Optimisation en dimension finie

G. Vinsard

15 juillet 2009

Ce polycopié correspond à la version 2004–2005 de l'enseignement que j'ai prodigué pendant près de dix ans.

M Antoine qui assure cet enseignement à partir de l'année 2005-2006 me fait l'honneur de désirer que ce texte soit distribué aux étudiant et je l'en remercie.

Toutefois, il est certain qu'il y aura des variations (positives) entre son cours et celui qui correspond au polycopié. Aussi je recommande aux étudiants la plus grande prudence dans la consultation de ce polycopié qui n'est fourni qu'à titre indicatif.

Résumé

Les méthodes de minimisation sans contrainte sont traitées ; notamment la question de la convergence des méthodes de Newton et du gradient ; les méthodes à métrique variable sont brièvement expliquées.

Les méthodes de minimisation avec contraintes égalité et inégalité sont traitées dans le cas d'une seule contrainte ; notamment les calculs sont poussés jusqu'à l'ordre deux.

Le cas de plusieurs contraintes égalités est également brièvement traité mais, caetera desunt*, pas celui de plusieurs contraintes inégalités.*

Un nombre important de références bibliographiques est fourni et on a essayé de munir ce texte d'un guide de lecture de ces références.

Une brève analyse du choix des logiciels à faire pour traiter des problèmes d'analyse numérique est également incluse.

Table des matières

Introduction

Les gens pressés lisent rarement les introductions. Ils ne prennent pas la peine de méditer sur ce petit paradoxe selon lequel on ne peut jamais rien trouver.

En effet si on considère une chose cherchée alors celui qui la cherche connaît cette chose ou ne la connaît pas. Dans le premier cas il n'a pas besoin de la chercher puisqu'il la connaît ; dans le second cas comment pourrait-il la reconnaître puisqu'il ne la connaît pas ?

C'est pourquoi les introductions contiennent généralement des indications préliminaires sur le sujet qui va être traité ; en quelque sorte un portrait robot de ce sujet qui ne sera connu cependant qu'après la lecture du corps du texte.

C'est donc important de lire une introduction ; et cela procure un avantage aux gens qui ne sont pas pressés sur ceux qui le sont.

Intérêts de l'enseignement

Voici déjà trois classes de raisons pour s'intéresser à cet enseignement.

Intérêt physique

On sait qu'un certain type de forces est représenté par le vecteur obtenu en faisant la variation première de l'énergie potentielle par rapport à ces arguments.

On sait aussi que là où la force disparaît, il suffit que la vitesse soit nulle pour que le mouvement puisse ne pas se produire.

Cet état s'appelle l'équilibre.

Déterminer les états possibles d'équilibre d'un système physique est sûrement utile ; on peut formuler cette recherche en terme mathématique par la recherche de l'argument minimisant d'une fonction, l'énergie potentielle ; ce qui est l'objet principal de l'enseignement.

Intérêt économique

Prenons un exemple.

On connaît la loi de distribution du nombre de clients susceptibles d'acheter un produit ; cette loi dépend d'un certain nombre de paramètres qui sont : le prix de vente, l'investissement publicitaire, l'investissement fait pour disposer d'un magasin doté de facilités d'accès et d'un personnel plus ou moins nombreux ...

Comment choisir ces paramètres de manière à maximiser les profits compte tenu d'autres paramètres qui sont : le prix d'achat en gros du produit, le prix de l'immobilier et les coûts salariaux ?

Voilà encore un problème qui peut être représenté en terme mathématique la recherche de l'argument minimisant d'une fonction, le gain net.

Intérêt métaphysique

Les deux premiers points sont de nature purement utilitaires et il n'est pas exclu que, pour cette raison, ils ne convainquent pas ceux qui étudient pour le plaisir d'étudier.

Il est possible de faire remarquer à ceux-ci qu'une question particulièrement intéressante est savoir si une description du monde sensible peut être entièrement formulée en terme mathématique.

Personne n'en sait rien tant que l'opération n'est pas réalisée et il est certain qu'elle ne le sera pas avant très longtemps.

Il n'en est pas moins vrai que les principes d'extremums (pas extrema) ont joué jusqu'ici un tel rôle en physique que négliger délibérément leur étude serait une omission de même importance que celles qui consisterait à ignorer Kant dans la théorie de la connaissance ; Piaget dans une doctrine pédagogique ; voire Maxwell dans la théorie de l'électricité.

Stratégie de rédaction

Qui se souvient de la formule d'extrapolation de Richardson ? — Les lecteurs forment une partition de deux classes comportant : ceux d'entre eux qui sont incapables de donner les termes de ce qu'on appelle la formule d'extrapolation de Richardson ; et ces autres qui en sont capables.

Le parti pris de ce texte est alors que les éléments de la seconde classe ne se sentiront pas insultés si la formule d'extrapolation de Richardson est rappelée au profit des éléments de la première classe.

Voici ce qui caractérise la tactique de rédaction ; toutes les techniques mathématiques sont rappelées.

Mais pour ne pas alourdir le texte par trop de considérations connexes, la stratégie de rédaction est elle d'éviter de développer un cadre mathématique trop strict.

Par exemple, plutôt que d'introduire des emboîtements d'espaces de fonctions classés selon leur dérivabilité, on se contentera de parler de "la classe des fonctions suffisamment dérivables pour l'usage qu'on en fait". Tant pis pour les autres ! Et le soin de trouver si tel ou tel autre problème particulier relève de ce texte est laissé à l'appréciation du lecteur.

De plus la lecture de la trilogie (l'enfant, le bachelier et l'insurgé) de Jules Vallès a convaincu l'auteur qu'il faut

> *"Des exemples ! Sans exemples je ne comprend rien"*
> *Jules Vallès*

Aussi ceux-ci sont nombreux.

Contenu du texte

Minimisation sans contrainte

On commence par discuter de la "minimisation sans contrainte de fonctions" f dont l'argument est un vecteur x de N nombres réel et la valeur un nombre réel.

La première classe de méthode de minimisation qui, partant d'un vecteur x_0 initial permet de construire un vecteur x_1 tel que $f(x_1) \leq f(x_0)$ en n'utilisant que x_0 est introduite sous les deux formes complémentaires que sont les méthodes de Newton et du gradient.

Ces deux méthodes sont étudiées dans leur principe et cette étude est complétée par celle de leur convergence. On insiste notamment sur la "convergence locale" de la méthode de Newton et la "convergence globale" de la méthode du gradient.

La seconde classe des méthodes à métrique variable est introduite en la présentant comme une imitation de la méthode de Newton qui peut être fait en utilisant le point x_0 mais aussi ceux qui le précèdent.

La description de la méthode du gradient conjugué pour les systèmes linéaires a été intercalée entre celles des deux classes pour fournir une illustration de la dernière.

Minimisation avec une seule contrainte égalité

Ensuite on discute de la "minimisation sans contrainte de fonctions" f dont l'argument est un vecteur x de N nombres réel assujettis à appartenir à une hypersurface de l'espace de tous les N nombres possibles et la valeur un nombre réel.

On insiste particulièrement sur le cas où l'hypersurface est différentiable et dépourvue de singularité afin de montrer la dualité entre une représentation par équation et par paramétrisation de cette hypersurface.

On utilise la possibilité de paramétrisation pour déduire les "méthodes lagrangiennes" de résolution.

Les investigations sont poussées jusqu'au second ordre de dérivation.

Enfin on étudie les "méthodes de pénalisation".

Minimisation avec une seule contrainte inégalité

Puis on discute de la 'minimisation sans contrainte de fonctions" f dont l'argument est un vecteur x de N nombres réel assujettis à être situés d'un seul côté d'une hypersurface de l'espace de tous les N nombres possibles, pour autant que ce soit possible, et la valeur un nombre réel.

On montre que ce problème se ramène soit au cas de la minimisation sans contrainte soit au cas de la minimisation avec une contrainte d'égalité ; puis on donne le formalisme qui permet de regrouper ces deux cas en un seul.

Les méthodes de pénalisations appropriée au cas de la contrainte inégalité sont également expliquées.

Minimisation avec plusieurs contraintes

Le cas des contraintes égalité est traité, un peu comme une recopie *mutatis mutandi* du cas où il n'y avait qu'une seule contrainte, mais en prenant en compte néanmoins sa spécificité.

Le cas de multiples contraintes inégalités n'est pas traité. Même pas dans sa version de programmation linéaire.

Les logiciels

Traiter de problèmes d'analyse numérique sans évoquer les logiciels de calcul c'est un peu étudier l'eau sèche (pour une explication de cette référence on pourra consulter [Fey95, tome 2, pp. 377–411]) : c'est nécessaire mais pas suffisant.

On a joint alors une partie traitant des logiciels de calculs en mettant un accent particulier sur un logiciel de calcul formel facile d'utilisation.

La bibliographie

Une bibliographie importante a été donnée. La majeure partie des titres dépassent le cadre de l'optimisation et même de l'analyse numérique en général mais tous ces titres traitent néanmoins du sujet même si c'est parfois de manière peu apparente.

De manière à reconnaître ceux des ouvrages qui intéressent le plus directement le sujet traité, l'optimisation, les titres et les pages de ces titres correspondant aux références sont marqué explicitement dans le texte.

De plus un petit guide de lecture est placé avant la liste de référence.

L'esprit de l'analyse numérique

Avant d'entrer dans la problématique de l'optimisation une recommandation peut être faite.

Généralement un problème de minimisation sous contraintes vient d'un problème physique ; la modélisation de ce problème conduit à un problème mathématique. On aurait grand tort de reporter toute la difficulté de l'ensemble sur la partie mathématique ; bien souvent une formulation maladroite conduit à une complexité mathématique importante alors qu'une formulation plus adroite peut conduire à un problème simple.

Pour illustrer ce dernier propos sans déflorer le sujet de l'optimisation, on peut rappeler ce qu'est la méthode d'intégration de Gauss.

Il s'agit de trouver une formule

$$\int_0^1 f(x) \quad dx \approx \sum_{n=0}^{N} w_n f(x_n)$$

qui serait exacte pour des polynômes de degré le plus grand possible.

Pour cela on peut déjà introduire un polynôme de degré P et écrire

$$\sum_{p=0}^{P} \alpha_p \int_0^1 x^p \quad dx = \sum_{n=0}^{N} \sum_{p=0}^{P} \alpha_p w_n x_n^p$$

ce qui permet déjà de trouver les w_n en fonction des x_n en choisissant de limiter le degré P à $P = N$.

Si maintenant on veut augmenter encore le degré pour lequel la formule est exacte en choisissant convenablement les x_n on est conduit à des équations non-linéaires en x_n dont il paraît difficile de trouver les solutions.

Qu'est-ce alors que la méthode de Gauss ? Loin de chercher à résoudre ces équations non-linéaires, on cherche au contraire à reformuler le problème en remarquant qu'un polynôme $P(x)$ de degré $P > N$ peut s'écrire l'écrivant

$$P(x) = Q(x)K(x) + R(x)$$

où $K(x)$ est un polynôme donné, $Q(x)$ et $R(x)$ le quotient et le reste de la division de $P(x)$ par $K(x)$.

