du voyageur de commerce) résistent depuis plus de 50 ans à l'assaut des travaux de recherche : malgré ces efforts, aucun algorithme polynomial n'a été trouvé.

La théorie de la complexité a été développée dans les années 1970 pour répondre à cette question. Le principal résultat est que tous les problèmes difficiles sont liées : la découverte d'un algorithme polynomial pour un seul problème difficile permettrait de déduire des algorithmes polynomiaux pour tous les autres.

La théorie de la complexité ne traite que des problèmes d'existence, à réponse oui/non. Ceci n'est pas gênant pour les problèmes d'optimisation. Un algorithme efficace pour le problème d'existence peut être utilisé pour résoudre efficacement son problème d'optimisation associé, par une simple dichotomie sur les valeurs de la fonction objectif.

2.1.4 La classe P et NP [Prins, 1994, Cori et al., 2001]

Etant donné une fonction $f : I\!N \rightarrow I\!N$, on dira qu'un problème appartient à la classe $DTIME(f)$ s'il existe une machine déterministe qui sur tout entrée de longueur n, résout le problème en $O(f(n))$ pas de calcul.

L'ensemble des problèmes d'existence qui admettent des algorithmes polynomiaux forment la classe P. $On peut poser P = \cup_{k\geq 0} DTIME(n \longmapsto n^k)$**.** Il faut pouvoir vérifier en temps polynomial une proposition de réponse Oui.

La classe NP est celles des problèmes d'existence dont une proposition de solution Oui est vérifiable polynomialement. On définit également la classe $NP = \cup_{k\geq 0} NTIME(n \longmapsto n^k)$, où N signifie non-déterministe. Le modèle de calcul non-déterministe est enrichi par une instruction de "choix" où il existe une façon immédiate pour conduire à la solution.

Les problèmes qui ne sont pas dans NP existent, mais ne présentent qu'un intérêt théorique pour la plupart. NP inclut P.

Pour un problème sans algorithme efficace, il faut procéder comme suit pour prouver l'appartenance à NP :

1. proposer un codage de la solution (appelé certificat) ;
2. proposer un algorithme qui va vérifier la solution au vu des données du certificat ;
3. montrer que cet algorithme a une complexité polynomiale.

Considérons le problème suivant : **étant donné un ensemble S de n nombres entiers et un entier b, existe-t-il un sous-ensemble T de S dont la somme des éléments est égale à b ?** On ne connait pas d'algorithme polynomial pour résoudre ce problème. Il n'empêche qu'il est dans NP, car **vérifier qu'une somme d'un ensemble d'entiers T, sous-ensemble d'un ensemble S de cardinalité n, est égale à b, est en $O(n^2)$**. Dans cette vérification, il faut s'assurer que tous les éléments de T sont dans S, qui nécessitera au plus n^2 tests.

2.1.5 Les problèmes NP-complets [Prins, 1994, Cori et al., 2001]

Il s'agit des problèmes les plus difficiles de NP, le "noyau dur". La notion de problème NP-complet est basée sur celle de transformation polynomiale d'un problème. Un problème d'existence P_1 se transforme polynomialement en un autre P_2 s'il existe un algorithme polynomial A transformant toute donnée pour P_1 en une pour P_2, en conservant la réponse Oui et Non. Par exemple, un stable d'un graphe simple G est une clique dans le graphe complémentaire G_c.

Un problème NP-complet est un problème de NP en lequel se transforme polynomialement tout autre problème de NP. La classe des problèmes NP-complets est notée NPC. On rencontre dans la littérature le terme NP-difficile pour les problèmes d'optimisation : un problème d'optimisation combinatoire est NP-difficile si le problème d'existence associé est NP-complet.

La conséquence pratique forte de la classe NPC : si on trouverait un algorithme polynomial A pour un seul problème NP-complet X, on pourrait en déduire un autre pour tout autre problème difficile Y de P. Il suffit de transformer polynomialement les données de Y en données pour X, puis exécuter l'algorithme pour X.

Une question immédiate est de savoir si de tels problèmes existent réellement dans NP. Le logicien Cook (et Levin) a montré en 1970 que **le problème SAT est NP-complet**. Depuis cette date, les travaux de recherche ont montré que la plupart des problèmes d'existence associés aux problèmes difficiles sans algorithmes polynomiaux connus sont NP-complets.

Problème SAT

Données Soit un ensemble de de variables $\{x_1, ..., x_n\}$. Soit une formule logique F sous forme normale conjonctive $F = C_1 \wedge C_2 \wedge ... \wedge C_l$, avec chaque clause $C_i = (y_{i,1} \vee y_{i,2} \vee ... \vee y_{i,k_i})$, où chaque $y_{i,j}$ est un littéral, c'est-à-dire $y_{i,j} = x_i$ ou $y_{i,j} = \neg x_i$.

Résultat "oui" ssi F est satisfaisable, qu'il existe une affectation de valeurs de vérités aux variables qui rende F vraie.

Proposition 1. *Le problème SAT est NP-complet.*

Preuve : Premier problème démontré NP-complet. Preuve réalisée par Cook et Levin [Cook, 1971, Levin, 1973]. □

2.2 Optimisation continue

2.2.1 Concepts de base d'optimisation continue

Définition 1 (problème d'optimisation continue (POC))**.** *Un problème général d'optimisation continue (POC) s'exprime comme suit : trouver des valeurs des variables de décision $x_1, ..., x_n$ telles que*

$$\begin{aligned} &minz = f(x_1, ..., x_n) \\ &g_i(x_1, ..., x_n) = 0, i = 1..m_e \\ &g_j(x_1, ..., x_n) \leq 0, j = m_e + 1..m. \end{aligned}$$

Définition 2 (région (ou espace) faisable). *L'ensemble des points $(x_1, ..., x_n)$ satisfaisant les m contraintes dans un POC est appelé espace faisable. Un point dans cet espace est dit point faisable, sinon on dit point infaisable.*

Définition 3 (minimum global et solution optimale). *Un point x^* de l'espace faisable pour lequel $f(x^*) \leq f(x)$ est vérifiés pour tous les points faisables x est dit minimum global (solution optimale) du POC.*

Définition 4 (minimum local et solution locale). *Un point faisable $x' = (x'_1, ..., x'_n)$ est un minimum local (solution locale) s'il existe un ϵ tel que $\forall x = (x_1, ..., x_n)$ avec*

$$\forall i, |x_i - x'_i| < \epsilon, f(x) \geq f(x').$$

2.2.2 Fonctions convexes

Définition 5 (ensemble convexe). *Un ensemble s est convexe si $\forall x', x'' \in s$, tout point du segment liant x' à x'' appartient à s ; c'est-à-dire*

$$\forall c \in [0, 1], cx' + (1 - c)x'' \in s.$$

Définition 6 (fonction convexe et fonction concave). *Soit $f(x_1, ..., x_n)$ une fonction définie sur un ensemble convexe s est convexe (resp. concave)si $\forall x' \in s, x'' \in s$*

$$f(cx' + (1 - c)x'') \leq cf(x') + (1 - c)f(x'')$$

$$(resp.\ f(cx' + (1 - c)x'') \geq cf(x') + (1 - c)f(x''))$$

pour tout $c \in [0, 1]$.

Théorème 1 (optimum d'un POC convexe). *Supposons que l'espace faisable s d'un POC est convexe. Si $f(x)$ est convexe dans s, alors tout minimum local du POC est un minimum global.*

Preuve : exercice !

Théorème 2 (convexité et concavité). *Supposons que $f''(x)$ existe pour tout x dans un espace convexe, alors $f(x)$ est convexe (resp. concave) dans s ssi $\forall x, f''(x) \geq 0$ (resp. $f''(x) \leq 0$).*

Définition 7 (matrice hessienne). *La matrice hessienne de la fonction $f : x = (x_1, ..., x_n) \rightarrow f(x) = f(x_1, ..., x_n)$ est la matrice carrée $n \times n$ où la (i, j)ième entrée est définie par $\frac{\partial^2}{\partial x_i \partial x_j}$.*

2.2.3 Illustration : résoudre un POC sans contraintes avec une seule variable

Soit le POC suivant

$$min_{x \in [a,b]} f(x).$$

Naïvement, on peut calculer l'ensemble de tous les minimums locaux puis on prend le plus petit. Nous avons trois cas :

1. points x où $a < x < b$ et $f'(x) = 0$ (appelés points stationnaires)
2. points x où $f'(x)$ n'existe pas
3. les points extremums a et b de $[a, b]$

Parmi les points stationnaires, on prendrait seulement ceux qui sont des minimums locaux, car on peut avoir $f'(x) = 0$ sans que le point soit un minimum local : $f'(x) = 0$ est une condition nécessaire mais insuffisante pour que le point x soit un minimum local. Nous avons donc besoin de propriétés mathématiques pour localiser les minimums locaux. Et en plus, nous devons aussi considérés les valeurs de f sur les points extremums a et b et aux points où les dérivées ne sont pas disponibles car ces derniers peuvent contenir le minimum global.

Voici un théorème qui donne quelques propriétés pour localiser un minimum local.

Théorème 3 (localisation des minimums locaux). *Si $f'(x_0) = 0$ et $f''(x_0) > 0$ alors x_0 est un minimum local.*

On veut résoudre le POC unidimensionnel sans contraintes

$$max_{a \leq x \leq b} f(x)$$

en se privant d'utiliser la dérivée $f'(x)$.

Dans cette section, on décrit comment résoudre ce POC si la fonction est unimodulaire.

Définition 8 (fonction unimodulaire). *La fonction $f : x \to f(x)$ est unimodulaire dans $[c, d]$ si $\exists x^* \in [a, b]$ où $f(x)$ est strictement croissante dans $[a, x^*)$ et décroissante dans $[x^*, b]$.*

Ce problème peut être résolu simplement par dichotomie. La procédure de résolution se présente comme suit :

1. Soient x_1 et x_2 des points distinct dans $[a, b]$.
2. Si $f(x_1) = f(x_2)$ alors nécessairement le maximum est dans l'intervalle $[a, x_2]$.
3. Si $f(x_1) < f(x_2)$ alors nécessairement le maximum est dans l'intervalle $[x_1, b]$.
4. Si $f(x_1) > f(x_2)$ alors nécessairement le maximum est dans l'intervalle $[a, x_2]$.
5. Relancer ce traitement sur l'intervalle restant jusqu'à ce que sa largeur soit inférieure à la précision exigée par l'utilisateur.

Cette recherche est dite méthode du nombre d'or si à chaque itération on prend $x_2 = a + (b - a)0.618$ et $x1 = b - (b - a)0.618$. Le nombre 0.618 est le nombre d'or. Il est appelé ainsi car il est la solution du problème originale suivant : Soit une tige de longueur 1. On veut trouver un point r de la tige tel que

$$\frac{\text{longueur de la tige}}{\text{longueur de la plus grande partie de la tige}} = \frac{\text{longueur de la plus grande partie}}{\text{longueur de la plus petite partie}}.$$

On a donc

$$\frac{1}{r} = \frac{r}{1 - r}.$$

La seule solution de ce problème est $r = \frac{5^{1/2}-1}{2} = 0.618$.

TP : mettre en oeuvre cet algo !

2.2.4 Algorithmes et convergence

La plupart des algorithmes que l'on va aborder sont des procédés itératifs. Un algorithme est donc vu comme une application F d'un espace U dans lui même. Le déroulement de cet algorithme à partir d'une valeur initiale u^0 sera de la forme

$$u^0 \in U, u^k \to u^{k+1} = F(u^k), k = 1, 2, ...$$

A F, on peut associer associer l'ensemble de ses points fixes $F^\infty(U)$ défini par

$$F^\infty(U) = \{u \in U | F(U) = U\}.$$

Un point fixe u^* est dit attractif s'il admet un voisinage $V(u^*)$ tel que

$$\forall u \in V(u^*), lim_{k\to\infty} F^k(u) = u^*.$$

On défini le bassin d'attraction associé à u^* l'ensemble des points u^0 tel que $lim_{k\to\infty} F^k(u^0) = u^*$.

Définition 9 (vitesse de convergence).
1. *si* $lim_{k\to\infty} \frac{\|u^{k+1}-u^*\|}{\|u^k-u^*\|} = \alpha < 1$, *on dit que la convergence est linéaire et* α *est la taux de convergence associé.*
2. *si* $lim_{k\to\infty} \frac{\|u^{k+1}-u^*\|}{\|u^k-u^*\|} = 0$, *on dit que la convergence est superlinéaire.*
3. *si* $\exists\gamma, lim_{k\to\infty} sup \frac{\|u^{k+1}-u^*\|}{\|u^k-u^*\|} = \alpha < 1$, *on dit que la convergence est superlinéaire d'ordre* γ, *et en particulier si* $\gamma = 2$, *on parle de vitesse de convergence quadratique.*

[?, ?] définissent

$$d_k = -\log_{10} |u^k - u^*|,$$

cette quantité, à une approximation près, représente le nombre de chiffres décimaux exacts de u_k.

Exemple 1. *[?]*
Soit la suite :

$$u_k = 12 + \frac{1}{n}$$

qui converge vers 12. *Nous avons :*

$$\begin{array}{rclrcl} u_{100} & = & 12.01 & d_{100} & = & 2 \\ u_{1000} & = & 12.001 & d_{1000} & = & 3 \\ u_{10000} & = & 12.0001 & d_{10000} & = & 4 \end{array}$$

Lorsque k est suffisamment grand, on peut trouver une constante C telle que :

$$|u_{n+1} - u^*| \approx C|u_k - u^*|^r$$

Nous avons donc $d_{n+1} = r \times d_n - \log_{10}(C)$. C'est-à-dire qu'à chaque itération, nous multiplions par environ r le nombre de chiffres exacts et nous ajoutons $R = -\log_{10}(C)$. Si $r = 1$, nous ne faisons qu'ajouter R chiffres décimaux par itération.

Exemple 2. *[?]*
Si $C = 0.999$ *alors* $R = 1/2500$. *Cela signifie que l'on obtient un nouveau chiffre exact toutes les 2500 itérations. Par contre, si* $r = 1.01$, *on multiplie par* 2 *le nombre de chiffres exacts toutes les* 70 *itérations ; car pour avoir* $2 = 1.01^n$, *on obtient* $n = log(2)/log(1.01) = 70$. *Si* $r = 2$, *on multiplie par 2 le nombre de chiffre significatifs à chaque itération* $n = 1$.

Ces exemples montrent bien l'intérêt d'utiliser les suites d'un ordre supérieur à 1. C'est pour cette raison que la convergence d'ordre deux ou *quadratique* est caractérisée de convergence *rapide*.

2.3 Récursivité et approche de résolution diviser pour régner [Gaudel et al., 1990, Cormen et al., 1994, Cormen et al., 2001]

L'expression d'algorithmes sous forme récursive permet des descriptions concises qui se prêtent bien à des démonstrations par récurrence. **Le principe est d'utiliser, pour décrire l'algorithme sur une donnée d, l'algorithme lui-même appliqué à un sous-ensemble de d ou à une donnée d' plus petite.** Le programme de calcul des nombres de FIBONACCI donné dans le chapitre précédent est une bonne illustation du procédé.

Nombre d'algorithmes utiles sont de structure récursive : pour résoudre un problème donné, ils s'appellent eux-mêmes récursivement une ou plusieurs fois sur des sous-problèmes très similaires. Ces algorithmes choisissent l'approche "diviser pour régner" : ils séparent le problème en plusieurs sous-problèmes similaires au problème initial, mais de taille moindre, résolvent les sous-problèmes de façon récursive, puis combinent ces solutions pour retrouver une solution au problème initial.

Ce paradigme de résolution donne lieu à trois étapes à chaque niveau de récursivité :

Diviser le problème en un certain nombre de sous-problèmes.

Régner sur les sous-problèmes en les résolvant récursivement. Par ailleurs, si la taille d'un sous-problème est assez réduite, on peut le résoudre directement.

Combiner les solutions aux sous-problèmes en une solution complète pour le problème initial.

2.3.1 Tri par fusion

L'algorithme de tri par fusion suit très fidèlement la règle du "diviser pour régner". Il agit de la manière suivante :

Diviser Diviser la séquence de n éléments à trier en deux sous-séquences de $n/2$ éléments.

Régner Trier les deux séquences récursivement à l'aide du tri par fusion.

Combiner Fusionner les deux sous-séquences triées pour produire la réponse triée.

Algorithm 1 TRI-FUSION(A, p, r)

Input: Une séquence de n nombres $A = \langle a_1, a_2, ..., a_n \rangle$.
Output: Une permutation $\langle a'_1, a'_2, ..., a'_n \rangle$ de la permutation de la suite d'entrée telle que $a_1 \leq a_2 \leq ... \leq a_n$.

```
1: if p < r then
2:    q ← ⌊(p + r)/2⌋
3:    TRI-FUSION(A, p, q)
4:    TRI-FUSION(A, q + 1, r)
5:    FUSIONNER(A, p, q, r)
6: end if
```

Lorsqu'un algorithme contient un appel récursif à lui même, son temps d'exécution peut souvent être décrit par une équation de récurrence ou récurrence, qui décrit le temps d'exécution global pour un problème de taille n en fonction du temps d'exécution pour des entrées de taille moindre. On peut se servir d'outils mathématiques pour résoudre la récurrence et trouver des bornes pour les performances de l'algorithme.

Ici, on appelle $T(n)$ le temps d'exécution d'un problème de taille n. Si la taille du problème est assez réduite, disons $n \leq c$ pour une certaine constante c, la solution directe consomme un temps constant, qu'on écrit $\Theta(1)$. Supposons qu'on divise le problème en a sous-problèmes, la taille de chacun étant $1/b$ de la taille du problème initial. Si l'on prend un temps $D(n)$ pour diviser le problème en sous-problèmes et un temps $C(n)$ pour construire la solution finale à partir des solutions aux sous-problèmes, on obtient la récurrence

$$T(n) = \begin{cases} \Theta(1) & \text{si } n \leq c, \\ aT(n/b) + D(n) + C(n) & \text{sinon.} \end{cases}$$

Notre analyse fondée sur la récurrence est simplifiée si nous supposons que la taille du problème initial est une puissance de deux. Chaque étape "diviser" génère alors deux sous-suites de taille $n/2$ exactement. Nous verrons plus loin que cette hypothèse n'affecte pas l'ordre de grandeur de la solution de la récurrence.

Le tri par fusion sur un seul élément prend un temps constant. Avec $n > 1$ éléments, on segmente le temps d'exécution comme suit.

Diviser Elle se contente de calculer le milieu du sous-tableau, ce qui consomme un temps constant, $D(n) = \Theta(1)$.

Régner On résoud récursivement deux sous problèmes, d'où $2T(n/2)$.

Combiner On peut démontrer que la procédure FUSIONNER sur un sous-problème à n éléments est de la forme $C(n) = an + b$, où a et b sont des constantes, on notera cette forme comme suit $C(n) = \Theta(n)$.

On a donc

$$T(n) = \begin{cases} \Theta(1) & \text{si } n \leq c, \\ 2T(n/2) + \Theta(n) & \text{si } n > 1. \end{cases}$$

On démontrera plus loin que $T(n)$ est de la forme $anlog(n) + bn + c$, où a, b, et c sont des constantes. On notera cette forme $\Theta(n\ log(n))$. On peut aussi remarquer que ce coût est nettement inférieur à celui du tri par insertion, car $\exists n_0, \forall n, n > n0, n^2 > nlog(n)$.

2.3.2 Exponentiation rapide

Il s'agit de calculer x^n pour x et n données, en calculant la complexité par rapport à n. La méthode naïve (multiplier n fois 1 par x) donne une complexité linéaire. En utilisant le fait que

$$\begin{cases} x^0 = 1 \\ x^n = (x^2)^{\mathbf{n/2}} \\ x^n = x(x^2)^{\mathbf{(n-1)/2}} \end{cases}$$

qui se traduit d'un point de vue algorithmique en :

— Si n est pair, x^n revient à calculer la puissance de x^2 à $n/2$.

— Si n est impair, x^n revient à multiplier x par la puissance de x^2 à $(n-1)/2$.

D'où l'algorithme :

Algorithm 2 PUISSDYN(x, n)

```
1: if (n = 0) then
2:    return 1
3: else if n est pair then
4:    return PuissDyn(x × x, n/2)
5: else
6:    return x×PuissDyn(x × x, (n − 1)/2)
7: end if
```

La complexité de l'exponentiation rapide est donc en $O(log\ n)$.

Chapitre 3

Optimisation linéaire et résolution exacte par séparation/évaluation

Les sections 3.1, 3.3, 3.2 sont extraits intégralement du document [Bastin, 2010] téléchargeable ici : https://www.iro.umontreal.ca/~bastin/Cours/IFT1575/IFT1575.pdf

3.1 Présentation générale [Bastin, 2010]

Certaines quantités ne peuvent s'écrire sous forme de nombres réels, issus d'un domaine continu. Au contraire, certaines décisions sont par nature discrètes, et doivent se représenter à l'aide de nombres entiers. Considérons par exemple une **entreprise de transport**, qui décide de renouvellement sa flotte de camions. Le **nombre de camions à acheter est un nombre naturel**.

La présence de telles variables entières modifie profondément la nature des programmes sous-jacents. Lorsqu'un problème est linéaire avec des variables entières, nous parlerons de *programmation mixte entière*. Si toutes les variables sont entières, nous utiliserons la terminologie de de *programmation (pure) en nombres entiers*. Si les variables entières sont à valeurs 0 ou 1 (binaires), nous parlerons de **programmation 0–1 (binaire)**.

Nous pourrions bien entendu considérer le cas **non-linéaire**, mais les complications sont telles qu'à ce jour, **aucune méthode pleinement satisfaisante n'existe**, même si d'intenses recherches sont actuellement conduites à ce niveau, notamment au sein de l'entreprise IBM.

Exemple 1 (Problème du sac à dos)**.** *Considérons un transporteur muni d'un sac (unique) pour transporter son butin. Son problème consiste à* **maximiser la valeur totale des objets** *qu'il emporte, sans toutefois dépasser une limite de* **poids** b *correspondant à ses capacités physiques. Supposons qu'il y a* n **type d'objets** *que le transporteurpourrait emporter, et que ceux-ci sont en nombre tel que quelle que soit la nature de l'objet considéré, la seule limite au nombre d'unités que le transporteur peut prendre est que le* **poids total reste inférieur à** b*. Si l'on associe* **à l'objet** j **une valeur** c_j **et un poids** w_j*, la* **combinaison optimale** *d'objets à emporter*

sera obtenue en résolvant le programme mathématique

$$\begin{aligned} \max_x \ & \sum_{j=1}^{n} c_j x_j \\ s.c. \ & \sum_{j=1}^{n} w_j x_j \leq b \\ & x_j \in \mathcal{N}, \ j = 1, \ldots, n. \end{aligned}$$

Ici, $x_j \in \mathcal{N}$ **signifie que** x_j **est un naturel**, *autrement dit un entier non négatif.* **Intuitivement, nous pourrions penser que la solution consiste à choisir en premier lieu les objets dont le rapport qualité-poids est le plus avantageux**, *quitte à* **tronquer** *pour obtenir une solution entière (nous ne pouvons pas diviser un objet). Cette solution peut malheureusement se révéler* **sous-optimale, voire mauvaise**, *comme on le constate sur l'exemple suivant.*

$$\begin{aligned} \max_x \ & 2x_1 + 3x_2 + 7x_3 \\ s.c. \ & 3x_1 + 4x_2 + 8x_3 \leq 14, \\ & x_1, \ x_2, \ x_3 \in \mathcal{N}. \end{aligned}$$

Si nous oublions la contrainte d'intégralité, la solution, de valeur 12.25, *est* $x = (0, 0, 14/8)$. *En* **tronquant**, *on obtient la solution entière* $x = (0, 0, 1)$ **de valeur 7**. *Il est facile de trouver de meilleures solutions, par exemple* $x = (1, 0, 1)$.

La solution optimale du problème du transporteur peut s'obtenir en **énumérant toutes les solutions admissibles** *et en conservant la meilleure (voir Table 3.1, , où les solutions inefficaces laissant la possibilité d'ajouter un objet n'ont pas été considérées).*

x_1	x_2	x_3	objectif
0	1	1	10
2	0	1	11
0	3	0	9
2	2	0	10
3	1	0	9
4	0	0	8

TABLE 3.1 – Problème du sac à dos : énumération des solutions

La solution optimale entière, $x = (2, 0, 1)$, **diffère** *passablement de la solution optimale linéaire. Cependant, il est clair que cette technique d'énumération* **ne peut s'appliquer à des problèmes de grande taille**.

Exemple 2. *Une entreprise doit* **choisir de nouveaux emplacements pour construire des usines et des entrepôts**. *Elle a le choix entre deux emplacements : Oran (LA) et Constantine (SF). Nous ne pouvons construire un entrepôt que dans une* **ville où nous disposons d'une usine**, *et nous ne pouvons* **pas**

	Valeur estimée (millions $)	Coût de construction (millions $)
Usine à LA	9	6
Usine à SF	5	3
Entrepôt à LA	6	5
Entrepôt à SF	4	2
Limite maximum	-	10

TABLE 3.2 – Données du problème

construire plus d'un entrepôt. *Nous associons à chaque construction (d'une usine ou d'un entrepôt dans chacun des lieux envisagés)* **sa valeur estimée et son coût** *de construction. L'objectif est de* **maximiser la valeur totale estimée**, *en ne dépassant pas une limite maximum sur les coûts. Les données du problème sont résumées dans la Table 3.2. Les* **variables** *sont*

$$x_j = \begin{cases} 1 & \textit{si la décision } j \textit{ est approuvée ;} \\ 0 & \textit{si la décision } j \textit{ n'est pas approuvée.} \end{cases}$$

L'objectif est de **maximiser la valeur estimée totale** *:*

$$\max z = 9x_1 + 5x_2 + 6x_3 + 4x_4.$$

Les contraintes fonctionnelles sont

1. *la* **limite maximum sur les coûts** *de construction :*

$$6x_1 + 3x_2 + 5x_3 + 2x_4 \leq 10;$$

2. *nous ne pouvons construire* **plus d'un entrepôt** *:*

$$x_3 + x_4 \leq 1;$$

3. *l'entrepôt ne sera* **à LA que si l'usine est à LA** *:*

$$x_3 \leq x_1;$$

4. *l'entrepôt* **ne sera à SF que si l'usine est à SF** *:*

$$x_4 \leq x_2 :$$

5. *contraintes 0–1 (intégralité) :*

$$x_j \in \{0, 1\}, \ j = 1, 2, 3, 4;$$

ou encore

$$0 \leq x_j \leq 1 \ \textit{et} \ x_j \ \textit{entier}, \ j = 1, 2, 3, 4.$$

Par conséquent, nous avons le **programme mathématique**

$$\begin{aligned} max\ z &= 9x_1 + 5x_2 + 6x_3 + 4x_4 \\ s.c.\ & 6x_1 + 3x_2 + 5x_3 + 2x_4 \leq 10; \\ & x_3 + x_4 \leq 1; \\ & -x_1 + x_3 \leq 0; \\ & -x_2 + x_4 \leq 0; \\ & x_1, x_2, x_3, x_4 \leq 1; \\ & x_1, x_2, x_3, x_4 \geq 0; \\ & x_1, x_2, x_3, x_4 \ entiers. \end{aligned}$$

Le modèle illustre deux cas classiques d'illustrations de **variables binaires** *:*

- *alternatives* **mutuellement exclusives** *: nous ne pouvons construire plus d'un entrepôt, i.e.*

$$x_3 + x_4 \leq 1;$$

- **décisions contingentes** *: nous ne pouvons construire un entrepôt que là où nous avons construit une usine :*

$$x_3 \leq x_1; x_4 \leq x_2 :$$

Exemple 3 (Localisation)**.** *Une entreprise envisage plusieurs sites de construction pour des usines qui serviront à approvisionner ses clients. A chaque site potentiel i correspond* **un coût de construction** a_i*, une* **capacité de production** u_i*, un* **coût de production unitaire** b_i *et des* **coûts de transport** c_{ij} *des usines vers les clients. Soit* y_i **une variable binaire** *prenant la valeur 1 si un entrepôt est construit sur le site i,* d_j **la demande de l'usine** j *et* x_{ij} **la quantité produite à l'usine** i *et destinée au marché j (flot de i à j). Un plan de construction optimal est obtenu en résolvant le programme*

$$\begin{aligned} \min_x\ & \sum_i a_i y_i + \sum_i b_i \sum_j x_{ij} + \sum_i \sum_j c_{ij} x_{ij} \\ s.c.\ & \sum_i x_{ij} = d_j, \\ & \sum_j x_{ij} \leq u_i y_i, \\ & x_{ij} \geq 0, \\ & y_i \in \{0, 1\}. \end{aligned}$$

Cette formulation contient deux éléments intéressants : un **coût fixe** *(construction) modélisé par une variable binaire y_i ainsi qu'une* **contrainte logique** *forçant les flots provenant d'un site à être nuls si aucune usine n'est construite en ce site. Notons aussi que certaines variables sont entières alors que* **d'autres (flots) sont réelles**.

Exemple 4 (**Contraintes logiques**)**. Des variables binaires peuvent servir à représenter des contraintes logiques**. *En voici quelques exemples, où* p_i **représente une proposition logique** *et x_i la variable logique (binaire)* **correspondante**.

contrainte logique	**forme algébrique**
$p_1 \oplus p_2 = vrai$	$x1 + x2 = 1$
$p_1 \vee p_2 \vee \cdots \vee p_n = vrai$	$x_1 + x_2 + \ldots + x_n \geq 1$
$p_1 \wedge p_2 \wedge \ldots \wedge p_n = vrai$	$x_1 + x_2 + \ldots + x_n \geq n$ $(ou = n)$
$p_1 \Rightarrow p_2$	$x_2 \geq x_1$
$p_1 \Leftrightarrow p_2$	$x_2 = x_1$

Exemple 5 (**Fonctions linéaires par morceaux**). *Considérons une fonction objectif à maximiser, pour laquelle dans chaque* **intervalle** $[a_{i-1}, a_i]$ **la fonction est linéaire**, *ce que nous pouvons exprimer par :*

$$\begin{aligned} &x = \lambda_{i-1}a_{i-1} + \lambda_i a_i \\ &\lambda_{i-1} + \lambda_i = 1, \\ &\lambda_{i-1},\ \lambda_i \geq 0, \\ &f(x) = \lambda_{i-1}f(a_{i-1}) + \lambda_i f(a_i). \end{aligned}$$

Car, il est bien établi que les points d'un segment d'extrémités A et B :

$$[A, B] = \{(1-t)A + tB | t \in [0, 1]\}.$$

Nous pouvons généraliser cette formule sur tout l'intervalle de définition de la fonction f **en contraignant les variables λ_i à ne prendre que deux valeurs non nulles, et ce pour deux indices consécutifs**. *Ceci se fait en introduisant des variables binaires y_i associées aux* **intervalles de linéarité** $[a_{i-1}, a_i]$ *:*

$$\begin{aligned} x &= \sum_{i=0}^{n} \lambda_i a_i, \\ f(x) &= \sum_{i=0}^{n} \lambda_i f(a_i), \\ \sum_{i=0}^{n} \lambda_i &= 1, \\ \lambda_i &\geq 0,\ i = 0, \ldots, n \\ \lambda_0 &\leq y_1, \\ \lambda_1 &\leq y1 + y2, \\ &\vdots \quad \vdots \quad \vdots \\ \lambda_{n-1} &\leq y_{n-1} + y_n, \\ \lambda_n &\leq y_n, \\ \sum_{i=1}^{n} y_i &= 1 (\textit{un seul intervalle "actif"}) \\ y_i &\in 0, 1,\ i = 1, \ldots, n. \end{aligned}$$

3.2 Contraintes mutuellement exclusives [Bastin, 2010]

3.2.1 Deux contraintes

Prenons l'exemple de deux contraintes. **L'une ou l'autre des deux contraintes doit être satisfaite, mais pas les deux simultanément.** Par exemple,

— soit $3x_1 + 2x_2 \leq 18$;
— soit $x_1 + 4x_2 \leq 16$.

Soit M **un très grand nombre** ; le système précédent **est équivalent à**

— soit

$$\begin{aligned} 3x_1 + 2x_2 &\leq 18, \\ x_1 + 4x_2 &\leq 16 + M; \end{aligned}$$

— soit

$$\begin{aligned} 3x_1 + 2x_2 &\leq 18 + M, \\ x_1 + 4x_2 &\leq 16. \end{aligned}$$

En introduisant une variable binaire y, nous obtenons le **système équivalent**

$$\begin{aligned} 3x_1 + 2x_2 &\leq 18 + M(1 - y), \\ x_1 + 4x_2 &\leq 16 + My. \end{aligned}$$

La signification de cette variable est

— $y = 1$, si la première contrainte est satisfaite ;
— $y = 0$, si la seconde contrainte est satisfaite.

Nous avons de la sorte construit deux alternatives mutuellement exclusives.

Nous aurions pu aussi introduire deux variables binaires :

— $y_1 = 1$, si la première contrainte est satisfaite ;
— $y_2 = 1$, si la seconde contrainte est satisfaite.

Nous devons avoir

$$y_1 + y_2 = 1.$$

Afin de se ramener au modèle précédent, il suffit de poser

$$\begin{aligned} y_1 &= y; \\ y_2 &= 1 - y. \end{aligned}$$

Il s'agit d'un cas particulier de la situation suivante : K parmi N contraintes doivent être satisfaites. Dans ce cas plus général, nous introduisons N variables binaires.

3.2.2 K contraintes parmi N

Soit les N **contraintes**

$$f_j(x_1, x_2, \ldots, x_n) \leq d_j, \ j = 1, 2, \ldots, N.$$

Nous introduisons N variables binaires, avec $y_j = 1$ **si la** j^e **contrainte est satisfaite** :

$$f_j(x_1, x_2, \ldots, x_n) \leq d_j + M(1 - y_j),\ j = 1, 2, \ldots, N.$$

Il reste à spécifier que seulement K **de ces contraintes peuvent êtres satisfaites** :

$$\sum_{j=1}^{N} y_j = K.$$

3.2.3 Fonction ayant N valeurs possibles

Soit la **contrainte**

$$f(x_1, x_2, \ldots, x_n) = d_1,\ \text{ou } d_2 \text{ ou } \ldots \text{ ou } d_N.$$

Nous introduisons N **variables binaires, avec** $y_j = 1$ **si la fonction vaut** d_j**.** La contrainte s'écrit alors

$$f(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{N} d_j y_j,$$

avec

$$\sum_{j=1}^{N} y_j = 1.$$

Exemple 6 (Wyndor Glass)**.** *Supposons que le temps de production maximum à l'usine 3 n'est pas toujours 18h, mais pourrait également être 6h ou 12h. Cette contrainte s'écrit alors*

$$3x_1 + 2x_2 = 6 \text{ ou } 12 \text{ ou } 18.$$

Nous introduisons alors trois variables binaires

$$3x_1 + 2x_2 = 6y_1 + 12y_2 + 18y_3,$$
$$y_1 + y_2 + y_3 = 1.$$

3.2.4 Objectif avec coûts fixes

Supposons que le coût associé à un produit j est composé de deux parties :

1. un coût fixe initial k_j encouru dès qu'une unité de j est produite ;
2. un coût c_j proportionnel au nombre d'unités de j produites.

Le coût total associé à la production de x_j **unités est**

$$f(x_j) = \begin{cases} k_j + c_j x_j & \text{si } x_j > 0, \\ 0 & \text{si } x_j = 0. \end{cases}$$

Supposons de plus que l'objectif consiste à minimiser la **somme de** n **fonctions avec coûts fixes**

$$\min z = \sum_{j=1}^{n} f_j(x_j).$$

Nous introduisons alors n variables binaires :

$$y_j = \begin{cases} 1 \text{ si } x_j > 0, \\ 0 \text{ si } x_j = 0. \end{cases}$$

L'objectif s'écrit alors

$$\min z = \sum_{j=1}^{n} c_j x_j + k_j y_j.$$

Les valeurs de x_j et de y_j dépendent l'une de l'autre : il s'agit d'un exemple de *décisions contingentes*. Nous devons en particulier avoir une contrainte qui précise que $x_j = 0$ **si** $y_j = 0$. Toutefois, les deux variables ne sont plus binaires, vu que x_j peut être quelconque. Soit M_j **une borne supérieure sur la valeur de** x_j. Nous pouvons écrire la **relation entre les deux variables** de cette manière :

$$x_j \leq M_j y_j.$$

Ainsi,

- si $y_j = 0$, alors $x_j = 0$;
- si $y_j = 1$, alors $x_j \leq M_j$;
- si $x_j > 0$, alors $y_j = 1$;
- si $x_j = 0$, alors toute soluton optimale satisfait $y_j = 0$ lorsque $k_j > 0$ (si $k_j = 0$, la variable y_j est inutile).

Nous obtenons par conséquent le programme

$$\begin{aligned} \min\ z &= \sum_{j=1}^{n} c_j x_j + k_j y_j \\ \text{s.c. } x_j &\leq M_j y_j, \\ y_j &\in \{0,1\},\ j = 1, 2, \ldots, n. \end{aligned}$$

3.2.5 Variables entières en variables 0–1

Soit x une **variable entière générale bornée** :

$$0 \leq x \leq u,$$

et soit N l'entier **tel que** $2^N \leq u \leq 2^{N+1}$. La **représentation binaire de** x est

$$x = \sum_{j=0}^{N} 2^j y_j.$$

L'intérêt de cette transformation est que les méthodes de programmation 0–1 sont souvent plus efficaces que les méthodes de programmation en nombres entiers. Elle engendre néanmoins une **multiplication du nombre de variables**.

Exemple 7. *Nous considérons* **trois types de produits, et deux usines** *; nous exprimons le* **profit par unité de produit** *en milliers de dollars. Nous connaissons les* **ventes potentielles par produit** *(unités/semaine), et la* **capacité de**

	Produit 1 temps de production (h/unité)	Produit 2 temps de production (h/unité)	Produit 3 temps de production (h/unité)	Capacité de production production (h/semaine)
Usine 1	3	4	2	30
Usine 2	4	6	2	40
Profit/unité (1000$)	5	7	3	–
Ventes potentielles (par semaine)	7	5	9	–

TABLE 3.3 – Exemple de production avec variables entières.

production par usine *(h/semaine). Nous avons toutefois comme contrainte que* **pas plus de deux produits ne peuvent être fabriqués**, *et* **une seule des deux usines doit être exploitée**. *Les données du problème sont résumées dans la Table 3.3. Les variables sont* x_j, **le nombre d'unités fabriquées du produit** j. *Pour représenter la* **contrainte "pas plus de deux produits"**, *nous devons introduire des variables 0–1 :*

$$y_j = \begin{cases} 1 & si\ x_j > 0; \\ 0 & si\ x_j = 0. \end{cases}$$

Afin de représenter la contrainte **"une seule des deux usines"**, *nous devons ajouter une variables 0–1 :*

$$y_4 = \begin{cases} 1 & si\ l'usine\ 1\ est\ choisie; \\ 0 & si\ sinon. \end{cases}$$

*L'***objectif** *est*

$$\max z = 5x_1 + 7x_2 + 3x_3.$$

Les **ventes potentielles** *sont*

$$x_1 \leq 7,\ x_2 \leq 5,\ x_3 \leq 9.$$

L'exigence **interdisant d'avoir plus de deux produits** *se traduit mathématiquement par*

$$y_1 + y_2 + y_3 \leq 2.$$

La **relation entre les variables continues et les variables 0–1** *s'exprime par*

$$x_1 \leq 7y_1,\ x_2 \leq 5y_2,\ x_3 \leq 9y_3.$$

La contrainte portant sur l'utilisation d'une seule usine est

- *soit* $3x_1 + 4x_2 + 2x_3 \leq 30$;
- *soit* $4x_1 + 6x_2 + 2x_3 \leq 40$.

En utilisant la variable 0–1 (et M très grand), elle se traduit par

$$3x_1 + 4x_2 + 2x_3 \leq 30 + M(1 - y_4),$$
$$4x_1 + 6x_2 + 2x_3 \leq 40 + My_4.$$

En résumé, nous avons le modèle

$$\begin{aligned} \max z &= 5x_1 + 7x_2 + 3x_3 \\ s.c.\ & x_1 \leq 7y_1,\ x_2 \leq 5y_2,\ x_3 \leq 9y_3 \\ & y_1 + y_2 + y_3 \leq 2, \\ & 3x_1 + 4x_2 + 2x_3 \leq 30 + M(1 - y_4), \\ & 4x_1 + 6x_2 + 2x_3 \leq 40 + My_4, \\ & x_1, x_2, x_3 \geq 0, \\ & y_j \in \{0, 1\}, j = 1, 2, 3, 4. \end{aligned}$$

Exemple 8. **Nous considérons à nouveau trois types de produits, pour lesquels nous pouvons placer cinq annonces publicitaires, avec un maximum de trois annonces par produit.** *L'estimation des revenus publicitaires est donnée dans la Table 3.4, où les profits sont exprimés en millions de dollars. Les*

Nombre d'annonces	Produit 1	Produit 2	Produit 3
0	0	0	0
1	1	0	-1
2	3	2	2
3	3	3	4

TABLE 3.4 – Revenus publicitaires.

variables du problèmes sont le nombre d'annonces pour le produit i, dénoté par x_i, mais l'hypothèse de proportionnalité est alors violée : nous ne pouvons représenter l'objectif sous forme linéaire uniquement avec ces variables.

Prenons tout d'abord comme variables

$$y_{ij} = \begin{cases} 1 & si\ x_i = j; \\ 0 & sinon. \end{cases}$$

L'objectif est

$$\max z = y_{11} + 3y_{12} + 3y_{13} + 2y_{22} + 3y_{23} - y_{31} + 2y_{32} + 4y_{33}.$$

Nous utiliserons les **5 annonces disponibles** *:*

$$\sum_{i=1}^{3} \sum_{j=1}^{3} jy_{ij} = 5.$$

Enfin, on a droit à **un seul type d'annonce par produit, la définition des variables 0–1** *donne*

$$\sum_{j=1}^{3} y_{ij} \leq 1, \ i = 1, 2, 3.$$

Soit une autre modélisation *en prenant comme variables*

$$y_{ij} = \begin{cases} 1 & si\ x_i \geq j; \\ 0 & sinon. \end{cases}$$

Autrement dit, nous avons remplacé l'égalité dans la première condition par une inégalité. Cette définition implique

$$\begin{aligned} x_i = 0 &\Rightarrow y_{i1} = 0, y_{i2} = 0, y_{i3} = 0; \\ x_i = 1 &\Rightarrow y_{i1} = 1, y_{i2} = 0, y_{i3} = 0; \\ x_i = 2 &\Rightarrow y_{i1} = 1, y_{i2} = 1, y_{i3} = 0; \\ x_i = 3 &\Rightarrow y_{i1} = 1, y_{i2} = 1, y_{i3} = 1. \end{aligned}$$

Ce qui peut encore s'énoncer somme

$$y_{i(j+1)} \leq y_{ij}, \ i = 1, 2, 3, \ j = 1, 2.$$

Supposons que $x_1 = 3$ (il y a trois annonces pour le produit 1). *Le profit associé à cette valeur doit être 3. Mais* $x_1 = 3$ **veut aussi dire que chaque variable binaire associée au produit 1 vaut 1** *; comment dès lors comptabiliser correctement la contribution de ces trois variables au profit ?* **La solution consiste à prendre comme profit associé à la variable** y_{ij} **la différence** $c_{ij+1} - c_{ij}$, *où* c_{ij} *est le revenu net si nous plaçons* j *annonce pour le produit* i. *Dans notre exemple, le profit associé à*

- y_{11} *est 1-0 = 1 ;*
- y_{12} *est 3-1 = 2 ;*
- y_{13} *est 3-3 = 0 ;*

Nous obtenons ainsi le programme mathématique suivant :

$$\begin{aligned} \max\ & z = y_{11} + 2y_{12} + 2y_{22} + y_{23} - y_{31} + 3y_{32} + 2y_{33} \\ s.c.\ & y_{i(j+1)} \leq y_{ij}, i = 1, 2, 3, \ j = 1, 2, \\ & \sum_{i=1}^{3} \sum_{j=1}^{3} y_{ij} = 5, \\ & y_{ij} \in \{0, 1\}, \ i = 1, 2, 3, \ j = 1, 2. \end{aligned}$$

3.2.6 Problème de recouvrement

Exemple 9 (**Affectation des équipages**). *Un problème important des compagnies aériennes consiste à* **constituer de façon efficace des équipages pour ses vols**. *Pour un équipage donné, une* **rotation consiste en une succession de services de vol débutant et se terminant en une même ville**. *Comme il y a un* **coût**

associé à chaque séquence de vols, *la compagnie cherche à* **minimiser les coûts d'affectation des équipages aux séquences** *tout en assurant le service sur chacun des vols.*

Considérons par exemple un problème avec 11 vols et 12 séquences de vols, dont les données sont décrites dans la Table 3.5. Les variables sont

Vol | Séquence	1	2	3	4	5	6	7	8	9	10	11	12
1	1			1			1			1		
2		1			1			1			1	
3			1			1			1			1
4				1			1		1	1		1
5	1					1				1	1	
6				1	1				1			
7							1	1		1	1	1
8		1		1	1				1			
9					1			1			1	
10			1				1	1				1
11						1			1	1	1	1
Coût	2	3	4	6	7	5	7	8	9	9	8	9

TABLE 3.5 – Affectation d'équipages.

$$x_j = \begin{cases} 1 & \textit{si la séquence de vols } j \textit{ est affectée;} \\ 0 & \textit{sinon.} \end{cases}$$

L'objectif est

$$\min z = 2x_1 + 3x_2 + 4x_3 + 6x_4 + 7x_5 + 5x_6 + 7x_7 + 8x_8 + 9x_9 + 9x_{10} + 8x_{11} + 9x_{12}.$$

Nous devons affecter **trois équipages**

$$\sum_{j=1}^{12} x_j = 3.$$

Le service doit être **assuré sur chacun des vols** *:*

$$\begin{aligned} x_1 + x_4 + x_7 + x_{10} &\geq 1 \\ x_2 + x_5 + x_8 + x_{11} &\geq 1 \\ x_3 + x_6 + x_9 + x_{12} &\geq 1 \\ x_4 + x_7 + x_9 + x_{10} + x_{12} &\geq 1 \\ &\dots \end{aligned}$$

Généralement, un **problème de recouvrement d'ensemble** met en oeuvre
— I : un ensemble d'objets (les vols dans l'exemple précédent) ;
— $\mathcal{J}$: une collection de sous-ensembles de I (e.g. les séquences de vols) ;

— J_i, $i \in I$: les sous-ensembles dans $\mathcal{J}$ qui contiennent i.