De cette façon on s'intéresse déjà au produit scalaire

$$< Q, K >= \int_0^1 Q(x)K(x) \quad dx$$

et notamment on remarque qu'on peut construire la suite de polynômes orthogonaux $K_0 = 1$, $K_1(x) = x - 1/2K_0(x)$,...par le procédé de Schmidt.

Si donc on choisissait pour $K(x)$ le polynôme $K_{N+1}(x)$, on pourrait écrire un polynôme quelconque de degré N comme

$$P(x) = Q(x)K_{N+1}(x) + R(x)$$

où le degré de $Q(x)$ est inférieur ou égal à N ainsi que le degré de $R(x)$.

De cette façon

$$\int_0^1 P(x) \quad dx = \int_0^1 R(x) \quad dx$$

puisque $Q(x)$ est alors nécessairement orthogonal à $K_{N+1}(x)$.

Il suffit alors de choisir les x_n de la formule d'intégration comme les zéros de $K_{N+1}(x)$ de manière que

$$P(x_n) = R(x_n)$$

et on dispose alors d'une formule d'intégration demandant le calcul de $N+1$ termes et qui est valable pour des polynômes de degré $2N + 1$.

Voilà une formulation adroite du problème ! La formulation maladroite eut été de vouloir résoudre de façon directe les équations non-linéaires.

Il n'y a pas de 'problèmes bourins' (sic ! Entendu lors d'une séance d'exercice où il était demandé de faire un calcul un peu sérieux.) mais seulement des formulations maladroites de ces problèmes qui rendent leur résolution 'bourine'.

Première partie
Rappels

Ces rappels ont deux objectifs : rappeler bien sûr, mais aussi habituer le lecteur aux notations utilisées dans la suite.

Il se peut que ce ne soit pas des rappels ; pour pallier ce cas, des indications bibliographiques ont été portées.

1 Éléments disparates

1.1 Analyse

1.1.1 Différentiabilité

La différentiabilité des fonctions de plusieurs variables est traitée dans la liste (non exhaustive) : [LS89, pp : 215–227], [Sch81, pp : 192–204], [KF94, pp : 475–489], [Car77, pp : 64–79], [Rud95b, pp : 197–204], [Rai93, pp : 88-128], [SG89, tome 2, pp : 191–229].

On donne ici un rappel des résultats qui sert également de familiarisation avec les notations utilisées.

Une fonction f est différentiable deux fois au point x si elle peut être écrite

$$f(x'') = f(x + \delta\lambda x') = f(x) + \delta\lambda \nabla f(x) \cdot x' + \frac{1}{2}(\delta\lambda)^2 \nabla^2 f(x) \cdot (x', x') + o(\delta\lambda^2) \tag{1}$$

où

- o est une fonction réelle quelconque (mais dépendant de f) telle que

$$\lim_{\epsilon \to 0} \frac{o(\epsilon)}{\epsilon} = 0 \tag{2}$$

- la notation $\nabla f(x) \cdot x'^n$ signifie :

$$\nabla f(x) \cdot x^n = \sum_{n=1}^{N} \frac{\partial f}{\partial x^n}(x) x^n \tag{3}$$

  si les composantes du vecteur x sont notées $x^1, \dots x^N$.
  $\nabla f(x)$ s'appelle le gradient de f.
- la notation $\nabla^2 f(x) \cdot (x', x')$ signifie :

$$\nabla^2 f(x) \cdot (x', x') = \sum_{n=1}^{N} \sum_{m=1}^{N} \frac{\partial^2 f}{\partial x^n x^m}(x) x'^n x'^m \tag{4}$$

  $\nabla^2 f(x)$ s'appelle le Hessien de f ou encore le Jacobien de $\nabla f(x)$.

Cette définition de la différentiabilité est celle de Fréchet :

- les dérivées premières sont représentées par une forme linéaire (notée $\nabla f(x)\cdot$) dont les composantes sont les dérivées partielles premières de la fonction $f(x)$ et qui s'applique à un vecteur quelconque x' de E_N.
- les dérivées secondes sont représentées par une forme bilinéaire symétrique (notée $\nabla^2 f(x)\cdot$) dont les composantes sont les dérivées partielles secondes de la fonction $f(x)$ et qui s'applique a priori à un couple de vecteurs quelconque (x', x'') de E_N. Ici on n'utilise cette forme bilinéaire symétrique que pour $x'' = x'$.

Il est possible d'avoir une autre définition de la différentiabilité (celle de Gâteaux) en considérant les dérivées dans la direction du vecteur x' et en n'assimilant pas nécessairement ces dérivées en des formes linéaires et bilinéaires comme précédemment.

La différence entre ces deux définitions est classiquement illustrée par la fonction

$$\begin{array}{rcccl} f & : & E_N & \longrightarrow & R \\ & & x & \longrightarrow & f(x) = |x| = \sqrt{\sum_{n=1}^{N} (x^n)^2} \end{array} \tag{5}$$

Cette fonction admet une dérivée première en $x = 0$ dans chacune des directions x' mais n'admet pas de développement (même limité à l'ordre 1) tel que (1) en ce point.

C'est du au manque de continuité de la dérivée dans une direction en $x = 0$. En effet s'il n'y a pas continuité de la dérivée dans une direction, il n'existe pas un opérateur linéaire unique qui traduit celle-ci ; il y en a un à gauche et un à droite mais ils ne coïncident pas[1]

Le lien entre les deux définitions de la différentiabilité est alors que celles ci coïncident si les dérivées dans toutes les directions sont continues, ce que ne vérifie pas la fonction 'valeur absolue'.

[1] toutefois si on est près à abandonner l'idée que la dérivée doit être un élément simple (un nombre, un vecteur, ...) et qu'on veuille accepter qu'elle puisse être plutôt un ensemble alors la notion est prête, c'est celle de sous-gradient.

Formule de Taylor avec reste Quand on utilise la formule de Taylor (1) avec la notation en $o()$, qui est très pratique dans les calculs où on sait qu'on finira par tronquer les expressions de cet $o()$, on oublie cependant un intermédiaire des raisonnements qui conduisent à elle et qui est la formule de Taylor avec reste (pour les démonstrations voir par exemple [SG89, tome 2, pp : 225–228]).

Si f est une fonction dérivable (indéfiniment pour ne pas être gêné) de $R \longrightarrow R$ alors on peut écrire que

$$\begin{aligned} &\exists\ \theta \in [0,1] \\ &f(x+x') = f(x) + \frac{df}{dx}(x)x' + \ldots + \frac{1}{n!}\frac{d^n}{dx^n}f(x)x'^n + \frac{1}{(n+1)!}\frac{d^{n+1}}{dx^{n+1}}f(x+\theta x')x'^{n+1} \end{aligned} \tag{6}$$

Cette formule peut être généralisée dans le cas des fonctions à variables dans E_N par

- pour l'ordre 0 (0 si on considère que le reste correspond à la fonction o)

$$\exists\ \theta \in [0,1] \quad f(x+\delta\lambda x') = f(x) + \delta\lambda \nabla f(x+\delta\lambda\theta x') \cdot x' \tag{7}$$

- pour l'ordre 1

$$\begin{aligned} &\exists\ \theta \in [0,1] \\ &f(x+\delta\lambda x') = f(x) + \delta\lambda\nabla f(x)\cdot x' + \frac{\delta\lambda^2}{2}\nabla^2 f(x+\delta\lambda\theta x')\cdot(x',x') \end{aligned} \tag{8}$$

Le cas des ordres supérieurs existe aussi mais ne sera pas utilisé ici.

1.1.2 Conditions nécessaire et suffisante de minimum local

Les conditions nécessaire et suffisante de minimum pour une fonction différentiable 2 fois sont traitées dans la liste (non exhaustive) : [LS89, pp : 215–227], [Sch81, pp : 350–392], [KF94, pp : 489–494], [Car77, pp : 96–102], [Rud95b, pp : 197–204], [Rai93, pp : 129–138].

On donne ici un rappel des résultats qui sert également de familiarisation avec les notations utilisées.

Si la fonction f est différentiable une fois au point x_∞, alors elle peut être écrite comme

$$f(x'') = f(x_\infty + (x'' - x_\infty)) = f(x_\infty) + \nabla f(x_\infty)\cdot(x'' - x_\infty) + o(|x'' - x_\infty|) \tag{9}$$

Si $|\nabla f(x_\infty)| \neq 0$, on peut choisir

$$x'' = x_\infty + \delta\lambda \nabla f(x_\infty) \tag{10}$$

et réécrire (9) comme

$$f(x_\infty + \delta\lambda\nabla f(x_\infty)) = f(x_\infty) + \delta\lambda(\nabla f(x_\infty))^2 + o(\delta\lambda|\nabla f(x_\infty)|) \tag{11}$$

Pour $\delta\lambda$ suffisamment proche de 0, on peut donc affirmer que le signe de $f(x_\infty + \delta\lambda\nabla f(x_\infty)) - f(x_\infty)$ est celui de $\delta\lambda$. En conséquence le point x_∞ n'est pas un minimum puisqu'il ne peut pas satisfaire (215). En renversant la proposition on obtient la

Condition nécessaire f étant différentiable une fois au point x_∞ , si x_∞ est un minimum local alors $|\nabla f(x_\infty)| = 0$, soit

$$\nabla f(x_\infty) = 0 \tag{12}$$

Si la fonction f est différentiable deux fois au point x_∞ et que $|\nabla f(x_\infty)| = 0$, alors elle peut être écrite comme

$$f(x'') = f(x_\infty + \delta\lambda x') = f(x_\infty) + \frac{1}{2}(\delta\lambda)^2\nabla^2 f(x_\infty)\cdot(x',x') + o(\delta\lambda^2) \tag{13}$$

Pour $\delta\lambda$ suffisamment proche de 0,

- si $\nabla^2 f(x_\infty)\cdot(x',x') > 0$ pour tout $x' \neq 0$ alors il est clair que $f(x_\infty) \leq f(x'')$ et on obtient la

Condition suffisante f étant différentiable deux fois au point x_∞, si le Hessien est défini positif alors x_∞ est un minimum local

- si $\nabla^2 f(x_\infty)\cdot(x',x') \geq 0$ et qu'il existe une ou plusieurs directions x' pour lesquelles 0 est atteint alors il faut pousser plus loin le développement de Taylor dans ces directions pour pouvoir conclure
- si $\nabla^2 f(x_\infty)\cdot(x',x')$ peut être indifféremment négatif ou positif suivant x' alors le point x_∞ n'est pas un minimum, c'est juste un point stationnaire

1.1.3 Théorème du point fixe

Démonstration [Sch81, pp : 101–102], [SM84, pp : II-28–30]

Si a est est une fonction de E_N dans E_N :

- telle que l'image $a(D)$ d'un domaine D fermé et borné de E_N est incluse dans ce domaine ($a(D) \subset D$
- lipschitzienne (et a est alors absolument continue puisque la constante k ne dépend pas de x ou y) soit

$$\forall x, y \in D \quad |a(x) - a(y)| \leq k|x - y| \tag{14}$$

- contractante, c'est à dire lipschitzienne avec un coefficient $k < 1$

alors l'itération

$$x_{n+1} = a(x_n) \tag{15}$$

démarrée d'un point quelconque $x_0 \in D$, converge vers un unique point fixe x_∞ (donc tel que $a(x_\infty) = x_\infty$) avec une vitesse de convergence satisfaisant à

$$|x_n - x_\infty| \leq \frac{k^n}{1-k}|x_1 - x_0| \tag{16}$$

ou encore

$$|x_{n+1} - x_\infty| \leq \frac{k}{1-k}|x_{n+1} - x_n| \tag{17}$$

Le théorème du point fixe sert essentiellement à fixer le cadre théorique de la résolution d'équations sous la forme d'itérations successives avec (15).

1.2 Algèbre linéaire

L'algèbre linéaire est traitée dans la liste (non exhaustive) : [SM84, Tome 1], [Cia82, pp : 3–134], [BBC+91, pp : 4–26].
On donne ici un rappel de résultats utiles pour la suite.

1.2.1 Résultats et définitions

Matrices particulières

- matrice adjointe et orthogonale : si les coefficients de A sont réels la matrice adjointe est la transposée ${}^tA$. Si A est inversible et à coefficients réels, elle est orthogonale si $A^{-1} = {}^tA$.
- matrice définie positive : A est positive si

$$\forall x \in E_N \quad {}^txAx \geq 0 \tag{18}$$

elle est définie si de plus

$$ {}^txAx = 0 \quad \Rightarrow \quad x = 0 \tag{19}$$

Il existe des critères pratiques pour savoir si une matrice est définie positive : notamment le critère de Silvester [Pos81a, p 100] qui correspond au critère qu'on peut déduire de l'algorithme de décomposition de Choleski.

Diagonalisation Si la matrice A est symétrique alors elle peut être diagonalisée ; la recherche des valeurs propres et de leurs vecteurs propres associés est difficile numériquement mais il ne faut pas oublier que

- la trace de la matrice est la somme des valeurs propres ;
- le déterminant de la matrice est le produit des valeurs propres (qu'on obtient alors quasiment directement dans le cas $N = 2$).

1.2.2 Résolution de systèmes linéaires

Le problème de trouver pratiquement x_∞ tel que

$$Ax_\infty = b \tag{20}$$

où x_∞ et b sont des vecteurs de E_N et A une matrice $N \times N$ symétrique peut être résolu par une méthode numérique :

- directe : il s'agit alors d'une factorisation de type LU (Gauss), de type Choleski pour les matrices définies positives ou encore de type Householder.
- itérative : il s'agit de méthodes de décomposition du type $A = M - n$ qui peuvent être de type Jacobi ou Gauss-Seidel (avec ou sans paramètre de relaxation). Les méthodes de type gradient et gradients conjuguées sont étudiées ici § 5.1, page 22.

Le choix d'une méthode de résolution dépend du type de matrice envisagé, c'est à dire de ses qualités (symétrique ou non, définie positive ou non) et aussi de sa dimension (grand ou petit système). Une synthèse très claire peut être trouvée dans [BBC+91].

1.2.3 Produit dyadique

Il est possible de fabriquer une matrice de terme général $u^n v^m$ (n pour la ligne m pour la colonne) sont les produits des composantes de deux vecteurs $u = (u^1 \dots u^N)$ et $v = (v^1 \dots v^N)$ de E_N. Cette matrice est le produit dyadique qui se note

$$]u, v[$$

et peut être redéfinie comme l'opérateur linéaire tel que pour un vecteur w de E_N

$$]v, w[u = ({}^twu)v \tag{21}$$

L'intérêt de ce produit dyadique[2] est qu'il permet des manipulations faciles de propriétés sur les matrices et de vecteurs.

Orthogonalité Si on a besoin de tous les vecteurs x orthogonaux, pour le produit scalaire euclidien, à un vecteur u, on peut écrire ceux-ci comme

$$\forall\ y \in E_N \quad x = y - \frac{{}^tyu}{{}^tuu}u \tag{22}$$

mais également

$$\forall\ y \in E_N \quad x = y - \frac{]u, u[}{{}^tuu}y \tag{23}$$

ce qui constitue une écriture plus souple, si le calcul mené est orienté "opérateur".

[2] On a préféré utiliser ce nom ancien dû à Gibbs plutôt que produit tensoriel afin de ne pas subir toutes les connotations de mot 'tenseur'

Inversibilité de proche en proche Si A est une matrice $N \times N$ inversible, u, v deux vecteurs de E_N tels que

$$1 + {}^t u A^{-1} v \neq 0 \tag{24}$$

alors on a

$$(A+]u,v[)^{-1} = A^{-1} - \frac{A^{-1}]u,v[A^{-1}}{1 + {}^t u A^{-1} v} \tag{25}$$

Cette formule peut être vérifiée facilement. La signification de la condition (24) est qu'une projection n'est pas inversible. En effet si on considère

$$I-]u,u[\tag{26}$$

alors

$$(I-]u,u[)^{-1} = I + \frac{]u,u[}{1-u^2} \tag{27}$$

qui n'existe que si $u^2 \neq 1$, c'est à dire que si $I-]u,u[$ n'est pas l'opérateur de projection sur l'hyperplan orthogonal à u. $I-]u,u[$ est en effet une homothétie de rapport $1-u^2$ dans la direction u et de rapport 1 dans les autres directions.

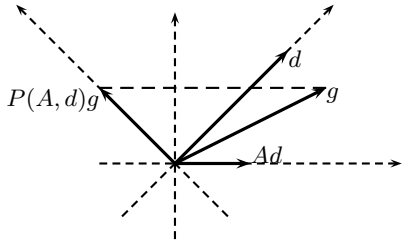
Produits Si A est une matrice symétrique

$$\begin{array}{lcl} A]v,w[& = &]Av,w[\\ ]v,w[A & = &]v,Aw[\\ ]v,w[]v',w'[& = & {}^t wv']v,w'[\end{array} \tag{28}$$

Projection oblique Si on introduit

$$P(A,d) = I - \frac{]Ad,d[}{{}^t dAd} \tag{29}$$

où A est une matrice définie positive (et alors l'angle entre les vecteur d et Ad est strictement compris entre $-\pi/2$) et $\pi/2$ et d est un vecteur, alors $P(A,d)$ est la projection oblique sur l'hyperplan orthogonal à d dans la direction Ad



On a, en utilisant (28),

$$P(A,d)A = \left(I - \frac{]Ad,d[}{{}^t dAd}\right) A = A\ {}^t P(A,d) \tag{30}$$

où ${}^t P(A,d)$ est la transposée de $P(A,d)$ ainsi que la projection oblique sur l'hyperplan orthogonal à Ad dans la direction d. et les propriétés

$$\begin{array}{lcl} P(A,d)Ad & = & 0 \\ {}^t P(A,d)d & = & 0 \end{array} \tag{31}$$

On a aussi

$${}^t dAd' = 0 \Longrightarrow P(A,d)P(A,d') = P(A,d')P(A,d) \tag{32}$$

En effet

$$\begin{array}{rcl} P(A,d)P(A,d') & = & \left(I - \frac{]Ad,d[}{{}^t dAd}\right)\left(I - \frac{]Ad',d'[}{{}^t d'Ad'}\right) \\ & = & I - \frac{]Ad,d[}{{}^t dAd} - \frac{]Ad',d'[}{{}^t d'Ad'} + \frac{]Ad,d[]Ad',d'[}{{}^t dAd\ {}^t d'Ad'} \\ & = & I - \frac{]Ad,d[}{{}^t dAd} - \frac{]Ad',d'[}{{}^t d'Ad'} + \frac{{}^t dAd']Ad,d'[}{{}^t dAd\ {}^t d'Ad'} \end{array} \tag{33}$$

d'où (32) : les projections obliques commutent pour des vecteurs d et d' tels que d et Ad' (ou d' et Ad) soient orthogonaux. De plus (démonstration analogue à la précédente, par inspection)

$${}^t dAd' = 0 \Longrightarrow P(A,d)Ad' = Ad' \tag{34}$$

1.3 Formules diverses

1.3.1 Inversion de $A + \delta\lambda B$

Si La matrice A est inversible et que $\delta\lambda$ est aussi petit qu'on veut alors $A + \delta\lambda B$ est inversible et son inverse est au premier ordre en $\delta\lambda$

$$(A + \delta\lambda B)^{-1} = A^{-1} - \delta\lambda A^{-1} B A^{-1} + \ldots \tag{35}$$

1.3.2 Minimisation de $\frac{1}{2}\ {}^t xAx - {}^t bx$

Si x_∞ est le vecteur qui minimise $f(x) = \frac{1}{2}\ {}^t xAx - {}^t bx$ alors le développement algébrique de

$$f(x_\infty + \delta\lambda x') = \frac{1}{2}\ {}^t(x_\infty + \delta\lambda x')A(x_\infty + \delta\lambda x') - {}^t b(x_\infty + \delta\lambda x') \tag{36}$$

peut être rangé en puissance croissante de $\delta\lambda$ comme

$$f(x_\infty + \delta\lambda x') = f(x_\infty) + \delta\lambda\ {}^t x' \left(\frac{1}{2}(A + {}^t A)x_\infty - b\right) + \frac{1}{2}\delta\lambda^2\ {}^t x' A x' \tag{37}$$

d'où on tire que nécessairement x_∞ est solution de

$$\frac{1}{2}(A + {}^t A)x_\infty = b \tag{38}$$

et que c'est suffisant si A est définie positive.

1.3.3 Extrapolation de Richardson

Si une fonction f différentiable deux fois dont l'argument est un vecteur x de E_N et la valeur un nombre réel est donnée, alors

$$f(x + \delta\lambda x') = f(x) + \delta\lambda \nabla f(x) \cdot x' + \frac{\delta\lambda^2}{2}\nabla^2 f(x) \cdot (x', x') + o(\delta\lambda^2)$$

on a aussi

$$f(x + k\delta\lambda x') = f(x) + k\delta\lambda \nabla f(x) \cdot x' + k^2\frac{\delta\lambda^2}{2}\nabla^2 f(x) \cdot (x', x') + o(\delta\lambda^2)$$

d'où il s'ensuit que

$$kf(x + \delta\lambda x') - f(x + k\delta\lambda x') = (k-1)f(x) + k(1-k)\frac{\delta\lambda^2}{2}\nabla^2 f(x) \cdot (x', x') + o(\delta\lambda^2)$$

et donc que si on approxime $f(x)$ par

$$\frac{f(x + k\delta\lambda x') - kf(x + \delta\lambda x')}{1-k}$$

plutôt que par

$$f(x + \delta\lambda x')$$

alors, si $\delta\lambda x' \neq 0$ quoique $\delta\lambda$ soit petit, la première approximation est meilleure que la première puisque la différence entre $f(x)$ et son approximation est un terme en $\delta\lambda^2$ plutôt qu'un terme en $\delta\lambda$.