Nous avons les variables binaires x_j, **prenant la valeur 1 si le sous-ensemble j est choisi, et 0 sinon**. En considérant un objectif linéaire, avec c_j le coût associé au sous-ensemble j. Nous obtenons le programme

$$\begin{aligned} \min_x & \sum_{j\in\mathcal{J}} c_j x_j \\ \text{s.c.} & \sum_{j\in J_i} x_j \geq 1, \ i \in I; \\ & x_j \in \{0,1\}, \ j \in \mathcal{J}. \end{aligned}$$

3.3 Stratégies de résolutions[Bastin, 2010]

3.3.1 Relaxation linéaire

Il est tentant "d'oublier" les contraintes d'intégralité, et de résoudre le problème en nombres continus ainsi produit. Nous parlerons alors de *relaxation*. Ainsi, nous pourrons construire la relaxtion en programme linéraire d'un programme mixte entier. **Une fois le programme relâché résolu, nous pourrions arrondir aux valeurs entières les plus proches.** Dans certains cas, cela peut fonctionner, mais l'exemple du sac à dos montre que **ce n'est pas toujours vérifié**. Cette méthode par arrondissement est même **parfois désastreuse**.

Exemple 10. *Considérons le programme*

$$\begin{aligned} \max\ & z = x_2 \\ s.c.\ & -x_1 + x_2 \leq \frac{1}{2}, \\ & x_1 + x_2 \leq \frac{7}{2}, \\ & x_1, x_2 \geq 0 \ \textit{et entiers}. \end{aligned}$$

La relaxation en programme linéaire donne

$$\begin{aligned} \max\ & z = x_2 \\ s.c.\ & -x_1 + x_2 \leq \frac{1}{2}, \\ & x_1 + x_2 \leq \frac{7}{2}, \\ & x_1, x_2 \geq 0. \end{aligned}$$

Ce nouveau programme a pour **solution** $\left(\frac{3}{2}, 2\right)$. *Que nous arrondissions cette solution à* $(1,2)$ *ou* $(2,2)$, **nous n'obtenons pas de solution réalisable**, *comme illustré sur la Figure 3.1*

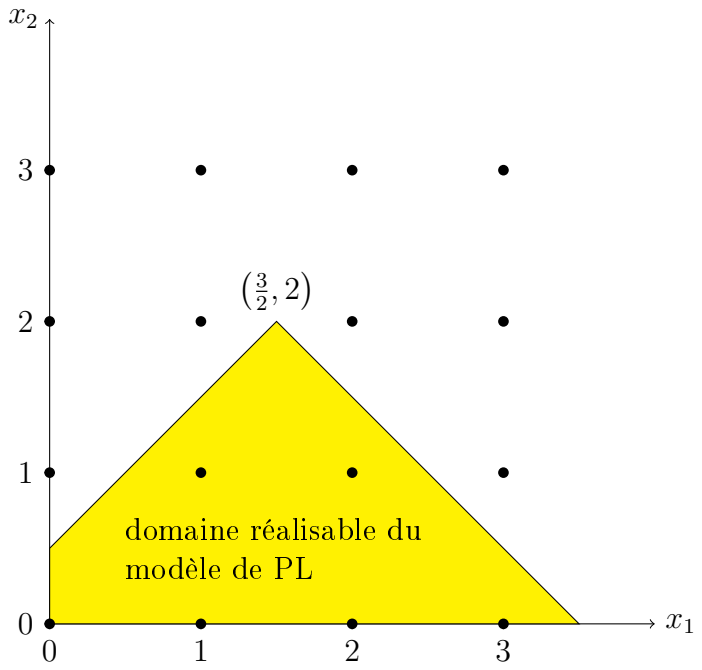


FIGURE 3.1 – Exemple de relaxation linéaire : problème d'admissibilité des solutions

Exemple 11. *Considérons le programme*

$$\begin{aligned} \max\ & z = x_1 + 5x_2 \\ s.c.\ & x_1 + 10x_2 \leq 20, \\ & x_2 \leq 2, \\ & x_1, x_2 \geq 0 \ \textit{et entiers.} \end{aligned}$$

La version relâchée de programme est

$$\begin{aligned} \max\ & z = x_1 + 5x_2 \\ s.c.\ & x_1 + 10x_2 \leq 20, \\ & x_2 \leq 2, \\ & x_1, x_2 \geq 0, \end{aligned}$$

qui a pour solution optimale $(2, 1.8)$. **En arrondissant à** $(2, 1)$ **afin de garantir l'admissibilité, nous obtenir la valeur 7 pour la fonction objectif, loin de la valeur optimale du programme mixte entier, avec pour valeur optimale 11**, *en* $(0, 2)$ *(voir Figure 3.2).*

La solution de la relaxation linéaire ne peut donc être exploité directement pour obtenir la solution exacte du problème en nombres entiers.

Cependant, la relaxation linéaire a les propriétés suivantes :

Borne supérieure La valeur de la solution optimale de la relaxation est une borne supérieure sur la valeur de la solution optimale du problème de maximisation en nombres entiers. C'est une borne inférieur sur un problème de minimisation.

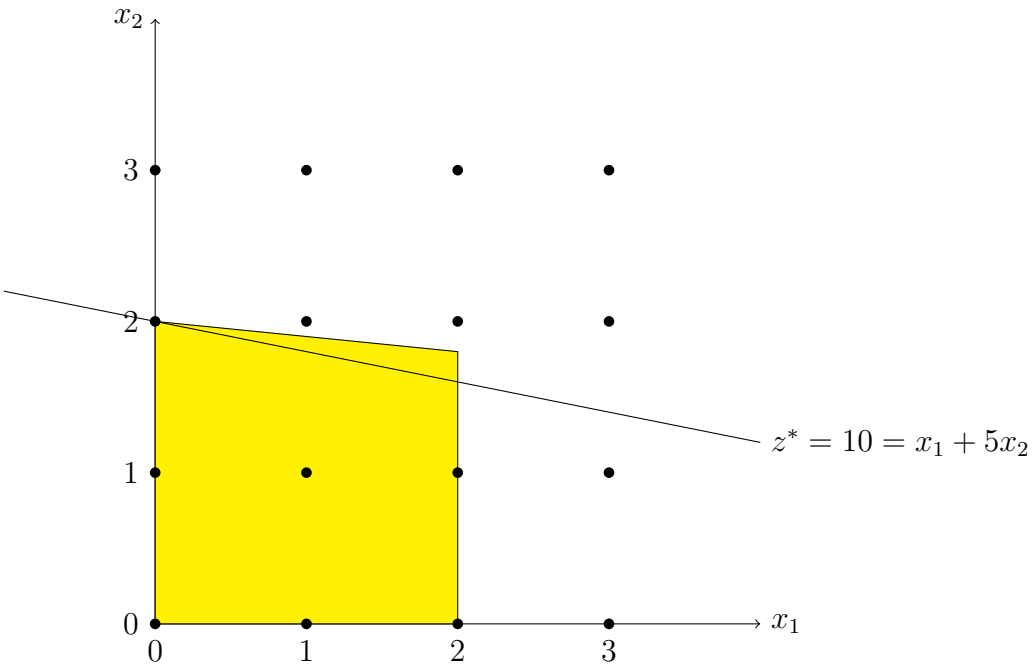


FIGURE 3.2 – Exemple de relaxation linéaire : solution médiocre

Solution entière Si la solution optimale de la relaxation est entière (donc admissible pour le problème en nombres entiers), elle est également la solution optimale du problème en nombres entiers.

Borne inférieure La valeur d'une solution admissible/faisable du problème en nombres entiers de maximisation fournit une borne inférieure sur la valeur de la solution optimale de ce problème en nombres entiers. C'est une borne supérieur sur un problème de minimisation.

D'une façon globale toute méthode de résolution qui garantit ces propriétés peut servir comme relaxation.

3.3.2 Approche par énumération

Un modèle en nombres entiers borné (par exemple, un modèle avec uniquement des variables 0–1) possède un nombre fini de solutions. Nous pourrions **envisager de toutes les énumérer**, mais le nombre de solutions explose rapidement. Pour $n = 20$ variables 0–1, il y a **plus d'un million de solutions possibles**. **Pour $n = 30$, c'est plus d'un milliard**. Comme il apparaît qu'il est vite déraisonnable de procéder à une énumération complète des solutions, nous allons essayer de mettre à profit la relaxation en programme linéaire pour éliminer certaines de ces solutions. Cette technique d'énumération partielle est connue sous le vocable de *branch-and-bound* **(B&B)**. Il s'agit d'une **approche** *diviser-pour-régner* :

— décomposition du problème en sous-problèmes plus simples ;
— combinaison de la résolution de ces sous-problèmes pour obtenir la solution du problème original.

Dans l'algorithme de branch-and-bound, chaque sous-problème correspond à un sommet dans l'arbre des solutions. Nous résolvons la relaxation linéaire de chaque sous-problème. L'information tirée de la relaxation linéaire nous permettra (peut-être) d'éliminer toutes les solutions pouvant être obtenues à partir de ce sommet.

3.3.2.1 Algorithme de branch & bound : cas 0–1

Un algorithme simple pour énumérer toutes les solutions d'un modèle 0–1 consiste à :

- choisir un sommet dans l'arbre des solutions ;
- choisir une variable x non encore fixée relativement à ce sommet ;
- générer les deux variables $x = 0$ et $x = 1$ (la variable x est dite *fixée*) : chaque alternative correspond à un sommet de l'arbre des solutions ;
- recommencer à partir d'un sommet pour lequel certaines variables ne sont pas encore fixées.

A la racine de l'arbre, aucune variable n'est encore fixée, tandis qu'aux feuilles de l'arbre, toutes les variables ont été fixées. Le nombre de feuilles est 2^n (pour n variables 0–1).

Le calcul de borne (ou évaluation) consiste à résoudre la relaxation linéaire en chaque sommet. L'élagage (ou élimination) consiste à utiliser l'information tirée de la résolution de la relaxation linéaire pour éliminer toutes les solutions émanant du sommet courant. Dés lors, le branch-and-bound est un algorithme de séparation et d'évaluation successives.

Exemple 12 (California Mfg)**.** *Rappelons le problème*

$$\begin{aligned} max\ & z = 9x_1 + 5x_2 + 6x_3 + 4x_4 \\ s.c.\ & 6x_1 + 3x_2 + 5x_3 + 2x_4 \leq 10; \\ & x_3 + x_4 \leq 1; \\ & -x_1 + x_3 \leq 0; \\ & -x_2 + x_4 \leq 0; \\ & x_1, x_2, x_3, x_4 \leq 1; \\ & x_1, x_2, x_3, x_4 \geq 0; \\ & x_1, x_2, x_3, x_4 \ \textit{entiers.} \end{aligned}$$

La relaxation linéaire permet aux variables de prendre les valeurs fractionnaires entre 0 et 1, ce qui conduit à la solution

$$\left(\frac{5}{6}, 1, 0, 1\right),$$

avec comme valeur $z = 16.5$*. Branchons sur la variable* x_1*.*

Dénotons sous-problème 1 celui obtenu avec $x_1 = 0$ *:*

$$\begin{aligned} max\ z &= 5x_2 + 6x_3 + 4x_4 \\ s.c.\ 3x_2 + 5x_3 + 2x_4 &\leq 10; \\ x_3 + x_4 &\leq 1; \\ x_3 &\leq 0; \\ -x_2 + x_4 &\leq 0; \\ x_2, x_3, x_4 &\ binaires. \end{aligned}$$

La solution de la relaxation linéaire est $(x_1, x_2, x_3, x_4) = (0, 1, 0, 1)$*, avec* $z = 9$*.*

Le sous-problème 2 est celui obtenu avec $x_1 = 1$ *:*

$$\begin{aligned} max\ z &= 5x_2 + 6x_3 + 4x_4 + 9 \\ s.c.\ 3x_2 + 5x_3 + 2x_4 &\leq 4; \\ x_3 + x_4 &\leq 1; \\ x_3 &\leq 1; \\ -x_2 + x_4 &\leq 0; \\ x_2, x_3, x_4 &\ binaires. \end{aligned}$$

La solution de la relaxation linéaire est alors

$$(x_1, x_2, x_3, x_4) = \left(1, \frac{4}{5}, 0, \frac{4}{5}\right),$$

avec $z = 16 + \frac{1}{5}$*.*

Nous obtenons dès lors les bornes suivantes :
— sous-problème 1 : $Z_1 \leq 9$;
— sous-problème 2 : $Z_2 \leq 16 + \frac{1}{5}$.
Notons que toutes les variables sont binaires et tous les paramètres dans l'objectif sont des valeurs entières. Dès lors, la borne supérieure pour le sous-problème 2 est 16. Pour le sous-problème 1, la solution obtenue est entiere : c'est la meilleure solution courante. Nous savons que la valeur optimale cherchée, Z, sera au moins

$$Z^* = 9 : Z \geq Z^*.$$

Quels sous-problèmes pouvons-nous à présent considérer afin de nous approcher de la solution optimale ? Tous les sous-problèmes actuellement traités doivent-ils donner naissance à d'autres problèmes. Si un sous-problème ne donne lieu à aucun autre problème, nous parlerons d'élagage, en référence avec l'idée de couper la branche correspondante dans l'arbre d'exploration.

Considérons tout d'abord le sous-probleme 1 : la solution optimale de la relaxation PL est entière. Il ne sert donc à rien de brancher sur les autres variables, puisque toutes les autres solutions entières (avec $x_1 = 0$) sont nécessairement de valeur inférieures ou égales à 9. Nous pouvons donc élaguer ce sommet.

Pour le sous-probleme 2, la solution optimale de la relaxation PL n'est pas entière :

$$Z^* = 9 \leq Z \leq 16.$$

La branche ($x_1 = 1$) peut encore contenir une solution optimale. Mais si nous avions eu $Z_2 \leq Z*$, nous aurions pu conclure que la branche ne pouvait améliorer la meilleure solution courante.

Un sous-problème est elagué si une des trois conditions suivantes est satisfaite :

— test 1 : sa borne supérieure (valeur optimale de la relaxation PL) est inférieure ou égale à Z^* (valeur de la meilleure solution courante) ;

— test 2 : sa relaxation PL n'a pas de solution réalisable ;

— test 3 : la solution optimale de sa relaxation PL est entière.

Lorsque le test 3 est verifié, nous testons si la valeur optimale de la relaxation PL du sous-problème, Z_i, est supérieure à Z^*. Si $Z_i > Z^*$, alors $Z^* := Z_i$, et nous conservons la solution, qui devient la meilleure solution courante. En résumé, nous obtenons l'algorithme ci-dessous.

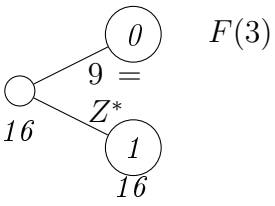
Algorithm 1. *Branch-and-Bound : cas binaire*

1. *Initialisation :*
   (a) *poser $Z^* = -\infty$;*
   (b) *appliquer le calcul de borne et les critères d'élagage à la racine (aucune variable fixée).*
2. *Critere d'arrêt : s'il n'y a plus de sous-problemes non élagués, arrêter.*
3. *Branchement :*
   (a) *parmi les sous-problèmes non encore élagués, choisir celui qui a été crée le plus récemment (s'il y a égalité, choisir celui de plus grande borne supérieure) ;*
   (b) *appliquer le Test 1 : si le sous-problème est élagué, retourner en 2.*
   (c) *brancher sur la prochaine variable non fixée.*
4. *Calcul de borne :*
   (a) *résoudre la relaxation PL de chaque sous-problème ;*
   (b) *arrondir la valeur optimale si tous les paramètres de l'objectif sont entiers.*
5. *Elagage : élaguer un sous-problème si*
   (a) *la borne superieure est inférieure ou égale à Z^* ;*
   (b) *la relaxation PL n'a pas de solution réalisable ;*
   (c) *la solution optimale de la relaxation PL est entière : si la borne supérieure est strictement supérieure à Z^*, Z^* est mise à jour et la solution de la relaxation PL devient la meilleure solution courante.*
6. *Retourner en 2.*

A partir de quel noeud devrions-nous brancher ? Il y a plusieurs choix possibles ; dans cette version, on propose comme règle de sélection de choisir le sous-problème le plus récemment créé. L'avantage est que cette approche facilite la réoptimisation lors du calcul de borne, car il n'y a que peu de changements apportés par rapport au dernier sous-probleme traité. Le désavantage est que cela peut créer un grand nombre de sous-probèmes. Une autre option est la règle de la meilleure borne : choisir le sous-problème ayant la plus grande borne supérieure.

Dans cette version, la règle de branchement consiste à choisir la prochaine variable non fixée. Il est souvent plus intéressant de choisir une variable à valeur fractionnaire. En branchant sur une telle variable, il est certain que les deux sous-problèmes crées mènent à des solutions differentes de la solution courante. De nombreux critères existent pour choisir une telle variable de facon à orienter la recherche vers un élagage rapide.

Exemple 13 (suite)**.** *Jusqu'à maintenant, voici l'arbre obtenu, en branchant sur la variable* x_1*.*



$F(3)$ *indique que le sous-probleme a été élagué (*fathomed*) en raison du Test 3.*

Sélection : nous choisissons le sous-problème 2, le seul qui n'a pas encore été élagué. Nous branchons sur la prochaine variable, soit x_2*. Deux nouveaux sous-problèmes sont crées :*

— *sous-problème 3 :* $x_1 = 1,\ x_2 = 0$ *;*

— *sous-problème 4 :* $x_1 = 1,\ x_2 = 1$*.*

Considérons tout d'abord le sous-problème 3 ($x_1 = 1,\ x_2 = 0$*). Nous obtenons le problème*

$$\begin{aligned} \max z_3 = {} & 6x_3 + 4x_4 + 9 \\ s.c.\ & 5x_3 + 2x_4 \leq 4 \\ & x_3 + x_4 \leq 1 \\ & x_3 \leq 1 \\ & x_4 \leq 0 \\ & x_3,\ x_4 \text{ binaire.} \end{aligned}$$

La solution de la relaxation PL est

$$(x_1, x_2, x_3, x_4) = \left(1, 0, \frac{4}{5}, 0\right).$$

et

$$Z = 13 + \frac{4}{5} :\ Z_3 \leq 13.$$

Le sous-problème 4 ($x_1 = 1,\ x_2 = 1$*) devient*

$$\begin{aligned} \max z_4 = {} & 6x_3 + 4x_4 + 14 \\ s.c.\ & 5x_3 + 2x_4 \leq 1 \\ & x_3 + x_4 \leq 1 \\ & x_3 \leq 1 \\ & x_4 \leq 1 \\ & x_3,\ x_4 \text{ binaire.} \end{aligned}$$

La solution de la relaxation PL est

$$(x_1, x_2, x_3, x_4) = \left(1, 1, 0, \frac{1}{2}\right).$$

et

$$Z = 16 : \ Z_4 \leq 16.$$

Aucun des tests d'élagage ne s'applique sur ces sous-problèmes. Nous devons dès lors choisir un des deux sous-problèmes pour effectuer un branchement, puisque ce sont ceux crées le plus récemment. Nous choisissons celui de plus grande borne supérieure, soit le sous-problème 4. Nous branchons sur x_3 et nous générons deux nouveaux sous-problèmes.

Le sous-problème 5, défini avec $x_1 = 1$, $x_2 = 1$, $x_4 = 0$, s'écrit La solution de la relaxation PL est

$$(x_1, x_2, x_3, x_4) = (1, 1, 0.2, 0)$$

et

$$Z = 14.$$

Le sous-problème 6, défini avec $x_1 = 1$, $x_2 = 1$, $x_4 = 1$, s'écrit La relaxation PL n'a pas de solution realisable : ce sous-probleme est élagué.

Le sous-problème 5 ne peut pas être élagué. Il est crée le plus récemment parmi les sous-problèmes non élagués (3 et 5), aussi choisissons-nous le pour effectuer un branchement. Nous branchons sur x_3 et générons les sous-problèmes suivants :

- *sous-problème 7 : $x_1 = 1$, $x_2 = 1$, $x_3 = 0$, $x_4 = 0$;*
- *sous-problème 8 : $x_1 = 1$, $x_2 = 1$, $x_3 = 1$, $x_4 = 0$.*

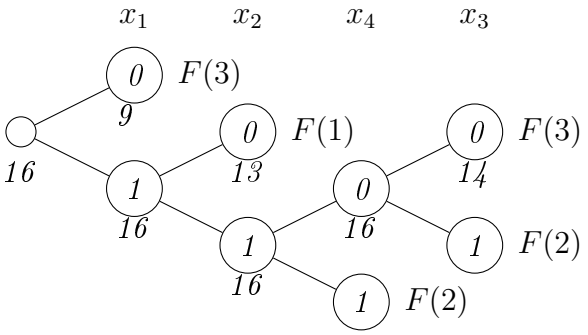
Toutes les variables sont fixées, aussi pouvons-nous directement résoudre ces sous-problèmes. Le sous-problème 7 a pour solution $(x_1, x_2, x_3, x_4) = (1, 1, 0, 0)$, pour $Z_7 = 14$. La solution est entière, aussi nous élaguons le sous-problème en vertu du Test 3. Puisque $Z_7 > Z^*$, $Z^* = 14$ et la solution du sous-problème devient la meilleure solution courante. Le sous-problème 8 a pour solution $(x_1, x_2, x_3, x_4) = (1, 1, 1, 0)$. Cette solution n'est pas réalisable. Le sous-problème est par conséquent élagué par le Test 2.

Le sous-problème 3 est le seul non encore élagué. Appliquons le Test 1 : $Z_3 = 13 \leq 14 = Z*$. Le sous-problème est donc élagué. Comme il n'y a plus de sous-problèmes non élagués, nous pouvons nous arrêter. La solution optimale est :

$$(x_1, x_2, x_3, x_4) = (1, 1, 0, 0),$$

et la valeur optimale est $Z^* = 14$.

L'arbre obtenu suite à l'exécution de l'algorithme se présente comme suit :



$F(j)$: le sous-problème est élagué par le Test j

3.3.2.2 Algorithme de branch & bound : cas général

Considérons à présent le cas général d'un modèle de programmation (mixte) en nombres entiers : variables entières générales et variables continues. Comme précédemment, nous ignorons dans un premier temps les contraintes d'intégralité (les valeurs des variables entières sont traitées comme variables continues), et résolvons le programme linéaire résultant. Si la solution optimale de ce programme satisfait aux contraintes d'intégralité, alors cette solution est aussi solution optimale du programme avec variables entières. Sinon, il doit exister au moins une variable x_j dont la valeur α est fractionnaire. La procédure de branchement se généralise alors comme suit : nous séparons alors le problème relaxé en deux sous-problèmes ; un sous-problème contiendra la contrainte $x_j \leq \lfloor \alpha \rfloor$ et le second la contrainte $x_j \geq \lceil \alpha \rceil = \lfloor \alpha \rfloor + 1$. Nous répétons le processus pour chacun des sous-problèmes. Cette procédure est habituellement représentée sous forme d'un arbre binaire où, à chaque niveau, une partition du sommet père s'effectue suivant la règle décrite précédemment. Il s'agit alors de parcourir cet arbre d'énumération afin d'y trouver la solution optimale. L'exploration d'un chemin de l'arbre peut prendre fin pour trois raisons :

- la solution devient entière ;
- le domaine admissible d'un sous-problème devient vide ;
- la valeur de l'objectif correspondant à la solution optimale du problème relaxé est inférieure (moins bonne) à celle d'une solution admissible connue, possiblement obtenue à un autre sommet de l'arbre.

Dans chacun de ces trois cas on dit que le sommet est sondé, et il est inutile de pousser plus loin dans cette direction. L'algorithme s'arrête lorsque tous les sommets sont sondés. La meilleure solution obtenue au cours du déroulement de l'algorithme est alors l'optimum global de notre problème.

Algorithm 2. *Algorithme de B&B : cas général*

1. *Initialisation :*
   (a) *Poser $Z^* = -\infty$.*
   (b) *Appliquer le calcul de borne et les critères d'élagage à la racine (aucune variable fixée).*
   (c) *Critere d'arrêt : s'il n'y a plus de sous-problèmes non élagués, arrêter.*
2. *Branchement :*
   (a) *Parmi les sous-problèmes non encore élagués, choisir celui qui a été crée le plus récemment (s'il y a égalité, choisir celui de plus grande borne supérieure).*
   (b) *Appliquer le Test 1 : si le sous-problème est élagué, retourner en 2.*
   (c) *Brancher sur la prochaine variable entière à valeur non entière dans la relaxation PL.*
3. *Calcul de borne : résoudre la relaxation PL de chaque sous-problème.*
4. *Elagage : élaguer un sous-problème si*

(a) La borne superieure est inférieure ou égale à Z^*.

(b) La relaxation PL n'a pas de solution réalisable.

(c) Dans la solution optimale de la relaxation PL, toutes les variables entières sont à valeurs entières : si la borne supérieure est strictement supérieure à Z^*, Z^* est mise à jour et la solution de la relaxation PL devient la meilleure solution courante.

5. Retourner en 2.

Un sous-problème est elagué si une des trois conditions suivantes est satisfaite :
— test 1 : sa borne supérieure (valeur optimale de la relaxation PL) est inférieure ou égale à Z^* (valeur de la meilleure solution courante) ;
— test 2 : sa relaxation PL n'a pas de solution réalisable ;
— test 3 : la solution optimale de sa relaxation PL est entière.

3.4 Branch and bound : exemple

Un programme linéaire mixte (MIP) se présente comme suit :

$$(MIP) \equiv \begin{cases} \text{minimize} & cx \\ \text{subject to} & Ax \leq b \\ & x_i \in \mathbb{Z}, i = 1..p \\ & x_i \in \mathbb{R}, i = p+1..n \end{cases} \tag{3.1}$$

Le problème : Comment le résoudre correctement et rapidement ?

Branch and Bound (BB) Diviser pour régner. "Diviser/Branch" : subdivise le problème. "Régner/Bound" : considérer la qualité de la solution des sous-problèmes.

Branch and Cut (BC) BB en renforçant la relaxation linéaire LP d'un MIP avec des inégalités avant tout branchement. On raisonne sur les contraintes !

Branch and Price (BP) BB en se concentrant sur le choix de la variable entrante (column generation/pricing) dans la résolution de la relaxation. On raisonne sur les variables (les colonnes) !

Branch and bound
— Diviser le problème en sous-problèmes
— Calculer la relaxation LPR du sous-problème en considérant "réelles" les variables entières
  — LPR infaisable : stop.
  — LPR a une solution faisable entière : résolu, solution optimale pour le sous-problème.
  — LPR a une solution moins bonne que la meilleure solution entière : stop.
  — Sinon LPR a des composantes réelles, Diviser en sous-problèmes.

Soit le problème MIP [1] :

1. http ://www.ie.bilkent.edu.tr/ mustafap/courses/bb.pdf

$$\begin{array}{ll} \text{maximize} & z = -x_1 + 4x_2 \\ \text{subject to} & -10x_1 + 20x_2 \leq 22 \\ & 5x_1 + 10x_2 \leq 49 \\ & x_1 \leq 5 \\ & x_i \geq 0, x_1 \text{ et } x_2 \text{ sont entiers.} \end{array} \tag{3.2}$$

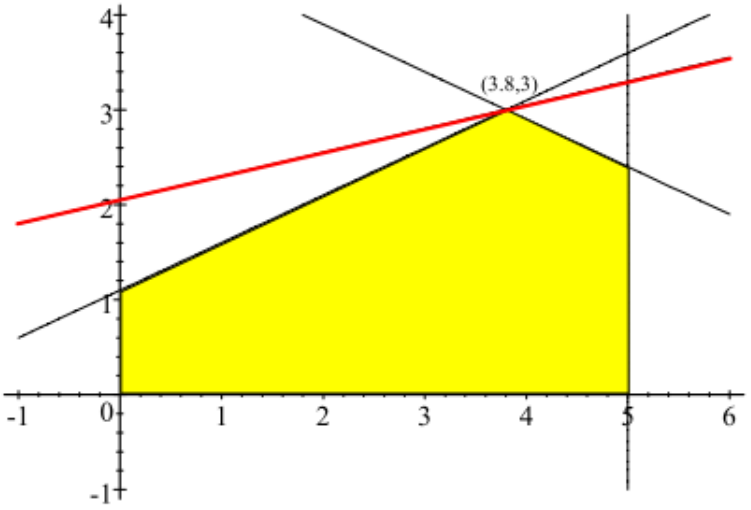
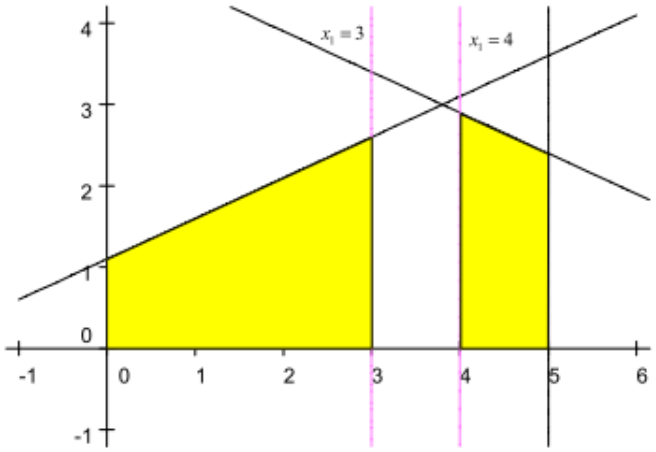


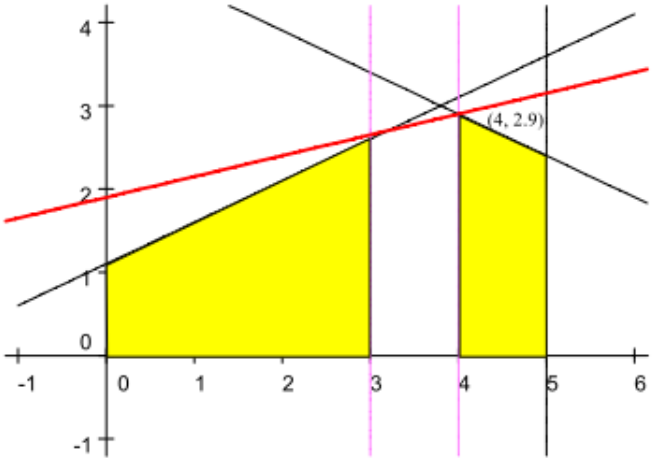
FIGURE 3.3 – Relaxation linéaire (3.3)

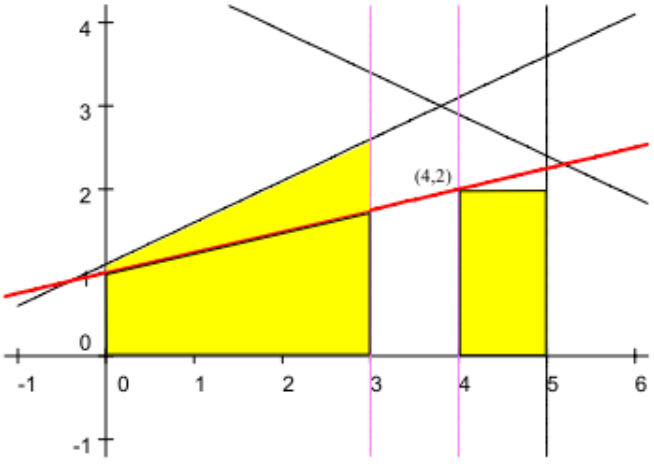
Au niveau de la relaxation

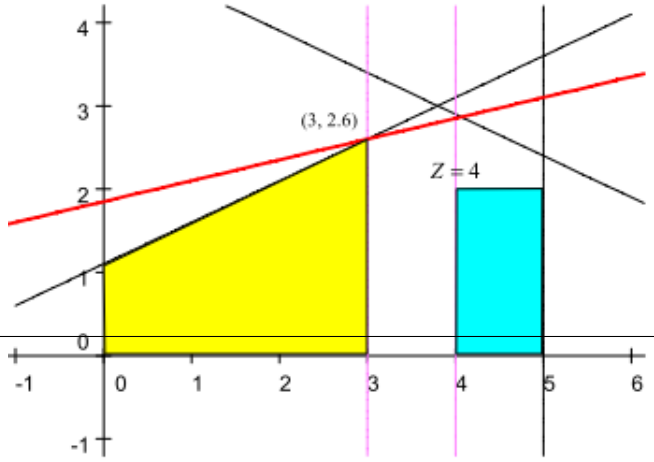
$$\begin{array}{ll} \text{maximize} & z = -x_1 + 4x_2 \\ \text{subject to} & -10x_1 + 20x_2 \leq 22 \\ & 5x_1 + 10x_2 \leq 49 \\ & x_1 \leq 5 \\ & x_i \geq 0, x_1 \text{ et } x_2 \text{ sont réelles.} \end{array} \tag{3.3}$$

La relaxation (3.3) est illustrée dans la Figure 3.3. **Sa solution est** $(3.8, 3)$ **avec** $z = 8.2$**.** Nous devons faire un "Branch" sur la variable fractionnaire x_1, en obtenant deux branches $x_1 \geq 4$ et $x_1 \leq 3$. Cette séparation (Branch) est illustrée dans la première image de la Figure 3.4.


$x_1 = 3$
$x_1 = 4$


(4, 2.9)


(4,2)


(3, 2.6)
Z = 4

Par la suite la procédure Branch and Bound procède comme suit :

1. Le sous-problème correspondant à $x_1 \geq 4$ est résolu sur $\mathbb{R}$. On obtient la solution $(4, 2.9)$ avec $z = 7.6$. On obtient ainsi deux sous-sous-problèmes $x_2 \geq 3$ et $x_2 \leq 2$.

2. Le sous-problème $x_2 \geq 3$ est infaisable.

3. Nous considérons donc maintenant le sous-problème $x_2 \leq 2$. Ce dernier a une solution optimale (donnée dans la deuxième image de la Figure 3.4, qui a comme solution $(4, 2)$ avec $z = 4$. Cette première solution va permettre d'avoir une meilleure borne (inférieure) $z^* = 4$ de notre problème avec une valeur $z = 4$.

4. En ce qui concerne la branche $x_1 \leq 3$, on obtient la solution $(3, 2.6)$ avec $z = 7.4$, donnant lieu à deux branches $x_2 \leq 2$ et $x_2 \geq 3$.

5. Le LPR de $x_2 \geq 3$ est infaisable.

6. En ce qui concerne $x_2 \leq 2$, on obtient la solution $(1.8, 2)$ avec $z = 6.2$, donnant lieu à deux branches $x_1 \geq 2$ et $x_1 \leq 1$.

7. Pour $x_1 \geq 2$, on obtient la solution $(2, 2)$ avec $z = 6$. La solution est entière, d'où la nouvelle borne $z^* = 6$.

8. Pour $x_1 \leq 1$, nous obtenons une solution $(1, 1.6)$ avec $z = 5.4$. 5.4 est inférieur à la meilleure borne courante z^*, d'où son élagage.

Soit un autre problème :

$$\begin{array}{ll} \text{maximize} & z = 9 + 5x_2 + 6x_3 + 4x_4 \\ \text{subject to} & 3x_2 + 5x_3 + 2x_4 \leq 4 \\ & x_3 + x_4 \leq 1 \\ & -x_1 + x_3 \leq 0 \\ & x_2 + x_4 \leq 0 \\ & x_i \in \{0, 1\} \end{array} \tag{3.4}$$

La procédure Branch and Bound est illustré dans la Figure 3.5

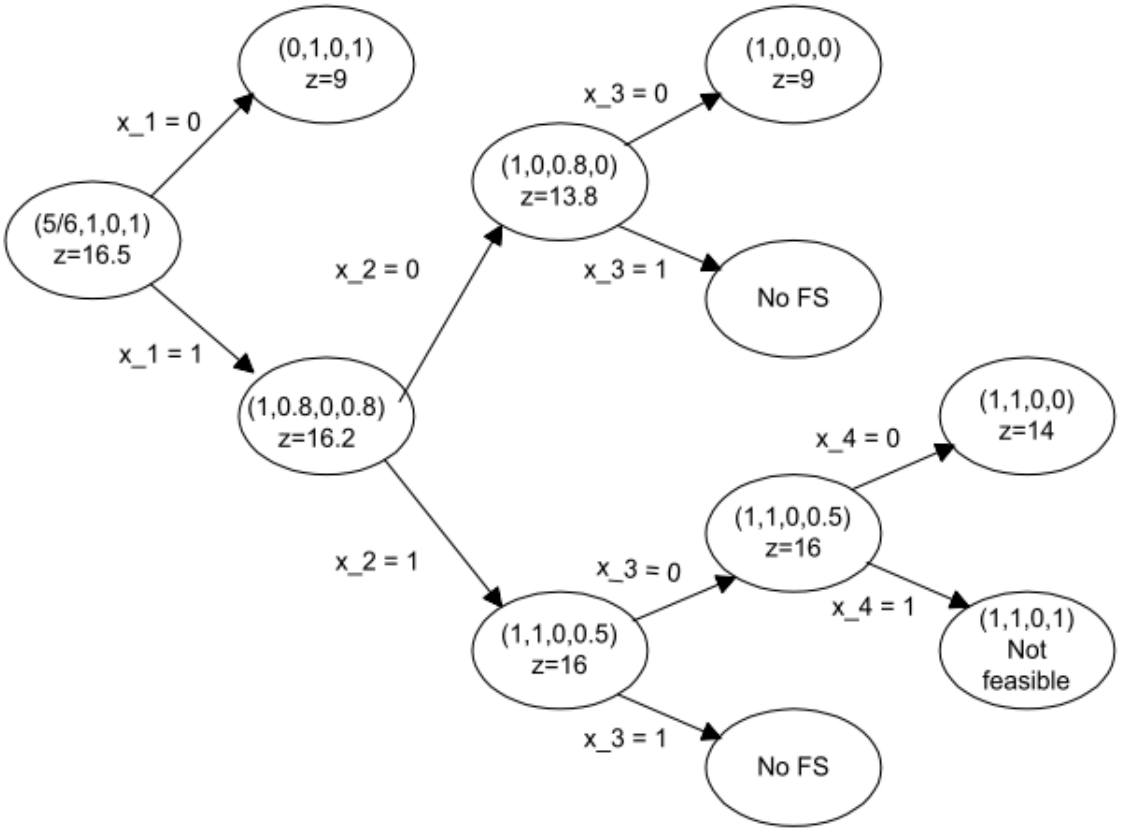


FIGURE 3.5 – Arbre de recherche de la procédure "Branche and Bound" sur le problème (3.3)

3.5 Branch and cut

Vu que l'énumération des variables entières peut être coûteuse, l'idée est d'ajouter des contraintes redondantes pour le modèle en nombres entiers, mais non pour la relaxation PL.

Exemple 14. *Considérons le programme mathématique*

$$\begin{aligned} \max_x \ & 4x_1 + \frac{5}{2}x_2 \\ & x_1 + x_2 \leq 6 \\ & 9x_1 + 5x_2 \leq 45 \\ & x_1,\ x_2 \in \mathcal{N}. \end{aligned}$$

Le dictionnaire optimal correspondant à la relaxation linéaire de ce programme contient

les deux contraintes

$$x_1 = \frac{15}{4} + \frac{5}{4}x_3 - \frac{1}{4}x_4,$$
$$x_2 = \frac{9}{4} - \frac{9}{4}x_3 + \frac{1}{4}x_4,$$

où x_3 et x_4 sont des variables d'écart. Puisque la variable de base x_1 n'est pas entière, cette solution de base n'est pas admissible. nous pouvons réécrire la première contrainte sous la forme

$$x_1 - \frac{5}{4}x_3 + \frac{1}{4}x_4 = \frac{15}{4}.$$

En utilisant l'identité

$$a = \lfloor a \rfloor + (a - \lfloor a \rfloor),$$

où $(a - \lfloor a \rfloor)$ représente la partie fractionnaire de a $(0 \leq a - \lfloor a \rfloor < 1)$, nous obtenons

$$x_1 + \left(\left\lfloor -\frac{5}{4} \right\rfloor + \frac{3}{4} \right) x3 + \left(\left\lfloor \frac{1}{4} \right\rfloor + \frac{1}{4} \right) x_4 = \left(\left\lfloor \frac{15}{4} \right\rfloor + \frac{3}{4} \right),$$

c'est-à-dire, en mettant tous les coefficients entiers à gauche et les coefficients fractionnaires à droite :

$$x_1 - 2x_3 - 3 = \frac{3}{4} - \frac{3}{4}x_3 - \frac{1}{4}x_4.$$

Puisque les variables x_3 et x_4 sont non négatives, la partie fractionnaire (constante du membre de droite) est inférieure à 1, le membre de droite est strictement inférieur à 1. Puisque le membre de gauche est entier, le membre de droite doit aussi être entier. Or un entier inférieur à 1 doit être inférieur ou égal à zéro. Nous en déduisons une contrainte additionnelle qui doit être satisfaite par toute solution admissible du problème originel, et que ne satisfait pas la solution de base courante :

$$\frac{3}{4} - \frac{3}{4}x_3 - \frac{1}{4}x_4 \leq 0.$$

En utilisant les identités $x_3 = 6 - x_1 - x_2$ et $x_4 = 45 - 9x_1 - 5x_2$, nous obtenons la coupe sous sa forme géométrique :

$$3x_1 + 2x_2 \leq 15.$$

Cette contrainte linéaire rend inadmissible la solution courante admissible, sans éliminer aucune autre solution entière. Si la solution du nouveau problème est entière, il s'agit de la solution optimale de notre problème. Sinon, nous construisons une nouvelle coupe et recommençons.

Exemple 15. *Reprenons le problème, illustré sur la Figure 3.6,*

$$\begin{aligned} \max\ & z = 3x_1 + 2x_2 \\ s.c.\ & 2x_1 + 3x_2 \leq 4, \\ & x_1, x_2 \text{ binaire.} \end{aligned}$$

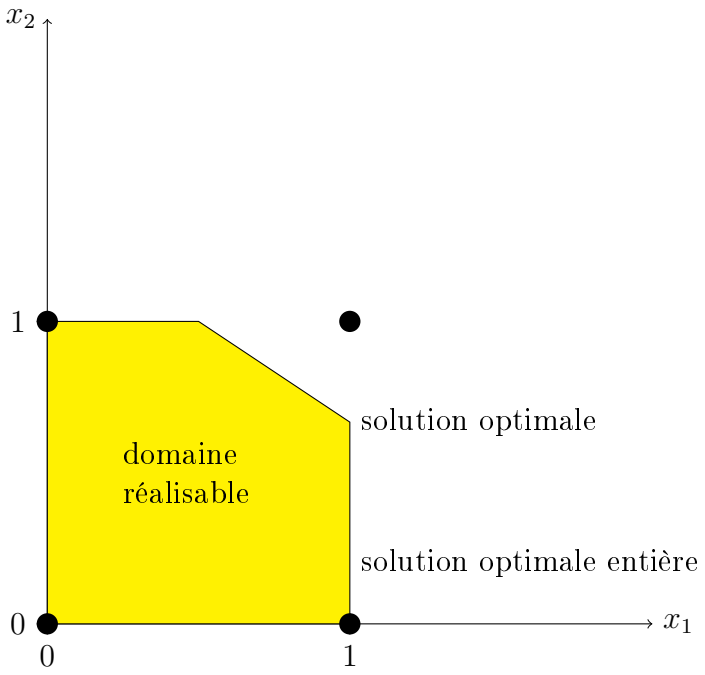


FIGURE 3.6 – Méthode de coupes

Les solutions réalisables sont $(0,0)$, $(1,0)$ *et* $(0,1)$.