Cela s'appelle une extrapolation de Richardson ; et c'est utile si $f(x)$ est une quantité finie difficile à calculer numériquement.

Par exemple si $N = 1$, $x = 0$ et $f(x) = \sin x/x + x - 1$ l'extrapolation de Richardson est très efficace.

x	$f(x) = \sin x/x + x - 1$	$(f(kx) - kf(x))/(1-k)$ pour $k = 1/2$
0.1	-0.09833416646828153	-0.0008326043588515741
0.01	-0.009983333416666351	-8.33326041682625e-06
0.001	-0.0009998333333416376	-8.333332601750953e-08
0.0001	-9.999833333340646e-05	-8.333334022836425e-10

avec deux calculs pour 0.1 et 0.05 on obtient la précision d'un calcul pour 0.001.

Cependant s'il se trouve que le point x annule ∇f alors l'extrapolation de Richardson donne des résultats décevants.

Par exemple pour $N = 1$, $x = 0$ et $f(x) = \sin x/x$, on trouve (la colonne de l'approximation de Richarson est évidemment la même que précédemment)

x	$f(x) = \sin x/x - 1$	$(f(kx) - kf(x))/(1-k)$ pour $k = 1/2$
0.1	0.0016658335317184525	-0.0008326043588517962
0.01	1.6666583333546647e-05	-8.33326041682625e-06
0.001	1.666666583632903e-07	-8.333332601750953e-08
0.0001	1.66666658252268e-09	-8.333334022836425e-10

on ne gagne qu'un facteur 2 en précision pour le prix de deux calculs.

Ce résultat est normal si on examine le détail de la construction de la formule d'extrapolation.

Mais là on peut fabriquer une extrapolation de Richardson adaptée.

$$\frac{f(x + k\delta\lambda x') - k^2 f(x + \delta\lambda x')}{1-k^2}$$

éradique le second ordre plutôt que le premier ordre qui est de toute façon nul

Par exemple toujours dans le cas $N = 1$, $x = 0$ et $f(x) = \sin x/x$, on trouve

x	$f(x) = \sin x/x - 1$	$(f(kx) - k^2 f(x))/(1-k^2)$ pour $k = 1/2$
0.1	0.0016658335317184525	2.0827133839773637e-07
0.01	1.6666583333546647e-05	2.0833446079393525e-11
0.001	1.666666583632903e-07	2.220446049250313e-15
0.0001	1.66666658252268e-09	0.0

où on obtient d'excellents résultats.

2 Les hypersurfaces et les hypervolumes

Si on se contentait de l'espace sensible ordinaire à 3 dimensions, le titre porterait le nom 'les surfaces et les volumes' ; on a préfixé ces notions (premières ?) par 'hyper' parce qu'on a l'intention de parler d'espaces de dimension finie mais quelconque.

Il ne faudrait cependant pas croire que l''espace sensible ordinaire à 3 dimensions' est simple et que parce qu'il est parfaitement compris alors il peut servir de modèle à des études portant sur des espaces de dimensions supérieures.

L''espace sensible ordinaire à 3 dimensions' est compliqué et s'il sert de modèle c'est parce que les sens permettent de l'explorer ; d'où l'adjectif sensible qu'on lui a accolé.

De toute façon ce n'est pas d'espace sensible qu'il s'agit mais plutôt des représentations d'objets dans cet espace.

Dans le cas de 3 dimensions, on a l'idée intuitive qu'une surface correspond à une équation et qu'un volume correspond à une inéquation. On le verra, ce n'est vrai qu'avec une précaution supplémentaire.

Et l'objet de cette partie est essentiellement d'introduire cette précaution.

2.1 Ce que peut représenter une équation

Une équation est une expression de la forme

$$G(X) = 0$$

où G est une fonction dont l'argument est pris dans E_N et la valeur est réelle ; on peut *a priori* lui associer le lieu D des points dans E_N qui satisfont à cette équation.

Ainsi l'équation représente symboliquement ce lieu (qui, si $N = 3$, peut être un vrai lieu dans l'espace) et on écrit

$$D = \{X \in E_N \text{ tel que } G(X) = 0\}$$

2.1.1 Un système d'équation se ramène à une seule équation

Si D est donné par les P équations

$$D = \{X \in E_N \text{ tel que } G_1(X) = 0, \ldots, G_P(X) = 0\} \tag{39}$$

alors en formant

$$\mathcal{G}(X) = (G_1(X))^2 + \ldots + (G_P(X))^2 \tag{40}$$

on peut donner à D la définition

$$D = \{X \in E_N \text{ tel que } \mathcal{G}(X) = 0\} \tag{41}$$

2.1.2 Une équation peut être une inéquation

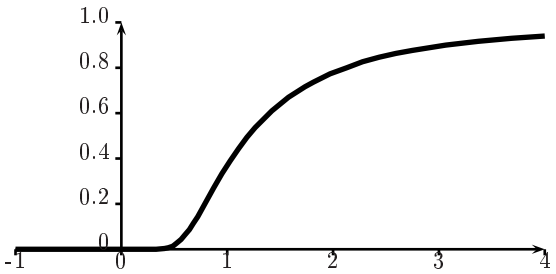
Si maintenant on revient à D donné par une seule équation, alors en choisissant

$$G(X) = W\left(\sum_{n=1}^{N} (X_n)^2 - 1\right) \tag{42}$$

où $W(u)$ est une fonction d'argument réel à valeur réelle d'expression

$$W(u) = \begin{cases} 0 & \text{si} \quad u \leq 0 \\ \exp -1/u^2 & \text{sinon} \end{cases} \tag{43}$$

et de graphe



G est indéfiniment différentiable et pourtant le domaine D de (39) pourrait tout aussi bien être donné par

$$D = \{X \in E_N \text{ tel que } (X_1)^2 + \ldots + (X_N)^2 \leq 1\} \tag{44}$$

c'est à dire par une inéquation.

Ce qui est amusant c'est qu'on a plus moins l'idée intuitive, dans le cas de la dimension $N = 3$, qu'une équation, si elle est fabriquée en utilisant une fonction indéfiniment différentiable, correspond à une surface et qu'une inéquation correspond à l'espace situé localement d'un côté de cette surface s'il existe.

Et cet exemple est emblématique de ce ce fait suivant lequel l'intuition peut être pris en défaut.

W est la fonction de Whitney que Whitney a spécialement conçue dans les années 40 pour l'usage qu'on en fait ici : montrer qu'il existe des fonctions indéfiniment différentiables dont le graphe n'est pas une surface [Pos81a].

2.1.3 Une équation correspond à d'étranges lieux

Pour illustrer qu'une équation puisse correspondre à des situations très différente il suffit de fournir des exemples.

Un premier lieu Dans le cas $N = 2$, si

$$G(X) = G(X_1, X_2) = (X_1)^2 + (X_2)^2 - R^2 \tag{45}$$

alors tant que $R \neq 0$ le domaine D correspondant est un cercle ; si $R = 0$ c'est un point du plan.

Un second lieu Dans le cas $N = 2$, si

$$G(X) = G(X_1, X_2) = (X_1)(X_2) - R^2 \tag{46}$$

alors tant que $R \neq 0$ le domaine D correspondant est composé de deux branches d'hyperboles ; si $R = 0$ il est composé de deux droites du plan.

Il y a une différence avec le premier lieu de (45) parce-que ici des points de D sont à l'infini ; de plus, si $R \neq 0$ D comporte deux composantes connexes.

Un tiers lieu

$$G(X) = G(X_1, X_2) = \sin(X_1 X_2) - a \tag{47}$$

alors si $|a| > 0$ le lieu est vide ; si $|a| = 1$ le lieu est la superposition des hyperboles $X_1 X_2 = \frac{2k+1}{2}\pi$ pour k entier relatif ; si $|a| < 1$, on pose $\sin\theta = a$ et c'est le lieu des hyperboles $X_1 X_2 = \theta + 2k\pi$ puis $X_1 X_2 = \pi - \theta + 2k\pi$.

Équation non algébrique Qu'est-ce qu'une expression algébrique ? C'est d'abord une combinaison finie de nombres, de variables et des signes '+', '-', '×' et '/' ; par extension les expressions algébriques peuvent également en être une combinaison infinie ; puis pour raccourcir (oublier ?) l'infini on peut introduire des fonctions comme 'exp', 'log', 'sin' comme termes légitimes d'expressions algébriques, même si on préfère souvent dire qu'il s'agit d'expressions analytiques.

La fonction G peut ne pas avoir d'expression algébrique mais être définie par un processus itératif.

Par exemple dans le cas où $P = 1$ on peut donner un système différentiel comme

$$\begin{cases} \dfrac{dY}{dt} & = & H(Y) \\ Y(t=0) & = & X \end{cases} \tag{48}$$

où H est une fonction de E_N dans E_N et choisir

$$G(X) = \max_{t>0} {}^t Y(t) Y(t) \tag{49}$$

Suivant la forme de H on trouvera une expression algébrique à G ou non.

On aurait tort de considérer qu'il s'agit là d'un cas limite, fabriqué par un esprit tortueux ; celle des optimisations qui ont un intérêt pratique se formule de manière similaire à (49) ; et souvent les autres ne sont que des cas d'écoles.

2.2 Équations sympathiques

Dans toute cette partie G sera une fonction à valeur réelle

Pour une fonction $G(X)$ indéterminée, même si elle est indéfiniment différentiable (voir la fonction de Whitney), il ne serait pas très commode d'introduire un ensemble

$$D = \{X \in E_N \text{ tel que } G(X) = 0\} \tag{50}$$

et de ne pas savoir si, connaissant un point X_0 élément de cet ensemble D, on reste encore dans cet ensemble dans un déplacement infinitésimal quelconque autour de ce point ; ou encore si on reste dans cet ensemble dans un déplacement infinitésimal soumis à certaines conditions autour de ce point (si c'est un ouvert ou un fermé).

Il est donc nécessaire de compléter la définition (50) pour éviter cet inconvénient et ne retenir des fonctions possibles qu'une classe de celles-ci dont chaque élément est tel que, dans le cas où $N = 3$ par exemple, l'ensemble

$$D = \{X \in E_3 \text{ tel que } G(X) = 0\} \tag{51}$$

s'il n'est pas vide, corresponde à une surface dans E_3 et l'ensemble

$$D = \{X \in E_3 \text{ tel que } G(X) < 0\} \tag{52}$$

s'il n'est pas vide, corresponde à une partie de E_3 telle qu'en chacun de ces points une variation infinitésimale autour de ce point appartienne encore à cette partie (un ouvert de E_3 ou encore un domaine de E_3).

2.2.1 Élimination des singularités

L'équation (40) a pour particularité que

$$\exists X \text{ tel que } \mathcal{G}(X) = 0 \wedge \nabla\mathcal{G}(X) = 0 \tag{53}$$

en effet

$$\nabla\mathcal{G}(X) = \frac{1}{2}(G_1(X)\nabla G_1(X) + \ldots + G_P(X)\nabla G_P(X)) \tag{54}$$

L'équation (42) a cette même particularité puisque toutes les dérivées de W en 0 sont nulles.

On serait alors porté à penser que (53) est peut-être une condition nécessaire pour qu'une équation recèle en son sein un monstre géométrique, une singularité.

Appelons donc singulier les objets géométriques définis par une équation (53) tels qu'ils soit possible d'y trouver des points X pour lesquels le gradient de la fonction soit nul.

Puis éliminons les objets géométriques singuliers des investigations.

Dans ces conditions, si

$$D = \{X \in E_N \text{ tel que } G(X) = 0 \text{ où } G \text{ est telle que si } G(X) = 0 \text{ alors } \nabla G(X) \neq 0\} \tag{55}$$

alors on peut appeler D une hypersurface régulière.

2.2.2 Paramétrisation des hypersurfaces régulières

Le théorème des fonctions implicites affirme que si G est une fonction différentiable telle que

$$G(0) = 0 \wedge \partial_1 G(0) \neq 0 \tag{56}$$

alors on peut remplacer

$$G(X) = 0 \tag{57}$$

dans un voisinage de $X = 0$ par

$$X_1 = \Phi(X_2, \ldots, X_3) \tag{58}$$

et que cette fonction Φ est elle même différentiable jusqu'à un ordre presqu'aussi grand que G.

Dans ces circonstances, l'équation (58) peut également être écrite

$$\left\{\begin{array}{lcl} X_1 & = & \Phi(X_2, \ldots, X_3) \\ X_2 & = & X_2 \\ & \vdots & \\ X_N & = & X_N \end{array}\right. \quad \text{qui peut être noté} \quad X = L(x) \quad \text{avec} \quad x = \left(\begin{array}{c} X_2 \\ \vdots \\ X_N \end{array}\right) \tag{59}$$

on retrouve une paramétrisation qui, on l'a vu dans l'introduction, est un cas favorable pour la minimisation sous contrainte.

Il est facile de généraliser : dans ce qui précède, il suffit de remplacer $G(0) = 0$ par $G(X_0) = 0$; de remplacer encore $\partial_1 G(0) \neq 0$ par $\nabla G(X_0) \neq 0$; d'introduire n tel que si $\nabla G(X_0) \neq 0$ alors $\partial_n G(0) \neq 0$; puis, en changeant ce qui doit être changé (mutatis mutandi), de réécrire (59).

On obtient alors, avec le théorème des fonctions implicites, un moyen d'affirmer que localement il est possible de remplacer une équation par une paramétrisation. Et c'est sur cela qu'on va jouer dans la suite[3].

Le mot local est important : chaque fois qu'on parle de minimiser une fonction et que cette fonction n'a pas de forme particulière il s'agit de minimum local, quand bien même cette clause ne serait pas stipulée.

Le cas de la minimisation sous contrainte ne fait pas exception : et c'est pourquoi le fait de ne pouvoir remplacer une équation par une paramétrisation que localement n'est en fait pas gênant.

2.2.3 Équation et paramétrisation

Si on suppose donc qu'un même ensemble D est défini de deux façons : d'abord, pour une fonction G satisfaisant à

$$\text{si } G(X) = 0 \text{ alors } \nabla G(X) \neq 0 \tag{60}$$

par une équation

$$D = \{X \in E_N \text{ tel que } G(X) = 0\} \tag{61}$$

puis, pour une application L de E_{N-1} dans E_N

$$D = \{X \in E_N \text{ tel que } X = L(x)\} \tag{62}$$

il faudra alors que

$$\forall x \in E_{N-1} : G(L(x)) = 0 \tag{63}$$

et donc, pour $\delta\lambda$ et x' quelconques

$$\forall x \in E_{N-1} : G(L(x + \delta\lambda x')) = 0 \tag{64}$$

[3] Il ne faudrait cependant pas en déduire qu'une fonction présentant une singularité n'admet pas de paramétrisation comme le montre l'exemple tiré de [AVGZ86, t. 1, p 24], et qui est appelé le parapluie de Whitney : pour $N = 3$, $G(X_1, X_2, X_3) = X_1^2 - X_3 X_2^2$ qui peut être paramétrisé par $X_1 = x_1 x_2$; $X_2 = x_2$; $X_3 = x_1^2$.

La raison pour laquelle les monstres sont éliminés est que leur traitement est trop difficile.

Il s'agit d'abord de faire un développement de Taylor de L autour de x.

$$L(x+\delta\lambda x') = L(x) + \delta\lambda \nabla L(x) x' + \frac{1}{2}\delta\lambda^2 [\nabla^2 L(x), x', x'] + \&c \tag{65}$$

où $\nabla L(x)$ est la matrice jacobienne de L au point x de dimension $N \times (N-1)$, soit

$$\nabla L(x) = \begin{pmatrix} \nabla L_1(x) \\ \vdots \\ \nabla L_N(x) \end{pmatrix} = \begin{pmatrix} \partial_1 L_1(x) & \dots & \partial_{N-1} L_1(x) \\ \vdots & & \vdots \\ \partial_1 L_N(x) & \dots & \partial_{N-1} L_N(x) \end{pmatrix} = (\partial_1 L(x) \dots \partial_{N-1} L(x)) \tag{66}$$

$[\nabla^2 L(x), x', x']$ est une notation pour désigner le terme quadratique en x' issue du développement de Taylor, soit,

$$[\nabla^2 L(x), x', x'] = \sum_{n=1}^{N-1} \sum_{m=1}^{N-1} \partial^2_{nm} L(x) x'^n x'^m \tag{67}$$

qui est un vecteur de dimension N ; et enfin &c contient tous les termes de puissance supérieure à 2 en $\delta\lambda$ (c'est à dire que $\&c = o(\delta\lambda^2)$ où $\lim_{u\to 0} o(u)/u = 0$).

Ensuite, en entrant (65) dans (64), il vient

$$\begin{array}{rcl} G(L(x+\delta\lambda x')) & = & G(L(x)) \\ & + & \delta\lambda \nabla G(L(x)) \nabla L(x) x' \\ & + & \frac{1}{2}\delta\lambda^2 \left({}^t x' \, {}^t \nabla L(x) \nabla^2 G(L(x)) \nabla L(x) x' + \nabla G(L(x)) [\nabla^2 L(x), x', x'] \right) \\ & + & \&c \\ & = & 0 \end{array} \tag{68}$$

égalité qui doit être valable pour tout $\delta\lambda$; d'où on déduit que nécessairement

$$\nabla G(L(x)) \nabla L(x) = 0 \tag{69}$$

ce qui signifie que ${}^t\nabla G(L(x))$ est orthogonal aux $N-1$ vecteurs colonnes $\partial_n L(x)$.

Ce résultat est satisfaisant parce que si X est un point de D, une variation $\delta\lambda X'$ autour de X qui serait dirigée suivant le gradient de G au point X conduirait à changer la valeur de G qui ne pourrait donc plus être de 0 : en conséquence, les vecteurs $\partial_n L(x)$ ne peuvent être dirigées suivant ce gradient ; ce qu'on retrouve ici de manière moins verbeuse.

Toutefois le verbe est un peu utile parce qu'il fait prendre conscience que les $N-1$ vecteurs colonnes $\partial_n L(x)$ doivent former une famille libre dans E_N de manière qu'une base de E_N puisse être :

$$\{\nabla G(L(x)), \partial_1 L(x), \dots, \partial_{N-1} L(x)\} \tag{70}$$

Cette description mène au repère mobile de Frenet utilisé en géométrie différentielle ; malgrès l'utilité qu'elle aurait pu avoir elle va cependant être laissée afin de ne pas introduire trop de formalisme.

Pour les même raisons que pour (69), on a

$$\forall x' : {}^t x' \, {}^t \nabla L(x) \nabla^2 G(L(x)) \nabla L(x) x' + \nabla G(L(x)) [\nabla^2 L(x), x', x'] = 0 \tag{71}$$

c'est à dire que la matrice $(N-1) \times (N-1)$

$${}^t\nabla L(x) \nabla^2 G(L(x)) \nabla L(x) + \begin{pmatrix} \text{terme général en } nm \\ \sum_{p=1}^{P} \partial_p G(L(x)) \partial^2_{nm} L_p(x) \end{pmatrix} = 0 \tag{72}$$

Si G était connue et qu'on cherchait L on pourrait espérer calculer L à partir des P équations de (69) ; ces $P \times P$ équations supplémentaires ne sont que les dérivées des P équations précédentes.

Mais il est illusoire d'espérer faire ce calcul effectivement dans tous les cas ; la résolution de systèmes de P équations aux dérivées partielles non linéaires est un problème très compliqué.

2.3 Système d'équations

Maintenant $\mathbf{G} = {}^{\mathbf{t}}(\mathbf{G_1}, \dots, \mathbf{G_P})$ est une application de $\mathbf{E_N}$ dans $\mathbf{E_P}$

Il y a donc plusieurs équations simultanées. On considère d'abord que chacune de ces équations satisfait à la condition d'être exempte de singularité

$$\text{si } G_p(X) = 0 \text{ alors } \nabla G_p(X) \neq 0 \tag{73}$$

et cet ensemble d'équations traduit l'intersection d'hypersurfaces régulières.

2.3.1 Paramétrisation

Pour chacune des équations, on peut introduire une paramétrisation L^p (p en exposant pour ne pas mélanger avec les indices : $L^p_n(x) = X_n$) comme précédemment

$$G_p(X) = 0 \Leftrightarrow X = L^p(x) \tag{74}$$

et alors le domaine D défini par

$$D = \{X \in E_N \text{ tel que } G_p(X) = 0 \text{ pour } p = 1 \dots P\} \tag{75}$$

peut être redéfini par

$$\begin{array}{rl} D = \{X \in E_N \text{ tel que} & \text{pour } p = 1 \dots P : \\ & \exists x^p \in E_{N-1} \text{ tel que} \\ & L^1(x^p) = L^2(x^p) = \dots = L^P(x^P) \\ & \text{et alors } X = L^1(x^1)\} \end{array} \tag{76}$$

c'est à dire qu'il faudrait d'abord résoudre le système $L^1(x^1) = L^2(x^2) = \dots = L^P(x^P)$ de $N \times (P-1)$ équations à $P(N-1)$ inconnue qui dans le meilleur des cas laisserait $N - P$ variables libres avec lesquels on construirait X.

C'est bien sûr une méthode à éviter. Il est préférable de traduire directement la condition

$$G(X) = 0 \tag{77}$$

par

$$X = L(x) \text{ où } x \text{ est pris dans } E_{N-P} \tag{78}$$

C'est à dire de paramétriser globalement les P équations. Mais évidemment il faut que cela soit possible.

Même sur un exemple simple, cette possibilité n'a rien d'évident.

2.3.2 Intersection de deux surfaces

Dans l'espace ordinaire ($N = 3$) on donne deux surfaces régulières par leurs équations $G_1(X) = 0$ et $G_2(X) = 0$; la question est de savoir si elles admettent une intersection.