Une contrainte redondante est :

$$x_1 + x_2 \leq 1.$$

Suite à l'ajout de la contrainte redondante, comme représenté sur la Figure 3.7, le problème est résolu à la racine.

Il y a plusieurs algorithmes permettant de générer de telles inégalités redondantes, appelées coupes. Mais il est rare que leur ajout permette de résoudre le probleme à la racine. L'ajout de coupes permet toutefois de réduire le nombre de sous-problèmes traités par l'algorithme de Branch-and-Bound. Nous pouvons même ajouter des coupes pour chaque sous-problème (pas seulement à la racine) : nous obtenons alors un algorithme de branch-and-cut.

Soit le problème [2] :

$$\begin{array}{ll} \text{minimize} & z = -6x_1 - 5x_2 \\ \text{subject to} & 3x_1 + x_2 \leq 11 \\ & -x_1 + 2x_2 \leq 5 \\ & x_1, x_2 \geq 0 \\ & x_1, x_2 \text{variables entières.} \end{array} \tag{3.5}$$

Ce problème est illustré dans la Figure 3.9

2. www.acsu.buffalo.edu/ nagi/courses/684/branch.pdf ?

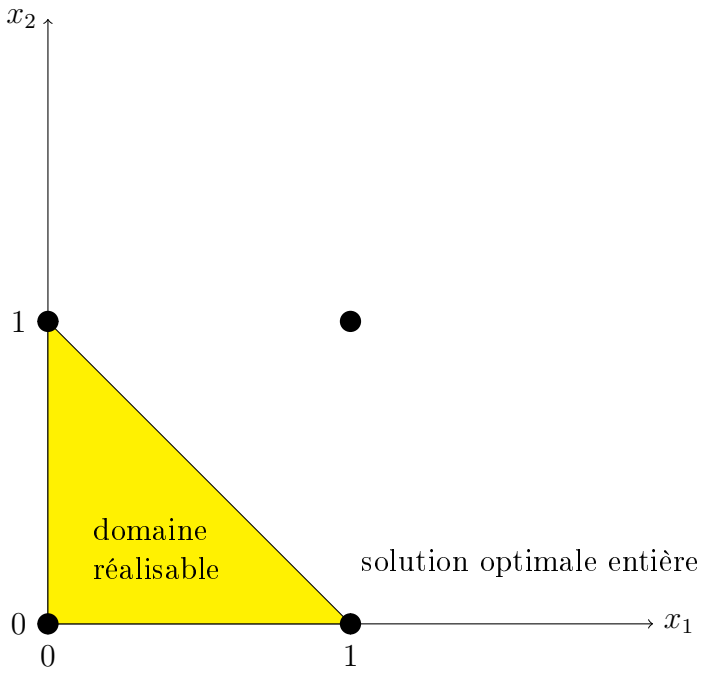


FIGURE 3.7 – Méthode de coupes : ajout d'une contrainte

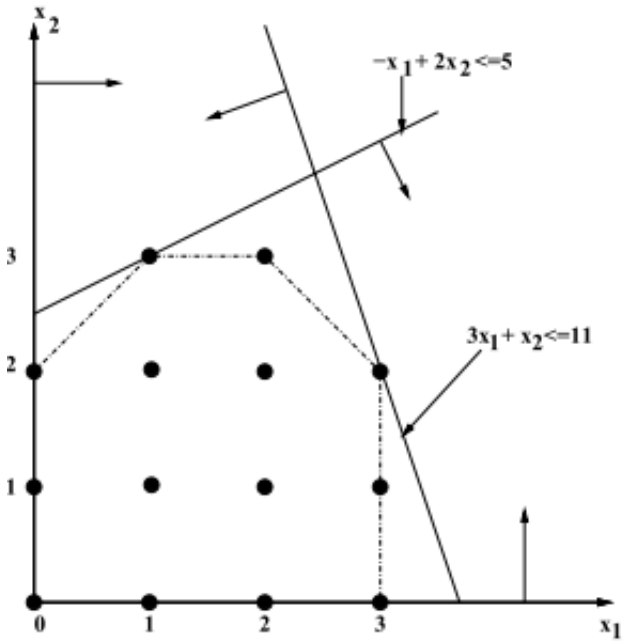


FIGURE 3.8 – Illustration du problème (3.5)

La procédure Branch and Bound sur le problème (3.5) est illustrée dans la Figure 3.9 avec la génération de la coupe $2x_1 + x_2 \leq 7$.

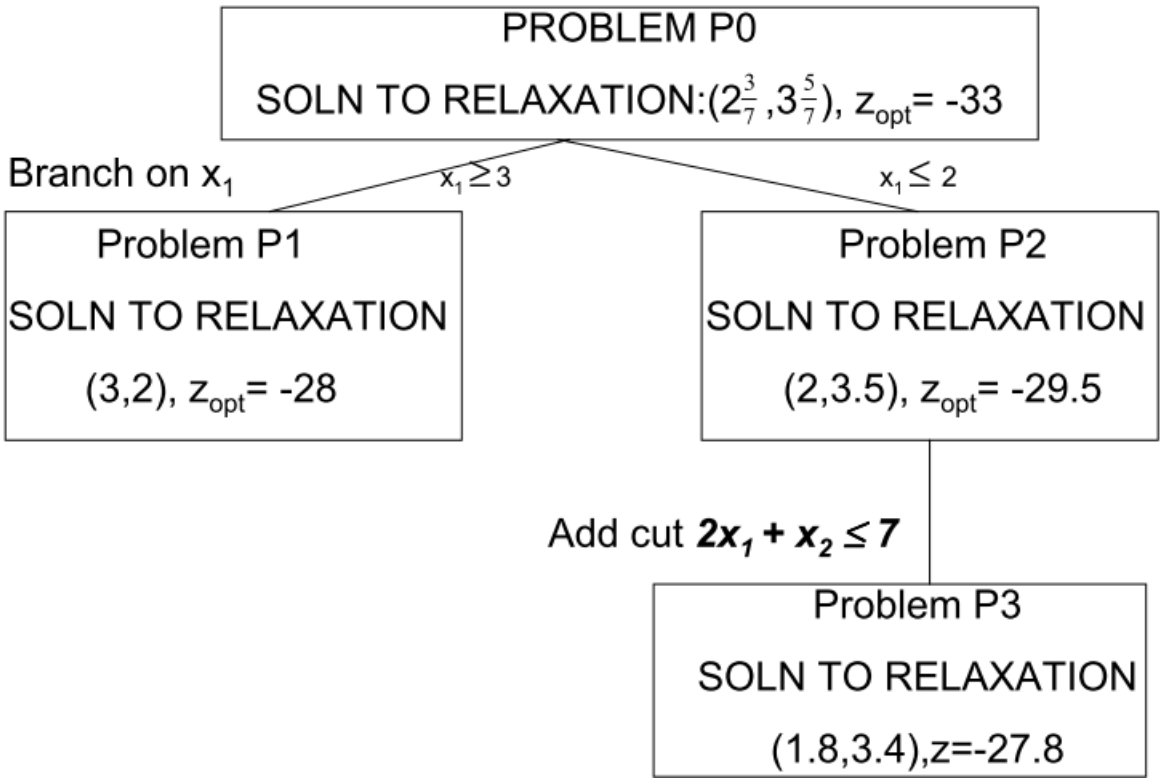


FIGURE 3.9 – Illustration de la procédure Branch and Bound sur le problème (3.5) avec la génération d'une coupe

Plus exactement la procédure Branch and Cut consiste en :

— La procédure est similaire à celle de Branch and Bound.

— Exception faite au niveau de l'étape de la résolution de la relaxation qui sera suivie d'une étape d'ajout de coupes valides.

Il existe plusieurs techniques de génération de coupes. Parmi lesquelles, on peut citer la suivante :

1. Prendre une combinaison linéaire et pondérée des inégalités à partir de la relaxation courante.
2. Exploiter le fait que les variables sont entières, en utilisant le mécanisme d'arrondi.
3. Les coupes générées ainsi, sont appelées les coupes de Chvatal-Gomory.

Par exemple, soit

$$\frac{1}{6}(3x_1 + x_2 \leq 11) + \frac{5}{12}(-x_1 + 2x_2 \leq 5$$
donnant $x_2 \leq 3\frac{11}{12}$

car $x_1 < 12$. La partie gauche doit être entière, d'où l'arrondi de la partie droite, donnant finalement :

$$x_2 \leq 3.$$

Soit l'exemple d'une coupe de Gomory. Soit le problème [3] :

3. http ://mat.gsia.cmu.edu/classes/integer/node15.html

$$\begin{array}{ll} \text{maximize} & z = 7x_1 + 9x_2 \\ \text{subject to} & -x_1 + 3x_2 \leq 6 \\ & 7x_1 + x_2 \leq 35 \\ & x_1, x_2 \geq 0 \\ & x_1, x_2 \text{variables entières.} \end{array} \tag{3.6}$$

Lors de la résolution de sa relaxation LPR, on obtient la tableau final du simplex donné dans la Figure 3.10

Variable	x_1	x_2	s_1	s_2	$-z$	RHS
x_2	0	1	7/22	1/22	0	7/2
x_1	1	0	-1/22	3/22	0	9/2
$-z$	0	0	28/11	15/11	1	63

FIGURE 3.10 – Tableau final du simplex du problème (3.6)

En considérant la première contrainte du tableau du simplex, on a :

$$x_2 + 7/22s_1 + 1/22s_2 = 7/2$$

qu'on peut réécrire en

$$x_2 - 3 = 1/2 - 7/22s_1 - 1/22s_2$$

La partie gauche est purement entière. La partie droite est composée de deux négatifs additionnés à un positif fractionnaire. D'où le fait que :

$$1/2 - 7/22s_1 - 1/22s_2 \leq 0.$$

Cette coupe est dire "coupe de Gomory".

L'idée des coupes est de se rapprocher le plus de l'idée illustrée dans la Figure 3.11.

http://www.more.ms.unimelb.edu.au/students/operationsresearch/lecturenotes/620362_Gomory.pdf

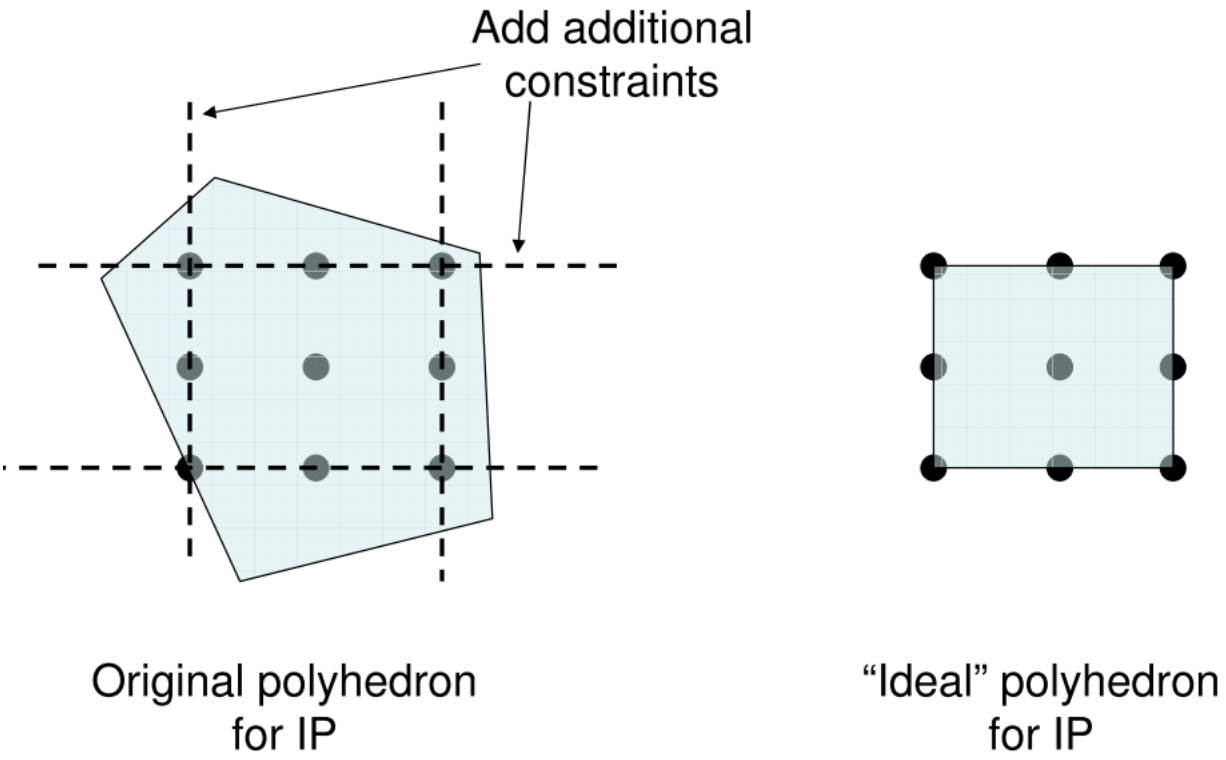


FIGURE 3.11 – Illustration des coupes idéales

Plus généralement, voici ces deux coupes formellement définies.

Proposition 2 (Coupe de Gomory). *Soit* $P = \{x \in \mathbb{Z}_+^n : Ax \leq b\}$. *Soit* $\alpha x \leq \beta$ *valide dans* P *alors*

$$\sum_{i \in 1..n} \lfloor \alpha_i \rfloor x_i \leq \lfloor \beta \rfloor$$

est aussi valide pour P.

Proposition 3 (Coupe de Chvatal-Gomory). *Soit* $P = \{x \in \mathbb{Z}_+^n : Ax \leq b, x \geq 0\}$, *avec* $A \in \mathbb{R}^{m \times n}$ *et* $u \in \mathbb{R}_+^m$, *alors*

$$\sum_{j \in 1..n} \lfloor \sum_{i \in 1..n} u_i A_{i,j} \rfloor x_j \leq \lfloor \sum_{i \in 1..m} u_i b_i \rfloor$$

est aussi valide pour P.

3.6 Travaux Dirigés I (Modélisation)

3.6.1 Exercice

Une entreprise de production de produits alimentaires désire orienter ses activités sur 3 lignes de produits I, II et III. Le profit moyen par produit est estimé à 300 DA par tonne pour le produit I, 200 DA pour le produit II, et 500 DA pour le produit III. Les équipements sont répartis en trois départements de production :

— Fabrication ;
— Mélange ;
— Emballage.

La durée maximale de charge pour chaque département est de 8 heures par jour. Les processus de production relatifs aux trois produits font l'objet des opérations successives suivantes :

— Produit I : Fabrication-Mélange. La production est enlevée par les utilisateurs dès qu'elle est réalisée. Chaque tonne ainsi produite exige 3 heures d'utilisation de la capacité de "fabrication" et 1 heures de capacité du département "mélange".
— Produit II : Mélange-Emballage. Le produit est réalisé à partir d'achats de composantes alimentaires non produites dans l'entreprise, fait l'objet des seules opérations de mélange et emballage. Chaque tonne produite exige 1 heure d'utilisation de capacité "mélange" et 2 heures de capacité "emballage".
— Produit III : Fabrication-Mélange-Emballage. Le produit subit les 3 séries d'opérations fabrication-mélange-emballage, chacune d'entre elles exigeant respectivement 2 heures, 1 heure et 1 heure de capacité des équipements.

1. Ecrire le programme linéaire modélisant ce problème ?

3.6.2 Exercice

Dans une entreprise de construction de matériel électrique, on dispose de 1200 heures/machine par mois et 1500 heures/ouvrier par mois. Les contacteurs nécessitent 15 heures/machine, 12 heures/ouvrier et rapportent 8 DA (par unité). Les disjoncteurs nécessitent 30 heures/machine, 30 heures/ouvrier et rapportent 20 DA (par unité). Les compteurs nécessitent 20 heures/machine, 25 heures/ouvrier et rapportent 18 DA (par unité).

1. Modéliser sous forme d'un programme linéaire PL la recherche des quantités optimales des contacteurs x_1, des disjoncteurs x_2 et des compteurs x_3 pour maximiser le gain ?
2. Formuler le dual du PL ?
3. Démontrer que la contrainte $17x_1 + 60x_2 + 45x_3 \leq 2700$ est non-redondante [4] dans le PL ?
4. Si des contraintes sont redondantes dans un PL, comment se comporterait l'algorithme du simplex ? Comment y remédier ?

4. Soit le système d'inéquations $l_i \equiv \sum_{j=1\ldots m} a_{i,j}x_j \geq c_i$, avec $i = 1\ldots n$. Une inéquation $l' \equiv \sum_{j=1\ldots m} b_j x_j \geq d$ est dite redondante si et seulement si $l' = \sum_{i=1\ldots n} \lambda_i l_i$ où $\forall i, \lambda_i$ est un réel positif.

3.6.3 Exercice

Etant donné deux variables x_1, x_2 binaires (i.e., $x_1 \in \{0,1\}$, $x_2 \in \{0,1\}$). Soit l'expression non linéaire de leur multiplication :

$$(P) \begin{cases} y = x_1 \times x_2 \\ x_1 \in \{0,1\}, x_2 \in \{0,1\}, y \in \{0,1\} \end{cases} \tag{3.7}$$

Nous proposons un système d'équations linéaires (Q)

$$(Q) \begin{cases} y \leq x_1 \\ y \leq x_2 \\ y \geq x_1 + x_2 - 1 \\ x_1 \in \{0,1\}, x_2 \in \{0,1\}, y \in \{0,1\} \end{cases} \tag{3.8}$$

Démontrer que (P) est équivalent à (Q) ?

3.6.4 Exercice

Soit le problème

$$(P) \begin{cases} \min \sum_{i=1..n} c_i x_i \\ \\ (\sum_{j=1..n} A_j x_j \leq b) \text{ OU } (\sum_{j=1..n} C_j x_j \leq d) \\ x_i \geq 0, i = 1..n \end{cases} \tag{3.9}$$

Le problème (P) n'est pas un PL car il contient la contrainte de disjonction "OU". Nous proposons un autre problème (Q)

$$(Q) \begin{cases} \min \sum_{i=1..n} c_i x_i \\ \sum_{j=1..n} A_j x_j \leq b + My \\ \sum_{j=1..n} C_j x_j \leq d + M(1-y) \\ x_i \geq 0, i = 1..n \\ y \in \{0,1\} \end{cases} \tag{3.10}$$

où M est un nombre très grand.

Démontrer que (P) est équivalent à (Q) ?

3.6.5 Exercice

Une firme *Penault* produit deux types de carrosseries C_1 et C_2. Nous nous intéressons à deux tâches indépendantes :

- *Peinture* : Cette firme ne pourra pas peindre C_2 à partir d'un maximum de 40 carrosseries C_1. Cette firme ne pourra pas aussi peindre C_1 à partir d'un maximum de 60 carrosseries C_2.
- *Fabrication* : Cette firme ne pourra pas produire C_2 à partir d'un maximum de 50 carrosseries C_1. Cette firme ne pourra pas aussi fabriquer C_1 à partir d'un maximum de 50 carrosseries C_2.

La carrosserie C_1 rapporte 300 DA et C_2 200 DA.

1. Modéliser sous forme d'un programme linéaire PL la recherche des quantités optimales x_1 de C_1 et x_2 de C_2 à produire pour maximiser le gain ? (N.B. Dans une tâche T donnée dans la firme, si la firme ne pourra pas produire C_2 à partir de k_1 C_1, et la firme ne pourra pas aussi produire C_1 à partir de k_2 C_2, la contrainte T peut être traduite en une seule contrainte linéaire $(x_1/k_1) + (x_2/k_2) \leq 1$.)
2. Tracer graphiquement le polyèdre décrivant le PL et extraire la solution optimale sans passer par l'algorithme du Simplexe ?
3. Soit la contrainte supplémentaire suivante : la firme *Penault* doit produire au moins 30 carrosseries C_1 et au moins 20 carrosseries C_2. Comment devient-il le PL ?

3.7 Travaux Dirigés II (séparation/évaluation)

3.7.1 Exercice [Problème d'affectation]

Soit un problème d'affectation ayant comme données :
— n tâches à affecter à n personnes
— Toute personne exécute 1 et 1 seule tâche.
— $c_{i,j}$ est le coût d'affectation de la tâche i à la personne j.
Le problème : trouver l'affectation des tâches aux personnes qui minimise le coût total ?
Soit l'instance du problème :

	Tâche 1	Tâche 2	Tâche 3	Tâche 4
Personne a	9	2	7	8
Personne b	6	4	3	7
Personne c	5	8	1	8
Personne d	7	6	9	4

Soit le calcul suivant :

$$LB1 = \sum_{i=1..n} \min_{j=1..m} c_{i,j}$$

Questions :

1. Démontrer que $LB1$ est une méthode de relaxation valide.
2. Dérouler l'algorithme par séparation évaluation sur l'instance donnée en exploitant $LB1$ comme relaxation.

3.7.2 Exercice [Problème de sac-à-dos]

Soit un problème de sac-à-dos ayant comme données :
— Un sac de volume V.
— Un ensemble de n objects $O = \{o_1, o_2, ..., o_n\}$.
— Tout objet o_i a un volume v_i et une utilité u_i.
— On suppose que $\sum_{i=1..n} v_i > V$.
Le problème : Trouver un sous-ensemble d'objets de O qui maximise l'utilité et qui soit borné par V.
Soit le calcul LB2 suivant :
— Soit le vecteur R des rapports entre les utilités et les volumes $[u_1/v_1, u_2/v_2, ..., u_n/v_n]$.
— Soit $x = (x_1, x_2, ..., x_n)$ les variables binaires associées aux n objets. $x_i = 0$ si l'objet i est mis dans le sac, sinon 1.
— Le calcul consiste en :
  — Trier R par ordre décroissant ;
  — $Disponible = V$;
  — Initialiser x à zéro ;
  — Pour $i = 1$ jusqu'à n faire :
    — Si $(v_i \leq Disponible)$ alors $x_i = 1$; sinon $x_i = Disponible/v_i$ finsi.
    — $Disponible = Disponible - v_i * x_i$;

Questions :

1. Démontrer que $LB2$ est une méthode de relaxation valide.

2. Dérouler l'algorithme par séparation évaluation sur l'instance donnée en exploitant $LB2$ comme relaxation sur les données suivantes :

Objets	1	2	3	4
u	16	22	12	8
v	5	7	4	3
V	14			

3. Soit maintenant un autre algorithme UB2 qui reprend l'algorithme LB2 mais en trouvant une solution faisable (sans garantie d'optimalité) dans une démarche gloutonne :
   — Soit le vecteur R des rapports entre les utilités et les volumes $[u_1/v_1, u_2/v_2, ..., u_n/v_n]$.
   — Soit $x = (x_1, x_2, ..., x_n)$ les variables binaires associées aux n objets. $x_i = 0$ si l'objet i est mis dans le sac, sinon 1.
   — Le calcul consiste en :
      — Trier R par ordre décroissant ;
      — $Disponible = V$;
      — Initialiser x à zéro ;
      — Pour $i = 1$ jusqu'à n faire :
         — Si $(v_i \leq Disponible)$ alors $x_i = 1$; sinon $x_i = 0$ finsi.
         — $Disponible = Disponible - v_i * x_i$;
4. Dérouler à nouveau l'algorithme par séparation évaluation sur l'instance donnée en exploitant $LB2$ et $UB2$ sur les données ci-dessus.
5. Formuler ce problème en terme d'un programme linéaire en nombres entiers ?
6. Résoudre l'instance du problème par séparation/évaluation en utilisant la relaxation linéaire. Faites vous aider avec le solveur ampl pour le calcul de borne via l'algorithme du simplexe. Comparer l'arbre de recherche utilisant la relaxation linéaire avec l'arbre via la borne $LB2$?

3.8 Travaux Pratiques (Optimisation)

3.8.1 Exercice [Un modèle simple]

Installer `AMPL` en décompressant "amplide-demo-linux32.tar.gz", la version "AMPL FOR STUDENTS" à partir du site officiel http://ampl.com, puis invoquer AMPL :

```
tar xzf amplide-demo-linux32.tar.gz
ample
```

Commençons par un exemple simple d'un problème d'optimisation.

Une compagnie de commercialisation de la peinture fournit deux types de couleurs : la couleur bleue et la couleur dorée. La couleur bleue est vendu 10 DA par gallon, et la dorée 15 DA. La compagnie a une unité de fabrication et fabrique une couleur par unité de temps. Cependant, la couleur bleue est plus facile à produire et l'unité de fabrication peut produire 40 gallons par heure, et 30 gallons pour la couleur dorée. L'unité de vente de cette compagnie informe qu'elle ne peut pas vendre plus de 860 gallons de la couleur dorée et 1000 gallons de la couleur bleue. Supposons que dans le volume horaire travaillé par semaine est de 40 heures, et le produit est toujours stocké dans la semaine qui suit. Nous voulons déterminer le nombre de gallons de couleur dorée et bleue à produire pour maximiser le gain.

Soit `PaintB` (resp. `PaintG`) le nombre de gallons à produire de la peinture de couleur bleue (resp. de couleur dorée). Le problème peut être formulé comme suit :

$$\begin{array}{ll} \max & 10PaintB + 15PaintG \\ s.t. & \frac{1}{40}PaintB + \frac{1}{30}PaintG \leq 40 \\ & 0 \leq PaintB \leq 1000 \\ & 0 \leq PaintG \leq 860 \end{array}$$

`AMPL` permet de concevoir des programmes mathématiques à saisir avec une syntaxe qui est très proche de la notation algébriques que l'on vient de voir. Pour utiliser `AMPL`, on doit créer un fichier texte qui va contenir le texte du programme mathématique (On utilisera un éditeur de texte).

Saisir le texte suivant dans l'éditeur :

```
## Example One
var PaintB; # amount of blue
var PaintG; # amount of gold

maximize profit: 10*PaintB + 15*PaintG;

subject to time: (1/40)*PaintB + (1/30)*PaintG <= 40;
subject to blue_limit: 0 <= PaintB <= 1000;
subject to gold_limit: 0 <= PaintG <= 860;
```

Sauvegarder l'exemple sous le nom `examp1.mod`.

Notez les différents opérateurs et expressions utilisés pour déclarer les variables, la fonction objectif et les contraintes.

Suivre les étapes suivantes pour manipuler notre exemple :

— `AMPL` est un environnement de programmation mathématique qui peut s'interfacer avec la plupart des solveurs connus (`Cplex`, `Minos`, `lpsolve`, ...). Précisons maintenant à `AMPL` le solveur que l'on veut utiliser :
`ample: option solver cplex;`
— Charger notre exemple
`ample: model examp1.mod;`
— Si vous avez des erreurs au niveau de la saisie, vous pouvez revenir à votre éditeur de texte, corriger puis recharger le fichier, en commençons par vider l'environnement :
`ample: reset;`
— Pour résoudre le problème, il suffit d'invoquer
`ample: solve;`
— `AMPL` affiche
`CPLEX 8.0.0: optimal solution; objective 17433.33333`
`2 simplex iterations (0 in phase I)`
— Pour afficher la solution, il suffit d'invoquer la fonction d'affichage `display` avec les variables à afficher
`ample: display PaintB, PaintG;`
— Vous pouvez rediriger la sortie vers un fichier
`ample: display PaintB, PaintG > examp1.out`

3.8.2 Exercice [Un modèle plus général]

Le problème que l'on vient de manipuler est figé ; cependant un nombre de problèmes économiques d'optimisation peuvent se formuler dans le même type de modèle mais se portant sur un nombre de produits, des coûts, et des volumes horaires quelconques. Soit ici ce modèle général :

Soient	n : le nombre de couleurs
	t : le temps total disponible
	p_i : le profit par gallon i, $i = 1..n$
	r_i : gallons par heure de la couleur i, $i = 1..n$
	m_i : la demande maximale de la couleur i, $i = 1..n$
Variables	x_i : le nombre de gallons de la couleur i, $i = 1..n$
Maximiser	$\sum_{i=1..n} p_i x_i$
Subject to	$\sum_{i=1..n} (1/r_i) x_i \leq t$
	$0 \leq x_i \leq m_i$ pour tout i, $i = 1..n$.

Ainsi, notre modèle simple se résume en le modèle général ci-dessus où $n = 2, t = 40, p_1 = 10, p_2 = 15, r_1 = 40, r_2 = 30, m_1 = 1000$ et $m_2 = 860$.

Justement `AMPL` permet de saisir des modèle généraux permettant de les paramétrer aisément. Le modèle général est toujours stocké dans un fichier à part, soit par exemple `model.mod`. Le paramétrage ou la fixation des paramètres sur une instance donnée est faite dans un fichier séparé, dit fichier des données, soit *model.dat*.

Dans notre cas, nous pouvons saisir le modèle général de la compagnie dans un fichier `examp1g.mod` comme suit :

```
param n;
param t;
param p{i in 1..n};
param r{i in 1..n};
param m{i in 1..n};
var paint{i in 1..n};

maximize profit: sum{i in 1..n} p[i]*paint[i];

subject to time: sum{i in 1..n} (1/r[i])*paint[i] <= t;
subject to capacity{i in 1..n}: 0 <= paint[i] <= m[i];
```

Soit le fichier de données `examp1g.dat` qui contient le jeu de données pour notre premier modèle simple :

```
## Example Two - Data
param n:= 2;
param t:= 40;

param p:= 1 10
          2 15;
param r:= 1 40
          2 30;
param m:= 1 1000
          2 860;
```

Chargeons maintenant ces fichiers pour résoudre encore le modèle simple dans cette nouvelle modélisation générale :

```
ample: reset
ample: model examp1g.mod;
ample: data examp1g.dat;
ample: solve;
```

Invoquons l'affichage de la solution :

```
ample: display paint;
```

3.8.3 Exercice [Une autre manière pour saisir les données]

On peut saisir tous les paramètres en une seule fois, avec :

```
## Example Two - Another way to write the data
param n:= 2;
param t:= 40;
param: p r m:=
  1 10 40 1000
  2 15 30 860;
```

3.8.4 Exercice [Les ensembles]

Dans la tâche de la modélisation, nous avons souvent besoin de garder la nomenclature de l'environnement dans lequel le problème a été posé. Soit par exemple la formulation

```
## Example Two - Model with sets
set P;
param t;
param p{i in P};
param r{i in P};
param m{i in P};
var paint{i in P};

maximize profit: sum{i in P} p[i]*paint[i];

subject to time: sum{i in P} (1/r[i])*paint[i] <= t;
subject to capacity{i in P}: 0 <= paint[i] <= m[i];
```

Ainsi, on voit que l'ensemble des couleurs est maintenant un ensemble que l'on peut énumérer en utilisant des noms réels :

```
## Example Two - Data with sets
set P:= blue gold;
param t:= 40;
param p:= blue 10 gold 15;
param r:= blue 40 gold 30;
param m:= blue 1000 gold 860;
```

Et aussi

```
## Example Two - Data with sets
set P:= blue gold;
param t:= 40;
param: p r m:=
blue 10 40 1000
gold 15 30 860;
```

3.8.5 Exercice [Des paramètres et des variables de deux dimensions]

Nous avons souvent besoin de faire appel à des modèles où les paramètres et les variables ont plusieurs dimensions, et tout particulièrement deux dimensions.

Soit le problème suivant :

La compagnie a fait une extension, et a maintenant trois dépôts pour stocker la peinture bleue. Dans une semaine, la peinture doit être acheminer à différents clients. Pour chaque dépôt avec chaque client, les coûts d'acheminement sont différents. Nous donnons ci-dessous ces coûts.

	Cust 1	Cust. 2	Cust. 3	Cust. 4
Warehouse 1	1	2	1	3
Warehouse 2	3	5	1	4
Warehouse 3	2	2	2	2

Note de traduction : Warehouse est le dépôt, Customer est le client.

Le nombre de gallons disponibles au niveau de chaque dépôt, et le nombre de gallons demandé par chaque client sont donnés respectivement comme suit :

```
Warehouse 1 250
Warehouse 2 800
Warehouse 3 760

Customer 1: 300
Customer 2: 320
Customer 3: 800
Customer 4: 390
```

Le modèle `AMPL` de ce problème est donné comme suit :

```
## Example Three - Model
param warehouse; # number of warehouses
param customer; # number of customers

#transportation cost from warehouse i
#to customer j
param cost{i in 1..warehouse, j in 1..customer};
param supply{i in 1..warehouse}; #supply at warehouse i
param demand{i in 1..customer}; #demand at customer j

var amount{i in 1..warehouse, j in 1..customer};

minimize Cost: sum{i in 1..warehouse, j in 1..customer}
cost[i,j]*amount[i,j];

subject to Supply {i in 1..warehouse}: sum{j in 1..customer} amount[i,j] = supply[i];
subject to Demand {j in 1..customer}: sum{i in 1..warehouse} amount[i,j] = demand[j];
subject to positive{i in 1..warehouse, j in 1..customer}: amount[i,j]>=0;
```

Stockons ce modèle dans un fichier `examp2.mod`.

Et le jeu de données qui suit dans `examp2.dat`.

```
## Example Three - Data
param warehouse:= 3;
param customer:= 4;
param cost: 1 2 3 4 :=
1 1 2 1 3
```

```
2 3 5 1 4
3 2 2 2 2;

param supply:=
1 250
2 800
3 760;

param demand:=
1 300
2 320
3 800
4 390;
```

On peut aussi saisir les noms propres des différents paramètres pour obtenir un modèle qui reprend les mêmes termes que ceux utilisés en pratique.

```
## Example Three - Model using sets
set Warehouses;
set Customers;

#transportation cost from warehouse i
#to customer j
param cost{i in Warehouses, j in Customers};
param supply{i in Warehouses}; #supply at warehouse i
param demand{j in Customers}; #demand at customer j
var amount{i in Warehouses, j in Customers};

minimize Cost: sum{i in Warehouses, j in Customers} cost[i,j]*amount[i,j];

subject to Supply {i in Warehouses}: sum{j in Customers} amount[i,j]=supply[i];
subject to Demand {j in Customers}: sum{i in Warehouses} amount[i,j]=demand[j];
subject to positive{i in Warehouses, j in Customers}: amount[i,j]>=0;
```

Et le jeu de données comme suit :

```
## Example Three - Data with sets
set Warehouses := Oakland San_Jose Albany;
set Customers:= Home_Depot K_mart Wal_mart Ace;

param cost: Home_Depot K_mart Wal_mart Ace :=
Oakland   1 2 1 3
San_Jose  3 5 1 4
Albany    2 2 2 2;

param supply:=
Oakland  250
```

```
San_Jose 800
Albany   760;

param demand:=
Home_Depot 300
K_mart     320
Wal_mart   800
Ace        390;
```

3.8.6 Exercice [Programmation en nombres entiers]

Souvent, dans les modèles mathématiques, certaines variables peuvent être entières. AMPL facilité la manipulation de ces variables via les mots clés `integer` et `binary` que l'on peut placer après le nom de la variable pour dire que la variable en question est respectivement entière ou prend ses valeurs dans le domaine binaire $\{0, 1\}$.

Soit une modification au niveau des capacités des dépôts :

```
Warehouse 1 550
Warehouse 2 1100
Warehouse 3 1060
```

Cette augmentation vient avec un coût donné ci-dessous

```
Warehouse 1 500
Warehouse 2 500
Warehouse 3 500
```

Nous voyons ici que la demande excède la produits disponibles.

Soit maintenant le modèle suivant qui prend en compte ces nouvelles spécificités :

```
## Example Four - Mixed-IP model file for the warehouse location
  problem
set Warehouses;
set Customers;
#transportation cost from warehouse i
#to customer j
param cost{i in Warehouses, j in Customers};
param supply{i in Warehouses}; #supply capacity at warehouse i
param demand{j in Customers}; #demand at customer j
param fixed_charge{i in Warehouses}; #cost of opening warehouse j

var amount{i in Warehouses, j in Customers};
var open{i in Warehouses} binary; # = 1 if warehouse i is opened, 0 otherwise

minimize Cost: sum{i in Warehouses, j in Customers}
                 cost[i,j]*amount[i,j] + sum{i in Warehouses} fixed_charge[i]*open[i];
```

```
subject to Supply {i in Warehouses}: sum{j in Customers} amount[i,j] <= supply[i]
subject to Demand {j in Customers}: sum{i in Warehouses} amount[i,j]=demand[j];
subject to positive{i in Warehouses, j in Customers}: amount[i,j]>=0;
```

Soit le jeu de données suivant :

```
## Example Four - Data file for the warehouse location problem
set Warehouses:= Oakland San_Jose Albany;
set Customers:= Home_Depot K_mart Wal_mart Ace;

param cost: Home_Depot K_mart Wal_mart Ace:=
Oakland  1 2 1 3
San_Jose 3 5 1 4
Albany   2 2 2 2;

param supply:=
Oakland  550
San_Jose 1100
Albany   1060;

param demand:=
Home_Depot 300
K_mart     320
Wal_mart   800
Ace        390;

param fixed_charge:=
Oakland  500
San_Jose 500
Albany   500;
```

Faites les tests de résolution !

3.8.7 Exercice [Mise en pratique des modèles étudiés]

1. Mettre en forme le problème de l'exemple 2 sous forme AMPL. Commentez !
2. Mettre en forme le problème de l'exemple 3 sous forme AMPL, en proposant des paramètres significatifs. Commentez !
3. Soit le problème de localisation de l'exemple 3 : modifier le modèle en supposant qu'un site a plusieurs capacités de production à décider !
4. Mettre en forme le problème de l'exemple d'un objectif avec coûts fixes sous forme AMPL. Commentez !
5. Mettre en forme le problème de l'exemple 7 sous forme AMPL, en mettant M à une très grande valeur. Comment se comporte le modèle si M est petit ? Commentez !
6. Mettre en forme le problème de l'exemple 8 sous forme AMPL. Commentez !

7. Mettre en forme le problème de l'exemple 9 sous forme AMPL, dans sa forme générique, en séparant entre le modèle et les données des paramètres.

3.8.8 Exercice [Séparation/Evaluation]

1. Déroulez à nouveau la résolution du cas 0-1 de l'algorithme par séparation/évaluation en exploitant le solveur AMPL sur le domaine réel.
2. Déroulez à nouveau la résolution du cas $I\!N$de l'algorithme par séparation/évaluation en exploitant le solveur AMPL sur le domaine réel.

3.8.9 Exercice [Programmation nonlinéaire]

Assez souvent, on fait appel dans les modèles à des fonction nonlinéaires qui donnent lieu à des programmes nonlinéaires.

Première chose, il faut maintenant faire appel à un solveur qui sait manipuler ces problèmes. `Cplex` sait résoudre des programmes linéaires classiques et les programmes linéaire en nombres entiers ou mixte. Pour les modèles nonlinéaire, on fera appel au solveur `Minos`.

```
option solver minos;
```

Soit le problème suivant :

Supposons que nous avons plusieurs alternatives d'investissement A. Et nous savons évaluer le retour sur investissement pour certaines années T. A partir de ces données, on peut calculer la matrice des covariance des investissements.

On veut donc :

— *maximiser le retour sur investissement sous le maximum des risques potentiels.*

— *minimiser les risques sous le minimum de retour sur investissement.*

Le modèle suivant reprend les termes de ce problème d'une façon générale.

```
##Example 5 - Nonlinear Portfolio
set A; # asset categories
set T := {1984..1994}; # years
param s_max default 0.00305; # i.e., a 5.522 percent std. deviation on reward
param R {T,A};
param mean {j in A} := ( sum{i in T} R[i, j] - mean[j];
param Rtilde {i in T, j in A} := R[i,j] - mean[j];

var alloc{A} >=0;

minimize reward: - sum{j in A} mean[j]*alloc[j];

subject to risk_bound:
  sum{i in T} (sum{j in A} Rtilde[i,j]*alloc[j])^2 / card{T} <= s_max;

subject to tot_mass: sum{j in A} alloc[j] = 1;
```

Notons :

— Nous avons introduit la variable `alloc` pour indiquer le taux de ressources que l'on veut utiliser au niveau de chaque investissement.
— Le retour sur investissement est donné comme suit $\sum_{j \in A} mean_j * alloc_j$.
— Le risque encouru est donné par $\sum_{i \in T}(\sum_{j \in A} \tilde{R}_{ij} * alloc_j)^2 / card(T)$; où $\tilde{R}_{ij} = R_{ij} - mean_j$.
— On voit que la fonction objectif du problème est linéaire, alors que les contraintes sont quadratiques.

Pour enfin compléter le modèle, soit le jeu de données suivant :

```
set A := US_3-MONTH_T-BILLS US_GOVN_LONG_BONDS SP_500 WILSHIRE_5000;

param R:
US_3-MONTH_T-BILLS US_GOVN_LONG_BONDS SP_500 WILSHIRE_5000 :=
1984 1.103 1.159 1.061 1.030
1985 1.080 1.366 1.316 1.326
1986 1.063 1.309 1.186 1.161
1987 1.061 0.925 1.052 1.023
1988 1.071 1.086 1.165 1.179
1989 1.087 1.212 1.316 1.292
1990 1.080 1.054 0.968 0.938
1991 1.057 1.193 1.304 1.342
1992 1.036 1.079 1.076 1.090
1993 1.031 1.217 1.100 1.113
1994 1.045 0.889 1.012 0.999 ;
```

Comme nous l'avons vu dans le cours, ces problèmes sont difficiles à résoudre. La plupart des solveurs fournissent des solution locales, c'est-à-dire des solutions qui sont optimales localement à un voisinage. Justement, `Minos` fournit des solutions locales.

Il existe des solveurs qui trouvent des solutions globales, telles que `Baron`, `GlobSol`, etc. mais qui ne sont pas encore interfacés avec l'environnement `AMPL`.

3.9 Exercices compléments

3.9.1 Exercice

Etant donné deux variables x_1, x_2 binaires (i.e., $x_1 \in \{0,1\}$, $x_2 \in \{0,1\}$). Soit l'expression non linéaire de leur multiplication :

$$(P) \begin{cases} y = x_1 \times x_2 \\ x_1 \in \{0,1\}, x_2 \in \{0,1\}, y \in \{0,1\} \end{cases} \tag{3.11}$$

Nous proposons un système d'équations linéaires (Q)

$$(Q) \begin{cases} y \leq x_1 \\ y \leq x_2 \\ y \geq x_1 + x_2 - 1 \\ x_1 \in \{0,1\}, x_2 \in \{0,1\}, y \in \{0,1\} \end{cases} \tag{3.12}$$

Démontrer que (P) est équivalent à (Q) ?

3.9.2 Exercice

Soit le problème

$$(P) \begin{cases} \min \sum_{i=1..n} c_i x_i \\ \\ (\sum_{j=1..n} A_j x_j \leq b) \text{ OU } (\sum_{j=1..n} C_j x_j \leq d) \\ x_i \geq 0, i = 1..n \end{cases} \tag{3.13}$$

Le problème (P) n'est pas un PL car il contient la contrainte de disjonction "OU". Nous proposons un autre problème (Q)

$$(Q) \begin{cases} \min \sum_{i=1..n} c_i x_i \\ \sum_{j=1..n} A_j x_j \leq b + My \\ \sum_{j=1..n} C_j x_j \leq d + M(1-y) \\ x_i \geq 0, i = 1..n \\ y \in \{0,1\} \end{cases} \tag{3.14}$$

où M est un nombre très grand.

Démontrer que (P) est équivalent à (Q) ?

3.9.3 Exercice

Une firme *Penault* produit deux types de carrosseries C_1 et C_2. Nous nous intéressons à deux tâches indépendantes :

- *Peinture* : Cette firme ne pourra pas peindre C_2 à partir d'un maximum de 40 carrosseries C_1. Cette firme ne pourra pas aussi peindre C_1 à partir d'un maximum de 60 carrosseries C_2.
- *Fabrication* : Cette firme ne pourra pas produire C_2 à partir d'un maximum de 50 carrosseries C_1. Cette firme ne pourra pas aussi fabriquer C_1 à partir d'un maximum de 50 carrosseries C_2.

La carrosserie C_1 rapporte 300 DA et C_2 200 DA.

1. Modéliser sous forme d'un programme linéaire PL la recherche des quantités optimales x_1 de C_1 et x_2 de C_2 à produire pour maximiser le gain ? (N.B. Dans une tâche T donnée dans la firme, si la firme ne pourra pas produire C_2 à partir de k_1 C_1, et la firme ne pourra pas aussi produire C_1 à partir de k_2 C_2, la contrainte T peut être traduite en une seule contrainte linéaire $(x_1/k_1) + (x_2/k_2) \leq 1$.)
2. Tracer graphiquement le polyèdre décrivant le PL et extraire la solution optimale sans passer par l'algorithme du Simplexe ?
3. Soit la contrainte supplémentaire suivante : la firme *Penault* doit produire au moins 30 carrosseries C_1 et au moins 20 carrosseries C_2. Comment devient-il le PL ?

3.9.4 Exercice

Soit le programme en nombres entiers :

$$(P) \begin{cases} \min & \sum_{j=1..n} c_j x_j \\ \text{s.c.} & \sum_{j=1..n} A_{i,j} x_j \leq V_i, i = 1..m \\ & x_j \in \{0,1\}, j = 1..n \end{cases} \tag{3.15}$$

— Définir ce qu'est une relaxation du problème (P) ?
— Proposer une méthode utilisant la méthode du Simplexe qui permet la relaxation du problème (P) ?

3.9.5 Exercice

Soit le programme en nombres entiers relatif au problème du sac-à-dos :

$$(P) \begin{cases} \max & \sum_{j=1..n} c_j x_j \\ \text{s.c.} & \sum_{j=1..n} p_j x_j \leq V \\ & x_j \in \{0,1\}, j = 1..n \end{cases} \tag{3.16}$$

— Définir ce qu'est une relaxation du problème (P) ?
— Proposer une heuristique rapide de résolution approchée de (P) sans faire appel à la méthode du Simplexe ?

3.9.6 Exercice

Soit le programme en nombres entiers :

$$(P) \begin{cases} \min & \sum_{j=1..n} c_j x_j \\ \text{s.c.} & \sum_{j=1..n} A_{i,j} x_j \leq V_i, i = 1..m \\ & x_j \in \{0,1\}, j = 1..n \end{cases} \tag{3.17}$$

— Définir le principe d'évaluation (bounding) dans l'algorithme de branch& bound résolvant (P) ?

— Définir le principe de séparation (branching) dans l'algorithme de branch& bound résolvant (P) ?
— Comment sont combinés ces deux principes (i.e., séparation et évaluation) afin de résoudre (P) ?

3.9.7 Exercice

Soit le programme linéaire suivant :

$$\begin{array}{ll} \min & 2x_1 + 3x_2 \\ \text{s.c.} & \frac{1}{2}x_1 + x_2 \geq 1 \\ & \frac{2}{3}x_1 - x_2 \geq -2 \\ & x_1, x_2 \geq 0 \end{array} \tag{3.18}$$

1. Dessiner le polyèdre qui définit la forme géométrique des inéquations du problème (i.e. l'espace faisable) ?
2. Donner l'ensemble des sommets du polyèdre ?
3. Donner le sommet qui correspond à la solution optimale du problème ; tout en justifiant son optimalité ?

Chapitre 4

Optimisation sans contraintes

Dans cette section, nous nous intéressons au problème suivant

$$min_{x \in \mathbb{R}^n} f(x) \tag{4.1}$$

où $x \in \mathbb{R}^n$ est un vecteur réel avec $n \geq 1$ et $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction continue et continûment dérivable.

Nous supposons que les premières et secondes dérivées de $f(x_1, x_2, ..., x_n)$ existent et sont continues sur tous les points. Soit $\frac{\partial f}{\partial \tilde{x}_i}$ la dérivée partielle de $f(x_1, x_2, ..., x_n)$ relativement à x_i, évaluée au point $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$.

Une condition nécessaire pour que $\tilde{x} = \tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}$ soit un minimum local de (7.2) est donnée par le théorème suivant :

Théorème 4. *Si $\tilde{x}$ est un minimum local de (7.2), alors $\frac{\partial f(\tilde{x})}{\partial x_i} = 0$.*

La preuve de ce théorème vient du fait que si $\tilde{x}$ est un minimum local alors il a un voisinnage où la fonction objectif prend son min en ce point. Supposons que le point en question n'annule pas la dérivée de la fonction objectif, alors nécessairement la fonction objectif est monotone en ce point faisant ainsi diminuer la fonction objectif ; ce qui contredit que le point est optimal dans son voisinnage.

Définition 10. *Un point $\tilde{x}$ ayant $\frac{\partial f}{\partial x_i} = 0$ pour tout $i = 1, 2, ..., n$ est dit un point stationnaire de f.*

Définition 11.

- *Le i-ème principal mineur (principal minor) d'une matrice $n \times n$ est le déterminant de la matrice $i \times i$ obtenue en supprimant $n - i$ lignes et les $n - i$ correspondantes colonnes de la matrice.*
- *Le k-ème principal mineur dominant (leading principal minor) d'une matrice $n \times n$ est le déterminant de la matrice $k \times k$ obtenue en supprimant les dernières $n - k$ lignes et colonnes de la matrice. On note $H_k(x_1, ..., x_n)$ le kième principal mineur dominant au point $(x_1, ..., x_n)$.*

Par exemple, soit $f(x_1, x_2) = x_1^2 + 2x_1x_2 + x_2^2$, alors $H_1(x_1, x_2) = 6x_1$ et $H_2(x_1, x_2) = 12x_1 - 4$.

Remarque 1 (compléments sur la convexité et la concavité)**.** *Supposons que $f(x_1, x_2, ..., x_n)$ a les dérivées secondes continues.*

- *$f(x_1, x_2, ..., x_n)$ est une fonction convexe dans S si et seulement si pour tout $x \in S$, tous les principaux mineurs de H sont non-négatifs.*
- *$f(x_1, x_2, ..., x_n)$ est une fonction concave dans S si et seulement si pour tout $x \in S$ et $k = 1, 2, ..., n$, tous les principaux mineurs non-nuls de H ont le même signe que -1^k.*

(exemples : page 655 du Winston)

Le théorème suivant montre quelques conditions suffisantes pour qu'un point soit un extrémum local à partir de la matrice hessienne.

Théorème 5.
- *Si $H_k(\tilde{x}) > 0, k = 1..n$, alors le point $\tilde{x}$ est un minimum local de (7.2).*
- *Si $H_k(\tilde{x}), k = 1..n$, est non-nul et a le même signe que -1^k, alors le point $\tilde{x}$ est un maximum local de (7.2).*
- *Si $H_k(\tilde{x}) \neq 0, k = 1..n$, et les deux conditions précédentes ne sont pas vérifiées, alors le point $\tilde{x}$ n'est pas un extrémum local de (7.2).*

(exemples page 670 du Winston)

Si un point stationnaire n'est pas un minimum local, il est dit un point d'inflexion (saddle point).

Nous pouvons aussi exploiter les propriétés de convexité et de concavité de la fonction objectif pour détecter des minimums globaux (voir la section 2.2.2).

Remarque 2 (rappel sur l'optimum d'un POC convexe)**.** *Supposons que l'espace faisable s d'un POC est convexe. Si $f(x)$ est convexe dans s, alors tout minimum local du POC est un minimum global.*

Exemple 16 ([?])**.** *Nous avons n points mesurant une grandeur physique m fois dans le temps $t_1, ..., t_m$. A partir de considérations théoriques nous savons que la fonction de cette grandeur physique est*

$$\Phi(t, x) = x_1 + x_2 e^{(x_3 - t)^2 / x_4} + x_5 cos(x_6 t).$$

$x_1 ... x_6$ sont des paramètres. Soit $r_j(x) = y_j - \phi(t_j, x)$ mesurant l'écart entre la mesure expérimentale et la valeur théorique. Trouver les $x_1, ..., x_6$ tels que

$$min_x f(x) = r_1^2(x) + ... + r_m^2(x).$$

Nous allons tout d'abord commencer par voir les méthodes locales de descente, puis nous allons aborder les méthodes globales par intervalles.

Les méthodes les plus efficaces d'optimisation continue sans contraintes utilisent le gradient de la fonction objectif. La méthode la plus simple est celle de la plus grande descente et les plus raffinées et performantes sont celles de quasi-Newton et Newton (secant methods).

Toutes ces méthodes sont basées sur la stratégie de recherche linéaire que nous introduisons ci-dessous.

FIGURE 4.1 – Recherche linéaire

$\lambda_k := 1$;
while $(f(x_k + \lambda_k/2d_k) < f(x_k + \lambda d_k))$ **do**
$\lambda_k := \lambda_k/2$;
end while

FIGURE 4.2 – Schéma des algorithmes de stratégie linéaire

Initialisation de x_0 et d_0 ;
k := 0 ;
do
Déterminer λ_k tel que $minf(x_k + \lambda_k d_k)$
Calculer le nouveau point $x_{k+1} := x_k + \lambda_k d_k$
Calculer la nouvelle direction d_{k+1} ;
while $(|x_{k+1} - x_k| < \epsilon)$

4.1 Recherche linéaire

L'algorithme choisi une direction d_k et cherche au long de cette direction à partir de l'itéré courant x_k un nouveau itéréx_{k+1} minimisant la fonction objectif. La structure générale de cet algorithme est donnée dans la Figure 4.1.

Nous voulons résoudre $minf(x_k + \lambda_k d_k)$ avec $\lambda_k \in [0, 1]$.

Une solution approchée de ce problème peut être obtenue avec la procédure donnée dans la Figure 4.1.

Cette recherche est de convergence linéaire. Elle peut être raffinée en construisant un modèle quadratique de $f(x_k + \lambda_k d_k)$ en fonction de λ_k. Ce modèle quadratique est de la forme

$$P(\lambda) = f_1 + h_1\lambda + h_3\lambda(\lambda - \lambda_2)$$

qui interpole la fonction au 3 points $\alpha = 0, \alpha = \alpha_2, \alpha = \alpha_3$; où

$$\begin{aligned} f_i &= f(x_k + \lambda_i d_k) \\ h_1 &= (f_2 - f_1)\lambda_2 \\ h_2 &= (f_3 - f_2)/(\lambda_3 - \lambda_2) \\ \lambda_2 &= \lambda_3/2 \\ h_3 &= (h_2 - h_1)/\lambda_3 \end{aligned}$$

On sait que la solution de ce modèle est $\lambda_3 = 0.5(\lambda_2 - h_1/h_3)$. Cette dernière solution peut donner une meilleure solution que que le choix naïf $\lambda/2$. Quand cette méthode est utilisée dans un quasi-Newton, on doit prendre en considération dans le choix de λ le fait que H doit rester définie positive.

4.2 Méthode de la plus grande descente

Ce sont les méthodes les plus simples. La recherche de la direction est faite suivant la décroissance de f qui est dans l'opposé de $\Delta f(x)$

$$x_{k+1} := x_k - \lambda_k \Delta f(x_k).$$

où $\Delta f(x) = [\frac{\partial f(x)}{\partial x_1}, ..., \frac{\partial f(x)}{\partial x_n}]$ est le vecteur gradient de $f(x)$.

On peut se poser la question : quel est la grandeur d'influence de λ_k ?

$$\frac{\partial f(x_{k+1})}{\partial \lambda_k} = \frac{\partial f(x_k - \lambda_k \Delta(f(x_k)))}{\partial \lambda_k} = 0$$
$$\Rightarrow -\Delta f(x_{k+1}) \partial f(x_k) = 0$$

λ_k est donc choisie telle que $\Delta f(x_{k+1})$ et $\Delta f(x_k)$ soient orthogonales.

Cette méthode est de convergence linéaire [**?**] ; mais elle a de bonne propriétés de convergence globale.

4.3 Méthode de quasi-Newton

La méthode la plus efficace et celle des méthodes de quasi-Newton. Ces méthodes construisent une approximation quadratique :

$$min_x \frac{1}{2} x^t H x + C^T x + b$$

où la matrice H est semi-définie positive.

Rappel 1 (matrice semi-définie positive (SDP)). *Une matrice $A \in I\!R^{n \times n}$ est dite semi-définie postive si*

$$\forall x \in I\!R^n \Rightarrow x^T A x \geq 0.$$

Elle est dite définie positive si

$$\forall x \in I\!R^n \Rightarrow x^T A x > 0.$$

Une condition nécessaire de l'optimalité d'une solution x^* de cette approximation quadratique est que les dérivées partielles soient toutes égales à zéro

$$\Delta f(x^*) = H x^* + C = 0.$$

Les méthodes de type Newton calculent H directement, ce qui est très couteux si le nombre de variables est important.

Plusieurs méthodes ont ainsi été développées pour approximer directement l'inverse de H avec seulement les valeurs du gradient. Deux formules ont été proposées, celle de DFP (Davidon, Fletcher et Powell) et BFGS (Broyden, Fletcher, Goldfard et Shanno).

La formule de BFGS est

$$H_{k+1} := H_k + \frac{q_k q_k^T}{q_k s_k} + \frac{H_k^T S_k^T S_k H_k}{S_k^T H_k s_k}$$

où $s_k := x_{k+1} - x_k$ et $q_k := \Delta f(x_{k+1}) - \Delta f(x_k)$.

DFP est similaire à BFGS en interchangant q_k et s_k. H_0 est fixée à n'importe quelle matrice symétrique semi-définie positive, par exemple la matrice identité. BFGS a les mêmes propriétés théoriques que DFP. Mais en pratique $BFGS$ est préférée à DFP car elle est plus stable numériquement (moins sensible aux erreurs d'arrondi).

L'algorithme général de quasi-Newton est celui de la stratégie linéaire où :

$$d_{k+1} := \ - H_{k+1}\Delta f(x_k)$$

Il a été prouvé que quasi-Newton est de convergence superlinéaire ; mais elle est moins bonne au niveau convergence globale que les méthodes de descente du gradient.

4.4 Moindres carrés

La stratégie de recherche linéaire de quasi-Newton peut être utilisée pour résoudre le problème d'optimisation particulier des moindres carrés

$$LS \quad minf(x) = \frac{1}{2}\|F(x)\| = \frac{1}{2}\sum F_i(x)^2.$$

Ce problème apparaît souvent dans plein d'applications pratiques particulières dans l'interpolation à partir de données expérimentales. Ce problèmes peut être résolu simplement avec un quasi-Newton. Les propriétés mathématiques de ce problème permettent une meilleure résolution.

La méthode qui tire profit de ces propriétés et qui est la plus utilisée sur les moindres carrés est celle de Levenberg-Marquardt (LM).

LM prend comme direction de recherche la solution d_k de l'équation

$$[J(x_k)^T J(x_k) + \lambda_k I]d_k = -J(x_k)F(x_k).$$

La solution recherche de (LS) est atteinte quand $F_i(x) = O, i = 1..m$.

La formule de Taylor permet de le ramener à :

$$F_i(x_k) + J(x_k)x = 0$$
$$\Rightarrow J_i(x_k).x = -F_i(x_k)$$

Puisque le système n'est pas carré, en le rend carré en le multipliant par la transposée

$$\Rightarrow J_i(x_k)^T J_i(x_k)x = -J_i(x_k)^T F_i(x_k).$$

Résoudre ce dernier système permet d'avoir la direction de quasi-Newton car $J_i(x_k)^T J_i(x_k)$ est le hessien de $\sum F_i(x)^2$.

En d'autres termes, plus λ_k augmente, LM prend comme direction de recherche la solution d_k de l'équation

$$[J(x_k)^T J(x_k) + \lambda_k I]d_k = -J(x_k)F(x_k).$$

plus on se rapproche de la méthode de descente et ainsi obtenir une meilleure convergence globale.

Donc quand on n'a pas la convergence de quasi-Newton, on augmente λ_k pour converger avec la méthode de la plus grande descente.

4.5 Optimisation sans fonction objectif : résolution locale des équations nonlinéaires

4.6 Algorithme de Newton

Les méthodes locales s'intéressent au problème suivant :

Définition 12 (local zero (LZ)). *Soit $f : I\!R^n \to I\!R^n$, trouver x^* une approximation locale du zéro de la fonction f ($y \in I\!R^n$ est un zéro de f si $f(y) = 0$).*

La méthode de Newton pour résoudre des systèmes d'équations consiste à appliquer le procédé itératif suivant en partant d'un point initial x^0 :

$$x^k = x^{k-1} - J(x^{k-1})^{-1} f(x^{k-1})$$

où $J(x)$ est la matrice jacobienne $n \times n$: $[\frac{\partial f_i(x)}{\partial x_j}, i = 1..m, j = 1..n]$.

Cette méthode est de convergence locale quadratique ; mais elle n'a pas de bonnes propriétés de convergence globale. C'est pour cette raison que l'on transforme le système d'équations en un problème d'optimisation pour bénéficier de ses méthodes de convergence globale. C'est l'objet de la sous-section suivante.

4.7 Transformation en un problèmes d'optimisation

Le système d'équations non-linéaires $f(x) = 0$ est transformé en un problème d'optimisation aux moindres carrés

$$min_x \sum_{i=1..m} f_i(x_1, ..., x_n)^2$$

Ainsi, on pourrait appliquer aisément les méthodes des moindres carrés du chapitre précédent. En raison de la particularité de ce problème, une méthode dédiée a été développée, l'algorithme de Levenberg-Marquardt-Method (LMM), pour exploiter ses propriétés.

LMM utilise la direction de recherche d_k qui est la solution du système linéaire :

$$(J(x_k)^T J(x) + \lambda_k I) d_k = -J(x_k) F(x_k)$$

où $J(x)$ est la matrice jacobienne$m \times n$: $[\frac{\partial f_i(x)}{\partial x_j}, i = 1..m, j = 1..n]$. Le scalaire λ_k contrôle la magnitude et la direction de d_k.

La valeur de λ_k est choisie pour garantir la convergence de l'ensemble du procédé. Quand λ_k est égale à zéro, l'algorithme se comporte comme un quasi-newton et

bénéficie ainsi de sa bonne convergence locale. Plus λ_k est grand, plus l'algorithme se comporte comme l'algorithme de la plus grande descente et bénéficie ainsi de sa convergence globale. λ_k est initialisée à zéro, et si l'algorithme ne converge pas, λ_k est augmenté progressivement pour pouvoir ainsi converger globalement ; par la suite λ_k est diminuer pour converger rapidement une fois que les itérations précédentes ont permis de se rapprocher suffisamment de la solution. Une fois λ_k fixée, le vecteur d_k est obtenu en résolvant le système linéaire ci-dessus. Finalement, une stratégie de recherche linéaire (line-search strategy) est utilisée pour fixer α_k pour que $x_{k+1} = x_k + \alpha_k d_k$ fasse décroître la fonction objectif (la fonction aux moindres carrés).

4.8 Optimisation avec des contraintes et une fonction objectif : le cas général

Un problème général d'optimisation continue (POC) s'exprime comme suit : trouver des valeurs des variables de décision $x_1, ..., x_n$ telles que :

$$\begin{array}{l} minz = f(x_1, ..., x_n) \\ g_i(x_1, ..., x_n) = b_i, i = 1..m_e \\ g_j(x_1, ..., x_n) \leq b_j, j = m_e + 1..m. \end{array} \tag{4.2}$$

4.9 Les multiplicateurs de Lagrange

Les multiplicateurs de Lagrange peuvent être utilisés quand les contraintes sont des égalités. Nous considérons donc le problème d'optimisation (POC) suivant :

$$\begin{array}{l} minz = f(x_1, ..., x_n) \\ c_i : g_i(x_1, ..., x_n) = b_i, i = 1..m_e \end{array} \tag{4.3}$$

Pour résoudre (7.3), on associe à chaque contrainte c_i un multiplicateur λ_i, et on forme ainsi le Lagrangien :

$$L(x_1, x_2, ..., x_n, \lambda_1, \lambda_2, ..., \lambda_m) = f(x_1, x_2, ..., x_n) + \sum_{i=1}^{m} \lambda_i [b_i - g_i(x_1, x_2, ..., x_n)] \tag{4.4}$$

Nous cherchons à trouver un point $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, , \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ qui minimise $L(x_1, x_2, ..., x_n, \lambda_1, \lambda_2, ..., \lambda_m)$. Souvent $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, , \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ résout aussi (7.3). Si $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, , \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ minimise L, alors

$$\frac{\partial L}{\partial \lambda_i} = b_i - g_i(x_1, x_2, ..., x_n) = 0$$

Ici $\frac{\partial L}{\partial \lambda_i}$ est la dérivée partielle de L par rapport à λ_i. Ceci montre bien que le point en question satisfait d'une façon optimale(7.3). Pour montrer que $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ résout (7.3), soit un point quelconque de l'espace faisable de (7.3). Puisque $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, \tilde{\lambda_2}, ..., \tilde{\lambda_m}$ minimise L, alors pour tout $\lambda'_i, i = 1..m$, nous avons :

$$L(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, , \tilde{\lambda_2}, ..., \tilde{\lambda_m}) \leq L(x'_1, x'_2, ..., x'_n, \lambda'_1, , \lambda'_2, ..., \lambda'_m)$$

Puisque $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ et $(x'_1, x'_2, ..., x'_n)$ sont tous les deux faisables, alors tous les facteurs des multiplicateurs sont nuls. On obtiens ainsi $f(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}) \leq f(x'_1, x'_2, ..., x'_n)$. Ceci montre que $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ est bien la solution de (7.3). En bref, si $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ résout le problème d'optimisation sans contraintes

$$minL(x_1, x_2, ..., x_n, \lambda_1, \lambda_2, ..., \lambda_m) \tag{4.5}$$

alors $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ résout aussi (7.3).

A partir du chapitre 7.1, nous savons qu'une condition nécessaire pour que $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ résolve (7.5) est qu'au point $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ on doit avoir

$$\frac{\partial L}{\partial x_1} = \frac{\partial L}{\partial x_2} = ... = \frac{\partial L}{\partial x_n} = \frac{\partial L}{\partial \lambda_1} = \frac{\partial L}{\partial \lambda_2} = ... = \frac{\partial L}{\partial \lambda_n} = 0 \tag{4.6}$$

Le théorème suivant donne des conditions suffisantes pour qu'un tel point soit un minimum global.

Théorème 6. *Si $f(x_1, x_2, ..., x_n)$ est une fonction convexe et chaque $g_i(x_1, ..., x_n)$ est une fonction linéaire, alors tout point $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ satisfaisant (7.6) est une solution optimale.*

Les variables λ_i ont une interprétation intéressante comme étant les coûts marginaux (shadow prices) associés à chacune des des contraintes. Si la partie gauche d'une contrainte $g_i(x_1, x_2, ..., x_n) = b$ croit de δ_i, alors la valeur de la fonction objectif croit de $\delta_i \lambda_i$.

4.10 Les conditions de Kuhn-Tucker

Nous discutons dans cette section des conditions nécessaires et suffisantes pour que $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ soit une solution optimale d'un POC avec des contraintes d'inégalité :

$$\begin{array}{l} minz = f(x_1, ..., x_n) \\ c_i : g_i(x_1, ..., x_n) \leq b_i, i = 1..m \end{array} \tag{4.7}$$

Pour que les résultats de cette section soient applicable, il qu'on ait seulement des contraintes d'inégalité inférieure. Les contraintes d'égalité ou d'inégalité supérieure peuvent être aisément ramenées en contraintes d'inégalité inférieure.

Le théorème suivant donne des conditions nécessaires pour qu'un point $\tilde{x} = (\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ soit optimal pour le POC (7.7).

Pour que ce théorème soit applicable il faut que les fonctions g_i satisfassent des conditions de régularité (voir [**?**]). Quand les contraintes sont linéaires, ces conditions sont toujours satisfaites. Nous supposons dans la suite que les contraintes satisfassent toujours les conditions de régularité.

Théorème 7.

Comme au niveau des multiplicateurs de Lagrange de la section précédente, les multiplicateurs λ_i associés au conditions KT correspondent aux coûts marginaux des contraintes.

4.11 SQP : Sequential Quadratique Programming

4.12 TP

Chapitre 5

Résolution des équations non-linéaires

5.1 Méthodes locales

5.1.1 Algorithme de Newton

Les méthodes locales s'intéressent au problème suivant :

Définition 13 (local zero (LZ)). *Soit $f : \mathbb{R}^n \to \mathbb{R}^n$, trouver x^* une approximation locale du zéro de la fonction f ($y \in \mathbb{R}^n$ est un zéro de f si $f(y) = 0$).*

La méthode de Newton pour résoudre des systèmes d'équations consiste à appliquer le procédé itératif suivant en partant d'un point initial x^0 :

$$x^k = x^{k-1} - J(x^{k-1})^{-1} f(x^{k-1})$$

où $J(x)$ est la matrice jacobienne $n \times n$: $[\frac{\partial f_i(x)}{\partial x_j}, i = 1..m, j = 1..n]$.

Cette méthode est de convergence locale quadratique ; mais elle n'a pas de bonnes propriétés de convergence globale. C'est pour cette raison que l'on transforme le système d'équations en un problème d'optimisation pour bénéficier de ses méthodes de convergence globale. C'est l'objet de la sous-section suivante.

5.1.2 Transformation en un problèmes d'optimisation

Le système d'équations non-linéaires $f(x) = 0$ est transformé en un problème d'optimisation aux moindres carrés

$$min_x \sum_{i=1..m} f_i(x_1, ..., x_n)^2$$

Ainsi, on pourrait appliquer aisément les méthodes des moindres carrés du chapitre précédent. En raison de la particularité de ce problème, une méthode dédiée a été développée, l'algorithme de Levenberg-Marquardt-Method (LMM), pour exploiter ses propriétés.

LMM utilise la direction de recherche d_k qui est la solution du système linéaire :

$$(J(x_k)^T J(x) + \lambda_k I)d_k = -J(x_k)F(x_k)$$

où $J(x)$ est la matrice jacobienne$m \times n$: $[\frac{\partial f_i(x)}{\partial x_j}, i = 1..m, j = 1..n]$. Le scalaire λ_k contrôle la magnitude et la direction de d_k.

La valeur de λ_k est choisie pour garantir la convergence de l'ensemble du procédé. Quand λ_k est égale à zéro, l'algorithme se comporte comme un quasi-newton et bénéficie ainsi de sa bonne convergence locale. Plus λ_k est grand, plus l'algorithme se comporte comme l'algorithme de la plus grande descente et bénéficie ainsi de sa convergence globale. λ_k est initialisée à zéro, et si l'algorithme ne converge pas, λ_k est augmenté progressivement pour pouvoir ainsi converger globalement ; par la suite λ_k est diminuer pour converger rapidement une fois que les itérations précédentes ont permis de se rapprocher suffisamment de la solution. Une fois λ_k fixée, le vecteur d_k est obtenu en résolvant le système linéaire ci-dessus. Finalement, une stratégie de recherche linéaire (line-search strategy) est utilisée pour fixer α_k pour que $x_{k+1} = x_k + \alpha_k d_k$ fasse décroître la fonction objectif (la fonction aux moindres carrés).

5.2 Méthodes globales : analyse par intervalles et satisfaction de contraintes [?, ?, ?]

Dans cette section, nous nous intéressons à la résolution globale des systèmes d'équations numériques continues quelconques. La résolution globale vise à trouver tous les zéro du système d'équations dans un domaine des variables défini par des intervalles. Nous adoptons la notation proposée dans la communauté de satisfaction de contraintes continues NCSP (Numeric Constraint Satisfaction Problems) [**?**].

Un NCSP est un triplet $< X, D, C >$ où :

— $\mathcal{X}$ est un ensemble de n variables $x_1, \ldots, x_n$

— $\mathcal{D} = \langle D_1, \ldots, D_n \rangle$ dénote un vecteur de domaines. La i^{ieme} composante de $\mathcal{D}$, D_i, est l'intervalle du domaine contenant toutes les valeurs possibles de x_i.

— $\mathcal{C} = \{C_1, \ldots C_m\}$ dénote un ensemble de contraintes numériques. $var(C_j)$ dénote l'ensemble des variables apparaissant dans C_j.

Nous allons nous focaliser sur les problèmes où le domaine est défini par intervalles : $\mathcal{D} \in I\!I(I\!R)^n$ où $I\!I(I\!R) = \{[a, b] | a, b \in I\!R \cup \{-\infty, +\infty\}\}$.

Un intervalle $[a, b]$ tel que $a > b$ est un intervalle vide. Un vecteur de domaines $\mathcal{D}$ où une composante D_i est vide est aussi vide et est dénoté $\emptyset$. La borne inférieure, la borne supérieure et le point milieu d'un intervalle D_i (resp. vecteur intervalles $\mathcal{D}$) sont respectivement dénotés par $\underline{D_i}$, $\overline{D_i}$, et $m(D_i)$ (resp. $\underline{\mathcal{D}}$, $\overline{\mathcal{D}}$, et $m(\mathcal{D})$).

La borne inférieure, la borne supérieure, le point milieu, la relation d'inclusion, l'opérateur d'union et d'intersection sont aussi définis sur les vecteurs. Ainsi, $\underline{\mathcal{D}}$ dénote $\langle \underline{D_1}, \ldots, \underline{D_n} \rangle$; $\mathcal{D}' \subset \mathcal{D}$ dénote $D'_i \subset D_i$ pour tout $i \in 1 \ldots n$; $\mathcal{D}' \cap \mathcal{D}''$ dénote $\langle D'_1 \cap D''_1, \ldots, D'_n \cap D''_n \rangle$.

Une contrainte d'arité k, $C_j(x_{j_1}, \ldots, x_{j_k})$, dénote conceptuellement une relation k-aire sur les nombres réels, c'est-à-dire, un sous-ensemble de $I\!R^k$.

Exemple 3. *Les zéros du problème*

$$\begin{cases} x^2 + y^2 - 2 = 0; \\ x^2 - y = 0; \\ x, y \in [-10^8, +10^8]; \end{cases}$$

correspondent au calcul des points d'intersection entre un cercle et une parabole (voir la figure 5.1).

Exemple 4. *Les zéros du problème*

$$\begin{cases} (x+1)*(x+2)*(x+3)*(x+4)*(x+5)*(x+6)*(x+7)*(x+8)* \\ \quad (x+9)*(x+10)*(x+11)*(x+12)*(x+13)*(x+14)*(x+15)* \\ \quad (x+16)*(x+17)*(x+18)*(x+19)*(x+20)+ \\ \qquad 0.11920928955078125e-6*x^{19} = 0; \\ x \in [-30, 0]; \end{cases}$$

correspondent au calcul des points d'intersection de la courbe de la figure 5.2 avec l'axe des abscisses.

FIGURE 5.1 – NCSP : intersection d'une parabole et d'un cercle.

FIGURE 5.2 – NCSP : intersection d'une courbe avec l'axe des abscisses.

Les méthodes globales de résolution des systèmes d'équations non-linéaires exploitent quatre concepts pour pouvoir trouver toutes les solutions :

— Le premier concept générique est la notion de contrainte : une relation que doit satisfaire un ensemble de variables. Ici, on exploitera le plus possible leurs différentes propriétés mathématiques et leurs différentes formes possibles pour mettre en oeuvre efficacement les trois concepts qui suivent.

— Le deuxième concept est celui de la projection d'une contrainte sur une variable. Projeter une contrainte sur une variable consiste à résoudre cette contrainte localement à cette variable. Le procédé de résolution est issu des développements mathématiques autour de cette contrainte. Par exemple, en analyse numérique par intervalles, plusieurs procédés existent pour manipuler les équations non-linéaires.

— Le troisième concept est celui de la propagation ou du filtrage qui permet de réduire l'espace de recherche du problème. Le filtrage consiste à appliquer itérativement les fonctions de projection de toutes les contraintes sur toutes les variables pour parvenir à une réduction conséquente des domaines des variables. Dans certains cas, ce filtrage peut être suffisant pour résoudre une classe de problèmes.

— Le dernier concept est celui de la décomposition qui consiste à diviser un problème, dont la résolution est inachevée par un filtrage, en plusieurs sous-problèmes indépendants. Ces sous-problèmes peuvent être à leur tour divisés, jusqu'à arriver à des problèmes solubles par simple filtrage.

Dans les sections suivantes, nous allons développer progressivement ces concepts.

5.2.1 Approximation des fonctions de projection

Les algorithmes de résolution globale des NCSPs se basent sur un processus de filtrage des domaines et ont besoin de calculer la projection — dénotée $\Pi_{C_j,x_i}(\mathcal{D})$ ou aussi $\Pi_{j,i}(\mathcal{D})$ — de la contrainte $C_j(x_{j_1},\ldots,x_{j_k})$ sur les variables x_i dans l'espace délimité par $D_{j_1}\times\ldots\times D_{j_k}$. La projection $\Pi_{j,i}(\mathcal{D})$ est définie comme suit.

— Si $x_i \notin var(C_j)$, $\Pi_{j,i}(\mathcal{D}) = D_i$.

— Si $x_i \in var(C_j)$ la projection est définie par l'ensemble de tous les éléments $d_i \in D_i$ tels qu'on trouverait des éléments $d_{j_1},\ldots,d_{i-1},d_{i+1},\ldots,d_{j_k}$ pour les $k-1$ variables restantes de $var(C_j)$ avec $\langle d_{j_1},\ldots,d_{i-1},d_i,d_{i+1}\ldots,d_{j_k}\rangle \in C_j$. Formellement :

$$\begin{aligned}\Pi_{j,i}(\mathcal{D}) = \{d_i | d_i \in D_i, \exists d_{j_1},\ldots,d_{i-1},d_{i+1},\ldots,d_{j_k} \\ (d_{j_1}\in D_{j_1},\ldots,d_{i-1}\in D_{i-1},d_{i+1}\in D_{i+1},\ldots,d_{j_k}\in D_{j_k}, \\ \langle d_{j_1},\ldots,d_i,\ldots,d_{j_k}\rangle \in C_j)\}.\end{aligned} \tag{5.1}$$

Pratiquement, cette projection ne peut être calculée exactement pour plusieurs raisons : (1) les nombres dans une machine sont des nombres flottants et non pas des nombres réels, les erreurs d'arrondi sont accumulées ; (2) la projection ne peut être représentée sous forme d'un nombre flottant ; (3) le calcul nécessaire pour obtenir une approximation fine de la projection est très coûteuse ; (4) la projection peut être discontinue alors qu'il est plus aisé de manipuler des intervalles clos au niveau des domaines des variables.

C'est pourquoi, la projection est approximée. Soit $\pi_{C_j,x_i}(\mathcal{D})$ ou aussi $\pi_{j,i}(\mathcal{D})$ une telle approximation. Pour garantir que toutes les solutions du CSP numérique sont préservées, l'algorithme de résolution qui utilise $\pi_{j,i}(\mathcal{D})$ nécessite que $\pi_{j,i}(\mathcal{D})$ contienne la projection exacte. Nous supposons aussi que l'opérateur $\pi_{j,i}(\mathcal{D})$ satisfait la propriété de contractance. Nous avons donc :

$$\Pi_{j,i}(\mathcal{D}) \subseteq \pi_{j,i}(\mathcal{D}) \subseteq \mathcal{D}.$$

$\pi_{j,i}(\mathcal{D})$ englobe et remédie à tous les problèmes cités ci-haut avec son mécanisme de sur-approximation. En particulier, il nous évite de détailler les dépendances entre les nombres réels et les nombres flottants qui ne font pas l'objet de cette section. Nous renvoyons le lecteur intéressé à la section détaillant les nombres dans le chapitre des préliminaires. Nous détaillons dans la suite la construction de $\pi_{j,i}$. L'analyse par intervalles [?] servira comme un outil de base pour construire ces fonctions.

Pour calculer la projection $\pi_{j,i}(\mathcal{D})$ de la contrainte C_j sur la variable x_i, nous avons besoin d'introduire la notion de *fonction-solution* qui exprime la variable x_i en fonction des autres variables de la contrainte. Par exemple, pour la contrainte $x+y=z$, les fonctions-solutions sont : $f_x = z-y, f_y = z-x, f_z = x+y$.

Si nous disposons de la fonction-solution qui exprime la variable x_i en fonction des autres variables de la contrainte, alors l'approximation de la projection de la contrainte sur x_i dans un domaine $\mathcal{D}$ peut être calculée avec une simple extension par intervalles de ces fonctions. C'est ainsi que nous allons obtenir $\pi_{j,i}(\mathcal{D})$.

Cependant, sur les contraintes complexes, il n'existe pas de telles fonctions analytiques ; par exemple $x + log(x) = 0$. L'objectif de cette section est justement de s'attaquer à ces cas complexes qui ne peuvent pas être résolus algébriquement. Trois approches principales ont été proposées :

— La première méthode exploite le fait que les fonctions analytiques existent toujours quand la variable à exprimer en fonctions des autres variables apparaît une fois seulement dans la contrainte. Dans l'exemple précédent, ce procédé donnerait : $x_{(1)} + \log(x_{(2)}) = 0$. De cette façon, il est aisé de calculer la fonction solution : il suffit d'exploiter les fonctions inverses des opérateurs de base. Dans notre exemple, on obtient $f_{x_{(1)}} = -\log(x_{(2)})$ and $f_{x_{(2)}} = \exp^{-x_{(1)}}$.

Une approximation de la projection de la contrainte sur x_i peut être calculée en intersectant les extensions naturelles des fonctions-solutions pour les différentes occurrences de x_i dans C_j. Pour l'exemple précédent, on peut prendre $\pi_{x+log(x)=0,x}(X) = -\log(X) \cap \exp^{-X}$. Les fonctions de projection obtenues de cette façon sont notées π^{nat}.

— La seconde stratégie utilise l'extension de Taylor pour transformer la contrainte en une contrainte linéaire. L'équation non-lineaire $f(X) = 0$ devient

$$f(c) + \sum_{i=1}^{n} nat(\frac{\partial f}{\partial x_i})(X) * (X_i - c_i) = 0$$

où $c = m(X)$. Supposons que les dérivées partielles sont évaluées sur la boite $\mathcal{D}$ qui contient X. $\mathcal{D}$ est considéré comme constant, et soit $c = m(\mathcal{D})$. L'équation devient :

$$f(c) + \sum_{i=1}^{n} nat(\frac{\partial f}{\partial x_i})(\mathcal{D}) * (X_i - c_i) = 0$$

C'est une équation linéaire par intervalles en X, qui ne contient pas des occurrences multiples. Les fonctions-solutions peuvent être extraites aisément. Cependant, au lieu de considérer les fonctions-solutions avec l'extension de Taylor séparément, on préfère regrouper ensemble les équations linéaires dans un seule système linéaire carré. La résolution du système linéaire permet une meilleure approximation des projections. (Voir la section suivante.) Les fonctions de projection obtenues de cette façon sont appelées π^{Tay}. Par exemple, soit la contrainte $x + \log(x) = 0$; avec la forme de Taylor sur la boite $\mathcal{D}$, on obtient l'équation linéaire

$$c + \log(c) + (1 + 1/\mathcal{D})(X - c) = 0$$

c'est-à-dire :

$$AX + B = 0$$

où $A = 1 + 1/\mathcal{D}$ et $B = \log(c) - c/\mathcal{D}$. La seule fonction solution de cette équation mono-dimensionnelle est : $X = -B/A$.

— La troisième approche [?] n'utilise aucune fonction-solution analytique. Elle transforme la contrainte $C_j(x_{j_1}, ...x_{j_k})$ en k contraintes mono-variables $C_{j,l}, l = 1 \dots k$. La contrainte mono-variable $C_{j,l}$ sur la variable x_{j_l} est obtenue en substituant les autres variables par leurs intervalles. La projection π_{j,j_l} est calculée grâce à $C_{j,l}$. Le plus petit zéro de $C_{j,l}$ dans l'intervalle de la variable considérée est la borne inférieure de la projection de C_j sur x_{j_l}. Et le plus grand zéro de $C_{j,l}$ est le borne supérieure pour la projection considérée. Ainsi, l'intervalle avec ces deux zéros comme bornes représente une approximation de la projection. Les fonctions de projection fonctions calculées de cette façon sont appelées π^{box}.

Dans [?], les deux zéros extremums de $C_{j,l}$ sont obtenus avec un algorithme dichotomique combinées avec la version mono-variable de Newton.

5.2.2 L'algorithme de filtrage

L'algorithme de filtrage est généralement vu comme un algorithme de point fixe. Dans la suite, une abstraction des algorithmes de filtrage est présentée : la suite $\{\mathcal{D}_k\}$ des domaines générée par l'application itérative de l'opérateur $Op : I\!I(I\!R)^n \longrightarrow I\!I(I\!R)^n$ (voir la figure 5.3).

$$\mathcal{D}_k = \begin{cases} \mathcal{D} & \text{if } k = 0 \\ Op(\mathcal{D}_{k-1}) & \text{if } k > 0 \end{cases}$$

FIGURE 5.3 – L'algorithme de filtrage comme un algorithme de point fixe

L'opérateur Op de filtrage satisfait généralement les trois propriétés suivantes :
— $Op(\mathcal{D}) \subseteq \mathcal{D}$ (contractance)
— Op est conservatif ; il ne supprime pas les solutions.
— $\mathcal{D}' \subseteq \mathcal{D} \Rightarrow Op(\mathcal{D}') \subseteq Op(\mathcal{D})$ (monotonicité)

Sous ces conditions, la limite de la suite $\{\mathcal{D}_k\}$, qui correspond au plus grand point fixe de l'opérateur Op, existe et est appelée *clôture*. Nous dénotons par $\Phi_{Op}(\mathcal{D})$.

Notre algorithme de filtrage est dérivé en appliquant le procédé de la figure 5.3 en instanciant Op à l'opérateur 14.

Définition 14 (opérateur du $2B$-filtrage).
$Op_{2B}(\mathcal{D}) = \cap_{C_j \in \mathcal{C}} \langle \pi_{j,1}(\mathcal{D}), \ldots, \pi_{j,n}(\mathcal{D}) \rangle$

La figure 5.4 montre comment les fonctions de projection sont utilisées par l'opérateur de $2B$-filtrage pour réduire les domaines des variables.

FIGURE 5.4 – $2B$-filtrage sur le système de contraintes $\{x^2 + y^2 = 1, y = x^2\}$

Suivant les fonctions de projection utilisées, on obtient différents algorithmes de $2B$-filtrage.

Op_{nat} L'opérateur Op_{nat} dénote Op_{2B} utilisant π^{nat}. Il abstrait l'algorithme de filtrage présenté dans [**?**, **?**, **?**].

Op_{box} Cet opérateur dénote Op_{2B} utilisant π^{box}. Il abstrait l'algorithme de filtrage présenté dans [**?**, **?**].

Op_{Tay} Cet opérateur dénote Op_{2B} utilisant π^{Tay}. Il abstrait la méthode de Newton par intervalles [**?**, **?**]. La méthode de Newton par intervalles contrôle dans un ordre précis les fonctions de projection à appliquer. Il est utilisé pour résoudre les équations non-linéaires carrées $\mathcal{C} = \{f_1(x_1, \ldots, x_n) = 0, \ldots, f_n(x_1, \ldots, x_n) = 0\}$.

La méthode de Newton par intervalles remplace la résolution du système carré d'équations par la résolution d'une suite de systèmes linéaires carrés. Chaque système linéaire est obtenu en évaluant la matrice Jacobienne sur le domaine courant. Le système qui en résulte est résolu souvent avec l'algorithme de Gauss-Seidel par intervalles. L'algorithme de Gauss-Seidel associe chaque contrainte C_i avec la variable x_i (après un renommage des variables), et boucle en appliquant seulement les fonctions de projection $\pi_{i,i}$ [1]

Définition 15 (Interval Newton operator).

$Op_{Tay}(\mathcal{D}) = \mathcal{D}'$ *where*
$\mathcal{D}' := \mathcal{D}$;
Let $A_{i,j} = [nat(\frac{\partial f_j}{\partial x_i})(\mathcal{D})], j = 1 \ldots n, i = 1 \ldots n$
for $i = 1 \ldots n$
$GS_i := (-f_i(mid(\mathcal{D})) - \sum_{j=1, j \neq i}^{n} A_{i,j}(D'_j - mid(D_j)))/A_{i,i} + mid(D_i)$;
$D'_i := D_i \cap GS_i$;
endfor

Remarquons que, en général, l'algorithme de Gauss-Seidel ne converge pas toujours vers la solution du système linéaire, mais il a de bonne propriétés de convergence pour les matrices de dominance diagonale. Ainsi, en pratique, avant la résolution du système linéaire, une étape de préconditionnement est

1. The for-loop corresponds to only one iteration of the Gauss-Seidel method and not to the complete solving of the interval linear system, which in pratice is not useful [**?**].

utilisée qui essaye de transformer la matrice Jacobienne en une matrice de dominance diagonale. Le préconditionnement consiste à multiplier le système d'équations linéaires par intervalles $A * X = B$ par une matrice M, donnant lieu à un nouveau système linéaire $A' * X = B'$ où $A' = MA$ et $B' = MB$. La matrice M est typiquement l'inverse de la matrice milieu de A.