Si domaine D est le lieu des points tels que ces deux équations soient satisfaites

$$D = \{X \in E_3 \text{ tel que } G_1(X) = 0 \text{ et } G_2(X) = 0\} \tag{79}$$

soit un point particulier X_0 de D tel que

$$G_1(X_0) = G_2(X_0) = 0 \tag{80}$$

le point $X_0 + \delta X$ voisin de X_0 appartiendra à D si

$$\text{pour } p = 1 \text{ ou } 2,\ G_p(X_0 + \delta X) = 0 \tag{81}$$

soit donc

$$\nabla G_p(X_0) \cdot \delta X + o(|\delta X|) = 0 \tag{82}$$

on peut donc construire une partie de D en résolvant le système d'équation différentielles

$$\left\{ \begin{array}{l} \dfrac{dX}{dt} = \nabla G_1(X(t)) \times \nabla G_2(X(t)) \quad \text{si} \quad \forall t \\ \\ X(t=0) = X_0 \end{array} \right. \tag{83}$$

où $\times$ est le produit vectoriel (on est dans E_3), et alors

$$\forall t, X(t) \subset D \tag{84}$$

Voilà tout ce qui peut être dit de général sur l'intersection de deux surfaces : il est possible de construire une telle intersection par résolution d'un système d'équation différentielle ordinaire.

Le cas de l'intersection de deux hypersurfaces est beaucoup plus compliqué : on ne dispose pas du produit vectoriel et il est nécessaire d'utiliser la théorie des formes différentielles extérieures pour faire ce genre de construction ; ce qui dépasse largement le cadre de ce texte.

2.3.3 Paramétrisation

Avec les réserves précédemment émises, on accepte de croire qu'un domaine défini par équation peut également être décrit localement par paramétrisation et il reste à recommencer dans le cas de plusieurs contraintes le travail déjà fait dans le cas d'une contrainte.

Il s'agit alors de répéter le développement de (63) avec

$$\forall x \in E_{N-P} : G(L(x)) = 0 \tag{85}$$

le développement (65) reste inchangé

$$L(x + \delta\lambda x') = L(x) + \delta\lambda \nabla L(x) x' + \frac{1}{2}\delta\lambda^2 [\nabla^2 L(x), x', x'] + \&c \tag{86}$$

à ceci près que ${}^t\nabla L(x)$ est maintenant une matrice de dimension $(N-P)\times N$ et que

$$[\nabla^2 L(x), x', x'] = \sum_{n=1}^{N-P}\sum_{m=1}^{N-P} \partial^2_{nm} L(x) x'^n x'^m \tag{87}$$

Ensuite on calcule $G(L(x+\delta\lambda x'))$ comme

$$\begin{array}{rcl} G(L(x+\delta\lambda x')) & = & G(L(x)) \\ & + & \delta\lambda \nabla G(L(x))\nabla L(x)x' \\ & + & \frac{1}{2}\delta\lambda^2 \left([\nabla^2 G(L(x)), \nabla L(x)x', \nabla L(x)x'] + \nabla G(L(x))[\nabla^2 L(x), x', x'] \right) \\ & + & \&c \\ & = & 0 \end{array} \tag{88}$$

en faisant attention que : $\nabla G(L(x))$ est une matrice $P\times N$; et que $[\nabla^2 G(L(x)), \nabla L(x)x', \nabla L(x)x']$ est un objet similaire à $[\nabla^2 L(x), x', x']$, c'est un vecteur de dimension N.

De la même façon que dans le cas où G était une fonction à valeur réelle, il faut que

$$\nabla G(L(x))\nabla L(x) = 0 \tag{89}$$

ce qui traduit l'orthogonalité de chacun des $N-P$ vecteurs colonnes $\partial_p L(x)$ (de dimensions N) avec chacun des P vecteurs lignes $\nabla G_p(x)$ (de dimensions N).

Et aussi, pour les mêmes raisons

$$\forall x' \in E_{N-P} : [\nabla^2 G(L(x)), \nabla L(x)x', \nabla L(x)x'] + \nabla G(L(x))[\nabla^2 L(x), x', x'] = 0 \tag{90}$$

On voit donc que les formules ne changent guère entre le cas où G est une fonction à valeur réelle et celui-ci où G est une fonction à valeurs dans E_P ; toutefois ce dernier cas demande une attention soutenue pour identifier les dimensions des objets.

Le cas des vecteurs liés

Il reste toutefois un problème en suspend : si le point X réalise les équations $G_p(X) = 0$ et que $\nabla G_p(X) \neq 0$; il est encore possible que les vecteurs ${}^t\nabla G_p(X)$ ne forment pas une famille libre dans E_N ; et dans ce cas on ne pourrait pas former une base de E_N en complétant les $N-P$ vecteurs $\partial_p L(x)$ par les P vecteurs ${}^t\nabla G_p(L(x))$.

On verra plus loin que cette situation est un peu gênante ; mais comme l'objectif est moins d'introduire des difficultés que de montrer comment traiter un problème quand ces difficultés n'arrivent pas, on supposera que la situation si elle existe en puissance n'existe pas en acte.

3 Exercices

3.1 Petites choses

Rien n'est petit, J'en suis convaincu tout autant qu'un autre [...]

Jean-Henri Fabre, Souvenirs entomologiques, Tome 1, La théorie du parasitisme, p. 562, édition Bouquin, Robert Laffont

3.1.1 Manipulation d'expressions

Si b est un vecteur de dimension N et A une matrice définie positive $N\times N$; si on considère le problème de minimiser la fonction $({}^t xAx)^2 - {}^t bx$ par rapport à sa variable x (un vecteur, le symbole ^{t} précédent un symbole signifie la transposée de ce dernier) alors résoudre ce problème (donner le nombre de solutions possibles, trouver ces solutions ...).

Solution proposée :

Si

$$f(x) = ({}^t xAx)^2 - {}^t bx$$

alors

$$\begin{array}{rcl} f(x+\delta x) & = & ({}^t(x+\delta x)A(x+\delta x))^2 - {}^t b(x+\delta x) \\ & = & ({}^t xAx + {}^t\delta xAx + {}^t xA\delta x + {}^t\delta xA\delta x)^2 - {}^t b(x+\delta x) \\ & = & ({}^t xAx + {}^t\delta x(A+{}^tA)x + {}^t\delta xA\delta x)^2 - {}^t b(x+\delta x) \\ & = & ({}^t xAx + 2\,{}^t\delta xAx + {}^t\delta xA\delta x)^2 - {}^t b(x+\delta x) \\ & = & ({}^t xAx)^2 + 4({}^t\delta xAx)^2 + ({}^t\delta xA\delta x)^2 + 4({}^t xAx)({}^t\delta xAx) \\ & + & 2({}^t xAx)({}^t\delta xA\delta x) + 4({}^t\delta xAx)({}^t\delta xA\delta x) - {}^t b(x+\delta x) \\ \text{ordre } 0 \longrightarrow & = & ({}^t xAx)^2 - {}^t bx \\ \text{ordre } 1 \longrightarrow & + & 4({}^t xAx)({}^t\delta xAx) - {}^t\delta xb \\ \text{ordre } 2 \longrightarrow & + & 2({}^t xAx)({}^t\delta xA\delta x) + 4({}^t\delta xAx)^2 \\ \text{ordre } 3 \longrightarrow & + & 4({}^t\delta xAx)({}^t\delta xA\delta x) \\ \text{ordre } 4 \longrightarrow & + & ({}^t\delta xA\delta x)^2 \end{array}$$

par conséquent pour que x_∞ soit un argument minimisant de f (le minimum est $f(x_\infty)$:

- Il est nécessaire que le terme d'ordre 1 soit nul (se demander pourquoi) pour tout δx, soit

$$4({}^t x_\infty Ax_\infty)(Ax_\infty) = b$$

qui peut se résoudre en considérant la solution de (se demander pourquoi cette solution existe)

$$Ax'_\infty = b$$

et en cherchant

$$x_\infty = \lambda x'_\infty$$

avec

$$\lambda^3 = \frac{1}{4({}^t b x'_\infty)}$$

- Si A est définie positive alors le terme d'ordre 2 est positif pour tout δx non nul. La solution précédente correspond donc à un minimum qui est alors proportionelle (préciser le coefficient) à

$${}^t b x'_\infty$$

Par delà la banalité de cet exercice, il est important de savoir reproduire les étapes de calcul et aussi les déductions.
Par exemple on peut se demander ce qui se passe si A n'est pas symétrique (il faut répondre qu'il suffit alors de remplacer A par sa partie symétrique puisque

$${}^t x A x = \frac{1}{2}({}^t x(A + {}^t A)x + {}^t x(A - {}^t A)x) = \frac{1}{2}\,{}^t x(A + {}^t A)x$$

ce qui permet de se ramener au cas précédent).
On peut aussi se demander ce qui se passe si A n'est pas définie positive (il faut répondre que

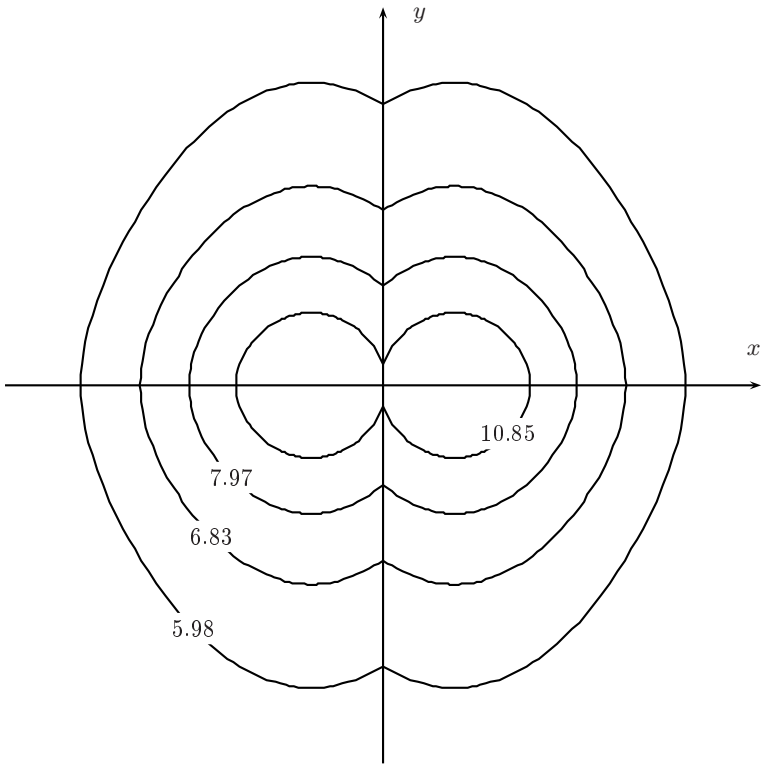
- si A a une valeur propre nulle alors si y est un vecteur propre associé à cette valeur propre $f(y) = -{}^t b y$ qui peut être aussi petit qu'on veut et donc il n'y a pas de minimum à f ;
- si A est définie négative $-A$ est définie positive et donc il y a un minimum dans les mêmes conditions que précédemment ;
- si A comporte des valeurs propres positives et négatives alors on peut prendre l'exemple bidimensionnel $f(x_1, x_2) = (x_1^2 - x_2^2)^2 - x_1$ pour voir que $f(t,t)$ peut devenir aussi petit qu'on veut et donc qu'il n'y a pas de minimum à cette fonction ; par contre avec l'exemple $f(x_1, x_2) = (x_1^2 - x_2^2)^2$ le minimum est atteint pour tous les points des droites $x_1 + x_2 = 0$ et $x_1 - x_2 = 0$.
  Donc, en revenant au cas général, si on peut trouver un argument stationnarisant f qui est le même que précédemment, le terme d'ordre 2 devient alors

$$2\lambda({}^t x'_\infty b)({}^t \delta x A \delta x) + 4\lambda^2({}^t \delta x b)^2$$

et son analyse peut certainement être faite sans que nécessairement ce soit simple à faire.

3.1.2 Lignes de niveau

Le tracé qui suit est celui des lignes de niveaux de la fonction h à valeurs dans R et dont les arguments sont $(x, y) \in E_2$ (x et y sont des réels) ; les valeurs de ces lignes de niveaux sont indiquées sur le tracé ; on donne de plus une fonction $f(x, y) = -x - y$: porter (approximativement) sur le dessin la position du point (u_∞, v_∞) tel que d'abord $h(u_\infty, v_\infty) \geq 6.83$; ensuite $\forall (u, v)$ tel que $h(u, v) \geq 6.83$, $f(u_\infty, v_\infty) \leq f(u, v)$ et tracer les gradients de f et de h en ce point.



3.1.3 La traversée d'une rivière

Voici pour finir un problème de minimisation qui se résoud très bien quasiment analytiquement (c'est à dire que le numérique ne sera nécessaire que pour minimiser une fonction à une variable).

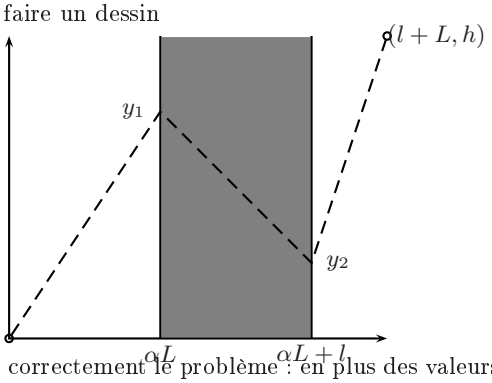
Sur une terre plate dans laquelle la position est repérée par les coordonnées cartésiennes (x, y) quelqu'un veut se rendre de $(0, 0)$ à $(l + L, h)$ en prenant le moins de temps possible.

L'ennui c'est qu'une rivière occupant le domaine $\alpha L < x < \alpha L + l$ sépare les deux endroits.

Sachant que ce quelqu'un marche à la vitesse constante v ; qu'il nage à la vitesse constante V ; que la vitesse de l'eau dans la rivière est constante, dirigée suivant l'axe des ordonnées et de valeur U ; comment pourra-t-il procéder ?

Solution proposée

Le lecteur de Polya [Pol89] sait qu'il faut faire un dessin



puis ce lecteur sait qu'il faut paramétriser correctement le problème : en plus des valeurs données α, L, l et h il introduit des valeurs inconnues y_1, y_2.

Mise en équation Si le problème était restreint à trouver le chemin parcouru en le moins de temps possible entre $(0, 0)$ et $(\alpha L, y1)$ alors celui-ci est certainement rectiligne, d'où le segment tracé. Et le temps mis est

$$t_1 = \frac{1}{v}\sqrt{\alpha^2 L^2 + y_1^2}$$

De même le chemin parcouru en le moins de temps possible entre $(\alpha L + l, y_2)$ et $(L + l, h)$ est aussi rectiligne.

$$t_3 = \frac{1}{v}\sqrt{(h - y_2)^2 + (1 - \alpha)^2 L^2}$$

Il y a une petite complication pour le chemin parcouru en le moins de temps possible entre $(\alpha L, y1)$ et $(\alpha L + l, y_2)$ puisque la vitesse de l'eau intervient. Dans le référentiel de l'eau pourtant le problème devient similaire deux précédents et donc le chemin y est une droite. Or le changement de référentiel conduit à des transformations linéaires et donc ce chemin pris dans le référentiel de la terre est encore une droite.

Le calcul du temps t_2 nécessite cependant de poser les équations : le chemin est déterminé par le système d'équations différentielles

$$\left\{ \begin{array}{lll} \frac{dx}{dt} = V\cos\theta & x(0) = \alpha L & x(t_2) = \alpha L + l \\ \frac{dy}{dt} = V\sin\theta + U & y(0) = y_1 & y(t_2) = y_2 \end{array} \right.$$

où θ est une variable à éliminer. On trouve

$$V^2 t_2^2 = l^2 + (y_2 - y_1 - U t_2)^2$$

t_2 n'est pas donnée explicitement.

Comme pour aller de $(0, 0)$ à $(l + L, h)$ il faut bien passer par des points d'ordonnées y_1 et y_2 en αL et $\alpha L + l$ le chemin qui est globalement parcouru en le moins de temps possible se décompose en ces trois chemins localement les plus courts[4] et le temps total de parcours est la fonction

$$\left\{ \begin{array}{lll} T(y_1, y_2) & = & \frac{1}{v}\left(\sqrt{\alpha^2 L^2 + y_1^2} + \sqrt{(h - y_2)^2 + (1 - \alpha)^2 L^2}\right) + t_2 \\ \\ \text{où} & & t_2 \text{ est solution (positive) de } V^2 t_2^2 = l^2 + (y_2 - y_1 - U t_2)^2 \end{array} \right.$$

qu'il faut minimiser par rapport à y_1 et y_2.

Traitement des équations On remarque qu'une paramétrisation de $V^2 t_2^2 = l^2 + (y_2 - y_1 - U t_2)^2$ (celle de la branche d'hyperbole pour laquelle $t_2 > 0$, ce qui est souhaité) est

$$\left\{ \begin{array}{lll} V t_2 & = & l\cosh\theta \\ y_2 - y_1 - U t_2 & = & l\sinh\theta \end{array} \right.$$

On remarque aussi que t_2 ne dépend que de la différence de y_2 et y_1 et donc que si on considère celle-ci fixée, une résolution partielle de minimisation de la partie variable de T conduit à voir que les deux chemins sur la terre sont nécessairement parallèles. Cela permet de réduire le problème au cas où $\alpha = 0$ (et donc $y_1 = 0$), les autres cas s'en déduisant.

Ces deux remarques permettent donc de ramener le problème à une fonction à un seul argument qui peut être θ. La minimisation de cette fonction n'est pas possible analytiquement mais on peut voir qu'elle possède un seul minimum et donc la recherche de ce dernier ne coûte presque rien numériquement.

[4] Une illustration du principe un peu vague selon lequel une portion de chemin 'optimal' est elle aussi 'optimale'

Deuxième partie

Minimisation en dimension finie de fonctions deux fois différentiables

Les méthodes numériques de base d'optimisation sans contraintes sont expliquées dans le cadre des fonctions 2 fois différentiables de l'espace euclidien E_n dans R.

Les méthodes de Newton, de gradient sont traitées avec des résultats de convergences obtenus en utilisant la condition de Lipschitz mais aucune référence à une éventuelle convexité.

La méthode du gradient conjugué est détaillée.

Le principe des méthodes quasi-Newtoniennes est expliqué mais on ne fournit ni algorithme optimisé de calcul ni résultats de convergence spécifiques.

La méthode de relaxation des directions n'est pas traitée.

Les difficultés liées aux longues vallées étroites et celles liées aux fonctions quasiment linéaires sont illustrées, dans E_2, sur l'exemple des ellipses et celui du problème de Steiner.

4 Méthodes à un pas

Si on donne un point de départ x_0, on cherche à trouver point x_1 tel que $f(x_1) \leq f(x_0)$. Le pas suivant (x_2) sera obtenu de la même manière en reprenant x_1 au lieu de x_0 et ainsi de suite.

4.1 Méthodes

Les méthodes introduites ici ont un long passé, elles sont développées de façon plus complète dans la liste (non exhaustive) et du point de vue

Numérique *[BBC+91, pp : 18–24], [Cia82, pp : 158–182], [PD77, pp : 40–133], [Cul94, pp : 37–62],*

Mathématique *[KF94, pp : 495–499], [SM84, pp : II-1–73]*

4.1.1 Méthode de Newton

Le développement de Taylor de f au point x_0 permet d'écrire

$$f(x) = f(x_0 + (x - x_0)) = f_2(x; x_0) + o(|x - x_0|^2) \tag{91}$$

avec

$$f_2(x; x_0) = f(x_0) + \nabla f(x_0) \cdot (x - x_0) + \frac{1}{2}\nabla^2 f(x_0) \cdot (x - x_0, x - x_0) \tag{92}$$

L'idée est de chercher à minimiser $f_2(x; x_0)$ au lieu de $f(x)$ par rapport à x, si c'est possible et dans l'espoir que le point x_1 solution satisfasse (39)

$$f(x_1) \leq f(x_0)$$

La condition nécessaire de minimum sur f_2 fournit

$$\nabla^2 f(x_0) \cdot (x_1 - x_0) + \nabla f(x_0) = 0 \tag{93}$$

qui soulève la discussion suivante :

1. Si la matrice Hessienne $\nabla^2 f(x_0)$ est inversible, on trouve un point x_1 unique. Cependant
   - si $\nabla^2 f(x_0)$ n'est pas positive, le point x_1 est seulement un point stationnaire de $f_2(x; x_0)$;
   - si $\nabla^2 f(x_0)$ est positive, le point x_1 est bien un minimum $f_2(x; x_0)$, il reste cependant à vérifier (39) sur f.
2. La matrice Hessienne n'est pas inversible
   - si $\nabla f(x_0)$ n'appartient pas au sous-espace des vecteurs propres associés à des valeurs propres non nulles de la matrice Hessienne, on ne trouve pas de point x_1 ;
   - dans le cas contraire, on en trouve trop.

4.1.2 Méthodes de gradient

Le développement de Taylor de f au point x_0 permet d'écrire

$$f(x) = f(x_0 + (x - x_0)) = f_1(x; x_0) + o(|x - x_0|) \tag{94}$$

avec

$$f_1(x; x_0) = f(x_0) + \nabla f(x_0) \cdot (x - x_0) \tag{95}$$

Si $\nabla f(x_0)$ est nul, on peut soupçonner le point x_0 d'être le minimum cherché (mais il faut vérifier que c'est avéré). Dans le cas contraire, on peut choisir

$$x_1 = x_0 - \alpha \nabla f(x_0) \tag{96}$$

où α, la longueur de pas (step length), est un réel positif dont la variété des modalités de détermination explique celle des méthodes de gradient.