Une propriété intéressante de l'opérateur de Newton est que dans certains cas, il est capable de prouver l'existence d'une solution. Quand $Op_{Tay}(\mathcal{D})$ est un sous-ensemble inclus (resp. strictement) dans $\mathcal{D}$, La théorème de point fixe de Brouwer s'applique et certifie la présence de solution (resp. d'une seule solution) dans $\mathcal{D}$ (cf. [?]).

5.2.3 Algorithme de décomposition

Quand un problème est difficile, nous le décomposons en deux ou plusieurs sous-problèmes potentiellement moins difficiles, dont l'union des solutions est égale aux solutions du problème d'origine. La décomposition se fait habituellement sur les domaines des variables, en divisant le domaine global en plusieurs sous-domaines, souvent suivant une seule dimension.

Sur les domaines continus, la décomposition s'effectue habituellement en *coupant* l'intervalle d'une variable en deux intervalles qui donneront lieu à deux sous-problèmes indépendants qui vont être résolus séparément.

Exemple 5. *La figure 5.5 illustre une décomposition du problème*

$$\begin{cases} x^2 + y^2 - 2 = 0; \\ x^2 - y = 0; \\ x, y \in [-10^8, +10^8]; \end{cases}$$

en deux sous-problèmes indépendants suivant l'axe des abscisses, où chacun des deux sous-problèmes isole une solution, c'est donc une bonne décomposition. Si la décomposition avait été faite suivant l'axe des ordonnées, nous aurions toujours un sous-problème avec les deux solutions : la décomposition suivant cet axe n'est pas intéressante.

FIGURE 5.5 – Décomposition du problème de l'intersection d'une parabole et d'un cercle.

5.3 SQP : Sequential Quadratique Programming

Chapitre 6

Méthodes d'optimisation approchées

Chapitre 7

Introduction à la programmation non-linéaire

Un problème général d'optimisation continue (POC) s'exprime comme suit : trouver des valeurs des variables de décision $x_1, ..., x_n$ telles que :

$$\begin{array}{l} minz = f(x_1, ..., x_n) \\ g_i(x_1, ..., x_n) = b_i, i = 1..m_e \\ g_j(x_1, ..., x_n) \leq b_j, j = m_e + 1..m. \end{array} \tag{7.1}$$

7.1 Compléments sur l'optimisation sans contraintes

Nous nous intéressons au problème suivant

$$min_{x \in I\!R^n} f(x) \tag{7.2}$$

où $x \in I\!R^n$ est un vecteur réel avec $n \geq 1$ et $f : I\!R^n \to I\!R$ est une fonction continue et continûment dérivable.

Nous supposons que les premières et secondes dérivées de $f(x_1, x_2, ..., x_n)$ existent et sont continues sur tous les points. Soit $\frac{\partial f}{\partial \tilde{x}_i}$ la dérivée partielle de $f(x_1, x_2, ..., x_n)$ relativement à x_i, évaluée au point $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$.

Une condition nécessaire pour que $\tilde{x} = \tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}$ soit un minimum local de (7.2) est donnée par le théorème suivant :

Théorème 8. *Si $\tilde{x}$ est un minimum local de (7.2), alors $\frac{\partial f(\tilde{x})}{\partial x_i} = 0$.*

La preuve de ce théorème vient du fait que si $\tilde{x}$ est un minimum local alors il a un voisinage où la fonction objectif prend son min en ce point. Supposons que le point en question n'annule pas la dérivée de la fonction objectif, alors nécessairement la fonction objectif est monotone en ce point faisant ainsi diminuer la fonction objectif ; ce qui contredit que le point est optimal dans son voisinage.

Définition 16. *Un point $\tilde{x}$ ayant $\frac{\partial f}{\partial x_i} = 0$ pour tout $i = 1, 2, ..., n$ est dit un point stationnaire de f.*

Le théorème suivant montre quelques conditions suffisantes pour qu'un point soit un extremum local à partir de la matrice hessienne.

Théorème 9. — *Si $H_k(\tilde{x}) > 0, k = 1..n$, alors le point $\tilde{x}$ est un minimum local de (7.2).*
— *Si $H_k(\tilde{x}), k = 1..n$, est non-nul et a le même signe que -1^k, alors le point $\tilde{x}$ est un maximum local de (7.2).*
— *Si $H_k(\tilde{x}) \neq 0, k = 1..n$, et les deux conditions précédentes ne sont pas vérifiées, alors le point $\tilde{x}$ n'est pas un extremum local de (7.2).*

Si un point stationnaire n'est pas un minimum local, il est dit un point d'inflexion (saddle point).

Nous pouvons aussi exploiter les propriétés de convexité et de concavité de la fonction objectif pour détecter des minimums globaux (voir aussi la section 2.2.2).

Nous présentons ci-dessous des stratégies supplémentaires pour qualifier la convexité et la concavité de toute fonction.

Définition 17. — *Le i-ème principal mineur (principal minor) d'une matrice $n \times n$ est le déterminant de la matrice $i \times i$ obtenue en supprimant $n-i$ lignes et les $n-i$ correspondantes colonnes de la matrice.*
— *Le k-ème principal mineur dominant (leading principal minor) d'une matrice $n \times n$ est le déterminant de la matrice $k \times k$ obtenue en supprimant les dernières $n-k$ lignes et colonnes de la matrice.*

Théorème 10. *Supposons que $f(x_1, x_2, ..., x_n)$ a les dérivées secondes continues.*
— *$f(x_1, x_2, ..., x_n)$ est une fonction convexe dans S si et seulement si pour tout $x \in S$, tous les principaux mineurs de H sont non-négatifs.*
— *$f(x_1, x_2, ..., x_n)$ est une fonction concave dans S si et seulement si pour tout $x \in S$ et $k = 1, 2, ..., n$, tous les principaux mineurs non-nuls de H ont le même signe que -1^k.*

7.2 Les multiplicateurs de Lagrange

Les multiplicateurs de Lagrange peuvent être utilisés quand les contraintes sont des égalités. Nous considérons donc le problème d'optimisation (POC) suivant :

$$\begin{array}{l} minz = f(x_1, ..., x_n) \\ c_i : g_i(x_1, ..., x_n) = b_i, i = 1..m_e \end{array} \tag{7.3}$$

Pour résoudre (7.3), on associe à chaque contrainte c_i un multiplicateur λ_i, et on forme ainsi le Lagrangien :

$$L(x_1, x_2, ..., x_n, \lambda_1, \lambda_2, ..., \lambda_m) = f(x_1, x_2, ..., x_n) + \sum_{i=1}^{m} \lambda_i [b_i - g_i(x_1, x_2, ..., x_n)] \tag{7.4}$$

Nous cherchons à trouver un point $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, , \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ qui minimise $L(x_1, x_2, ..., x_n, \lambda_1, \lambda_2, ..., \lambda_m)$. Souvent $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, , \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ résout aussi (7.3). Si $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, , \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ minimise L, alors

$$\frac{\partial L}{\partial \lambda_i} = b_i - g_i(x_1, x_2, ..., x_n) = 0$$

Ici $\frac{\partial L}{\partial \lambda_i}$ est la dérivée partielle de L par rapport à λ_i. Ceci montre bien que le point en question satisfait d'une façon optimale(7.3). Pour montrer que $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ résout (7.3), soit un point quelconque de l'espace faisable de (7.3). Puisque $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, \tilde{\lambda_2}, ..., \tilde{\lambda_m}$ minimise L, alors pour tout $\lambda'_i, i = 1..m$, nous avons :

$$L(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, , \tilde{\lambda_2}, ..., \tilde{\lambda_m}) \leq L(x'_1, x'_2, ..., x'_n, \lambda'_1, , \lambda'_2, ..., \lambda'_m)$$

Puisque $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ et $(x'_1, x'_2, ..., x'_n)$ sont tous les deux faisables, alors tous les facteurs des multiplicateurs sont nuls. On obtiens ainsi $f(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}) \leq f(x'_1, x'_2, ..., x'_n)$. Ceci montre que $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ est bien la solution de (7.3). En bref, si $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ résout le problème d'optimisation sans contraintes

$$minL(x_1, x_2, ..., x_n, \lambda_1, \lambda_2, ..., \lambda_m) \tag{7.5}$$

alors $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ résout aussi (7.3).

A partir du chapitre 7.1, nous savons qu'une condition nécessaire pour que $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ résolve (7.5) est qu'au point $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n}, \tilde{\lambda_1}, \tilde{\lambda_2}, ..., \tilde{\lambda_m})$ on doit avoir

$$\frac{\partial L}{\partial x_1} = \frac{\partial L}{\partial x_2} = ... = \frac{\partial L}{\partial x_n} = \frac{\partial L}{\partial \lambda_1} = \frac{\partial L}{\partial \lambda_2} = ... = \frac{\partial L}{\partial \lambda_n} = 0 \tag{7.6}$$

Le théorème suivant donne des conditions suffisantes pour qu'un tel point soit un minimum global.

Théorème 11. *Si $f(x_1, x_2, ..., x_n)$ est une fonction convexe et chaque $g_i(x_1, ..., x_n)$ est une fonction linéaire, alors tout point $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ satisfaisant (7.6) est une solution optimale.*

Les variables λ_i ont une interprétation intéressante comme étant les coûts marginaux (shadow prices) associés à chacune des des contraintes. Si la partie gauche d'une contrainte $g_i(x_1, x_2, ..., x_n) = b$ croit de δ_i, alors la valeur de la fonction objectif croit de $\delta_i \lambda_i$.

7.3 Les conditions de Kuhn-Tucker

Nous discutons dans cette section des conditions nécessaires et suffisantes pour que $(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ soit une solution optimale d'un POC avec des contraintes d'inégalité :

$$\begin{array}{l} minz = f(x_1, ..., x_n) \\ c_i : g_i(x_1, ..., x_n) \leq b_i, i = 1..m \end{array} \tag{7.7}$$

Pour que les résultats de cette section soient applicable, il qu'on ait seulement des contraintes d'inégalité inférieure. Les contraintes d'égalité ou d'inégalité supérieure peuvent être aisément ramenées en contraintes d'inégalité inférieure.

Le théorème suivant donne des conditions nécessaires pour qu'un point $\tilde{x} = (\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$ soit optimal pour le POC (7.7).

Pour que ce théorème soit applicable il faut que les fonctions g_i satisfassent des conditions de régularité (voir [?]). Quand les contraintes sont linéaires, ces conditions sont toujours satisfaites. Nous supposons dans la suite que les contraintes satisfassent toujours les conditions de régularité.

Théorème 12.

Comme au niveau des multiplicateurs de Lagrange de la section précédente, les multiplicateurs λ_i associés au conditions KT correspondent aux coûts marginaux des contraintes.

Chapitre 8

Méthodes d'optimisation pour le deep-learning

8.1 Introduction aux architectures neuronales

8.2 Problématique d'optimisation dans l'apprentissage automatique

Chapitre 9

Optimisation multicritère [Aribi, 2014]

Le texte de ce chapitre est pris intégralement du premier chapitre du mémoire de Noureddine Aribi [Aribi, 2014].

9.1 Composantes d'un problème d'optimisation multicritère

La plupart des problèmes d'optimisation combinatoire issus des problèmes réels impliquent plusieurs critères (objectifs), on parle alors de problèmes de décision multicritères ou d'optimisation multicritère. De plus, ces critères sont généralement incommensurables et contradictoires (e.g. minimiser les coûts et maximiser la productivité). L'optimisation multicritère apparaît donc comme un outil d'aide à la décision, conçu pour satisfaire des besoins conflictuels.

Commençons par détailler les composantes [?] d'un problème d'optimisation multicritère. L'ensemble des **alternatives** (objets ou actions) est l'ensemble des solutions potentielles d'intérêt pour le décideur. Cet ensemble peut être décrit en extension en énumérant les solutions dans une liste finie, ou en compréhension en explicitant un ensemble de contraintes caractérisant les solutions. Chaque alternative est décrite par un ensemble d'**attributs** (descripteurs ou points de vue) spécifiant ses caractéristiques mesurables. Un **critère** est l'association d'une **préférence** sur les valeurs possibles de l'attribut. Différentes approches de décision multicritère ont vu le jour, en combinant une opération de comparaison et une opération d'**agrégation**. Nous détaillerons ces éléments de base dans la suite de cette section.

9.1.1 Notion de préférence et de fonction d'utilité

Le concept de *décision* est difficilement séparable de celui de *préférence*. Les préférences sont l'expression, de la part d'une personne confrontée à une situation de choix, de l'attrait que présente chaque alternative envisageable [?]. Les préférences apparaissent dans plusieurs domaines, comme l'économie, la psychologie, les sciences politiques, les statistiques, etc.

La notion de préférence dépasse bien le cadre d'une simple décision. En économie par exemple, les préférences reflètent le prix qu'une personne peut payer afin d'obtenir un objet [?]. En psychologie, elle s'interprète comme l'attitude d'une personne face à un ensemble d'objets.

L'un des principaux sujets qui ont émergé dans le domaine de la recherche de décision comportementale [?], au cours des trois dernières décennies, est le traitement des préférences des décideurs. Ces préférences sont souvent construites suivant un processus d'*élicitation* utilisant des algorithmes et des procédures cognitives de recherche des préférences du décideur.

Les préférences d'un décideur (définies sur un ensemble fini d'alternatives) portent une sémantique particulière. Ces préférences peuvent être représentées de différentes manières, exprimées sous la forme d'une relation mathématique binaire sur les alternatives.

Définition 18 (Relation de préférence)**.** *Soit $\mathcal{A}$ l'ensemble des alternatives. Soient $a, b \in \mathcal{A}$.*

Préférence large *Une relation de préférence est notée $\succsim$. L'expression $a \succsim b$ signifie que l'alternative a est au moins aussi bonne que l'alternative b. Nous supposons que $\succsim$ est binaire, prenant ses valeurs dans $\{0, 1\}$.*

Préférence stricte *La relation de préférence stricte $\succ$ est définie : $a \succ b$ ssi $a \succsim b$ et non$(b \succsim a)$.*

Indifférence *Nous noterons $\sim$ la relation d'indifférence définie comme suit : $a \sim b$ ssi $a \succsim b$ et $b \succsim a$.*

Incomparabilité *Si pour deux alternatives différentes a et b, nous n'avons ni $a \succsim b$, ni $b \succsim a$, nous dirons que a et b sont incomparables.*

La relation $\succ$ est la partie asymétrique de $\succsim$, et $\sim$ est sa partie symétrique.

La relation de préférence que nous avons introduite permet d'établir un classement sur l'ensemble des alternatives, sans pour autant donner une information sur l'*intensité* des préférences exprimées. Pour contourner cette faiblesse, on fera appel à la notion de fonction d'utilité, en imposant des hypothèses qui sont assez naturelles.

Proposition 4 (Définition d'une fonction d'utilité)**.** *Soit $\succsim$ une relation de préférence sur les alternatives $\mathcal{A}$, telle que $\succsim$ soit un préordre complet*[1]*. Il existe une fonction $u : \mathcal{A} \to \mathbb{R}$ telle que :*

$$\forall a, b \in \mathcal{A}, a \succsim b \Leftrightarrow u(a) \geq u(b).$$

u est dite fonction d'utilité de la relation de préférence $\succsim$.

Nous supposons, sans perte de généralité, que chaque fonction d'utilité est à maximiser du point de vue du preneur de décision (DM). La relation de préférence peut être traduite de façon quantitative par une fonction d'utilité, en associant à chaque alternative un nombre réel. Les préférences seront donc établies via des comparaisons sur $\mathbb{R}$. Concernant les relations de préférence stricte et d'indifférence, nous avons :

$$\forall a, b \in \mathcal{A}, a \succ b \Leftrightarrow u(a) > u(b),$$
$$\forall a, b \in \mathcal{A}, a \sim b \Leftrightarrow u(a) = u(b).$$

1. Un préordre complet est une relation réflexive, transitive et complète.

9.1.2 Décision multicritère

La décision multicritère traite des problèmes liés à plusieurs points de vue ou attributs. Par exemple, considérons une situation d'un décideur face à une problématique d'achat d'un appareil photo numérique. L'ensemble des alternatives est celui des appareils photo numérique disponibles dans le commerce. Au fait, un appareil photo numérique est caractérisé par plusieurs attributs, à savoir le nombre de méga-pixels, le poids, la plage de zoom maximal, etc. Ainsi, le décideur est amené à raisonner sur ces attributs pour se fixer sur son achat. Plus généralement, nous avons besoin de définir la notion d'attributs.

Définition 19 (Attributs). *Soit $\mathcal{A}$ l'ensemble des alternatives. Les ensembles $\mathcal{A}^1, ..., \mathcal{A}^n$ sont des attributs si et seulement si $\mathcal{A} = \mathcal{A}^1 \times \mathcal{A}^2 \times ... \times \mathcal{A}^n$. Autrement dit, $\mathcal{A}$ est le produit cartésien des attributs. Toute alternative a est un n-uplet $(a^1, ..., a^n)$, avec $\forall i \in 1..n, a^i \in \mathcal{A}^i$. La fonction d'utilité u appliquée à une alternative a, portera aussi sur ses attributs $u(a^1, ..., a^n)$.*

A un décideur, nous associons n relations de préférence $\succsim^1, ..., \succsim^n$, où $\succsim^i$ correspond à la relation de préférence du décideur associée à l'attribut i d'utilité u_i :

$$\forall a^i, b^i \in \mathcal{A}^i, a^i \succsim^i b^i \Leftrightarrow u_i(a) \geq u_i(b).$$

On appelle *critère* l'association d'un attribut (ou point de vue) et d'une préférence : le décideur déclare une préférence sur les différentes valeurs possibles de l'attribut.

Soit $N = \{1, ..., n\}$ l'ensemble des indices des critères. La fonction d'utilité (individuelle) sur le critère i est la fonction $u_i : \mathcal{A} \mapsto I\!R$. La valeur a_i est la valeur d'utilité de l'alternative a selon le critère i. En d'autres termes, nous avons [2] $a_i = u_i(a)$. La fonction $u : \mathcal{A} \to I\!R^n$, dite fonction d'utilité multicritère, est définie comme suit : pour toute alternative a, $u(a) = (u_1(a), ..., u_n(a)) = (a_1, ..., a_n)$.

L'espace des critères $\mathcal{Y} \subset I\!R^n$ est défini comme étant l'image de $\mathcal{A}$ par u :

$$\forall y \in \mathcal{Y}, \exists a \in \mathcal{A}, u(a) = y.$$

Nous désignerons l'espace des critères comme étant l'espace des solutions (réalisables). Soit $y \in \mathcal{Y}$ une solution réalisable associée à l'alternative $a \in \mathcal{A}$, nous notons $y_i = a_i = u_i(a)$ où $i \in 1..n$. Dans cette thèse, nous travaillerons dans l'espace des critères. Toute alternative désignera désormais une solution (réalisable).

On peut énumérer plusieurs types de problèmes [?] de décision :

Choix : choisir la ou les meilleures alternatives ;

Classification, tri : mettre des catégories pré-définies ordonnées sur les alternatives ;

Rangement : produire un ordre partiel ou total sur les alternatives ;

Scorage/utilité : attribuer une valeur d'utilité à chaque alternative ;

2. Certaines références utilisent la notation $a_i = u_i(a^i)$ pour bien préciser que la valeur du critère dépend de la valeur de l'attribut associé.

Pour appréhender un problème de décision multicritère, la méthodologie générale combine une opération de comparaison et une opération d'agrégation :

Agréger puis comparer : consiste à construire une relation d'agrégation entre les n critères permettant par la suite d'effectuer des comparaisons en utilisant cette relation. Le principal représentant de cette approche est la théorie multi-attributs MAUT (*Multi-Attribute Utility Theory*) [**?**].

Comparer puis agréger : qui exploite les relations $\succsim^1, ..., \succsim^n$ pour comparer les alternatives afin de construire la relation de préférence globale $\succsim$. Le principal représentant de cette approche est l'ensemble des méthodes par surclassement, et plus particulièrement les différentes versions de la méthode Electre [**?**].

Dans cette thèse, nous considérons les méthodes de type "**agréger puis comparer**". Nous supposons, sans perte de généralité, que chaque critère est à **maximiser** du point de vue du preneur de décision (DM). Dans le processus de décision multicritère de type "agréger puis comparer", le problème multicritère est transformé en un problème monocritère, en utilisant une *fonction d'agrégation*, afin de bénéficier des méthodes classiques d'optimisation mono-critère. Ainsi, nous supposons l'existence d'une fonction d'utilité globale U qui représente la relation de préférence globale $\succsim$, que tout DM tente de maximiser

$$\forall a, b \in \mathcal{A}, a \succsim b \Leftrightarrow U(a) \geq U(b).$$

Cette fonction d'utilité globale U doit dépendre seulement des critères $U(a) = f(a_1, ..., a_n)$. Cette fonction d'agrégation permet au décideur de trouver une bonne solution de compromis selon ses préférences. On se pose ainsi la question clé : quelle est la fonction d'agrégation la mieux adaptée ? et comment la paramétrer ? La réponse à cette question dépendra nécessairement des propriétés des critères et de la relation de préférence globale.

Les préférences du décideur sont implicitement ou explicitement traduites en termes de paramètres de la méthode multi-objectif choisie (cf., [**?**, **?**, **?**, **?**]). Ensuite, la fonction d'agrégation est combinée avec les paramètres acquis pour résoudre le problème multi-objectif, et obtenir la solution de meilleur compromis du point de vue du décideur.

Il existe de nombreuses méthodes d'agrégation multicritère pour résoudre des problèmes de décision. Cependant, le choix le plus approprié d'une méthode multicritère bien adaptée à un problème multicritère est subjectif (cf. [**?**]), et est lui même un problème compliqué.

De la littérature on peut relever deux méthodes principales liées à deux points de vue opposés : la fonction SOMME et la fonction MINIMUM, correspondant respectivement aux concepts d'*utilitarisme classique* et *égalitarisme*[3]. L'approche utilitariste traite l'aspect multicritère du problème avec une approche de *scalarisation* simple et efficace, avec laquelle on agrège toutes les fonctions objectifs pour former un problème avec une seule fonction objectif (appelée fonction d'agrégation). Ainsi, une

3. Des compromis entre ces deux approches existent, e.g., les opérateurs d'agrégation OWA [**?**], et l'intégrale de CHOQUET [**?**].

décision optimale est l'une de celles qui maximisent cette fonction. Cependant, ce genre de fonctions d'agrégation n'est pas très pertinent dans le contexte du partage "équitable" [?, ?].

Par contre, la deuxième approche (égalitariste) est spécialement bien adaptée aux problèmes pour lesquels la propriété d'équité joue un rôle central. Cette approche est utilisée surtout en décision multi-agents [?], où les critères représentent les préférences de différents agents (e.g., problème de partage d'un ensemble fini de ressources entre plusieurs agents). Suivant la fonction MINIMUM, une décision optimale est celle qui *maximise* la satisfaction de l'agent le moins satisfait. Néanmoins, cette fonction d'agrégation est entravée par l'*effet de noyade* [?], puisqu'elle ne peut pas distinguer entre des solutions ayant la même composante minimale. Ainsi, par exemple, les deux solutions (alternatives) $\langle 0, 1, 1, 1\rangle$ et $\langle 1000, 1000, 1000, 0\rangle$, pourtant très différentes, sont indiscernables suivant la fonction MINIMUM. Cette fonction peut toutefois être modifiée pour avoir un comportement plus expressif. L'idée clé consiste à proposer des raffinements du MINIMUM, e.g., les opérateurs DISCRIMIN et LEXIMIN, la norme de Chebycheff, etc. Le but étant de réduire l'ensemble des alternatives indifférentes et d'assurer la propriété d'efficacité. On reviendra sur ce point avec plus de détails dans la section 9.3.

9.2 Comparaison des solutions

9.2.1 Relation de dominance et PARETO-optimalité

Les problèmes d'optimisation multicritère se ramènent à un problème central de comparaison. Cette comparaison est fondée sur la relation de dominance.

Définition 20 (Dominance faible de PARETO)**.** *Soient deux solutions réalisables $y, y' \in \mathcal{Y}$. La dominance faible entre y et y' est définie par :*

$$y \succsim_p y' \Leftrightarrow [\forall i \in \{1, ..., n\},\ y_i \geq y'_i] \tag{9.1}$$

La relation $y \succsim_p y'$ signifie que y est au moins aussi bonne que y' sur tous les critères.

Définition 21 (Dominance de PARETO)**.** *Soient deux solutions réalisables $y, y' \in \mathcal{Y}$. La relation de dominance de PARETO est définie comme la partie asymétrique de la relation $\succsim_p$:*

$$y \succ_p y' \Leftrightarrow [y \succsim_p y' \ \textit{et} \ \textsc{non}(y' \succsim_p y)] \tag{9.2}$$

La relation $y \succ_p y'$ signifie que y est au moins aussi bonne que y' sur tous les critères, tout en étant strictement meilleure sur au moins un critère. Si y est meilleure que y' sur tous les critères, alors on dit que y domine *fortement* y'.

La principale caractéristique des problèmes d'optimisation multicritère est l'existence de plusieurs critères, d'où la nécessité de revoir la notion d'optimalité qui est inspirée de l'optimisation combinatoire mono-critère.

Définition 22 (PARETO optimalité). *Une solution $y^* \in \mathcal{Y}$ est PARETO-optimale (efficace, non-dominée, non-inférieure) si et seulement s'il n'existe pas une solution y telle que y domine y^*. L'ensemble des solutions PARETO-optimales forment le front de Pareto P défini par :*

$$P = \{y \in \mathcal{Y} \mid \nexists y' \in \mathcal{Y}, \quad y' \succ_p y\} \tag{9.3}$$

On note que la relation de dominance de PARETO a un faible pouvoir discriminant. Ainsi, beaucoup de solutions efficaces restent incomparables (i.e., $\exists y, y' \in \mathcal{Y}$, NON$(y \succ_p y')$ et NON$(y' \succ_p y)$). L'ordre induit par la relation de dominance de PARETO est donc partiel.

Exemple 17 (Problème du sac-à-dos multiobjectif). *Le problème du sac-à-dos multiobjectif*[4] *[?, ?] est une généralisation du problème classique du sac-à-dos mono-objectif. Ce problème est connu pour être NP-complet*[5]*. La difficulté consiste à choisir un sous-ensemble de m objets maximisant les n fonctions objectifs (profits) tout en ne dépassant pas le poids maximal W autorisé pour le sac. Chaque objet i, a un poids w_i et un profit c_i^j pour chaque fonction objectif. Formellement,*

$$\begin{aligned} \texttt{Maximize} \quad & u_j(x) = \sum_{i=1}^{m} c_i^j x_i \quad j = 1, ..., n \\ \texttt{subject to} \quad & \sum_{i=1}^{m} x_i w_i \leq W \\ & x_i \in \{0, 1\} \qquad\qquad i = 1, ..., m \end{aligned} \tag{9.4}$$

On définit la variable x_i associée à un objet i de la façon suivante : $x_i = 1$ si l'objet i est mis dans le sac, et $x_i = 0$ sinon. Les variables $x_i, i = 1..m$ permettent de définir les fonctions d'utilité en compréhension. En fait, l'espace fini des solutions des contraintes de (9.4) définit l'espace des alternatives. En outre, la contrainte $w(x) = \sum_{i=1}^{m} x_i w_i \leq W$ garantit que la somme des objets choisis ne dépasse pas la capacité du sac-à-dos. Pour quatre objets ($m = 4$), deux fonctions objectifs ($n = 2$), et un sac-à-dos d'un poids maximal de 10 kg *($W = 10$), on a par exemple les données suivantes :*

TABLE 9.1 – Jeu de données pour le problème du sac-à-dos multiobjectif

i	1	2	3	4
c_i^1	18	12	17	2
c_i^2	3	11	7	15
w_i	4	5	6	5

4. Conférence plénière "*Algorithmic Decision Theory and Preference-based Optimization*", Patrice Perny, JFPC'2011.

5. On peut donc raisonnablement penser qu'il est inutile d'en chercher une solution sous forme d'un algorithme de complexité polynomiale.

A partir du jeu de données du tableau 9.1, on peut poser le problème du sac-à-dos bi-objectif suivant :

$$\begin{array}{rl} \texttt{Maximize} & u_1(x) = 18x_1 + 12x_2 + 17x_3 + 2x_4 \\ \texttt{Maximize} & u_2(x) = 3x_1 + 11x_2 + 7x_3 + 15x_4 \\ \texttt{subject to} & 4x_1 + 5x_2 + 6x_3 + 5x_4 \leq 10 \\ & x_i \in \{0,1\}, i = 1, ..., 4 \end{array} \tag{9.5}$$

L'ensemble des solutions du problème (9.5) est donné dans la table 9.2. Les solution $\{s_1..s_4\}$ ne sont dominées par aucune autre solution. Elles constituent donc les points du front de Pareto *pour l'exemple 17. Par contre, les quatre solutions qui restent $\{s_5, s_6, s_7, s_8\}$ sont des solutions dominées au sens de* Pareto, *car $s_4 \succ_p s_5, s_1 \succ_p s_6, s_2 \succ_p s_7$ et $s_1 \succ_p s_8$.*

Table 9.2 – Solutions de l'instance du problème du sac-à-dos multiobjectif (Les solutions $s_1..s_4$ sont efficaces. Les solutions $s_5..s_8$ sont dominées. L'ensemble $\{...\}$ précise les indices des variables booléennes qui sont à un (les autres sont à zéro). Par exemple, l'ensemble $\{2,4\}$ précise que $x_2 = x_4 = 1$ et $x_1 = x_3 = 0$.)

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
	$\{2,4\}$	$\{1,4\}$	$\{1,2\}$	$\{1,3\}$	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$
$\hat{u_1}$	14	20	30	35	18	12	17	2
$\hat{u_2}$	26	18	14	10	3	11	7	15

Deux types de solutions Pareto-optimales peuvent être différenciées : les solutions *supportées* et les solutions *non-supportées*. Les premières solutions sont celles situées sur l'enveloppe convexe[6] de l'ensemble des solutions de l'espace des critères (voir la figure 9.2). Ces dernières peuvent être trouvées à l'aide d'une optimisation d'une agrégation linéaire des critères [?] utilisant différents vecteurs de poids. Les deuxièmes solutions représentent l'ensemble des solutions non-supportées, i.e., les solutions non-dominées et qui n'appartiennent pas à la fermeture convexe.

Il faut noter que les solutions non-supportées sont plus difficile à obtenir par rapport aux solutions supportées [?]. Cependant, la recherche des solutions non-supportées est motivée par deux principales raisons. D'un coté, les solutions supportées peuvent ne représenter qu'un petit sous-ensemble de solutions efficaces. D'un autre coté, les solutions supportées peuvent ne pas avoir une répartition uniforme sur le front de Pareto, et elles ne garantissent donc pas un bon compromis (voir la figure 9.3).

6. L'enveloppe convexe [?] d'un ensemble Q de points est le plus petit polygone convexe P tel que chaque point de Q est soit sur le contour de P, soit à l'intérieur. Nous rappelons qu'un polygone est convexe si, étant donnés deux points quelconques situés sur le contour ou à l'intérieur, tous les points du segment de droite reliant ces deux points se trouvent sur le contour ou à l'intérieur du polygone. Un polygone est une courbe plane, refermée sur elle-même et composée d'une suite de segments de droite appelés côtés du polygone.

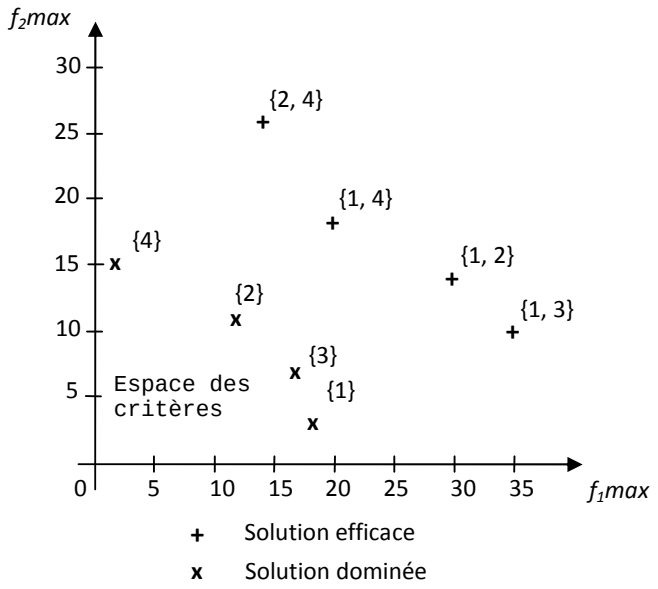


FIGURE 9.1 – Espace de critères du problème du sac-à-dos bi-objectif (9.5). L'ensemble $\{...\}$ précise les indices des variables booléennes qui sont à un (les autres sont à zéro). Par exemple, l'ensemble $\{2, 4\}$ précise que $x_2 = x_4 = 1$ et $x_1 = x_3 = 0$.)

9.2.2 Détermination des frontières

L'ensemble des solutions supportées et non-supportées (si elles existent) constitue ce qu'on appelle le front de PARETO (voir la figure 9.2).

Définition 23 (Frontière de PARETO). *La frontière*[7] *de Pareto (ou ensemble de solutions non-dominées au sens de Pareto) est l'ensemble des solutions réalisables défini par* $ND_P = \{y \in \mathcal{Y} | \nexists y' \in \mathcal{Y}, y' \succ_p y\}$. *L'ensemble des alternatives engendrant la frontière de Pareto est défini par* $ArgND_P = \{a \in \mathcal{A} | \nexists b \in \mathcal{A}, u(b) \succ_p u(a)\}$.

La recherche des solutions PARETO-efficaces est basée sur le critère de dominance pour pouvoir comparer les solutions. Dans cette optique, il ne faut pas omettre le fait que deux solutions peuvent être équivalentes dans l'espace des critères $\mathcal{Y}$, i.e., ayant exactement les mêmes valeurs sur l'ensemble des critères, alors qu'elles ont des coordonnées totalement différentes dans l'espace de décision. Ainsi, on distingue deux types de front de PARETO :

1. *Front minimal :* ce type de front est généré sans considérer les solutions équivalentes ;
2. *Front maximal :* ce type de front est généré en considérant toutes les solutions équivalentes de l'espace de décision pour chaque solution non-dominée dans l'espace des critères.

Pour caractériser les solutions réalisables de compromis, on pourrait faire appel à une définition d'équilibre entre les critères. La théorie du choix social a formalisé ces solutions d'équilibre en faisant appel au principe de transfert [?] défini comme suit.

7. La notation est inspirée de [?].

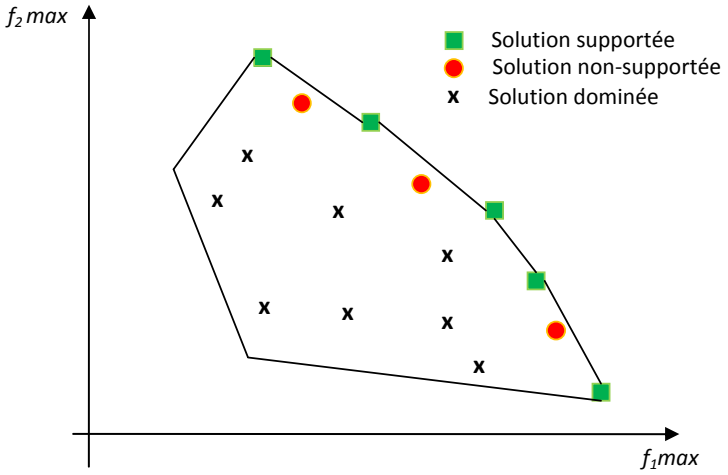


FIGURE 9.2 – Représentation des différents types de solutions en optimisation bi-objectif.

Définition 24 (Propriétés de compatibilité, P-monotonie, principe de transfert)**.**

— *Soient deux solutions* $y, y' \in \mathbb{R}^n$*, et* $\succsim_?$ *une relation sur* $\mathbb{R}^n$*. On dit que* $\succsim_?$ *est compatible avec la dominance de Pareto, ou que* $\succsim_?$ *est P-monotone si et seulement si on a* $y \succsim_p y' \Longrightarrow y \succsim_? y'$*.*

— *Soit* $y \in \mathbb{R}^n$ *et* $\succsim_?$ *une relation de dominance sur* $\mathbb{R}^n$*. Soit l'existence de* $i, j \in 1..n$ *tels que* $y_i > y_j$*. Soit la notation d'un vecteur* e^z *tel que* $\forall i \neq z, e^z_i = 0$ *et* $e^z_z = 1$*. La relation* $\succsim_?$ *respecte le principe de transfert si et seulement si pour tout* ϵ *vérifiant* $0 \leq \epsilon \leq \frac{y_i - y_j}{2}$*, on a* $y - \epsilon e^i + \epsilon e^j \succsim_? y$*.*

Définition 25 (Transformée de Lorenz et Dominance généralisée de Lorenz)**.**

— *Soit* $y \in \mathbb{R}^n$ *une solution. La transformée de Lorenz de la solution* y*, notée* $L(y)$*, est le vecteur des sommes cumulées du tri en ordre croissant de* y *:*

$$L(y) = (y_1, y_1 + y_2, ..., \sum_{i=1..n} y_i).$$

— *Soient* $y, y' \in \mathbb{R}^n$ *deux solutions réalisables.* y *domine* y' *au sens de Lorenz, noté* $y \succsim_L y'$*, si et seulement si* $L(y) \succsim_p L(y')$*.*

Proposition 5. *La relation de dominance généralisée de Lorenz est compatible avec la dominance de Pareto et respecte le principe de transfert.*

Dans un contexte multi-agents, la P-monotonie exprime la préférence pour une situation plus favorable à tous les agents, et le principe de transfert traduit l'envie de répartir les coûts de manière équitable entre les agents.

Définition 26 (Frontière de Lorenz)**.** *La frontière de Lorenz (ou ensemble de non-dominés au sens de Lorenz) est l'ensemble des solutions réalisables défini par* $ND_L = \{y \in \mathcal{Y} | \nexists y' \in \mathcal{Y}, y' \succ_L y\}$*. L'ensemble des alternatives engendrant la frontière de Lorenz est défini par* $ArgND_L = \{a \in \mathcal{A} | \nexists b \in \mathcal{A}, u(b) \succ_L u(a)\}$*.*

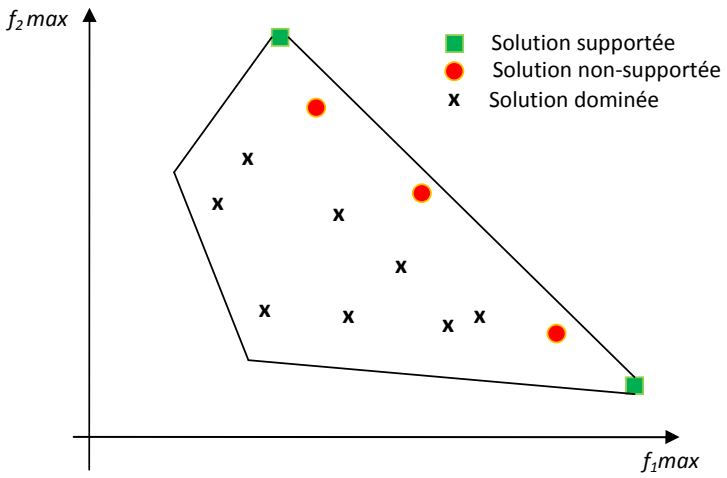


FIGURE 9.3 – Importance des solutions non supportées.

Définition 27 (Courbe de LORENZ)**.** *On considère un vecteur objectif de n éléments supposé numéroté par utilité croissante, i.e., $y_1 \leq y_2 \leq ... \leq y_i \leq ... \leq y_n$, où y_i désigne l'utilité de l'élément i. On note que les x_i représentent des gains à maximiser.*

On pose :

— $m = \frac{1}{n}\sum_{i=1}^{n} y_i$ *: l'utilité moyenne.*

— $y'_i = \frac{y_i}{m}$ *: l'utilité relative de l'élément i par rapport à la moyenne m.*

On note maintenant q_k la proportion de l'utilité totale reçue par les k éléments les plus "pauvres" :

$$\begin{aligned} q_k &= \frac{y_1 + ... + y_k}{v_1 + ... + y_n} \\ &= \frac{1}{n}\sum_{i=1}^{k} y'_i, \qquad 1 \leq k \leq n \end{aligned} \tag{9.6}$$

La courbe de LORENZ est la courbe reliant les points de coordonnées $(p_k = \frac{k}{n}, q_k), k = 0..n$, avec $q_0 = 0$.

Exemple 18. *Soient $n = 3, y = \langle y_1, y_2, y_3\rangle = \langle 2, 3, 5\rangle$. Suivant la formule (9.6), on obtient les points $(p_1, q_1), (p_2, q_2)$ et (p_3, q_3) ayant comme coordonnées :*

$$\begin{bmatrix} p_1 & p_2 & p_3 \\ q_1 & q_2 & q_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 1 \\ 0.2 & 0.5 & 1 \end{bmatrix} \tag{9.7}$$

La courbe de LORENZ de cet exemple est donnée sur la figure 9.4. Cette courbe est toujours convexe, et le degré de sa convexité donne une information sur l'inégalité

entre les différents éléments d'un profil d'utilité. Si tous les éléments d'un profil d'utilité sont complètement égaux, la courbe de LORENZ *devient linéaire.*

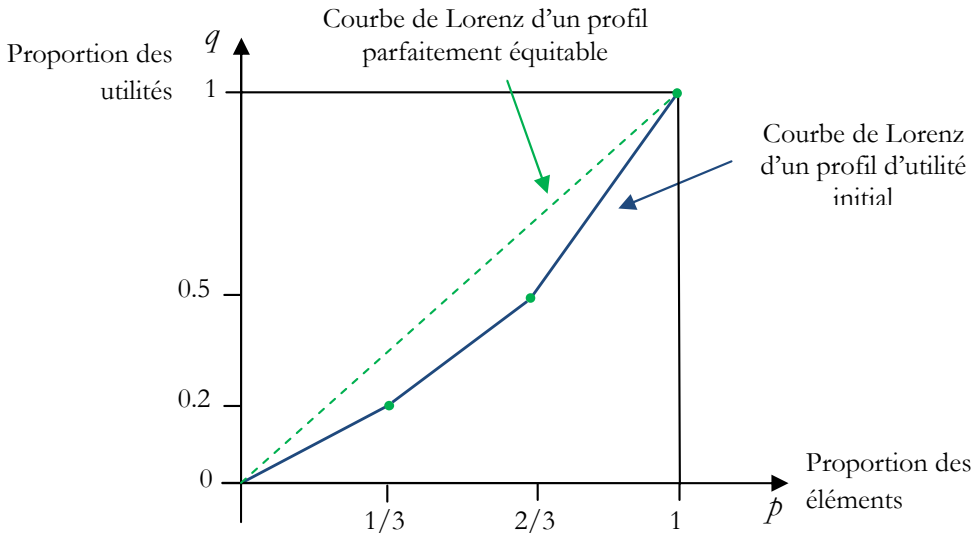


FIGURE 9.4 – Courbe de LORENZ.

9.3 Méthodes d'agrégation

9.3.1 La somme pondérée

La méthode de la somme pondérée est basée sur une combinaison linéaire des critères. Cette méthode est sans doute la méthode multicritère la plus simple, et la plus couramment utilisée. Elle est définie comme suit :

$$\begin{aligned} \underset{x\in\mathcal{Y}}{\texttt{Maximize}} \quad & U(x) = \sum_{i=1}^{n} w_i x_i \\ \texttt{subject to} \quad & \sum_{i=1}^{n} w_i = 1 \qquad (9.8) \\ & w_i \in [0,1], \qquad i = 1, ..., n. \qquad (9.9) \end{aligned}$$

où w_i est le poids (coefficient) affecté au i^{e} critère. Ce coefficient représente l'importance relative[8] que le décideur attribue au critère. Si tous les poids sont positifs, le maximum du modèle (9.8) est PARETO-optimal[9] [?, ?]. Ainsi, maximiser

8. La contrainte de normalisation des poids $\sum_{i=1}^{n} w_i = 1$ justifie le terme "moyenne pondérée".
9. C'est une propriété majeure des fonctions d'agrégation additives.

(9.8) est suffisant pour garantir l'efficacité des solutions. Cependant, il est impossible d'obtenir les solutions non-supportées qui se trouvent sur des portions concaves du front de PARETO dans l'espace des critères (cf., [?, ?]).

Il existe d'autres types de moyenne (e.g., géométrique, harmonique, etc.), qui peuvent s'écrire sous la forme :

$$U_f(x) = f^{-1}(\sum_{i=1}^{n} w_i f(x_i)),$$

où f est une fonction continue strictement croissante. Elles sont dites moyennes généralisées. Si f est la fonction identité, on retrouve la moyenne pondérée classique. Si $f = \log$ (resp. $f(x) = 1/x$), on retrouve la moyenne géométrique (resp. moyenne harmonique).

9.3.2 Méthode d'ordre lexicographique

Il est habituellement difficile de déterminer les coefficients de la fonction de la somme pondérée, car dans le cas général :

- Les critères ne sont pas basés sur une échelle commune ;
- Les critères sont mutuellement conflictuels ;
- Les conséquences d'une différence donnée (compromis) ne peuvent pas être quantitativement connues avant l'optimisation.

La méthode d'ordre lexicographique semble être une alternative intéressante pour éviter l'utilisation des poids. Cette méthode optimise séquentiellement les critères par ordre d'importance. Ainsi, l'ordre de priorité entre les critères doit être fixé explicitement.

Formellement, la méthode lexicographique résout une séquence de problèmes d'optimisation dans l'ordre de priorité comme suit :

$$\begin{array}{lll} \underset{x\in\mathcal{A}}{\texttt{Maximize}} & u_i(x), & i = 1, 2, ..., n \\ \texttt{subject to} \quad u_j(x) \geq u_j(x^*), & j = 1, ..., i-1 \quad \text{if} \quad i > 1. \end{array} \tag{9.10}$$

A partir du deuxième problème d'optimisation (i.e., pour $i = 2, ..., n$), les critères précédents sont convertis en contraintes d'inégalité [?]. Ces nouvelles contraintes sont formulées en utilisant les valeurs objectifs optimales $u_j(x^*)$ avec $j = 1..i-1$, et elles sont ajoutées au système initial de contraintes.

Cette méthode a le mérite d'être simple à mettre en œuvre, le temps d'exécution étant raisonnable (cf. [?, ?]). De plus, les solutions calculées sont nécessairement PARETO-optimales. En revanche, cette méthode fournit souvent des solutions extrêmes (eg., prendre un avion, ou marcher à pied). Aussi, les critères les plus importants sont favorisés et optimisés (même faiblement) au détriment des autres critères les moins prioritaires. Ce qui constitue un défaut important dans les applications qui cherchent des solutions équilibrées.

Ordre lexicographique d'une paire d'alternatives

Définition 28 (Ordre lexicographique)**.** *[?] Les critères sont classés par le décideur en fonction de leur importance. La solution sélectionnée est celle qui a la meilleure valeur pour le critère le plus important. Les solutions indifférentes par rapport au premier critère seront différenciées à l'aide du deuxième critère le plus important, et ainsi de suite. Formellement, l'opérateur d'ordre lexicographique peut être défini par récurrence :*

$$x \succ_{lex} y \Leftrightarrow (x_1 > y_1) \vee ((x_1 = y_1) \wedge \langle x_2, ..., x_n \rangle \succ_{lex} \langle y_2, ..., y_n \rangle) \tag{9.11}$$

où $x, y \in \mathcal{Y}$.

9.3.3 Méthodes basées sur des raffinements de l'ordre Min

Dans le monde réel, on peut distinguer une classe particulière de problèmes multicritères, dont la résolution doit tenir compte de deux propriétés importantes, qui sont :

1. L'équilibre entre les différents critères. Dans le cadre de la théorie du choix social, on parle plutôt de la notion de partage (allocation, division) équitable d'une ressource divisible, ou d'un ensemble de ressources indivisibles entre plusieurs agents. Ce critère est aussi connu sous le nom du *Welfare* [?].
2. L'efficacité de la solution calculée. Cette propriété assure que la solution d'équilibre calculée n'est pas dominée au sens de Pareto.

Parmi les problèmes exigeant une ou les deux propriétés à la fois, on peut citer les problèmes suivants :

— Tournées de véhicules ;
— Planification des infirmières (*nurse rostering*) ;
— Élaboration des emplois du temps ;
— etc.

Dans un contexte multi-agents, à chaque alternative a correspond un profil d'utilité $\langle u_1(a), ..., u_n(a) \rangle$, où u_i est l'utilité individuelle de l'agent i pour l'alternative a. Cette utilité individuelle mesure son niveau de "bien être".

Dans la section 9.1.2, on a vu qu'une combinaison linéaire des utilités individuelles ne capture pas bien l'idée d'équité d'une solution. Ceci suggère la nécessité de recourir à des fonctions non-linéaires. Dans cette optique, l'approche MaxMin, qui consiste à maximiser la fonction Minimum [10], constitue l'alternative la plus intuitive au modèle linéaire. L'idée consiste à maximiser la satisfaction de l'agent le moins satisfait, ce qui garantit une meilleure valeur dans le pire des cas (approche égalitariste [?]). Cependant, à partir de deux agents, la recherche d'une allocation équitable suivant le critère MaxMin, est un problème NP-difficile [?, ?]

Soient $x = \langle x_1, ..., x_n \rangle$ et $y = \langle y_1, ..., y_n \rangle$ deux solutions de l'espace des critères $\mathcal{Y}$. Dans ce qui suit, on suppose que les éléments x_i et y_i appartiennent à une échelle linéairement ordonnée (e.g., $[0, 9]$), ou à un sous-ensemble fini de cette échelle.

10. Bien que la fonction Minimum n'est pas linéaire, le problème de recherche d'allocation MaxMin peut être écrit sous la forme d'un programme linéaire en nombres entiers (cf. [?]).

9.3.3.1 Ordre Min

L'ordre Min (minimum) [?] peut être défini sur l'ensemble des solutions $\mathcal{Y}$ comme suit :

$$x \succsim_{min} y \Leftrightarrow min(\{x_1, ..., x_n\}) \geq min(\{y_1, ..., y_n\}) \tag{9.12}$$

Il faut noter que toutes les composantes des deux alternatives x et y sont *comparables* (i.e., $\forall x, y \in \mathcal{Y}, x \succsim_{min} y$ ou $y \succsim_{min} x$). L'inconvénient est que cet opérateur génère beaucoup de cas d'indifférence (i.e., $x \sim_{min} y$) [11]. Par ailleurs, la maximisation de la fonction Min peut conduire à des solutions (décisions) non-efficaces (voir l'exemple 19).

Exemple 19. *Soient quatre alternatives x, y, z et t évaluées sur trois critères, telles que $x = \langle 9, 8, 1\rangle, y = \langle 1, 3, 9\rangle, z = \langle 5, 6, 6\rangle$ et $t = \langle 5, 7, 8\rangle$. On remarque que les deux alternatives z et t maximisent la fonction Min. Cependant, l'alternative z n'est pas Pareto-optimale, car elle est dominée par la solution t.*

Dans la littérature on trouve deux améliorations notables de l'opérateur Min, à savoir les ordres Discrimin et Leximin (c.f., [?, ?]), qui permettent de faire la distinction entre deux vecteurs ayant la même valeur minimale.

9.3.3.2 Ordre Discrimin

L'ordre Discrimin [?, ?] entre deux alternatives de tailles égales, consiste à appliquer la fonction Min sur les deux vecteurs, après avoir ôté les composantes identiques ayant le même rang (i.e., pour le même critère). Formellement,

$$x \succsim_{disc} y \Leftrightarrow min_{i \in D(x,y)}(x_i) \geq min_{i \in D(x,y)}(y_i) \tag{9.13}$$

où $D(x, y) = \{i | x_i \neq y_i\}$.

Le Discrimin est un raffinement de l'ordre Min et de l'ordre de Pareto. La partie asymétrique de cet ordre (i.e., $\succ_{disc}$) est une relation irréflexive et transitive, par contre, sa partie symétrique (i.e., $\sim_{disc}$) n'est pas transitive.

Exemple 20. *Soient les deux alternatives $x = \langle 5, 1, 4\rangle$, $y = \langle 6, 1, 3\rangle$. On a,*

- *$x \sim_{min} y$, cependant $x \succ_{disc} y$.*
- *x et y sont incomparables suivant l'ordre de Pareto, alors que $x \succ_{disc} y$.*

9.3.3.3 Leximin

L'ordre Leximin [?] est proposé comme un raffinement de l'ordre Discrimin. Il est également basé sur l'idée d'éliminer les éléments égaux, mais une fois que chaque alternative a été triée [12] par ordre non-décroissant (problème de maximisation). Leximin applique un ordre lexicographique sur les vecteurs triés des alternatives. Une définition formelle de l'ordre Leximin [13] est donnée comme suit :

11. Lorsque $x \succsim_{op} y$ et $y \succsim_{op} x$, on écrit $x \sim_{op} y$ tel que $op =$ Min, Discrimin ou Leximin.
12. Cela suppose que les critères sont à priori *commensurables*.
13. Il est possible de définir de façon analogue le leximax.

$$x \succ_{leximin} y \Leftrightarrow \exists k \leq n, \forall i \in \{1, ..., k-1\}, x_{(i)} = y_{(i)} \text{ et } x_{(k)} > y_{(k)}, \quad (9.14)$$

où $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$, la même chose pour y.

Les alternatives x et y sont dites indifférentes ($\sim_{leximin}$) si les vecteurs triés sont identiques. Pour illustrer le principe de cet ordre, on considère l'exemple 21 ci-dessous.

Exemple 21. *Soient les deux alternatives suivantes :*
— $x = \langle 2, 3, 5, 6, 8 \rangle$, *et*
— $y = \langle 2, 6, 3, 4, 8 \rangle$.
Suivant l'ordre DISCRIMIN, *comparer x et y revient à comparer x' et y' tel que*
— $x' = \langle 3, 5, 6 \rangle$, *et*
— $y' = \langle 6, 3, 4 \rangle$.
car $x_1 = y_1 = 2$ et $x_5 = y_5 = 8$, et donc on supprime les éléments $\{2, 8\}$ des deux vecteurs x et y, ce qui permet d'avoir $x \sim_{disc} y$ (lorsque $x' \sim_{disc} y'$). Toutefois, $x \succ_{leximin} y$ car $x_{(1)} = y_{(1)} = 2$, $x_{(2)} = y_{(2)} = 3$ et $x_{(3)} > y_{(3)}$ (i.e., $5 > 4$).

LEXIMIN permet de distinguer entre les alternatives indifférentes suivant l'ordre DISCRIMIN. En outre, l'ordre LEXIMIN ne contredit pas une préférence stricte exprimée par l'ordre DISCRIMIN, i.e., si $x \succ_{disc} y$ alors $x \succ_{leximin} y$. On souligne que [?] ont proposé un modèle linéaire de l'agrégateur LEXIMIN qui a des propriétés d'équité (ou d'équilibre) que nous introduisons ci-dessous.

Équilibre sans compensation entre les critères : L'ordre LEXIMIN n'autorise aucune concession sur la composante minimum des alternatives, et ce même si une légère diminution de cette composante minimum permettait d'augmenter de manière considérable les utilités des autres composantes [?]. Ceci met en évidence la grande sensibilité du préordre LEXIMIN vis-à-vis des composantes faibles du vecteur objectif. Par exemple, $\langle 10, 10, 10, 10 \rangle \succ_{leximin} \langle 7, x, y, z \rangle$ quelles que soient les valeurs de x, y, et z, avec $7 \leq x \leq y \leq z$. Cette absence de compensation justifie le choix d'autres méthodes multicritères, autorisant une compensation entre les composantes d'une alternative (e.g., OWA, Intégrale de CHOQUET).

La recherche de solutions équilibrées a été abordée par d'autres critères comme : COURNOT-NASH [?], ENVY-FREENESS (absence d'envie) [?], DEVIATION [?], etc). Par exemple, dans le cadre de la théorie des jeux dite à *stratégies mixtes*[14], on cherche souvent un équilibre appelé équilibre de NASH. Cet équilibre caractérise une situation où "*aucun joueur n'a intérêt à dévier individuellement de sa stratégie*"[15]. Cependant ce critère n'assure pas toujours le principe d'efficacité, et conduit parfois à des équilibres non PARETO-optimaux (e.g., dilemme du prisonnier). Il est intéressant de noter que dans le contexte des problèmes de partage équitable[16] avec

14. Dans un jeux à stratégie mixte une probabilité est attribuée à chacune des *stratégies pures* d'un joueur. Ainsi, cette transformation en *stratégies mixtes* permet de garantir (par le biais du théorème de NASH [?]) l'existence d'au moins un point d'équilibre.

15. De façon similaire, on peut interpréter ce critère comme une situation dans laquelle les joueurs n'ont pas de regrets après avoir vu les stratégies des autres joueurs.

16. Une discussion plus détaillée sur les problèmes de partage équitable peuvent être trouvées dans [?].

le critère d'absence d'envie (*envy-freeness*), la solution optimale n'est pas nécessairement PARETO-optimale. De plus, le problème de la recherche d'une allocation avec un minimum possible d'envie (*envy*) n'est (en général) pas solvable ou même approximable en temps polynomial (cf. [?]).

9.3.3.4 MINIMUM augmenté

Il n'est pas toujours suffisant de se focaliser sur les pires des cas (composantes minimales) pour assurer l'équité. En effet, on peut avoir des alternatives très différentes, alors qu'elles sont indiscernables suivant l'ordre MIN (e.g., $\langle 1,1,1,1\rangle$ et $\langle 1,5,5,5\rangle$). Ce comportement est connu sous le nom d'*effet de noyade*. Pour remédier à cet effet indésirable, on considère un raffinement du MIN avec une somme (pondérée éventuellement), appelé MINIMUM *augmenté* :

$$x \succsim_{min_+} y \Leftrightarrow min(\{x_1,...,x_n\}) + \epsilon \sum_{i=1}^{n} x_i \geq min(\{y_1,...,y_n\}) + \epsilon \sum_{i=1}^{n} y_i \quad (9.15)$$

où ϵ représente une valeur réelle strictement positive et suffisamment petite. Ce nouveau critère peut être considéré comme une agrégation lexicographique de la fonction égalitariste MIN (en priorité), et la fonction utilitariste SOMME. On Précise que le MINIMUM augmenté peut aussi être vu comme un cas particulier de la distance de Tchebycheff (augmentée et non pondérée) au point idéal $\langle 1,1,...,1\rangle$ dans $[0,1]^n$ (voir la section 9.3.4).

Exemple 22. *Les deux alternatives* $x = \langle 1,2,8,4\rangle$, $y = \langle 1,6,4,2\rangle$ *sont indiscernables par le* MIN *et le* DISCRIMIN. *Par contre, on a* $x \succ_{min_+} y$ *car* $\sum_{i=1}^{n} x_i = 15$ *et* $\sum_{i=1}^{n} x_i = 13$.

On note que le MINIMUM augmenté ne résout pas complètement le problème d'effet de noyade. Par exemple, les deux alternatives $x = \langle 1,1,1,13\rangle$ et $y = \langle 1,5,5,5\rangle$ restent toujours indifférentes suivant ce critère. Pour cela, on peut suggérer une autre possibilité comme le LEXIMIN. Ainsi, l'alternative y est préférée par rapport à x suivant le critère LEXIMIN (i.e., $y \succ_{leximin} x$).

9.3.4 La norme de Tchebycheff

La norme de Tchebycheff permet de caractériser des solutions réalisables équilibrées. On appelle point idéal le vecteur des valeurs optimales pour chaque critère. Ce dernier ne correspond pas nécessairement à une solution réalisable. Nous adoptons la formulation donnée par Dubus [?].

Définition 29 (Point idéal et point Nadir)**.**

— *Le point idéal* $I = (I_1,...,I_n) \in \mathbb{R}^n$ *est un vecteur tel que :*

$$\forall i \in 1..n, I_i = max\{y'_i | y' \in \mathcal{Y}\}.$$

— *Soit* $\forall j \in 1..n, A^*_j = argmax_{a\in\mathcal{A}} u_j(a)$ *et* $(y^j_1,...,y^j_n) = u(A^*_j)$. *Le point Nadir selon* u *est le vecteur* $N = (N_1,...,N_m) \in \mathbb{R}^m$ *vérifiant* $\forall j \in 1..n, N_j = min_{i\in 1..n} y^i_j$.

Définition 30 (Norme de Tchebycheff)**.** *Soient I le point idéal, et N le point Nadir. Une norme de Tchebycheff est un agrégateur défini par :*

$$Tcheb(y_1, ..., y_n) = -max_{j\in 1..n}\frac{I_j - y_j}{I_j - N_j}.$$

Une solution y est préférée à une solution y' au sens de Tchebycheff si et seulement si

$$Tcheb(y) \geq Tcheb(y').$$

Le point idéal définit les coordonnées du meilleur point que l'on pourrait atteindre. Comme expliqué dans [?], le point Nadir identifie une zone de l'espace contenant les solutions intéressantes. La valeur $I_j - y_j$ définit la distance par rapport au point idéal. La division sur $I_j - N_j$ est une simple mise à l'échelle. D'autre part, les valeurs $\frac{I_j - y_j}{I_j - N_j}$ reflète une sorte de *regret* de choisir la valeur y_i. La norme de Tchebycheff d'une solution renvoie le minimum des regrets. Quand les regrets sont pondérés, on obtient la norme de Tchebycheff pondérée.

La proposition suivante (voir par exemple [?]) montre la compatibilité de la dominance de Tchebycheff par rapport à celle de Pareto.

Proposition 6. *La norme de Tchebycheff est compatible avec la dominance faible de Pareto :*

$$\forall y \in \mathcal{Y}, y' \in \mathcal{Y}, y \succsim_p y' \Longrightarrow Tcheb(y) \geq Tcheb(y').$$

Steuer et Choo [?] ont proposé une procédure interactive pour trouver des solutions non-dominées à des programmes mathématiques avec plusieurs objectifs en utilisant la norme de Tchebycheff pondérée. Cette procédure a été améliorée dans [?].

Par ailleurs, la norme de Tchebycheff (pondérée) *augmentée* constitue un raffinement intéressant de la norme de Tchebycheff (pondérée). L'idée consiste à hybrider cette norme avec la fonction Somme en faisant une agrégation lexicographique. Cette idée est très similaire à l'idée du Minimum augmenté développée dans la section 9.3.3.4.

9.3.5 Opérateurs OWA

La fonction d'agrégation OWA (*Ordered Weighted Average*) [?, ?], est une généralisation de la fonction *moyenne*. OWA permet de découvrir des solutions non-supportées en cas de non-convexité du front de Pareto. Cette fonction est définie de $\mathbb{R}^n$ dans $\mathbb{R}$, où n est la taille du vecteur sur lequel elle opère. L'idée est de pondérer les critères relativement à leur rang.

$$U(y) = O_w(y) = \sum_{i=1}^{n} w_i y_{\sigma(i)} \tag{9.16}$$

avec $y \in \mathcal{Y}$, $w = \langle w_1, \ldots, w_n\rangle \in [0,1]^n$ et $\sum_i w_i = 1$.

On note $\langle y_{\sigma(1)}, \ldots, y_{\sigma(n)}\rangle$ le vecteur $\langle y_1, \ldots, y_n\rangle$ obtenu par réarrangement des éléments du vecteur par ordre non-décroissant (i.e., $y_{\sigma(1)} \leq \cdots \leq y_{\sigma(i)} \leq \cdots \leq y_{\sigma(n)}$).

Le poids w_1 est associé à la plus petite valeur $y_{\sigma(1)}$, alors que le poids w_n est associé à la plus grande valeur $y_{\sigma(n)}$. Avec OWA il est possible d'exprimer les opérateurs d'agrégation suivants :

— *Le minimum :* $\text{MIN}(y_{\sigma(1)}, \ldots, y_{\sigma(n)})$, avec $w = \langle 1, 0, \ldots, 0\rangle$.
— *Le maximum :* $\text{MAX}(y_{\sigma(1)}, \ldots, y_{\sigma(n)})$, avec $w = \langle 0, 0, \ldots, 1\rangle$.
— *La moyenne :* $\text{AVG}(y_{\sigma(1)}, \ldots, y_{\sigma(n)})$, avec $w_1 = \cdots = w_n = 1/n$.
— La moyenne en ignorant la meilleure composante et la plus mauvaise composante, i.e., avec $w_1 = 0, w_2 = 1/(n-1), \ldots, w_{n-1} = 1/(n-1), w_n = 0$.
— *La médiane :* $(y_{(n/2)} + y_{(n/2)+1})/2$, si $w_{(n/2)} = w_{(n/2)+1} = 1/2$, quand n est pair.
— *La médiane :* $y_{(n+1)/2}$, si $w_{(n+1)/2} = 1$, quand n est impaire.

Des formulations en programmation linéaire de OWA ont été proposées par Ogryczak et Sliwinnski [?].

On note que OWA peut être vue comme une mesure d'inégalité plus compensatoire utilisée dans la théorie du choix social. En effet, lorsque les éléments du vecteur de pondération w sont strictement décroissants (i.e., $w_i > w_i + 1, i = 1..n-1$), et quand les différences de $w_i - w_{i+1}$ ont tendance à être arbitrairement grandes, alors la fonction OWA tend à représenter le critère LEXIMIN [17] (cf. [?]).

[?] ont proposé l'opérateur OWA garantissant le principe d'équité entre les solutions. Ogryczak et al. [?] ont exploité le modèle équitable OWA pour résoudre un problème d'allocation équitable de ressources pour le dimensionnement d'un réseau. Kostreva et al. [?] ont fait une synthèse intéressante sur des modèles linéaires des agrégateurs équitables OWA et LEXIMIN.

En outre, il est intéressant de noter que OWA peut être reformulée en fonction du vecteur de Lorenz [?]. On a alors :

$$O_w(y) = \omega \cdot L(y) \tag{9.17}$$

où $\omega = \langle w_1 - w_2, w_2 - w_3, ..., w_{n-1} - w_n, w_n\rangle$ est un vecteur de pondération positif ; $L(y) = \langle L_1(y), ..., L_n(y)\rangle$, est le vecteur de Lorenz associé à y défini par $L_k(y) = \sum_{i=1}^{k} y_{\sigma(i)}$ (i.e., l'utilité des k agents les plus pauvres), et $\sigma(i)$ est la permutation triant les y_i par ordre non-décroissant.

Exemple 23. *Étant données trois alternatives ayant les vecteurs d'utilité suivants :*

— $x = \langle 7, 8, 9\rangle$
— $y = \langle 5, 8, 10\rangle$
— $z = \langle 15, 11, 2\rangle$

Les vecteurs de Lorenz sont donnés par :

— $L(x) = \langle 7, 15, 24\rangle$
— $L(y) = \langle 5, 13, 23\rangle$
— $L(z) = \langle 2, 13, 28\rangle$

Il est facile de remarquer que $L(x)$ domine $L(y)$ au sens de PARETO (i.e., $L(x) \succ_P L(y)$), et donc $x \succ_L y$. On remarque aussi que le vecteur $L(z)$ est incomparable avec les vecteurs $L(x)$ et $L(y)$ au sens de PARETO, et donc l'alternative z est incomparable avec les alternatives x et y au sens de Lorenz.

17. Dans le cas où les différences entre les utilités peuvent tendre vers 0, la représentation de l'ordre LEXIMIN à l'aide d'OWA n'est pas possible.

Or, si on utilise la formulation OWA (9.17) avec le vecteur de pondération $w = \langle \frac{6}{9}, \frac{2}{9}, \frac{1}{9} \rangle$*, on obtient :*

$$\begin{aligned} O_w(x) &= \frac{(6-2)}{9} \times 7 + \frac{(2-1)}{9} \times 15 + \frac{1}{9} \times 24 \\ &= \frac{28+15+24}{9} \\ &= \frac{67}{9} \\ O_w(y) &= \frac{56}{9} \\ O_w(z) &= \frac{49}{9} \end{aligned}$$

Ce qui entraîne l'ordre total de préférence : $x \succ y \succ z$

Le critère OWA est manifestement plus riche que la dominance de Lorenz. Cependant, OWA impose le choix d'un jeu de poids qui peut s'avérer difficile à déterminer sans informations préférentielles.

Contrainte Orness En raison de l'importance du degré de compensation d'OWA, une mesure a été définie pour évaluer le seuil de compensation d'un vecteur de pondération. Cette mesure est connue sous le nom d'Orness. Plus cette mesure est grande, plus la compensation est importante, et inversement. Cette contrainte est formellement définie par :

$$\sum_{i=1}^{n} w_i(n-i) = (n-1) \times OC \tag{9.18}$$

où $OC \in \mathbb{I} = [0,1]$ est la valeur de la mesure d'Orness. Par exemple, les valeurs proches de 1 correspondent au minimum alors que les valeurs proches de 0 correspondent au maximum, et $Orness_{AVG} = 1/2$.

Contrainte d'entropie Une autre propriété intéressante caractérisant l'opérateur OWA, est l'entropie H. Ce facteur reflète le taux d'information impliqué dans le calcul de la valeur agrégée. L'entropie est exprimée par la contrainte suivante :

$$H(w) = -\sum_{i=1}^{n} w_i ln(w_i) \tag{9.19}$$

Si pour un certain rang k, $w_k = 1$ et $w_i = 0, i \neq k$ alors $H(w) = 0$ (le minimum de l'entropie), ce qui signifie que OWA utilise un minimum de composantes (critères) dans le calcul de la valeur agrégée. Par contre, si $w_i = 1/n, i = 1, \ldots, n$ alors $H(w) = ln(n)$ prend la valeur maximale, ce qui signifie que OWA utilise un maximum de critères dans le calcul de la valeur agrégée.

9.3.6 Intégrale de CHOQUET

L'intégrale de CHOQUET [?, ?] recouvre toute une gamme de fonctions d'agrégation (e.g., MIN, MAX, OWA, etc.) dans laquelle les poids dépendent non seulement du rang des critères (comme dans la moyenne pondérée ordonnée), mais aussi d'une *mesure floue* (appelée *capacité*). Cette mesure permet d'étendre la notion de poids à un sous-ensemble de critères, afin d'exprimer leur degré d'interaction.

Définition 31 (Capacité). *Étant donnée une suite $N = \{1, ..., n\}$ désignant un ensemble de critères. Une capacité est une fonction $\mu : 2^N \to [0, 1]$, telle que $\mu(\emptyset) = 0, \mu(N) = 1$ et $\forall A, B \in 2^N, A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$. La capacité est dite additive si $\mu(S \cup T) = \mu(S) + \mu(T)$ pour n'importe quels deux sous-ensembles disjoints $S, T \subseteq N$. La capacité est dite basée sur la cardinalité si et seulement si pour tout $T \subseteq N$, $\mu(T)$ dépend seulement de la cardinalité de T.*

Définition 32 (Intégrale de CHOQUET). *Étant donnée une capacité μ, l'intégrale de CHOQUET (voir par exemple [?]) est une intégrale, par rapport à une capacité μ sur une solution réalisable $y = \langle y_1, y_2, ..., y_n \rangle$, est donné par :*

$$U(y) = C_\mu(y) = \sum_{i=1}^{n} \left[y_{\sigma(i)} - y_{\sigma(i-1)} \right] \mu(X_{\sigma(i)}) \tag{9.20}$$

où $\sigma(.)$ est une permutation sur $\{1, ..., n\}$ telle que $0 = y_{\sigma(0)} \leq y_{\sigma(1)} \leq y_{\sigma(2)} \leq ... \leq y_{\sigma(n)}$; $X_{\sigma(i)} = \{j \in N | \ y_j \geq y_{\sigma(i)}\} = \{\sigma(i), \sigma(i+1), ..., \sigma(n)\}$, $\forall i \in N$ et $X_{\sigma(n+1)} = \emptyset$.

Dans le contexte de l'agrégation, $\mu(A)$ peut être vu comme étant l'importance du sous-ensemble $A \subseteq N$ de critères dans la décision. On note que $X_{\sigma(i+1)} \subseteq X_{\sigma(i)}$ et donc $\mu(X_{\sigma(i+1)}) \leq \mu(X_{\sigma(i)})$, $\forall i \in N$. Ainsi, les poids w_i sont toujours positifs ou nuls. Lorsque la capacité μ est *additive* sur N, i.e., $\forall A, B \in 2^N$ tels que $A \cap B = \emptyset$ on a $\mu(A \cup B) = \mu(A) + \mu(B)$, l'intégrale de CHOQUET correspond à une moyenne pondérée [?] avec $w_i = \mu(\{X_{\sigma(i)}\})$.

Exemple 24. *L'intégrale de CHOQUET sur l'alternative $y = \langle 7, 3, 20 \rangle$ est donnée par :*

$$C_\mu(a) = 3\mu(\{1, 2, 3\}) + (7 - 3)\mu(\{1, 3\}) + (20 - 7)\mu(\{3\})$$

L'intégrale de CHOQUET est utilisée lorsqu'on veut modéliser à la fois l'importance de chaque critère, et l'interaction entre les critères. Considérons, par exemple, le cas d'un choix entre quatre alternatives a, b, c et d évaluées sur deux critères (voir la figure 9.6). Suivant le critère de dominance, il est clair que la solution d est meilleure que la solution a. Néanmoins, le choix entre les autres paires de solutions dépend des préférences du décideur. A travers cette figure trois situations d'interaction peuvent être distinguées.

1. **Redondance :** La satisfaction de l'un des deux critères suffit pour qu'une alternative soit satisfaisante (voir la figure 9.6 - (a)). On a $\mu(\{1\}) = \mu(\{2\}) = 1$, et $\mu(\{1, 2\}) < \mu(\{1\}) + \mu(\{2\})$.

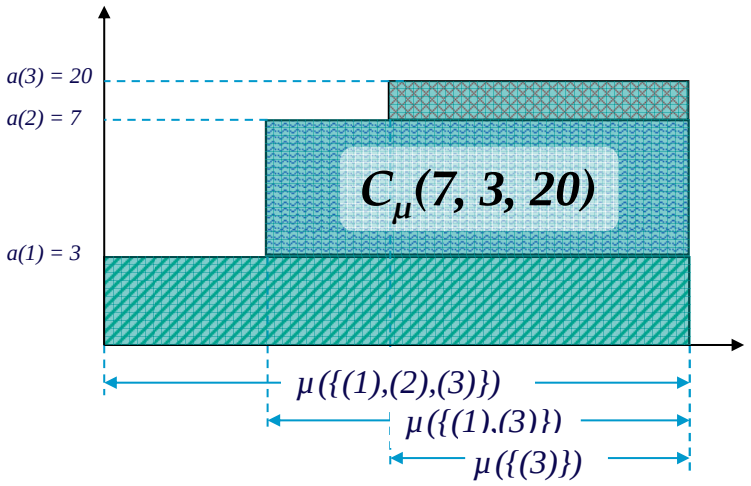


FIGURE 9.5 – L'intégrale de CHOQUET sur l'alternative $a = \langle 7, 3, 20\rangle$ de l'exemple 24, calculée avec la formule (32).

2. **Complémentarité :** Pour qu'une alternative soit globalement satisfaisante, il faut qu'elle soit bonne sur les deux critères à la fois (voir la figure 9.6 - (b)). Ce qui équivaut à $\mu(\{1\}) = \mu(\{2\}) = 0$, et $\mu(\{1, 2\}) > \mu(\{1\}) + \mu(\{2\})$.
3. **Indépendance :** Une situation d'absence d'interaction signifie que chaque critère apporte sa propre contribution dans l'évaluation de la satisfaction globale de l'alternative, indépendamment des autres critères (voir la figure 9.6 - (c)). Ce comportement se traduit par $\mu(\{1, 2\}) = \mu(\{1\}) + \mu(\{2\})$.

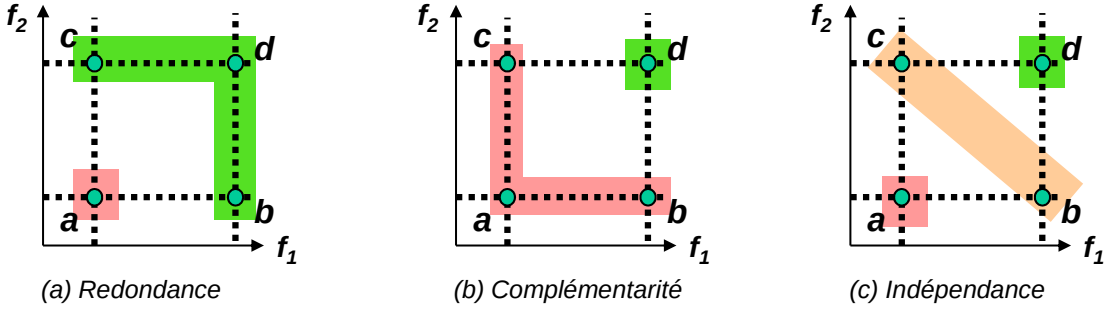


FIGURE 9.6 – Différents cas d'interaction : (*a*) $b \sim c \sim d \succ a$, (*b*) $d \succ a \sim b \sim c$ et (*c*) $d \succ b \sim c \succ a$.

L'inconvénient de l'intégrale de CHOQUET est qu'elle nécessite (dans le cas général) la connaissance de $2^n - 2$ coefficients.

Exemple 25. *Ici, on voudrait repérer les critères importants et ceux qui sont négligeables. Soit par exemple [?] les valeurs de capacité :*

A	1	2	3
$\mu(A)$	0	0.2	0.2
A	$\{1, 2\}$	$\{1, 3\}$	$\{2, 3\}$
$\mu(A)$	0.8	0.8	0.4

A partir de $\mu(1) = 0$, on peut déduire que le critère 1 est inutile. Par contre, le critère 1 devient important une fois combiné avec les autres critères, notamment les critères 2 et 3. On se pose la question sur comment mesurer le degré d'importance d'un critère ? Shapley (cf., [?]) a montré que la seule définition possible est :

$$\phi(i) = \sum_{A \subseteq N \setminus i} \frac{(n-a-1)!a!}{n!}[\mu(A \cup \{i\}) - \mu(A)]$$

avec $a = |A|$, le cardinal de A. En appliquant cet indice à l'exemple, on obtient : $\phi(1) = 0.4$, $\phi(2) = \phi(3) = 0.3$. Avec un tel indice, on peut constater que des capacités différentes peuvent avoir les mêmes indices d'importance. On se pose ainsi la question : comment distinguer deux capacités ayant le même indice d'importance ? La réponse à cette question réside dans la définition d'un nouveau indice, celui de l'interaction. En reprenant l'exemple 25, on constate que les deux critères 1 et 2, pris individuellement, ne sont pas importants ; par contre, leur réunion est importante. On dit qu'il y a un phénomène de synergie [?] entre ces deux critères ; les deux critères sont complémentaires. On peut définir la quantité de synergie entre deux critères i et j par la formule $\mu(\{i,j\}) - \mu(\{i\}) - \mu(\{j\})$. Dans notre exemple, la synergie entre 1 et 2 est 0.6, alors qu'entre 2 et 3, elle est nulle. On pourrait donc avoir le phénomène inverse où les critères sont individuellement importants, mais leur réunion ne serait pas plus importante. On parle dans ce cas, de critères redondants ou substituables.

L'indice d'interaction [?] entre deux critères i et j est la moyenne de la quantité de synergie entre i et j en présence d'un groupe A, pour tous les groupes A possibles :

$$I_{i,j} = \sum_{A \subseteq N \setminus \{i,j\}} \frac{(n-a-2)!a!}{n!}[\mu(A \cup \{i,j\}) - \mu(A \cup \{i\}) - \mu(A \cup \{j\}) + \mu(A)].$$

Dans l'exemple 25, on a $I_{1,2} = I_{1,3} = 0.3$ alors que $I_{2,3} = -0.3$ montrant que la synergie entre les critères 2 et 3 est négative. Avec ces définitions d'indices d'importance et d'interaction, il a été constaté qu'en pratique [?], il est difficile d'avoir des capacités ayant les mêmes indices d'importance et d'interaction. On peut démontré que pour $n = 2$, il est impossible d'avoir deux capacités ayant les mêmes indices. Pour $n > 2$, si les capacités ont les mêmes indices, on pourrait les distinguer en faisant appel à la synergie entre 3 critères

$$\mu(\{i,j,k\}) - \mu(\{i,j\}) - \mu(\{i,k\}) - \mu(\{j,k\}) + \mu(\{i\}) + \mu(\{j\}) + \mu(\{k\}).$$

Pour $n \leq 3$, on ne peut pas avoir deux capacités ayant les mêmes indices entre 2 et 3 critères. D'une façon plus générale : étant donné un problème à n critères, une capacité est déterminée de façon unique par ses indices d'importance et d'interaction entre 2, 3 et jusqu'à n critères.

La puissance de l'intégrale de CHOQUET qui réside dans sa prise en charge précise de l'interaction entre les critères, est pénalisée par la nécessité de définir $2^n - 2$ paramètres. Cependant, il existe des capacités particulières qui nécessitent moins de coefficients. C'est le cas tout particulièrement des capacités k-additives. Une capacité est dite k-additives, si tous ses indices d'interaction sont nuls à partir de k critères. Ainsi, dans une capacité 1-additive, l'intégrale de CHOQUET correspond à une simple somme pondérée. Une capacité 2-additives nécessite seulement $\frac{n(n+1)}{2} - 1$ coefficients.

En pratique [?], il a été constaté que l'on gagne peu en précision en passant d'une capacité 2-additives à une capacité supérieure.

9.4 Approches d'élicitation

Sans être exhaustif, nous exposons un ensemble de méthodes d'élicitation relatives aux différentes composantes de la décision multicritère, et tout particulièrement des méthodes d'agrégation. La littérature de l'élicitation s'est énormément développée récemment. Cependant, les références exposées dans cette section sont parmi les plus souvent citées.

UTA et élicitation des fonctions d'utilité dans un modèle additif

La méthode UTA (UTilité Additives) de Jacquet-Lagrèze et Siskos [?] a été proposée pour construire les fonctions d'utilité additive de la forme $U(a) = \sum_{i=1..n} u_i(a^i)$, où u_i est continue, non-décroissante et affine par morceaux dans l'intervalle $[t_i, T_i]$ de variation de l'attribut $a^i : \forall i \in 1..n, a \in \mathcal{A}, t_i \leq a^i \leq T_i$. En raison du fait que cette méthode a été à l'origine de plusieurs méthodes d'élicitation, nous exposons son modèle linéaire en détail. UTA procède en plusieurs étapes :

1. Soit B un sous-ensemble d'alternatives. Pour toute paire $a, b \in B$, nous avons un des deux cas :
   (a) a est préférée à b : $a \succ b$,
   (b) a est indifférente de b : $a \sim b$.
2. Subdiviser uniformément l'intervalle $[t_i, T_i]$ en $\pi_i - 1$ sous-intervalles $[z_{i,j}, z_{i,j+1}]$, avec $j = 1..\pi_i - 1$, $i = 1..n$, $z_{i,1} = t_i, z_{i,\pi_i} = T_i$.
3. Pour toute composante $b_i (i = 1..n)$ du vecteur $b \in B$, il existe j tel que $b_i \in [z_{i,j}, z_{i,j+1}]$. $u_i(b_i)$ est approximée par

$$u_i(b_i) = u_i(z_i) + \frac{u_i(z_{i,j+1}) - u_i(z_{i,j})}{z_{i,j+1} - z_{i,j}}(b^i - z_{i,j}).$$

   Pour tout $b \in B$, $U(b)$ est évaluée comme suit :

$$U(b) = \sum_{i=1..n} u_i(b^i) + \epsilon_b$$

   où ϵ_b est une erreur d'approximation.
4. Nous avons les contraintes linéaires issues des préférences :

$$\begin{aligned} a \sim b &\Leftrightarrow U(a) - U(b) = 0 \\ a \succ b &\Leftrightarrow U(a) - U(b) \geq \delta, \end{aligned}$$

   où δ est une valeur positive, représentant une valeur d'écart.

5. Notons $u_{i,j} = u_i(z_{i,j}), i = 1..n, j = 1..\pi_i$. On obtient finalement le PL suivant :

$$\begin{cases} min & \sum_{b \in B} \epsilon_b \\ s.t. & \\ & \sum_{i=1..n}[u_i(a^i) - u_i(b^i)] + \epsilon_a - \epsilon_b \geq \delta, \forall a, b \in B, a \succ b \\ & \sum_{i=1..n}[u_i(a^i) - u_i(b^i)] + \epsilon_a - \epsilon_b = 0, \forall a, b \in B, a \sim b \\ & u_{i,j+1} \geq u_{i,j}, \forall i = 1..n, j = 1..\pi_i \\ & u_{i,1} = 0, \forall i = 1..n \\ & \sum_{i=1..k} u_{i,\pi_i} = 1 \\ & u_{i,j} \geq 0, \forall i \in 1..n, j = 1..\pi_i \\ & \epsilon_a \geq 0, \forall a \in B \end{cases} \quad (9.21)$$

L'objectif est donc de construire les fonctions d'utilité individuelles u_i (et donc des relations de préférence associées aux critères) reflétant au mieux la préférence globale disponible sur les alternatives B. C'est pour cette raison que cette méthode est dite "par désagrégation". La solution du programme linéaire final (9.21) fournira les valeurs des utilités individuelles des alternatives minimisant l'erreur, et donc s'approchant le mieux de la forme additive supposée.

Plusieurs variantes (e.g., UTASTAR, UTAMP1, UTAMP2, Stochastic UTA, etc.) [?] de UTA ont été proposées dans la littérature. Il existe des variantes modélisant autrement la relation de préférence globale ou posant d'autres fonctions objectifs du programme linéaire. On peut citer des modèles utilisant le critère de Kendall, ou encore supposant un ordre lexicographique. UTA a aussi été à l'origine de plusieurs méthodes par désagrégation, comme par exemple PREFCALC. La littérature abondante sur UTA est suffisamment exposée dans [?]. Il existe plusieurs travaux récents sur UTA, notamment celui de [?] proposant une reformulation de UTA résolue avec un algorithme par séparation/évaluation donnant de bons résultats si le nombre de critères est largement inférieur à celui des variables. UTA^{GMS} [?, ?] est une extension de UTA, qui considère notamment plusieurs types de fonctions additives, et les fonctions d'utilité des critères sont considérées monotone non-décroissantes (alors que UTA les suppose linéaires par morceaux). La méthode GRIP [?] étend UTA^{GMS} en prenant en compte des informations préférentielles additionnelles sur l'intensité des préférences entre des paires d'alternatives.

AHP et élicitation hiérarchique

La démarche AHP (*Analytic Hierarchy Process*) [?] comprend plusieurs étapes. La première étape consiste à décomposer le problème multicritère en sous-problèmes organisés sous une forme hiérarchique. Les feuilles représentent les alternatives et les nœuds internes les critères. Pour évaluer l'importance d'une alternative, une comparaison entre toutes les paires d'alternatives est quantifiée et stockée dans une matrice A. Ainsi, la matrice A contient les élément $A_{i,j}$ quantifiant l'importance relative de l'alternative i par rapport à l'alternative j. Si on suppose que la fonction d'agrégation [?, ?] est une somme pondérée $U(a) = \sum_{i=1..n} w_i a_i$, alors w_i reflète l'importance de a par

rapport au critère i. Les poids w_i sont calculés en exploitant les matrices de comparaisons, en supposant notamment que $A_{i,j} \approx \frac{w_i}{w_j}$. Il a été démontré que le calcul de w peut se ramener au calcul d'un vecteur propre.

KAPPALAB et élicitation de la capacité

L'utilisation de l'intégrale de CHOQUET nécessite les valeurs de la fonction de capacité. Le système KAPPALAB [?] propose une panoplie de méthodes pour éliciter la fonction de capacité. Pour identifier la capacité, nous supposons la disponibilité d'une information préférentielle sur la fonction d'utilité globale ou encore sur la relation de préférence entre un sous-ensembles d'alternatives $\mathcal{O} \subseteq \mathcal{A}$. Les modèles d'élicitation disponibles dans KAPPALAB sont des modèles d'optimisation qui diffèrent suivant la forme de la fonction objectif ou encore suivant la nature de l'information préférentielle. En d'autres termes, on voudrait apprendre la fonction capacité à partir de l'information préférentielle dans $\mathcal{O}$. L'information partielle peut être un ordre partiel entre les alternatives de $\mathcal{O}$, un ordre partiel entre les critères, ou encore un ordre partiel entre les valeurs de la capacité d'un certain nombre de sous-ensembles de critères. Cette information préférentielle est automatiquement traduite en un système de contraintes sur les valeurs des capacités. Bien évidemment, le nombre de variables du modèle est de l'ordre de 2^n, qui devient non réaliste si le nombre de critères est important. C'est pour cette raison que le modèle d'élicitation pour l'intégrale de CHOQUET nécessite une hypothèse de k-additivité des capacités. Suivant la nature du modèle d'élicitation, plusieurs approches de résolution sont proposées, comme par exemple la programmation linéaire, ou encore la méthode des moindres carrées. Pour plus de détails, nous recommandons au lecteur de consulter la référence [?].

Décision et élicitation dans un modèle additif avec des paramètres incertains

Plusieurs démarches sont proposées dans la littérature pour la décision et l'élicitation dans un environnement incertain. Sage et White en 1984 [?] ont initié ISMAUT (*imprecisely specified multiattribute utility theory*) qui est une extension du cadre classique de décision multicritère en supposant des valeurs incertaines au niveau des valeurs d'utilité ou des poids.

La démarche ARIADNE (*Aid based on DomiNance structural information Elicitation*) [?] permet d'inférer la relation de préférence globale en supposant des valeurs d'utilité incertaines en utilisant les intervalles dans un modèle additif pondéré. Chaque valeur incertaine est manipulée sous forme d'un intervalle. Au niveau de chaque paire d'alternatives (a, b), ARIADNE essaye de trouver une pondération montrant la dominance $a \succ b$ en résolvant un programme linéaire. Ainsi, plusieurs programmes linéaires sont résolus inférant les poids sous forme hiérarchique, et générant ainsi la structure de préférence.

Elicitation des utilités avec min/max regret

Wang et Boutillier [?] ont proposé une approche d'élicitation de la fonction d'utilité incomplète dans une démarche min-max regret, que nous exposons brièvement. Cette approche suppose l'existence d'une fonction $Pr_d(s)$ mesu-

rant la probabilité qu'une valeur agrégée (outcome) s soit réalisée quand le système adopte la décision $d \in D$. La fonction d'utilité $u : S \rightarrow [0, 1]$ associe la valeur $u(s)$ pour toute valeur agrégée s. Soit n le nombre de valeurs agrégées dans S, on a $u_i = u(s_i), s_i \in S, i = 1..n$. La valeur agrégée attendue EU d'une décision d par rapport à une fonction d'utilité u est définie comme suit :

$$EU(d, u) = \sum_{s_i \in S} Pr_d(s_i)u_i.$$

Notons que EU est linéaire en u. Sur la base de cette formulation probabiliste, un modèle min-max regret est établi. La démarche nécessite la résolution de $O(|D|^2)$ programmes linéaires résonnant sur toutes les paires de décision. Enfin, cette solution min-max regret a été étendue, dans un cadre itératif et incrémental, pour tenir compte des cas où la solution min-max regret a un niveau de regret non acceptable par l'utilisateur.

Elicitation de l'opérateur d'ordre lexicographique

L'élicitation de l'opérateur lexicographique a été abordée dans les travaux suivants : [?] construisent la relation de préférence en supposant un ordre lexicographique entre les critères ; [?] proposent un algorithme dédié pour éliciter l'ordre entre les critères ; [?] présentent une analyse détaillée sur l'opérateur lexicographique, avec des algorithmes dédiés pour inférer l'ordre entre les critères.

Elicitation de préférences

Le problème de l'élicitation des préférences a été abordé, et concrétisé par de nombreux travaux [?, ?, ?, ?, ?] développés au sein de la communauté d'aide multicritère à la décision. Ces travaux englobent à la fois des algorithmes et des procédures cognitives et itératives de recherche, de découverte, d'acquisition et d'apprentissage des préférences du décideur, et elles sont préconisés comme moyens pour restreindre le domaine de préférences admissibles qui permet de prendre une bonne décision.

Chapitre 10

Problèmes de transport, d'affectation et d'ordonnancement

10.1 Problèmes d'ordonancement [R. Faure, 1995]

On a affaire à des problèmes d'ordonnancement lorsque, en vue de la réalisation d'un objectif quelconque, il faut accomplir un ensemble de tâches (ou opérations), elles mêmes soumises à un ensemble de contraintes. Les contraintes auxquelles sont soumises les diverses tâches qui concourent à la réalisation de l'objectif sont de divers types. On peut en citer :

les contraintes du type potentiel qui sont : des contraintes temporelles. (e.g., la tâche i ne dois pas commencer avant telle date), des contraintes de succession (e.g., la tâche i doit commencer après la tâche j, ...).

les contraintes du type dijonctif qui impose la disjonction entre des tâches, qui ne peuvent être réalisées en simultanément.

les contraintes du type cumulatif qui concerne le non dépassement des ressources disponibles durant la durée du projet.

Plusieurs méthodes sont proposées dans la littérature pour appréhender les différents problèmes d'ordonnancement. Ici, nous allons nous intéresser à des problèmes d'ordonnancement où il existe seulement les contraintes de type potentiel. Les données des problèmes d'ordonnancement abordés ici sont :

— l'ensemble des tâches $T = \{t_1, ..., t_m\}$,
— la durée d_i de chaque tâche t_i,
— l'ensemble des précédences $t_i \rightarrow t_j$ où on exprime que t_i doit s'achever avant t_j.

On exposera les méthodes PERT et MPM qui permettent :

1. d'établir un ordonnancement,
2. de déterminer les tâches critiques, i.e., celles dont l'exécution ne peut être ni retardée, ni ralentie,
3. et de déterminer les marges des tâches non critiques.

On cherche donc :

— un arrangement séquentiel des tâches qui respecte les contraintes de précédence,
— pour chaque tâche t_i, une date de début au plus tôt dd_i,
— pour chaque tâche t_i, une date de fin au plus tard df_i.

10.1.1 Méthode PERT

On fait correspondre à chaque tâche t_i un arc d'un graphe orienté (dit ici digraphe), sa durée d'exécution étant égale au poids de cet arc. Le digraphe reflète les précédences requises dans l'exécution du projet. Ainsi, la tâche correspondant à l'arc (i, j) ne peut commencer que si toutes les tâches correspondant à des arcs (k, i) ont été complétées. Le digraphe peut contenir des tâches fictives de durée nulle afin de forcer certaines précédences. Les sommets du digraphe représentent des événements, début (fin) des activités correspondant aux arcs dont ils sont l'extrémité initiale (finale). Le fait que le digraphe est sans circuit est garant de la faisabilité du projet. En effet, l'existence d'un circuit impliquerait une contradiction dans les précédences : une tâche devant en même temps précéder et succéder une autre ! On supposera dorénavant que les sommets ont déjà été numérotés de 1 à n de manière compatible avec leurs rangs, c'est-à-dire que $r(j) > r(i)$ implique $j > i$ (Dans une arborescence, les sommets à la même distance de la racine sont dits être au même rang. La racine est par convention au rang 0 et la hauteur de l'arbre est le rang maximum.). En plus, si le digraphe possède plusieurs sommets sans prédécesseurs, on supposera avoir introduit un sommet 1 relié par un arc de durée nulle à chacun de ces sommets. Ce sommet indique le début du projet. De même, si le digraphe possède plusieurs sommets sans successeurs, ceux-ci seront reliés par un arc de durée nulle à un dernier sommet n (fin du projet). Enfin, on supposera éliminés les arcs parallèles par l'introduction de tâches fictives.

L'algorithme 3 contient les étapes de la méthode PERT, nécessitant la construction au préalable le graphe de précédences entre les tâches.

Algorithm 3 PERT

Input: Un graphe orienté $G = (V, E)$, sans circuit, des tâches $T = \{t_1, ..., t_m\}$ avec leurs durées $d_1, d_2, ..., d_m$. Soit $P(i) = \{k \in V | ki \in E\}$. Soit $S(i) = \{k \in V | ik \in E\}$.

Output: Dates au plus tôt dd_i, dates au plus tard df_i, et durée du chemin critique des tâches T.

1: $dd_1 \leftarrow 0$
2: **for** $k \leftarrow 2$ to m **do**
3: $\quad dd_k \leftarrow \max\{dd_j + d_{jk} | j \in P(k)\}$
4: **end for**
5: $df_n \leftarrow dd_n$
6: **for** $k \leftarrow m - 1$ downto 1 **do**
7: $\quad dd_k \leftarrow \min\{df_j - d_{kj} | j \in S(k)\}$
8: **end for**

Soit le problème d'ordonnancement :

Tâches	Précédences	Durée (jours)
A	-	3
B	-	9
C	-	5
D	A	8
E	B	4
F	B	7
G	B	20
H	C, F	6
I	D, E	5

Sa solution avec PERT est donnée dans la Figure 10.1.

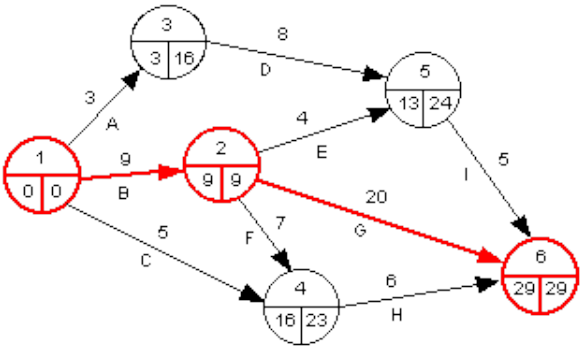


FIGURE 10.1 – Solution PERT du problème d'ordonnancement (http ://www.apprendre-en-ligne.net/graphes/pert/index.html)

10.1.2 Méthode MPM

Contrairement à la méthode PERT, cette méthode ne nécessite aucune définition d'événements. Conceptuellement, elle repose sur un graphe $G = (X, U)$, dans lequel les sommets représentent les tâches du projet (ainsi qu'une tâche début D, et une tâche fin F), et les arcs représentant les contraintes.

Soient les données ci-dessous d'un problème d'ordonancement :

Tâches	Précédences	Durée (mois)
a	-	4
b	a	6
c	-	4
d	-	12
e	b,c,d	10
f	b,c	24
g	a	7

```
h              e,g                        10
i               f,h                         3
-------------------------------------------------
```

Le graphe correspondant est donné dans la Figure 10.2.

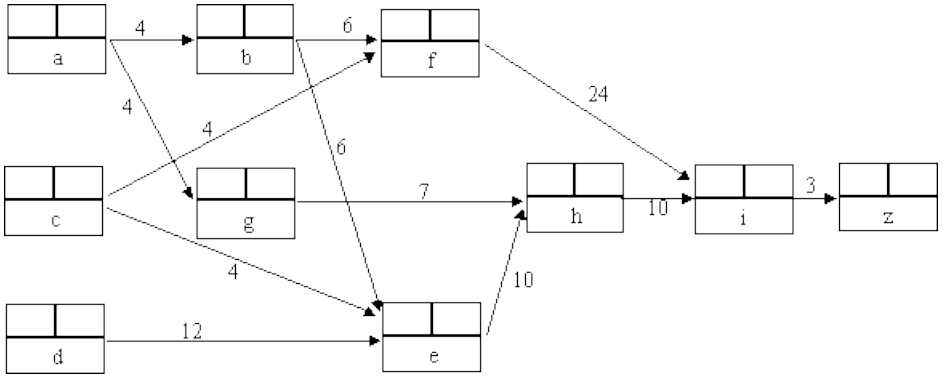


FIGURE 10.2 – Graphe initial MPM (http ://www.iutbayonne.univ-pau.fr/ grau/D/chapitre2.html)

Dates au plus tôt Les dates dd sont calculées avec la formule :

$$dd_i = \max\{dd_j + d_{ji} | j \in P(i)\}, i = 1..m.$$

Dates au plus tard Les dates df au plus tard sont calculées avec la formule :

$$df_m = dd_m,$$

$$df_i = \min\{df_j - d_{ij} | j \in S(i)\}, i = m - 1..1.$$

Marges totales C'est le retard mt_i maximum que l'on peut prendre dans la mise en route d'une tâche t_i sans remettre en cause les dates au plus tard des tâches suivantes.

$$mt_i = df_i - dd_i, , i = 1..m.$$

La marge totale est la différence entre le début au plus tard et le début au plus tôt. On ne peut accorder un délai correspondant à la marge totale que si aucune des tâches précédentes n'a utilisé de marge non libre.

Marges libres C'est le retard maximum ml_i que l'on peut prendre dans la mise en route d'une tâche t_i sans remettre en cause les dates au plus tôt des tâches suivantes.

$$ml_i = \min\{dd_j - dd_i - d_{ij} | j \in S(i)\}, i = m - 1..1.$$

La marge libre d'une tâche est donc le délai pouvant être accordé pour son début sans pénaliser la marge totale des tâches qui suivent.

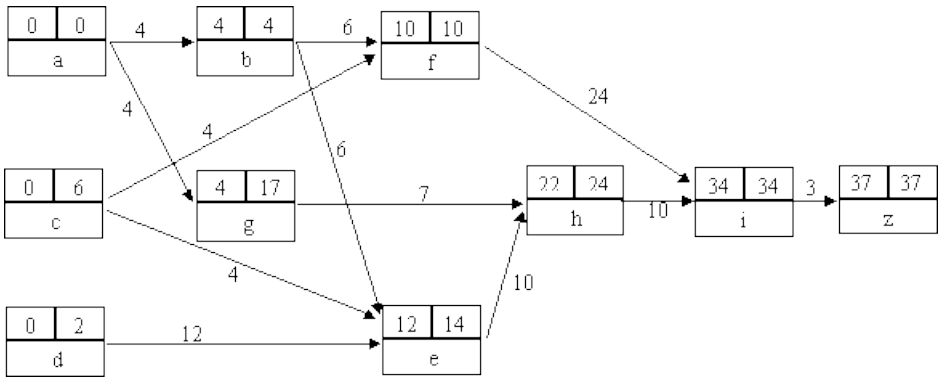


FIGURE 10.3 – Graphe MPM (http ://www.iutbayonne.univ-pau.fr/ grau/D/chapitre2.html)

La Figure 10.3 illuste les dates *dd* et *df* du projet.

Calcul des propriétés de l'ordonnancement :

$dd_a = dd_c = dd_d = 0$
$dd_b = dd_a + 4 = 4$
$dd_g = dd_a + 4 = 4$
$dd_f = Max(dd_b + 6; dd_c + 4) = Max(10; 4) = 10$
$dd_e = Max(dd_b + 6; dd_c + 4; dd_d + 12) = Max(10; 4; 12) = 12$
$dd_h = Max(dd_e + 10; dd_g + 7) = Max(22; 11) = 22$
$dd_i = Max(dd_f + 24; dd_h + 10) = Max(34; 32) = 34$
$dd_z = dd_i + 3 = 37$

$df_i = df_z - V(i, z) = 37 - 3 = 34$
$df_h = df_i - V(h, i) = 34 - 10 = 24$
$df_f = df_i - V(f, i) = 34 - 24 = 10$
$df_e = df_h - V(e, h) = 24 - 10 = 14$
$df_g = df_h - V(g, h) = 24 - 7 = 17$
$df_b = Min[df_e - V(b, e); df_f - V(b, f) = Min[14 - 6; 10 - 6] = 4$
$df_a = Min[df_b - V(a, b); df_g - V(a, g) = Min[4 - 4; 17 - 4] = 0$
$df_c = Min[df_f - V(c, f); df_e - V(c, e) = Min[10 - 4; 14 - 4] = 6$
$df_d = df_e - V(d, e) = 14 - 12 = 2$

$mt(a) = df_a - dd_a = 0 - 0 = 0$
$mt(b) = df_b - dd_b = 4 - 4 = 0$
$mt(c) = df_c - dd_c = 6 - 0 = 6$
$mt(d) = df_d - dd_d = 2 - 0 = 2$
$mt(e) = df_e - dd_e = 14 - 12 = 2$
$mt(f) = df_f - dd_f = 10 - 10 = 0$
$mt(g) = df_g - dd_g = 17 - 4 = 13$
$mt(h) = df_h - dd_h = 24 - 22 = 2$
$mt(i) = df_i - dd_i = 34 - 34 = 0$

$ml(a) = Min[dd_b - dd_a - V(a,b); dd_g - dd_a - V(a,b)] = Min(0;0) = 0$
$ml(b) = Min[dd_f - dd_b - V(b,f); dd_e - dd_b - V(b,e)] = Min(0;2) = 0$
$ml(c) = Min[dd_f - dd_c - V(c,f); dd_e - dd_c - V(c,e)] = Min(6;8) = 6$
$ml(d) = dd_e - dd_d - V(d,e) = 0$
$ml(e) = dd_h - dd_e - V(e,h) = 0$
$ml(f) = dd_i - dd_f - V(f,i) = 0$
$ml(g) = dd_h - dd_g - V(g,h) = 11$
$ml(h) = dd_i - dd_h - V(h,i) = 2$
$ml(i) = dd_z - dd_i - V(i,z) = 0$

Chapitre 11

Modèles stochastiques

Chapitre 12

Théorie des jeux

Chapitre 13

Annexes

13.1 Branch and Price et génération de colonnes

Soit le problème de plus courts chemins avec contraintes [?]. Soit le graphe de la Figure 13.1. En plus du cout c_{ij} attribué à chaque arc, nous disposons aussi d'un temps maximal t_{ij}. L'objectif est de trouver le plus court chemin du noeud 1 au noeud 6 sans excéder 14 unités de temps.

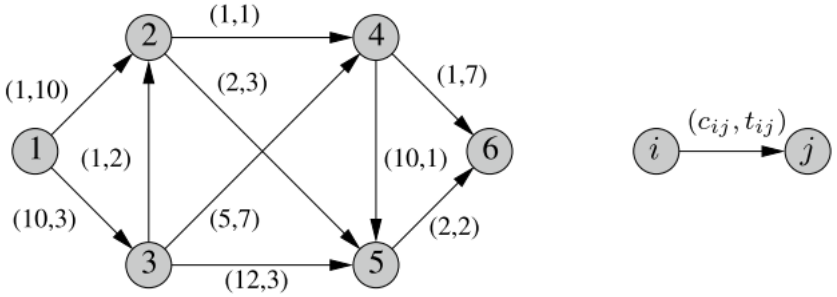


FIGURE 13.1 – Problème de plus courts chemins avec une contrainte de capacité [?]

On peut formaliser ce problème dans le problème IP (13.1) sur le domaine 0-1.

$$\begin{array}{ll} \text{minmisze} & z = \sum_{(i,j)\in A} c_{ij}x_{ij} \\ \text{subject to} & \sum_{j:(1,j)\in A} x_{1j} = 1 \\ & \sum_{j:(i,j)\in A} x_{ij} - \sum_{j:(j,i)\in A} x_{ji} = 0, i = 2,3,4,5 \\ & \sum_{i:(i,6)\in A} x_{i6} = 1 \\ & \sum_{(i,j)\in A} t_{ij}x_{ij} \leq 14 \\ & x_{ij} \in \{0,1\}, (i,j) \in A \end{array} \tag{13.1}$$

En décortiquant le problème, on peut montrer qu'il existe neuf chemins possibles. Le chemin optimal est 13246. Comment obtenir une telle solution ? Au fait, le problème du plus court chemin est un problème facile, mais en l'augmentant avec la contrainte de capacité $\sum_{(i,j)\in A} t_{ij}x_{ij} \leq 14$, le problème devient NP-difficile.

Ignorons la contrainte de capacité $\sum_{(i,j)\in A} t_{ij}x_{ij} \leq 14$. Le problème devient un simple problème de plus courts chemins. Soit

$$X = \{x_{ij} \in 0..1 | (13.1) \text{sauf la contrainte de capacité} \leq 14\}.$$

Il est connu dans la théorie des flots que les solutions (sommets du polyèdre) définies par X correspondent aux chemins P du graphe allant de 1 à 6. Nous pouvons ainsi reformuler les contraintes du problème, mise à part la contrainte de capacité, comme suit :

$$\begin{array}{l} x_{ij} = \sum_{p\in P} x_{pij}\lambda_p, (i,j) \in A \\ \sum_{p\in P} \lambda_p = 1 \\ \lambda_p \geq 0, p \in P \end{array} \tag{13.2}$$

où x_{pij} correspond à une constante booléenne égale à 1 si l'arc (ij) appartient au chemin p, sinon 0.

En remplaçant (13.2) dans (13.1), nous obtenons :

$$\begin{array}{ll} \text{minmisze} & z = \sum_{p\in P}(\sum_{(i,j)\in A} c_{ij}x_{pij})\lambda_p \\ \text{subject to} & \sum_{p\in P}(\sum_{(i,j)\in A} t_{ij}x_{pij})\lambda_p \leq 14 \\ & \sum_{p\in P}\lambda_p = 1 \\ & \lambda_p \geq 0, p \in P \\ & x_{ij} = \sum_{p\in P} x_{pij}\lambda_p, (i,j) \in A \\ & x_{ij} \in \{0,1\}, (i,j) \in A \end{array} \tag{13.3}$$

Le problème (13.3) est dit le maître (master problème). Il y a autant de variables λ_p que de chemins P dans le graphe.

La contrainte

$$x_{ij} = \sum_{p\in P} x_{pij}\lambda_p, (i,j) \in A$$

est ignorée, car c'est une contrainte indépendante du problème, qui est satisfaite une fois le meilleure chemin est trouvé. Nous obtenons un problème avec 2 contraintes et 9 variables. Nous avons donc :

$$\begin{array}{ll} \text{minmisze} & z = \sum_{p\in P}(\sum_{(i,j)\in A} c_{ij}x_{pij})\lambda_p \\ \text{subject to} & \sum_{p\in P}(\sum_{(i,j)\in A} t_{ij}x_{pij})\lambda_p \leq 14 \\ & \sum_{p\in P}\lambda_p = 1 \\ & \lambda_p \geq 0, p \in P \end{array} \tag{13.4}$$

Dans le dual du problème (13.4), nous avons les deux variables π_1 et π_0 associées aux deux contraintes du problème. Ainsi, sur de très grands graphes, le nombre de variables devient prohibitif. D'où la nécessité à faire appel à la mécanique de la génération de colonnes. En conséquence, nous travaillerons que sur un ensemble restreint de colonnes, formant le problème maître restreint (restricted master problem) (MRP). Les variables hors (MRP) seront intégrées comme dans l'algorithme du simplex. Tant qu'il existe une variable qui n'est pas dans le (MRP) et qui pourrait l'améliorer, nous l'intégrerons. Ce processus itératif termine une fois qu'on ne trouve plus une colonne améliorante (c'est exactement le mécanisme de la colonne entrante et la colonne sortante du simplex).

On doit donc repérer la varibale entrante x_p ayant un cout réduit négatif :

$$\overline{c_p} = \sum_{(i,j)\in A} c_{ij}x_{pij} - \pi_1(\sum_{(i,j)\in A} t_{ij}x_{pij}) - \pi_0 < 0. \tag{13.5}$$

L'origine de cette formulation vient du tableau du simplex (**??**). En revoyant le déroulement de l'algorithme du simplex (voir la section **??**), nous avons :

$$Z = c_B B^{-1} b + (c_N - c_B B^{-1} A_N) X_N$$

Etant donnée la variable hors-base λ_p, son cout réduit est :

$$(c_N - c_B B^{-1} A_N)$$

où encore en utilisant la solution duale :

$$(c_N - \pi A_N)$$

Par analogie, on a le résultat final :

$$C_p = \sum_{(i,j)\in A} c_{ij} x_{pij}$$
$$A_0 = 1$$
$$A_1 = \pi_1(\sum_{(i,j)\in A} t_{ij}$$

La recherche de cette colonne entrante ayant le plus petit cout réduit revient à optimiser ce qu'on appelle le sous-problème (the subproblem). Dans notre cas, ce sera trouver le plus court chemin dans le graphe, avec un autre cout :

$$\overline{c_p}^* = min_{CC} \sum_{(i,j)\in A} (c_{ij} - \pi_1 t_{ij}) x_{ij} - \pi_0. \tag{13.6}$$

où CC correspond aux quatre contraintes :

$$\sum_{j:(1,j)\in A} x_{1j} = 1$$
$$\sum_{j:(i,j)\in A} x_{ij} - \sum_{j:(j,i)\in A} x_{ji} = 0, i = 2,3,4,5$$
$$\sum_{i:(i,6)\in A} x_{i6} = 1$$
$$x_{ij} \in \{0,1\}, (i,j) \in A$$

Si $\overline{c_p}^* \geq 0$, alors il n'existe aucune colonne améliorante, d'où l'atteinte de l'optimalité. Sinon, la colonne trouvée est injectée dans le (RMP), et on réitère le processus. L'ensemble du processus itère 5 fois, comme illustré dans le Figure 13.2.

Pour trouver une solution entière, ce processus est intégré dans une procédure de type Branch and Bound, décrite au début de cette note.

Iteration	Master Solution	$\bar{z}$	π_0	π_1	$\bar{c}^\star$	p	c_p	t_p
BB0.1	$y_0 = 1$	100.0	100.00	0.00	−97.0	1246	3	18
BB0.2	$y_0 = 0.22, \lambda_{1246} = 0.78$	24.6	100.00	−5.39	−32.9	1356	24	8
BB0.3	$\lambda_{1246} = 0.6, \lambda_{1356} = 0.4$	11.4	40.80	−2.10	−4.8	13256	15	10
BB0.4	$\lambda_{1246} = \lambda_{13256} = 0.5$	9.0	30.00	−1.50	−2.5	1256	5	15
BB0.5	$\lambda_{13256} = 0.2, \lambda_{1256} = 0.8$	7.0	35.00	−2.00	0			
Arc flows: $x_{12} = 0.8, x_{13} = x_{32} = 0.2, x_{25} = x_{56} = 1$								

FIGURE 13.2 – Itérations de la génération de colonne sur le premier noeud du BB [?]

13.2 Correction [Cormen et al., 2001]

On dit qu'un algorithme est correct si, pour chaque instance d'entrée, il se termine avec la sortie désirée correcte. Dans ce document, nous décrirons généralement les algorithmes au moyen de programmes écrits en pseudo-code, très proche de C, Java, ou aussi Pascal.

Nous rappelons que l'invariant d'une boucle, dans un algorithme, est une propriété qui est tout le temps vraie à :

1. l'entrée du corps de la boucle,
2. la sortie du corps de la boucle,
3. à la fin de la boucle.

Pour prouver l'invariant, il est nécessaire de procéder comme suit :

Initialisation La propriété invariante est vraie à l'entrée de la première itération.

Maintenance Si la propriété est vraie dans une itération, l'invariant reste vraie dans l'itération suivante.

Terminaison A la fin de la boucle, l'invariant est suffisant pour démontrer la correction.

L'établissement de cette preuve permet systématiquement d'avoir la preuve de correction de l'algorithme.

Exemple

L'invariant de la boucle de l'algorithme est :

Invariant 1 (Invariant de la boucle TRI-INSERTION)**.** *Le sous-tableau* $A[1..j-1]$ *contient tous les éléments originaux de* $A[1..j-1]$*, et dans un ordre croissant.*

On exploitera cette propriété pour démontrer que l'algorithme est correct.
La preuve se présente comme suit :

Initialisation Il est évident que $A[1..1]$ est bien trié.

Maintenance Supposons que l'algorithme est arrivé à la $j$ième itération, d'où $A[1..j-1]$ est trié. A la fin de la $j$ième itération, l'algorithme aurait placé $A[j]$ à la première cas i, $i < j$ où $A[i] > A[j]$. Ceci va faire en sorte à ce que $A[1..j]$ soit aussi trié.

Terminaison A la fin de la boucle, on aurait $A[1..n]$ trié. D'où la correction de l'algorithme.

13.3 Algorithmique et complexité

Une des premières références historiques sur les algorithmes date du neuvième siècle, à Baghdad, dont l'auteur est Al Khwarizmi qui a proposé les méthodes de base pour additionner, multiplier, et diviser des nombres entiers - et aussi le calcul des racines d'une équation, et les décimales du nombre π. Ces procédures étaient précises, non-ambigües, mécaniques, efficaces, et correctes : elles étaient simplement des *algorithmes*, un terme qui a honoré son concepteur El Khwarizmi.

Depuis cet avènement, et celui de l'adoption du système décimal, les scientifiques ont développé, et développent des algorithmes complexes pour résoudre des problèmes variés.

13.3.1 Exemple de motivation [Papadimitriou et al., 2006]

Soit la suite de Fibonacci (F_n)

$$F_n = \begin{cases} F_{n-1} + F_{n-2} & \text{si n>1,} \\ 1 & \text{si n=1,} \\ 0 & \text{si n=0.} \end{cases}$$

La suite génère les nombres suivants

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, ...$$

Cette suite a aussi la caractéristique d'être de nature exponentielle, d'ailleurs il existe l'approximation $F_n \approx 2^{0.694n}$. Mais, comment pourrons nous calculer des nombres aussi grands que F_{100} ou même F_{200} ?

13.3.1.1 Une première version

Algorithm 4 FIB1(n)

```
1: if n = 0 then
2:    return 0
3: else if n = 1 then
4:    return 1
5: else
6:    return fib1(n − 1) + fib1(n − 2)
7: end if
```

Trois questions essentielles sont posées :

1. **L'algorithme est-il correct ?**
2. **Combien de temps exige-t-il pour terminer, en fonction de n ?**
3. **Peut-on l'améliorer ?**

La réponse à la première question est plutôt évidente : l'algorithme est bien correct, car il met en oeuvre exactement la définition récurrente de la suite de Fibonacci. La deuxième question est délicate. Soit $T(n)$ le nombre de pas dont a besoin l'algorithme pour calculer F_n. Une première évidence :

$$T(n) \leq 2 \text{ pour } n \leq 1.$$

Pour des nombres n plus grands, on a la relation suivante :

$$T(n) = T(n-1) + T(n-2) + 3 \text{ pour } n > 1.$$

Nous disposons d'une solution approchée $T_n \approx 2^{0.694n}$. Ceci veut dire que le nombre de pas croit aussi vite que la suite de Fibonacci. $T(n)$ a évidemment une **croissance exponentielle**.

Soit par exemple le calcul de F_{200}, le calcul de FIB1(n) exécute $T(200) \geq F_{200} \geq 2^{138}$ pas élémentaires. On se pose la question sur le temps nécessaire pour ce calcul sur un ordinateur ? La réponse est immédiate : prenons un ordinateur puissant à l'heure actuelle, exécutant 40 mille milliards 40×10^{12} instructions par seconde (un ordinateur personnel, un PC, avec un processeur i7-3770 peut exécuter 20×10^{10} opérations flottantes par seconde - on dit 200 gigaFLOPS). Nous aurons besoin sur cet ordinateur de 2^{92} **secondes** ! Ce qui veut dire que l'on doit attendre jusqu'à l'extinction théorique du soleil !

D'après l'hypothèse de Moore qui suppose que les ordinateurs doublent leur puissance tous les 18 mois, alors, on pourrait arriver au fait que : si on arrivait à calculer raisonnablement F_{100}, alors avec l'hypothèse de Moore, on pourrait vraisemblablement, calculer avec la même puissance $F(101)$, ... Voulant dire qu'on gagnera chaque année un chiffre. C'est ainsi qu'on se voit totalement découragé face à la réalité terrible de la croissance exponentielle qui **ne peut être cassée par aucune machine de nos jours !**

La question reste posée : **peut on mieux faire en terme de coût de calcul ?** Et sans passer par l'amélioration des machines, dont on vient de voir l'impuissance !

13.3.1.2 Une version polynomiale

On montre dans la Figure 13.3, le comportement de l'algorithme FIB1.

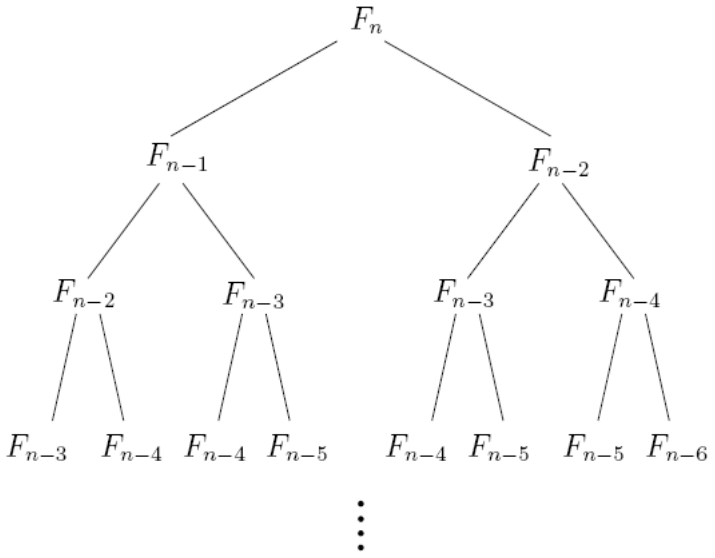


FIGURE 13.3 – Trace de FIB1(n) [Papadimitriou et al., 2006]

En analysant cette trace, on voit bien que plein d'appels sont répétés. On peut donc adopter l'astuce qui consiste à stocker les calculs déjà faits. On obtient la version FIB2.

Algorithm 5 FIB2(n)

```
1: if n = 0 then
2:    return 0
3: end if
4: créer un tableau F[0..n]
5: F[0] ← 0
6: F[1] ← 1
7: for i ← 2..n do
8:    F[i] ← F[i − 1] + F[i − 2]
9: end for
10: return F[n]
```

L'algorithme est évidemment correct, car, encore une fois, il est similaire à la définition récurrente de Fibonacci. La question : combien requiert cet algorithme pour calculer F_n en fonction de n ? La boucle de l'algorithme nécessite $n-1$ pas de calculs. De ce fait, le temps d'exécution de **cet algorithme est de nature linéaire en** n. Il devient maintenant possible de calculer F_{200000} **avec peu d'efforts !**

Bien concevoir un algorithme fait gagner des exponentielles !

13.3.2 Complexité [Cormen et al., 2001]

Un algorithme est un ensemble d'opérations de calcul élémentaires, organisées selon des règles précises dans le but de résoudre un problème donné. Pour chaque donnée du problème, l'algorithme retourne une réponse après un nombre fini d'opérations. Les opérations élémentaires sont par exemple les opérations arithmétiques usuelles, les transferts de données, les comparaisons entre données, etc.

Il apparaît utile de ne **considérer comme véritablement élémentaires que les opérations dont le temps de calcul est constant**, c'est-à-dire ne dépend pas de la taille des opérandes. Par exemple :

— l'addition d'entiers de taille bornée à priori (les int en C) est une opération élémentaire ;
— l'addition d'entiers de taille quelconque ne l'est pas.
— De même, le test d'appartenance d'un élément à un ensemble n'est pas une opération élémentaire en ce sens, parce que son temps d'exécution dépend de la taille de l'ensemble, et ceci même si dans certains langages de programmation, il existe des instructions de base qui permettent de réaliser cette opération.

Nous avons déjà pris goût aux algorithmes avec l'algorithme de calcul des termes de la suite de Fibonacci pour lequel nous avons introduit une première analyse simple. Pour illustrer notre démarche, nous aborderons un problème posé fréquemment en pratique, celui du tri d'une suite de nombres en ordre croissant.

Nous pouvons définir formellement ce problème comme suit :

Entrée Une séquence de n nombres $\langle a_1, a_2, ..., a_n \rangle$.

Sortie Une permutation $\langle a'_1, a'_2, ..., a'_n \rangle$ de la permutation de la suite d'entrée telle que $a_1 \leq a_2 \leq ... \leq a_n$.

Etant donnée une suite d'entrée telle que $\langle 5, 90, 45, 89 \rangle$, un algorithme de tri fournira en sortie $\langle 5, 45, 89, 90 \rangle$. On dira d'une telle suite d'entrée que c'est une instance du problème.

Nous introduirons une première solution au problème du tri avec la version du TRI PAR INSERTION.

Algorithm 6 TRI-INSERTION(A)

Input: Une séquence de n nombres $A = \langle a_1, a_2, ..., a_n \rangle$.
Output: Une permutation $\langle a'_1, a'_2, ..., a'_n \rangle$ de la permutation de la suite d'entrée telle que $a_1 \leq a_2 \leq ... \leq a_n$.

```
1: for j ← 2 to longueur(A) do
2:    pivot ← A[j]
3:    {insertion de A[j] dans la suite triée A[1..j − 1]}
4:    i ← j − 1
5:    while (i > 0) ∧ (A[i] > pivot) do
6:       A[i + 1] ← A[i]
7:       i ← i − 1
8:    end while
9:    A[i + 1] ← pivot
10: end for
```

Le temps pris par la procédure TRI-INSERTION dépend de la **taille de son entrée** : le tri d'un millier de nombres prend plus de temps que le tri de trois nombres. La notion de taille d'entrée dépend du problème étudié, parfois il est plus approprié de décrire la taille de l'entrée avec deux nombres. Par exemple, la taille de l'entrée dans un problème sur les graphes, peut être le nombre de sommets et le nombres d'arêtes. **Pour chaque problème étudié, nous préciserons la taille (mesure) utilisée.**

Le temps d'exécution d'un algorithme sur une entrée particulière est le nombre d'opérations élémentaires exécutées.

Le temps d'exécution de l'algorithme est la somme des temps d'exécution de chaque instruction exécutée ; une instruction qui s'exécute en un temps c_i et qui est exécutée n fois interviendra pour $c_i n$. Pour calculer $T(n)$, le temps d'exécution de TRI-INSERTION, on additionne les produits des coûts par le nombre de fois, et on obtient :

$$T(n) = c_1 n + c_2(n-1) + c_4(n-1) + c_5 \sum_{j=2..n} t_j + c_6 \sum_{j=2..n} (t_j - 1) + c_7 \sum_{j=2..n} (t_j - 1) + c_9(n-1),$$

où t_j est le nombre d'itérations de la deuxième boucle à l'itération j de la première boucle. Si la suite est déjà ordonnée, on a

$$T(n) = c_1 n + c_2(n-1) + c_4(n-1) + c_5(n-1) + c_9(n-1) = (c_1 + c_2 + c_4 + c_5 + c_9)n + (c_2 + c_4 + c_5 + c_9).$$

On peut exprimer cette dernière formule sous la forme $an + b$, où a et b sont des constantes. On dit que l'algorithme est en $O(n)$.

Si la suite est dans un ordre inverse, on se trouve dans le pire des cas. On doit comparer chaque élément $A[j]$ avec chaque élément de sous-tableau trié $A[1..j-1]$, et donc $t_j = j$ pour $j = 2, 3, ..., n$. Sachant que

$$\sum_{j=2..n} j = n(n+1)/2 - 1$$

et

$$\sum_{j=2..n} (j-1) = n(n-1)/2.$$

On obtient ainsi

$$T(n) = c_1 n + c_2(n-1) + c_4(n-1) + c_5(n(n+1)/2-1) + c_6(n(n-1)/2) + c_7(n(n-1)/2) + c_9(n-1)$$

$$T(n) = (c_5/2 + c_6/2 + c_7/2)n^2 + (c_1 + c_2 + c_4 + c_5/2 + c_6/2 + c_7/2 + c_9)n.$$

Ce dernier coût est de la forme an^2+bn+c où a, b et c sont des constantes. On dit que l'algorithme est en $O(n^2)$. Nous venons donc de calculer le coût de l'algorithme du TRI-INSERTION dans le pire des cas, c'est-à-dire le temps d'exécution le plus long pour une entrée quelconque de taille n. C'est la mesure de coût la plus utilisée en pratique. Nous verrons dans la suite un ensemble de procédés permettant de calculer ce coût.

13.4 Mesure de complexité

Analyser un algorithme consiste à prévoir **les ressources à cet algorithme**. Parfois, les ressources pertinentes sont la **mémoire** utilisée, la **largeur de bande d'une communication**, ou les **portes logiques**, mais le plus souvent, on souhaite mesurer le **temps de calcul**. Dans ce cours, on prendra comme modèle de calcul, **une machine à accès aléatoire, à processeur unique**. Dans ce modèle, les instructions sont exécutées l'une après l'autre, sans opérations simultanées.

Considérons un problème donné, et un algorithme pour le résoudre. Sur une donnée x de taille n, l'algorithme requiert un certain temps, mesuré en nombre d'opérations élémentaires, soit $c(x)$. Le coût en temps varie évidemment avec la taille de la donnée, mais peut aussi varier sur les différentes données de même taille n.

Par exemple, considérons l'algorithme de tri qui, partant d'une suite $(a_1, ..., a_n)$ de nombres réels distincts à trier en ordre croissant, cherche la première descente, c'est-à-dire le plus petit entier i tel que $a_i > a_{i+1}$, échange ces deux éléments, et recommence sur la suite obtenue. Si l'on compte le nombre d'inversions ainsi réalisées, il varie de 0 pour une suite triée à $n(n-1)/2$ pour une suite décroissante. Notre but est d'**évaluer le coût d'un algorithme, selon certains critères, et en fonction de la taille n des données**.

Pour certains problèmes, on peut **mettre en évidence une ou plusieurs opérations qui sont fondamentales** au sens où le temps d'exécution d'un algorithme résolvant ce problème est toujours proportionnel au nombre de ces opérations. Il est alors possible de comparer des algorithmes traitant ce problème selon cette mesure simplifiée.

Donnons quelques exemples d'**opérations fondamentales** :

1. pour la recherche d'un élément dans une liste en mémoire centrale : le nombre de comparaisons entre cet élément et les entrées de la liste ;
2. pour la recherche d'un élément sur un disque : le nombre d'accès à la mémoire secondaire ;
3. pour trier une liste d'éléments : on peut considérer deux opérations fondamentales : le nombre de comparaisons entre des éléments et le nombre de déplacements d'éléments ;
4. pour multiplier deux matrices de nombres : le nombre de multiplications et le nombre d'additions.

Remarquons que si l'on choisit plusieurs opérations fondamentales, on **peut les décompter séparément** puis, si besoin est, on les affecte chacune d'un poids qui tient compte des temps d'exécution différents.

1. En faisant **varier le nombre d'opérations fondamentales**, on fait varier le **degré de précision de l'analyse**, et aussi son degré d'abstraction, i.e. d'indépendance par rapport à l'implémentation. A la limite, si l'on veut **faire une microanalyse très précise du temps d'exécution du programme, il suffit de décider que toutes les opérations du programme sont fondamentales**.
2. On a fait l'hypothèse que le temps d'exécution est proportionnel à la mesure choisie. On ne peut pas comparer deux algorithmes utilisant des mesures différentes.

Après avoir déterminé les opérations fondamentales, il s'agit de compter le nombre d'opérations de chaque type. Il n'existe pas de systèmes complet de règles permettant de compter le nombre d'opérations en fonction de la syntaxe des algorithmes mais l'on peut faire **quelques remarques** :

Séquence Lorsque les opérations sont dans une séquence d'instructions, leurs nombres s'ajoutent.

if-then-else Pour les branchements conditionnels, il est en général difficile de déterminer quelle branche de la condition est exécutée, et donc quelles sont les opérations à compter. Cependant, on peut majorer ce nombre d'opérations.

Boucles Pour les boucles, le nombre d'opérations dans la boucle est $\sum P(i)$, où i est la variable de contrôle de la boucle, et $P(i)$ le nombre d'opérations fondamentales lors de l'exécution de la ième itération. Si le nombre d'itérations est difficile à calculer, on peut se contenter d'une bonne majoration.

Appels procédures Pour les appels de procédures et fonctions non récursives, on peut s'arranger à calculer la complexité de ces appels, et les prendre en compte suivant l'imbrication de l'appel dans l'algorithme.

Appels récursifs Pour les appels de procédures et fonctions récursives, compter le nombre d'opérations fondamentales donne en général lieu à la **résolution de relations de récurrence**. En effet le nombre $T(n)$ d'opérations dans l'appel de la procédure avec un argument de taille n s'écrit, selon la récursion, en **fonction de divers $T(k)$, pour $k < n$. L'exemple de Fibonacci** donné dans le chapitre d'introduction illustre bien ce cas.

Il est évident que **le calcul du coût d'un algorithme dépend de la donnée sur laquelle il opère**.

1. Il faut d'abord **définir une mesure de taille sur les données** qui reflète la quantité d'information contenue. **Par exemple, si l'on additionne ou on multiplie des entiers, une mesure significative est le nombre de chiffres des nombres**.
2. Pour certains algorithmes, le temps d'exécution ne dépend que de la taille des données ; mais la plupart du temps la complexité varie aussi, pour une taille fixée des données, en fonction de la donnée elle-même.

13.4.1 La complexité dans le meilleur des cas

Le coût $Min_A(n)$ d'un algorithme A dans le meilleur des cas

$$Min_A(n) = min_{|x|=n} c(x).$$

13.4.2 La complexité dans le pire des cas

Le coût $Max_A(n)$ d'un algorithme A dans le cas le plus défavorable ou dans le cas le pire [1] est par définition le maximum des coûts, sur toutes les données de taille n :

$$Max_A(n) = max_{|x|=n} c(x).$$

13.4.3 La complexité en moyenne

Dans des situations où l'on pense que **le cas le plus défavorable ne se présente que rarement, on est plutôt intéressé par le coût moyen de l'algorithme**. Une formulation correcte de ce coût moyen suppose que l'on connaisse une distribution de probabilités sur les données de taille n. **Si $p(x)$ est la probabilité de la donnée x**, le coût moyen $Moy_A(n)$ d'un algorithme A sur les données de taille n est par définition

$$Moy_A(n) = \sum_{|x|=n} p(x)c(x).$$

Le plus souvent, on suppose que la distribution est uniforme, c'est-à-dire que $p(x) = 1/T(n)$, où $T(n)$ est le nombre de données de taille n. Alors, l'expression du coût moyen prend la forme

$$Moy_A(n) = \frac{1}{T(n)} \sum_{|x|=n} c(x).$$

En pratique, la complexité en moyenne est souvent beaucoup plus difficile à déterminer que la complexité dans le pire des cas , d'une part parceque l'analyse devient mathématiquement difficile, et d'autre part parcequ'il n'est pas toujours facile de déterminer un modèle de probabilités adéquat au problème.

En claire, il existe entre les complexités en moyenne et les complexités extrêmales la relation suivante :

$$Min_A(n) \leq Moy_A(n) \leq Max_A(n).$$

Si le comportement de l'algorithme dépend uniquement de la taille des données (comme dans l'exemple de la multiplication de matrices), alors ces trois quantités sont confondues. Mais en général, ce n'est pas le cas et l'on ne sait même pas si le coût moyen est plus proche du coût minimal ou du coût maximal.

1. "worst-case" en anglais.

13.4.4 Grandeurs des fonctions et notations de Landau : O, ω, ... [Gaudel et al., 1990]

On a déterminé **la complexité d'un algorithme comme une fonction de la taille des données** ; il est très important de connaitre la **rapidité de croissance** de cette fonction lorsque la **taille des données croît**. En effet, pour traiter un problème de petite taille la méthode employée importe peu, alors que pour un problème de grande taille, les différences de performance entre algorithmes peuvent être énormes.

Souvent une simple approximation de la fonction de complexité suffit pour savoir si un algorithme est utilisable ou non, ou pour **comparer entre différents algorithmes**.

Par exemple, pour n grand, il est souvent secondaire de savoir si un algorithme fait $n+1$ ou $n+2$ opérations.

Parfois les constantes multiplicatives ont, elles aussi, peu d'importance.

- Supposons que l'on ait à comparer l'algorithme A_1 de complexité $M_1(n) = n^2$ et l'algorithme A_2 de complexité $M_2(n) = 2n$. A_2 **est meilleur que** A_1 pour presque tous les $n(n > 2)$;
- De même si $M_1(n) = 3n^2$ et $M_2(n) = 25n$; A_2 **est meilleur que** A_1 pour $n > 8$.
- [*Négliger un terme dans une addition*] **Quelles que soient les constantes multiplicatives k_1 et k_2 telles que $M_1(n) = k_1 n^2$ et $M_2(n) = k_2 n$, l'algorithme A_2 est toujours meilleur que A_1 à partir d'un certain n, car la fonction $f(n) = n^2$ croît beaucoup plus vite que la fonction $g(n) = n$. En effet**

$$\lim_{n\to\infty} g(n)/f(n) = 0.$$

On dit que l'ordre de grandeur asymptotique de $f(n)$ est strictement plus grand que celui de $g(n)$.

- [*Négliger une constante*] **D'autre part, quelles que soient les constantes multiplicatives k_1 et k_2 telles que $M_1(n) = k_1 f(n)$ et $M_2(n) = k_2 f(n)$, l'algorithme A_2 est toujours du meme ordre que A_1, en effet**

$$\lim_{n\to\infty} k_1 f(n)/k_2 f(n) = k_1/k_2.$$

La Figure 13.4 met en évidence la différence de rapidité de croissance de certaines fonctions usuelles : les ordres de grandeur asymptotique des fonctions $1, log_2(n), nlog_2(n), n^2, n^3, 2^n$ vont en croissant strictement ; ces fonctions forment une échelle de comparaison.

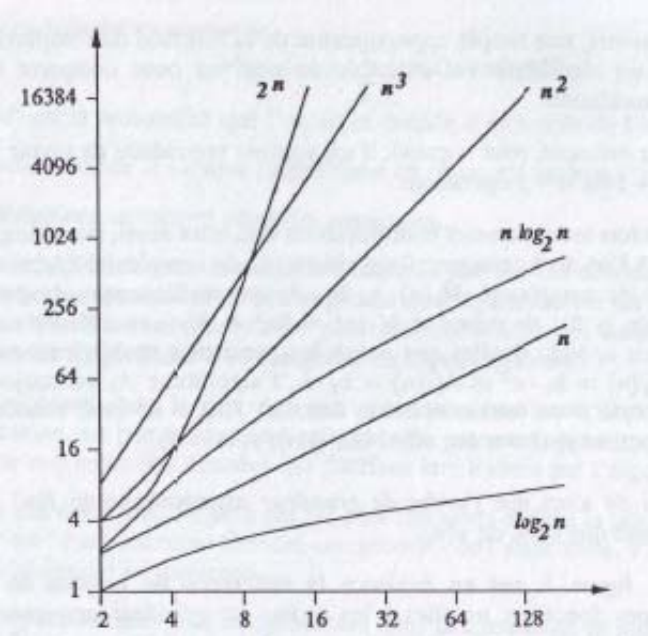


FIGURE 13.4 – Croissance de certaines fonctions usuelles [Gaudel et al., 1990]

Pour analyser la complexité $M_A(n)$ d'un algorithme A, on s'attache d'abord à déterminer l'ordre de grandeur asymptotique de $M_A(n)$: **on cherche dans une échelle de comparaison, éventuellement plus complète que celle qui est formée par les fonctions de la Figure 13.4, une fonction qui a une rapidité de croissance voisine de $M_A(n)$.**

Supposons que l'on ait à comparer deux algorithmes A_1 et A_2 de complexités $M_{A_1}(n)$ et $M_{A_2}(n)$. **Si l'ordre de grandeur de $M_{A_1}(n)$ est strictement plus grand que l'ordre de grandeur de $M_{A_2}(n)$, alors on peut conclure immédiatement que A_1 est meilleure que A_2 pour n grand.** Par contre, si $M_{A_1}(n)$ et $M_{A_2}(n)$ ont même ordre de grandeur asymptotique, il faut faire une analyse plus fine pour pouvoir comparer A_1 et A_2.

Pour comparer l'odre de grandeur asymptotique des fonctions, on a l'habitude d'utiliser la notion suivante :

Définition 33. *Etant donné deux fonctions f et g de $\mathbb{N}$ dans $\mathbb{R}^+$,*

$$f = O(g)$$

si et seulement si $\exists c \in \mathbb{R}^{+*}, \exists n_0 \in \mathbb{N}$ tel que

$$\forall n > n_0, f(n) \leq c.g(n).$$

On dit aussi que

$$g = \Omega(f)$$

si et seulement si $f = 0(g)$.

Ainsi $f = O(g)$ **veut dire que l'ordre de grandeur asymptotique de f, est inférieur ou égal à celui de g**, on dit que f **est dominée asymptotiquement par g**; par exemple $2n = O(n^2)$, mais aussi $2n = O(n)$.

Cette notion qui donne un majorant de l'ordre de grandeur asymptotique de f, est très utile pour de nombreuses applications, **mais elle n'est pas suffisante pour comparer entre elles les performances de différents algorithmes**, car il faut connaitre les ordres de grandeurs exacts, et non des majorants. Lorsque l'on dit que la complexité $M_A(n)$ d'un algorithme A est en $h(n)$, on veut dire que son ordre de grandeur asymptotique est exactement $h(n)$ (i.e. $h(n)$ est le plus petit majorant).

On est donc amené à introduire la définition suivante :

Définition 34. *Etant donné deux fonctions f et g de $\mathbb{N}$ dans $\mathbb{R}^+$,*

$$f = \Theta(g)$$

si et seulement si $\exists c, d \in \mathbb{R}^{+*}, \exists n_0 \in \mathbb{N}$ tel que

$$\forall n > n_0, d.g(n) \leq f(n) \leq c.g(n).$$

Ou d'une façon équivalente,

$$f = \Theta(g)$$

si et seulement si $f = 0(g)$ et $g = O(f)$.

On dit que f **et** g **ont même ordre de grandeur asymptotique**. **La notion Θ est plus précise que la notion O**. Par exemple $2n = \Theta(n)$, mais $2n$ **n'est pas en** $\Theta(n^2)$. Cependant, dans la plupart des ouvrages d'algorithmique, les résultats des analyses sont mis sous la forme $O(f(n))$, alors qu'un décompte précis des opérations fondamentales permet souvent de conclure que la complexité est exactement $\Theta(f(n))$. Par exemple, on dit souvent que le tri par tas est en $O(nlog(n))$, alors qu'en fait il est en $\Theta(nlog(n))$.

Soulignons un point fondamental : les définitions de O et Θ reposent sur l'existence de certaines constantes finies, mais il n'est rien précisé sur la valeur de ces constantes. **Cela n'a pas d'importance pour obtenir des résultats asymptotiques lorsque les fonctions ont des ordres de grandeur différents**. Par exemple si $f(n) = 2n$ et $g(n) = n^2$, alors $f(n) < g(n)$ pour $n > 2$. Si $f(n) = 1000n$ et $g(n) = n^2$, alors $f(n) < g(n)$ **pour** $n > 10^4$.

Ainsi, si l'ordre de grandeur de f est plus petit que celui de g alors il existe un seuil à partir duquel la valeur de f est c fois plus petite que celle de g, mais on ne sait pas quel est ce seuil.

Par contre, si f et g ont même ordre de grandeur, il devient beaucoup plus difficile de les comparer : la détermination des constantes, et éventuellement des termes d'ordre inférieur nécessite en général des techniques mathématiques beaucoup plus complexes. Il faut bien être conscient de ce que l'obtention de résultats tels que : "l'algorithme A va deux fois plus vite que l'algorithme B sur un ordinateur standard", est en général très difficile.

La notion d'ordre de grandeur de la complexité des algorithmes a une grande importance pratique. Supposons que l'on dispose pour résoudre un problème donné de sept algorithmes dont les complexités dans le cas le pire ont respectivement pour ordre de grandeur 1 (c'est-à-dire une fonction constante, qui ne dépend pas de la taille des données), $log_2(n)$ (c'est-à-dire une fonction polynomiale d'ordre 3), 2^n (c'est-à-dire une fonction exponentielle).

Complexité / Taille	1	$\log_2 n$	n	$n\log_2 n$	n^2	n^3	2^n
$n = 10^2$	$\simeq 1\mu s$	6,6 μs	0,1 ms	0,6 ms	10 ms	1 s	4×10^{16} a
$n = 10^3$	$\simeq 1\mu s$	9,9 μs	1 ms	9,9 ms	1 s	16,6 mn	∞
$n = 10^4$	$\simeq 1\mu s$	13,3 μs	10 ms	0,1 s	100 s	11,5 j	∞
$n = 10^5$	$\simeq 1\mu s$	16,6 μs	0,1 s	1,6 s	2,7 h	31,7 a	∞
$n = 10^6$	$\simeq 1\mu s$	19,9 μs	1 s	19,9 s	11,5 j	$31,7 \times 10^3$ a	∞

FIGURE 13.5 – Temps d'exécution et taille des données [Gaudel et al., 1990]

Le tableau de la Figure 13.5 donne une estimation du temps d'exéction de chacun de ces algorithmes pour différentes tailles n des données du problème sur un ordinateur pouvant effectuer 10^6 opérations par seconde. **Il montre bien que, plus la taille des données est grande, plus les écarts entre les différents temps d'exécution se creusent.**

Complexité / Temps calcul	1	$\log_2 n$	n	$n\log_2 n$	n^2	n^3	2^n
1s	∞	∞	10^6	63×10^3	10^3	100	19
1 mn	∞	∞	6×10^7	28×10^5	77×10^2	390	25
1 h	∞	∞	36×10^8	13×10^7	60×10^3	15×10^2	31
1 jour	∞	∞	86×10^9	27×10^8	29×10^4	44×10^2	36

FIGURE 13.6 – Temps d'exécution et taille des données [Gaudel et al., 1990]

Le tableau de la Figure 13.6 donne une estimation de la taille maximale des données que l'on peut traiter par chacun des algorithmes en un temps d'exécution fixé (et toujours sur un ordinateur effectuant 0^6 opérations par seconde).

D'après ces deux tableaux, il est clair que certains algorithmes sont utilisables pour résoudre des problèmes sur ordinateurs, et que d'autres ne sont pas, ou peu utilisables.

Les algorithmes utilisables pour des données de grande taille sont ceux qui s'exécutent en temps :

constant (c'est le cas de la complexité en moyenne de certaines méthodes de hachage) ;

logatithmique (par exemple la recherche dichotomique, ou les arbres binaires de recherche) ;

linéaire (par exemple la recherche séquentielle) ;

$n.log(n)$ (par exemple les bons algorithmes de tri).

Les algorithmes qui prennent un temps polynomial, c'est-à-dire en $\Theta(n^k)$ avec $k > 0$, ne sont vraiment utilisables que pour $k < 2$.

Lorsque $2 \leq k \leq 3$, on peut traiter que des problèmes de taille moyenne, et lorsque k dépasse 3 on ne peut traiter que des petits problèmes.

Les algorithmes en temps exponentiel, c'est-à-dire en $\Theta(2^n)$ par exemple, sont à peu près inutilisables, sauf pour des problèmes de très petit taille. Ce sont de tels algorithmes que l'on a qualifiés d'inefficaces.

Le tableau de la Figure 13.6 montre comment la taille des données et le temps d'exécution varient en fonction l'un de l'autre. On voit en particulier que si l'on multiplie par 10 la vitesse de calcul de l'ordinateur, on ne modifie quasiment pas la taille maximale des données que l'on peut traiter avec un algorithme exponentiel, alors que l'on multiplie évidemment par 10 la taille des données traitables par un algorithme linéaire. **Il est donc toujours d'actualité de rechercher des algorithmes efficaces, même si les projets technologiques accroissent les performances du matériel.**

Complexité	1	$\log_2 n$	n	$n\log_2 n$	n^2	n^3	2^n
Evolution du temps quand la taille est multipliée par 10	t	t+3,32	10×t	(10+ε)×t	100×t	1000×t	t^{10}
Evolution de la taille quand le temps est multiplé par 10	∞	n^{10}	10×n	(10−ε)×n	3,16×n	2,15×n	n+3,32

FIGURE 13.7 – Temps d'exécution et taille des données [Gaudel et al., 1990]

Taille → / Complexité ↓	20	50	100	200	500	1 000
$10^3.n$	0.02 s	0.05 s	0.1 s	0.2 s	0.5 s	1 s
$10^3.n.\log_2 n$	0.09 s	0.3 s	0.6 s	1.5 s	4.5 s	10 s
$100.n^2$	0.04 s	0.25 s	1 s	4 s	25 s	2 mn
$10.n^3$	0.02 s	1 s	10 s	1 mn	21 mn	27 h
$n^{\log_2 n}$	0.4 s	1.1 h	220 jours	12 500 ans	5.10^{10} ans	--
$2^{n/3}$	0.0001 s	0.1 s	2.7 h	3.10^6 ans	--	--
2^n	1 s	36 ans	--	--	--	--
3^n	58 mn	2.10^{11} ans	--	--	--	--
$n!$	77 100 ans	--	--	--	--	--

FIGURE 13.8 – Temps d'exécution et taille des données [Prins, 1994]

13.5 Nombre et arithmétique des intervalles

Les nombres naturels, $\mathbb{N}$, forment l'ensemble de base qui est la pierre angulaire de la définition des autres structures de nombres. L'ensemble des nombres relatifs est noté $\mathbb{Z}$. L'ensemble des nombres rationnels, $\mathbb{Q}$, est l'ensemble des fractions, ou de couples d'entiers, $\mathbb{Q} = \{\frac{x}{y} | x \in \mathbb{Z} \wedge y \in \mathbb{Z}^*\}$. L'ensemble des nombres réels, $\mathbb{R}$, est une extension des nombres rationnels. Un nombre réel est défini comme la limite d'une suite infinie de nombre rationnels. L'ensemble des nombres réels qui ne peuvent pas être mis sous forme rationnelle sont appelés nombres irrationnels, on les notera $\overline{\mathbb{Q}}$. Les nombres complexes, $\mathbb{C}$, sont un ensemble de couples de nombres réels. La propriété fondamentale des nombres complexes est que toute équation polynomiale, à coefficients complexes, a des solutions. On dispose dans chacun des ensembles "$\mathbb{N}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$" d'un ensemble de fonctions mathématiques pour faire des calculs.

Les nombres flottants : l'ensemble des réels est infini. En pratique, on dispose de moyens finis pour représenter les nombres et les calculs. Ce qui a donné lieu à l'ensemble des nombres flottants qui est un ensemble fini pour approximer les réels. Un flottant est sous la forme $a \times b^c$, ou a, b et c appartiennent à un sous-domaine fini de $\mathbb{Z}$. Tout nombre d'un système particulier de nombres flottants est spécifié avec une base b fixe[2]. La norme internationale IEEE754 [?] représente le flottant sur 64 bits, $b = 2$, a est stocké sur 53 bits (un bit est utilisé pour représenter le signe) et c sur 11 bits (avec la représentation binaire biaisée). Pour avoir plus d'informations sur un système de flottants, [?] utilise la notation $\mathbb{F}[b, a, n \ldots M]$, avec [n,M] l'intervalle des nombres représentables dans le système flottant utilisé. La norme [?] est notée $\mathbb{F}[2, 53, -2^{10} + 2 \ldots 2^{10} - 1]$. Tout réel ne peut pas être représenté exactement dans le système flottant, et de même pour le calcul sur les flottants, car une opération sur les flottants peut donner un nombre qui n'est pas représentable dans le système flottant utilisé. En conséquence, pour représenter l'approximation d'un nombre réel

2. La base b des systèmes flottants est souvent la base binaire.

x, tout système flottant[3] fournit les quatre routines suivantes :

1. *roundnear* : impose au système de prendre le flottant le plus proche du nombre x ;
2. *roundup* : impose au système de prendre le nombre le plus petit parmi les flottants qui sont plus grands que le nombre x ;
3. *rounddown* : impose au système de prendre le nombre le plus grand parmi les flottants qui sont plus petits que le nombre x.
4. *roundzero* : impose au système de prendre parmi les deux flottants qui encadrent x celui qui est le plus proche de zéro.

Les machines actuelles utilisent l'arithmétique en virgule flottante. Dans cette arithmétique, les nombres réels sont approximés par un sous-ensemble fini de réels appelés nombres machine. À cause de la finitude de cet ensemble, deux erreurs sont générées immanquablement. La première erreur est produite lors de l'approximation d'un nombre réel par un nombre flottant. La deuxième est générée par les calculs intermédiaires réels approximés sur les nombres flottants. En pratique, on peut tomber sur des calculs qui donnent des résultats qui sont éloignés de la bonne valeur.

Exemple 6. *[?]*
Soit l'expression

$$333.75 \times y^6 + x^2 \times (11 \times x^2 \times y^2 - y^6 - 121 \times y^4 - 2) + 5.5 \times y^8 + \frac{x}{2 \times y}.$$

L'évaluation de cette expression au point $(x = 77617, y = 33096)$ *donne avec la représentation double* $1.17260\ldots$*, avec la représentation étendue* $1.17260\ldots$*, alors que la valeur exacte est* $-0.82739\ldots$*!*

D'ailleurs, en analyse numérique, les algorithmes sont souvent présentés avec une étude de stabilité numérique pour caractériser la sensibilité de l'algorithme aux erreurs d'approximation.

La raison principale de l'instabilité numérique, en utilisant l'arithmétique standard, est que l'arithmétique flottante ne respecte pas les propriétés de l'arithmétique réelle. L'arithmétique des intervalles donne les moyens de calcul pour estimer et contrôler automatiquement les erreurs d'approximation. Contrairement à l'arithmétique standard, l'arithmétique flottante des intervalles respecte les propriétés de l'arithmétique réelle des intervalles.

L'arithmétique des intervalles a été introduite par [?]. Nous présenterons brièvement cette arithmétique en montrant ses avantages par rapport à l'arithmétique standard, dans le même ordre d'idée que [?].

Dans l'arithmétique réelle des intervalles, un nombre réel e est approximé par deux nombres réels l et r tel que $l \leq e \leq r$. Le nombre e est donc représenté par l'intervalle $[l, r] = \{x | l \leq x \leq r\}$. La taille de cet intervalle donne une mesure qualitative de l'approximation. Les calculs sont faits sur les intervalles avec la garantie d'encadrement des résultats intermédiaires.

3. [?] présente une bonne synthèse des propriétés des systèmes flottants.

Soit E un ensemble partiellement ordonné par la relation $\leq$, on note par $I\!I(E) = \{[x,y]|x \in E, y \in E, x \leq y\}$ l'ensemble des intervalles construits sur E. L'intervalle contenant tous les éléments de l'ensemble $Y \subseteq E$, est dénoté $\Box Y$. Si $\min(Y)$ et $\max(Y)$ existent, alors $\Box Y = [\min(Y), \max(Y)]$. La borne inférieure et la borne supérieure d'un intervalle X sont respectivement dénotées par $\underline{X}$ et $\overline{X}$.

$I\!I(I\!R)$ est l'ensemble des intervalles sur $I\!R$. $wid(X) = \overline{X} - \underline{X}$ est la largeur de l'intervalle X. La largeur d'un vecteur d'intervalles X est

$$wid(X) = max(wid(X_1), \ldots, wid(X_n)).$$

Nous définissons de la même façon la largeur d'une matrice d'intervalles.

$mid(X) = (\overline{X} - \underline{X})/2$ est le point milieu de l'intervalle X. Le milieu d'un vecteur d'intervalles X est

$$mid(X) = \langle mid(X_1), \ldots, mid(X_n)\rangle.$$

Nous définissons de la même façon le milieu d'une matrice d'intervalles.

La distance $dist(X,Y)$, avec $X \in I\!I(I\!R), Y \in I\!I(I\!R)$, (resp. $X \in I\!I(I\!R)^n, Y \in I\!I(I\!R)^n$) est définie par $\max(|\underline{X} - \underline{Y}|, |\overline{X} - \overline{Y}|)$ (resp.

$$\max\{dist(X_1, Y_1), \ldots, dist(X_n, Y_n)\}).$$

De la même façon, on définit la distance entre deux matrice.

On peut démontrer [?] que cette distance donne lieu à une topologie d'espace métrique $(I\!I(I\!R), dist)$. On trouve dans [?], une définition de la continuité, de la continuité uniforme et de propriétés analytiques sur les intervalles.

Soient A et B des nombres intervalles de $I\!I(I\!R)$. Les quatre opérations $\{\times, +, -, /\}$ sur $I\!I(I\!R)$ doivent satisfaire :

$$A\ op\ B \supseteq \{a\ op\ b | a \in A, b \in B\},\ \text{avec}\ op \in \{\times, +, -, /\}$$

Nous garderons la même propriété sur l'extension des autres opérations unaires, trigonométriques, etc.

La multiplication, l'addition, la soustraction et la division sont définies dans (13.7). [?, ?, ?, ?] ont proposé plusieurs améliorations pour avoir des bornes plus précises du calcul de la division par un intervalle contenant *zéro*.

$$\begin{array}{rcl} [a,b] + [c,d] & = & [a+c, b+d] \\ [a,b] - [c,d] & = & [a-d, b-c] \\ [a,b] \times [c,d] & = & \\ & & [\\ & & \quad \min(a\times c, a\times d, b\times c, b\times d), \\ & & \quad \max(a\times c, a\times d, b\times c, b\times d) \\ & &] \\ [a,b]/[c,d] & = & [a,b] \times [1/d, 1/c]\ \textbf{if}\ 0 \notin [c,d] \\ & & [-\infty, +\infty]\ \textbf{if}\ 0 \in [c,d] \end{array} \qquad (13.7)$$

La soustraction (resp. la division) n'est pas l'opération inverse de l'addition (resp. multiplication). La plupart des propriétés de distributivité de l'arithmétique réelle sont absentes de l'arithmétique des intervalles. [?] a démontré que les quatre opérations arithmétiques sur les intervalles définies par (13.7) sont continues, sauf la

division par des intervalles contenant *zéro*. Les fonctions rationnelles sur les intervalles sont aussi continues.

L'ensemble $I\!I(I\!R)$ est muni de la relation binaire $\leq$ définie par :

$$\begin{array}{l} X \in I\!I(R), Y \in I\!I(R), X \leq Y \equiv \overline{X} \leq \underline{Y} \\ X \in I\!I(R)^n, Y \in I\!I(R)^n, X \leq Y \equiv (\overline{X_i} \leq \underline{Y_i}, i = 1 \ldots n) \\ X \in I\!I(R)^{m \times n}, Y \in I\!I(R)^{m \times n}, \\ \quad X \leq Y \equiv (\overline{X_{i,j}} \leq \underline{Y_{i,j}}, i = 1 \ldots m, j = 1 \ldots n) \end{array} \tag{13.8}$$

L'ensemble des intervalles (resp. vecteurs d'intervalles et matrices d'intervalles) munis de la relation (13.8) a une structure de treillis.

En pratique, nous utiliserons l'arithmétique *flottante* des intervalles qui est une approximation discrète de l'arithmétique réelle des intervalles. Soient α et β deux inconnues réelles d'un calcul intermédiaire dans un calcul donné. Nous pouvons encadrer les deux inconnues $\alpha \in A$ et $\beta \in B$ par deux intervalles réels A et B. Les deux bornes des deux intervalles A et B sont des réels, donc ne peuvent être représentées sur machine. L'idée est de les représenter par les deux plus petits intervalles flottants [4] qui contiennent ces deux intervalles, soit A_f et B_f avec $A \subseteq A_f$ et $B \subseteq B_f$. Pour toute opération *op* de l'arithmétique réelle nous disposons de son équivalent en arithmétique réelle des intervalles, qui garantit que : $(A\ op\ B) \subseteq (A_f\ op\ B_f)$. Ce dernier intervalle n'est pas nécessairement représentable sur machine, nous le représenterons par le plus petit intervalle flottant $(A_f\ op\ B_f)_f$ qui contient cet intervalle réel. Ces deux passages à l'arithmétique flottante des intervalles nous mènent vers le principe de l'arithmétique flottante des intervalles [**?**] :

$$\alpha \in A, \beta \in B \Rightarrow \alpha\ op\ \beta \in (A_f\ op\ B_f)_f \tag{13.9}$$

Le principe de base de l'arithmétique réelle des intervalles et aussi l'arithmétique flottante des intervalles est le fait que le résultat soit toujours encadré par un intervalle. Cette propriété permet, dans les algorithmes de résolution d'équations, d'avoir des solutions *encadrées par intervalles* et non pas des solutions approchées.

En conclusion, en analyse par intervalles, un problème est résolu en trois étapes [**?**] :

1. la théorie est faite sur l'arithmétique des intervalles ;
2. le calcul est fait sur l'arithmétique flottante des intervalles ;
3. le principe d'inclusion (13.9) garantit la validité de la transition de l'arithmétique réelle des intervalles à l'arithmétique flottante des intervalles.

Plusieurs bibliothèques de l'arithmétique flottante des intervalles ont vu le jour, par exemple [**?**, **?**].

13.5.1 Extension des fonctions sur les intervalles

Soit $f : I\!R^n \rightarrow I\!R^m$.

4. En pratique, l'encadrement est implémenté en exploitant les deux fonctions *roundnear* et *rounddown*.

$exact(f)(X)$, avec $X \in I\!I(I\!R)^n$, est l'intervalle exact de variation de la fonction f dans l'intervalle X :

$$exact(f)(X) = \square\{f(x)|x \in X\}.$$

Une fonction $F : I\!I(I\!R)^n \to I\!I(I\!R)^m$ est appelée fonction d'inclusion de f si et seulement si :

$$\{f(x)|x \in X\} \subseteq F(X) \text{ pour tout } X \in I\!I(I\!R)^n \tag{13.10}$$

Les fonctions d'inclusion pour le cas matriciel sont définies de la même façon.

Le calcul de $exact(f)(X)$ est précis s'il n'existe pas d'occurrence multiple de variable (voir la section 13.5.2). Mais dans le cas général, le calcul de $exact(f)(X)$ de $f(x), x \in X$ est un problème indécidable (voir [?]). Calculer l'intervalle $\tilde{X}$, une approximation de $exact(f)(X) = Xe, X \in I\!I(I\!R)$ à ϵ près, défini par :

$$|\underline{\tilde{X}} - \underline{Xe}| \leq \epsilon \text{ et } |\overline{\tilde{X}} - \overline{Xe}| \leq \epsilon,$$

est un problème NP-difficile (voir [?]). En conséquence, les différentes fonctions d'inclusion que nous allons proposer sont des approximations *larges* de $exact(f)(X)$.

La définition 35 nous donne un moyen pour caractériser le comportement asymptotique des fonctions d'inclusion.

Définition 35 (ordre de convergence d'une extension)**.** *[?]*

Soit $F(X)$ l'extension sur les intervalles de $f : I\!R^n \to I\!R^m$ évaluée sur un domaine X. S'il existe une constante K, indépendante du domaine $X \in I\!I(I\!R)$ tel que

$$wid(F(X)) - wid(exact(f)(X)) \leq K wid(X)^\alpha$$

pour tous les domaines X avec $wid(X)$ suffisamment petit et un $\alpha > 0$ fixe, nous dirons que F est une fonction d'inclusion d'ordre α. Si α est égal à 1 (resp. à 2), alors nous dirons que l'inclusion est linéaire ou de premier ordre (resp. quadratique ou de deuxième ordre).

Exemple 7. *Voir les exemples donnés dans la sous-section* **??** *sur l'ordre de convergence d'une suite de nombres dans le cas général.*

13.5.2 Extension naturelle des fonctions

Soit $f : I\!R^n \to I\!R$ une fonction dont le calcul est défini sous forme d'une expression $f(x), x \in I\!R^n$, qui est composée des éléments suivants :

1. La variable x (ou une de ses composantes $\{x_1, \ldots, x_n\}$).
2. Des constantes réelles.
3. Les quatre opérations $\{+, -, \times, /\}$ de base de $I\!R$.
4. Des fonctions prédéclarées, $\{g_i\}_{i=1}^q$.
5. Des symboles syntaxiques (parenthèses, etc.).

L'extension naturelle sur les intervalles de la fonction f sur $X \in I\!I(I\!R)$ est définie [?] comme l'expression obtenue a partir de $f(x)$ en remplaçant :

1. chaque occurrence du vecteur réel x par le vecteur d'intervalles X,
2. chaque composante x_i par la composante intervalle X_i,
3. les opérations arithmétiques sur $\mathbb{R}$ par leurs correspondants dans l'arithmétique des intervalles,
4. et chaque occurrence d'une fonction prédéclarée g_i par son extension sur les intervalles G_i.

L'extension naturelle de $f(x)$, avec $x = (x_1, \ldots, x_n)$, de variables réelles $\{x_1, \ldots, x_n\}$ sur les intervalles est dénotée par $nat(f)(X)$, avec $X = (X_1, \ldots, X_n)$, de variables intervalles $\{X_1,\ldots,X_n\}$. Par construction, nous avons la propriété :

Théorème 13 (théorème fondamental de l'arithmétique des intervalles)**.**
[?] Soit $nat(f)(X)$ l'extension naturelle sur les intervalles de $f(x)$. Alors $nat(f)(X)$ contient toutes les valeurs de $f(x)$ pour tout $x \in X$.

Le théorème 13 est qualifié par [?, ?] de théorème fondamental de l'arithmétique des intervalles.

Il existe plusieurs formes syntaxiques possibles de l'extension naturelle d'une fonction, qui donnent lieu à différents résultats.

Exemple 8. *[?]*
Soit l'expression :

$$e(x) = \frac{x+1}{x}, \ \textit{avec } x \in \mathbb{R}$$

nous pouvons extraire les deux extensions naturelles :

$$e_1(X) = 1 + \tfrac{1}{X}, \ \textit{avec } X \in \mathbb{I}(\mathbb{R})$$
$$e_2(X) = \tfrac{X+1}{X}, \ \textit{avec } X \in \mathbb{I}(\mathbb{R})$$

L'évaluation des deux extensions naturelles sur l'intervalle $[1,2]$ donne :

$$e_1([1,2]) = 1 + \tfrac{1}{[1,2]} = [1.5, 2]$$
$$e_2([1,2]) = \tfrac{[1,2]+1}{[1,2]} = [1, 3]$$

La valeur donnée par $e_1([1,2])$ est plus précise que celle de $e_2([1,2])$, car e_2 contient une occurrence multiple de la variable X.

Quand il n'y a pas d'occurrence multiple de variables, l'évaluation donne des intervalles précis, plus généralement nous avons :

Théorème 14 (expression sans occurrence multiple de variables)**.**
Soit $f : \mathbb{R}^n \to \mathbb{R}$ dont l'expression est une suite d'opérations utilisant seulement les quatre opérateurs standards. Supposons que dans cette expression chaque variable a une seule occurrence. Alors, l'évaluation de l'extension naturelle $nat(f)(X)$ de l'expression de f dans un domaine X est égale à $exact(f)(X)$.

Lors de l'établissement d'une extension naturelle, il est souhaitable d'éviter des occurrences multiples de variables.

Dans le cas général, l'établissement de l'extension naturelle optimale est un problème ouvert dans l'analyse par intervalles. Il est donc difficile de choisir, parmi un ensemble d'extensions naturelles, l'extension qui donnera le résultat le plus précis. Le choix est fait généralement à partir de considérations heuristiques.

Le choix de l'extension naturelle a des conséquences décisives sur les performances du calcul scientifique utilisant l'arithmétique des intervalles. Ainsi, il est important de disposer d'un prétraitement pour essayer de transformer l'extension naturelle en une autre extension naturelle plus précise.

Théorème 15 (convergence linéaire de l'extension naturelle)**.**
[?]
L'extension naturelle d'une fonction sur les intervalles est d'ordre 1.

Le théorème 15 montre que la surestimation de l'extension naturelle $nat(f)(X)$ est bornée par $wid(X)$.

13.5.3 Extension de Taylor

Soit $f : \mathcal{D}' \subseteq I\!R^n \to I\!R$ une fonction dérivable à dérivées continues ou de classe C1. Soient $\mathcal{D} \subseteq \mathcal{D}'$ et le point $c \in \mathcal{D}$. L'extension de la valeur intermédiaire de f sur le vecteur d'intervalles $\mathcal{D} = \langle D_1, \ldots, D_n \rangle$ en un point $X \in I\!I(I\!R)^n$ et centrée sur c est définie par

$$tay(f)(\mathcal{D}, X) = \begin{cases} f(c) + F'(\mathcal{D})(X - c) \\ \textbf{ou} \\ f(c_1, \ldots, c_n) + \sum_{i=1}^{n} \frac{\partial F}{\partial X_i}(\mathcal{D})(X - c)_i \end{cases} \tag{13.11}$$

avec $F'(\mathcal{D})$ l'évaluation du gradient de f sur le vecteur d'intervalles $\mathcal{D}$.

Le gradient $F'(\mathcal{D})$ est calculé en utilisant une extension (naturelle par exemple) sur les intervalles de la dérivée scalaire. Soit $x \in X$, la formule (13.11) est aussi l'extension sur les intervalles de la formule de Taylor d'ordre 1 :

$$f(c) + (x - c)^T f'(\xi) \in f(c) + F'(\mathcal{D})(X - c)$$

où ξ est un point entre x et c.

La propriété fondamentale de cette extension est issue du théorème 16.

Théorème 16 (convergence quadratique de l'extension de Taylor)**.**
[?]
Supposons que les composantes de F' sont des extensions de f' d'ordre un. Alors $tay(F)(\mathcal{D}, X)$ est une extension d'ordre 2 de f.

Le théorème 16 montre que la surestimation de l'extension de Taylor $tay(F)(\mathcal{D}, X)$ est bornée par $wid(X)^2$. La formule des valeurs intermédiaires est donc efficace sur des domaines de petite taille. Sur des domaines de grande taille — plus précisément de largeur supérieure à un —, la formule des valeurs intermédiaires croit plus rapidement vers ∞ que l'extension naturelle sur les intervalles quand $wid(X) \to \infty$. En conséquence, il est préférable d'utiliser l'extension naturelle sur les domaines de taille importante, et la formule de Taylor sur les domaines de taille réduite. [**?**] proposent

l'heuristique suivante : utiliser la formule de Taylor d'une fonction sur un intervalle X seulement quand $wid(X) \leq 1/2$.

Nous pouvons définir une extension généralisée de la formule de Taylor d'ordre k par :

Définition 36 (extension de Taylor d'ordre k)**.** *La formule (fonction) de Taylor d'ordre $k \geq 1$ dans un domaine $\mathcal{D}$ de la fonction réelle $f(x)$, à n variables x_i, avec $i = 1 \ldots n$, est définie par*

$$tay^k(f)(X) = f(c) + \sum_{\lambda=1}^{k-1} \sum_{i_1,\ldots,i_\lambda} \frac{1}{\lambda!} \frac{\partial^\lambda f}{\partial x_{i_1} \ldots \partial x_{i_\lambda}}(c)(X-c)_{i_1} \ldots (X-c)_{i_\lambda} + \sum_{i_1,\ldots,i_k} \frac{1}{k!} \frac{\partial^k F}{\partial X_{i_1} \ldots \partial X_{i_k}}(\mathcal{D})(X-c)_{i_1} \ldots (X-c)_{i_k}$$

avec $X \in I\!I(\mathcal{D})$, $c = mid(X)$ et $x \in X$.

L'instance d'ordre 1, $tay^1(f)(X)$, est la formule des valeurs intermédiaires (13.11). L'extension de Taylor d'ordre $k > 1$ est aussi de convergence quadratique (voir [?]).

En pratique, à partir de l'ordre 2, l'extension sur les intervalles de la formule de Taylor devient trop coûteuse.

13.6 Analyse par intervalles

Cette section est une introduction aux méthodes d'analyse par intervalles (voir [?, ?, ?, ?] pour plus de détails). On limitera notre présentation aux méthodes basées sur Newton par intervalles pour résoudre des systèmes multi-variés d'équations nonlinéaires.

L'objectif est de trouver les zéros d'un système de n équations $f_i(x_1, \ldots, x_n)$ avec n inconnues x_i dans un vecteur d'intervalles $X = \{X_1, \ldots, X_n\}$ où $x_i \in X_i$ pour $i = 1, \ldots, n$.

En premier, soit le cas des équations linéaires définies comme suit :

$$\mathbf{A}.x = \mathbf{b} \tag{13.12}$$

avec $\mathbf{A}$ une matrice d'intervalles et $\mathbf{b}$ un vecteur d'intervalles. Résoudre ce système linéaire sur les intervalles nécessite la détermination d'un vecteur d'intervalles X_0 contenant toutes les solutions pour tous les systèmes linéaires classiques (scalaires) dénotés $A.x = b$ où $A \in \mathbf{A}$ et $b \in \mathbf{b}$. Trouver X_0 est un problème difficile, deux méthodes basiques par intervalles fournissent une surestimation du vecteur par intervalles X_1 contenant X_0.

Notons que l'algorithme d'élimination de Gauss fonctionne sur les systèmes linéaires sur les intervalles. De ce fait, il est possible d'obtenir X_1 si les pivots de Gauss ne contiennent pas zéro dans le processus de triangulation. C'est pourquoi, en général, l'interval calculé est assez large et l'étape de préconditionnement est nécessaire. En d'autres termes, on doit multiplier les deux cotés de l'équation 13.12 avec l'inverse de la matrice milieu de $\mathbf{A}$. La matrice $m(\mathbf{A})^{-1}\mathbf{A}$ est alors "très proche" de la matrice identité et la largeur de X_1 est assez petite [?].

Une autre méthodes bien connue dans la résolution des systèmes linéaires sur les intervalles, est la méthode itérative de Gauss-Seidel. Pour toute inconnue X_i, l'algorithme [?] est défini avec le processus itératif suivant :

$$X_i^{k+1} = (\mathbf{b}_i - \sum_{j=1}^{i-1} \mathbf{A}_{i,j} X_j^{k+1} - \sum_{j=i+1}^{n} \mathbf{A}_{i,j} X_j^k)/\mathbf{A}_{i,i} \cap X_i^k \tag{13.13}$$

Ici aussi, une étape de préconditionnement permettrait de réduire la taille des intervalles. Les détails sur l'implémentation de cet algorithme sont explicités dans [?]. Remarquons que cette méthode est très proche des méthodes de filtrage proposées en programmation par contraintes (voir le chapitre sur l'optimization globale).

Une autre alternative consiste à utiliser la méthode de Krawczyk. Cette méthode nécessite aussi un préconditonnement du système (see [?, ?]).

Pour résoudre un système nonlinéaire, l'algorithme de Newton est le plus souvent utilisé. Ci-dessous le schéma général :

$$X_{k+1} = N(\tilde{x}_k, X_k) \cap X_k \quad \text{with} \quad N(\tilde{x}_k, X_k) = \tilde{x}_k - A.f(\tilde{x}_k)) \tag{13.14}$$

A est une matrice d'intervalles qui contient toutes les inverses de la matrice Jacobienne du système F. Le scalaire $\tilde{x}_k$ doit être choisie dans X_k (par exemple le milieu X_k). Ainsi, les propriétés suivantes[5] sont vérifiées :

— Si $N(\tilde{x}_k, X_k) \cap X_k = \emptyset$, alors le système F a une seule solution dans X_k.

— Si $N(\tilde{x}_k, X_k)_k \cap X_k \subset X_k$, il existe une seule ou plusieurs solutions dans X_{k+1}.

La détermination de la matrice A est une étape délicate dans les algorithmes proposés. La matrice A est l'inverse de la matrice jacobienne $A = [F^{'}(X_k)]^{-1}$ évaluée sur le vecteur des intervalles X_k. C'est pourquoi, inverser la matrice sur les intervalles peut être couteuse en temps et en espace. L'alternative consiste à résoudre le système linéaire $F^{'}(X_k)(N(\tilde{x}_k, X_k) - \tilde{x}_k) = -f(\tilde{x}_k)$ pour déterminer $N(\tilde{x}_k, X_k)$. Ce travail peut être entretenu par l'un des algorithmes cités précédemment (see [?]).

Le schéma de Krawczyk offre des alternatives intéressantes. Il est défini par le schéma itératif suivant :

$$X_{k+1} = K(\tilde{x}_k, X_k) \cap X_k \text{ with } K(\tilde{x}_k, X_k) = \tilde{x}_k - [f(\tilde{x}_k)]^{-1} f(\tilde{x}_k) + (I - [f(\tilde{x}_k)]^{-1} F^{'}(X_k))(X_k - \tilde{x}_k) \tag{13.15}$$

Les propriétés de ce schéma sont utilisées par Moore [?] pour vérifier l'existance et l'unicité du zéro et la convergence de ce schéma. Notons que ce schéma utilise l'inverse de la matrice scalaire. Cette méthode est particulièrement rapide et efficace sur des intervalles réduits.

5. Pour les autres propriétés et l'implementation, voir [?]

Bibliographie

[Aribi, 2014] Aribi, N. (2014). Contribution à l'élicitation des paramètres en optimisation multicritère. Technical report, Thèse Docteur en Sciences, Université Oran1, Université de Nice-Sophia Antipolis.

[Bastin, 2010] Bastin, F. (2010). Modèles de recherche opérationnelle. Technical report, Support de cours, Département d'Informatique et de Recherche Opérationnelle https://www.iro.umontreal.ca/~bastin/Cours/IFT1575/IFT1575.pdf.

[Cook, 1971] Cook, S. A. (1971). The complexity of theorem-proving procedures. In Harrison, M. A., Banerji, R. B., and Ullman, J. D., editors, *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing, May 3-5, 1971, Shaker Heights, Ohio, USA*, pages 151–158. ACM.

[Cori et al., 2001] Cori, R., Hanrot, G., and Steyaert, J.-M. (2001). Conception et analyse des algorithmes. Technical report, Ecole polytechnique.

[Cormen et al., 1994] Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (1994). *Introduction à l'algorithmique (traduit par Xavier Cazin)*. Dunod.

[Cormen et al., 2001] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms, Second Edition*. The MIT Press.

[Gaudel et al., 1990] Gaudel, M.-C., Froidevaux, C., and Soria, M. (1990). *Types de données et algorithmes*. McGraw-Hill, Paris.

[Levin, 1973] Levin, L. A. (1973). Universal search problems (??????????????????????????). *Problems of Information Transmission (??????????????????????????)*, 9(3).

[Papadimitriou et al., 2006] Papadimitriou, C. H., Dasgupta, S., and Vazirani, U. (2006). *Algorithms*. McGraw Hill.

[Prins, 1994] Prins, C. (1994). *Algorithmes de graphes*. Eyrolles.

[R. Faure, 1995] R. Faure, B. Lemaire, C. P. (1995). *Précis de recherche opérationnelle : Méthodes et exercices d'application*. Eyrolles.