Gradient à pas fractionné C'est probablement la plus simple des modalités de choix de la longueur de pas α :

1. on fixe au départ $\alpha = \alpha_0$;
2. on trouve un point x_1^* par (96) ;
3. si $f(x_1^*) < f(x_0)$, on multiplie α par un facteur quelconque (toujours le même et > 1) on recommence l'opération 1 et on s'arrête dès que le cas contraire arrive ;
4. dans le cas contraire, on multiplie α par un facteur quelconque (toujours le même et < 1) et on s'arrête dès l'obtention du cas contraire.

Gradient à pas optimum On traite sérieusement le problème unidimensionnel de minimiser la fonction à une variable réelle

$$f_1^0(\alpha) = f(x_0 - \alpha \nabla f(x_0)) \tag{97}$$

et donc la longueur de pas de (96) est α_∞ qui minimise (97).

Il est intéressant de considérer ce cas pour des raisons théoriques mais évidemment on ne sait pas minimiser exactement une fonction à une variable dans le cas général ; d'où les méthodes approximées qui suivent.

Longueur de pas obtenu par une méthode dichotomique Dans le cas de fonction à une seule variable, on connaît un intervalle dans lequel on est sur qu'il y a un minimum et un seul il est alors possible d'utiliser des méthodes de type dichotomie (simple, de Fibonacci ou du nombre d'or)[Kar77].

Longueur de pas obtenu par la méthode de Newton Le développement à l'ordre 2 de $f_1^0(\alpha)$ au voisinage de 0 est

$$f_1^0(\delta\alpha) = f(x_0) - \delta\alpha \nabla f(x_0) \cdot \nabla f(x_0) + \frac{1}{2}\delta\alpha^2 \nabla^2 f(x_0)(\nabla f(x_0), \nabla f(x_0)) + \dots \tag{98}$$

d'où on tire la valeur α_{opt} optimal en remplaçant $f_1^0(\alpha)$ par son expression tronquée à l'ordre 2

$$f(x_0) - \alpha \nabla f(x_0) \cdot \nabla f(x_0) + \frac{1}{2}\alpha^2 \nabla^2 f(x_0)(\nabla f(x_0), \nabla f(x_0)) \tag{99}$$

qu'on maximise avec

$$\alpha_{opt} = \frac{\nabla f(x_0) \cdot \nabla f(x_0)}{\nabla^2 f(x_0)(\nabla f(x_0), \nabla f(x_0))} \tag{100}$$

On voit que le calcul du hessien est nécessaire pour cette méthode.

L'algorithme induit par (96) peut être noté

$$x_1 = x_0 - \frac{\nabla f(x_0) \cdot \nabla f(x_0)}{\nabla^2 f(x_0)(\nabla f(x_0), \nabla f(x_0))} \nabla f(x_0) = x_0 - \frac{]\nabla f(x_0), \nabla f(x_0)[}{{}^t\nabla f(x_0) \nabla^2 f(x_0) \nabla f(x_0)} \nabla f(x_0) \tag{101}$$

Paramètre de relaxation Il existe une technique d'appoint utilisable pour faire converger des algorithmes qui ne convergeraient pas autrement ou encore pour faire converger plus vite des algorithmes qui convergent lentement : celle du paramètre de relaxation (qu'il ne faut pas confondre avec la méthode de relaxation qui n'est pas traitée ici, voir [Cia82, pp : 185–189]).

Cette technique n'est pas propre au gradient, elle peut être utilisée en accompagnement d'un peu toutes les méthodes, mais son explication est très claire avec le gradient.

On suppose qu'une méthode de détermination de α a été choisie et donc que l'algorithme est (96) muni d'un procécédé de détermination de α. On corrige cet algorithme en introduisant un cœefficient $\omega > 0$ comme

$$x_1 = x_0 - \omega\alpha \nabla f(x_0) \tag{102}$$

Si on choisit

- $\omega < 1$ on dit que l'algorithme est sous-relaxé ; on améliore les chances de convergence de l'algorithme.
- $\omega > 1$ on dit que l'algorithme est sur-relaxé ; on diminue les chances de convergence de l'algorithme mais on l'accélère

Le choix de ω dépend du problème traité. En fait on ne peut pas chercher un ω car cela reviendrait à chercher un α optimal.

Une stratégie possible est de prendre un α optimal par une technique quelconque (en dehors de celle du pas fractionné) et de traiter le produit $\alpha\omega$ comme on le ferait pour la technique du pas fractionnée.

4.2 Convergence

La section précédente a mis en évidence que les méthodes à un pas se mettent sous la forme

$$x_{n+1} = a(x_n) \tag{103}$$

où a est une fonction de E_N dans E_N dont l'expression dépend de la fonction f et de la méthode choisie.

Une question importante est celle de la convergence de ces itérations.

4.2.1 Convergence locale de la méthode de Newton

Il est possible d'appliquer le théorème du point fixe (cf. § 1.1.3, p. 5) à la méthode de Newton dans laquelle on suppose la matrice hessienne $\nabla^2 f(x)$ inversible pour les x considérés. Dans ce cas

$$a(x) = x - \left(\nabla^2 f(x)\right)^{-1} \nabla f(x) \tag{104}$$

Si on suppose l'existence du point fixe x_∞ alors on choisit le domaine D comme la boule centrée sur x_∞ et de rayon d

$$D = \{x \in E_n / |x - x_\infty| \leq d\} \tag{105}$$

et alors

1. $\nabla f(x_\infty) = 0$ (puisque $\nabla^2 f(x)$ est inversible) ;
2. on choisit d tel que a soit contractante dans D, ce qui est toujours possible puisque ($\nabla f(x_\infty) = 0$ par le point précédent) si $x = x_\infty + \delta\lambda x'$

$$\begin{array}{rcl} \nabla f(x_\infty + \delta\lambda x') & = & 0 + \delta\lambda \nabla^2 f(x_\infty) \cdot x' + o(\delta\lambda) \\ \nabla^2 f(x_\infty + \delta\lambda x') & = & \nabla^2 f(x_\infty) + \delta\lambda \nabla^3 f(x_\infty) \cdot x' + o(\delta\lambda) \end{array} \tag{106}$$

où $\nabla^3 f(x_\infty) \cdot x'$ est une matrice fabriquée avec les dérivées troisièmes de f qu'on suppose exister.
On a (cf. 35)

$$\begin{array}{rcl} (\nabla^2 f(x_\infty) + \delta\lambda(\nabla^3 f(x_\infty) \cdot x'))^{-1} & = & (\nabla^2 f(x_\infty))^{-1} \\ & - & \delta\lambda(\nabla^2 f(x_\infty))^{-1}(\nabla^3 f(x_\infty) \cdot x')(\nabla^2 f(x_\infty))^{-1} \\ & + & o(\delta\lambda) \end{array} \tag{107}$$

et donc on obtient pour tout $x = x_\infty + \delta\lambda x'$

$$\begin{array}{rcl} a(x_\infty + \delta\lambda x') & = & a(x_\infty) + \delta\lambda x' - \delta\lambda(\nabla^2 f(x_\infty))^{-1}(\nabla^2 f(x_\infty))x' + o(\delta\lambda) \\ & = & a(x_\infty) + o(\delta\lambda) \end{array} \tag{108}$$

A partir de (108) et en introduisant un $y = x_\infty + \delta\lambda y'$, on déduit

$$a(x) - a(y) = a(x_\infty + \delta\lambda x') - a(x_\infty + \delta\lambda y') = o(\delta\lambda) \tag{109}$$

et comme

$$x - y = \delta\lambda(x' - y') \tag{110}$$

il sera toujours possible (en prenant $\delta\lambda < d$, d une limite supérieure) de rendre $|a(x) - a(y)|$ plus petit que $x - y$ et donc à fortiori $\exists\ k < 1$ tel que

$$\text{si } |x - x_\infty|, |y - x_\infty| < d \text{ alors } |a(x) - a(y)| \leq k|x - y| \tag{111}$$

L'application a est contractante dans un voisinage de x_∞.

3. l'image de D par a est incluse dans D
   En effet

$$|a(x) - x_\infty| = |a(x) - a(x_\infty)| \leq |x - x_\infty| \leq d \tag{112}$$

Ce qui permet de conclure que :

- la méthode de Newton converge au voisinage du point x_∞ minimisant f et avec une vitesse de convergence meilleure que celle prévue par (16), (17) puisque k est d'autant plus petit que l'on est proche de x_∞ ; ce point peut être précisé en introduisant des degrés dans la convergence (linéaire, quadratique ...) ;
- cependant il ne s'agit que d'une convergence locale et non globale puisque la démonstration s'appuie fortement sur le développement de Taylor au voisinage de x_∞ qui est d'ailleurs supposé exister (Voir [BBC+91, pp : 22–24]).
- comme on n'a jamais fait référence à la condition suffisante de minimum ($\nabla^2 f(x)$ est supposée inversible mais pas spécialement définie positive) la méthode de Newton ne garantit pas l'obtention d'un minimum mais seulement d'un point stationnaire.

4.2.2 Convergence globale des méthodes de gradient

L'itération du gradient (96)

$$x_{n+1} = x_n - \alpha_n \nabla f(x_n)$$

où la stratégie de choix des α_n n'est pas spécifiée peut converger vers un x_∞ pour lequel

$$\nabla f(x_\infty) \neq 0$$

Il suffit que la suite α_n converge vers 0.

L'objectif est de montrer que sous certaines conditions cette situation peut être évitée. Cela ne voudra pas dire que l'une ou l'autre des méthodes du gradient qui ont été spécifiées converge, mais qu'il est possible de trouver une méthode du gradient qui converge.

Dit autrement, on ne cherche pas un résultat particulier de convergence mais un résultat général.

De plus on cherche un résultat global dans le sens qu'il ne contienne pas de conditions selon laquelle ń si on n'est pas trop loin de la solution alors l'algorithme converge mais plutôt des assurances du type ń ∀ le point de départ x_0 l'algorithme converge sans d'ailleurs présager de la vitesse de convergence.

La suite $\mathbf{f(x_0), f(x_1 = x_0 - \alpha_0 \nabla f(x_0)), \ldots}$ converge. Tout d'abord puisque on cherche un point x_∞ qui minimise f alors on peut supposer celle ci bornée inférieurement et donc la suite

$$f(x_0), f(x_1 = x_0 - \alpha_0 \nabla f(x_0)), \ldots \tag{113}$$

est bornée inférieurement. De plus les α_n sont tels que

$$f(x_{n+1}) \leq f(x_n) \tag{114}$$

C'est en effet toujours ainsi puisque, sans que ce soit forcément nécessaire pour le procédé de choix des α_n et ce ne serait d'ailleurs pas souhaitable, le cas α_n arbitrairement petit satisfait à (114).

On a donc une suite décroissante et bornée inférieurement dans R, celle ci converge nécessairement ([Rud95b, pp : 50–51]).

On utilise la formule de Taylor avec reste La fonction réelle d'une variable réelle u choisie dans $[0, 1]$

$$u \longrightarrow f(x_n - u\alpha_n \nabla f(x_n)) \tag{115}$$

peut, via la formule de Taylor avec reste à l'ordre 1, être écrite en $u = 1$ comme

$$f(x_n - \alpha_n \nabla f(x_n)) = f(x_n) - \alpha_n \nabla f(x_n - \theta \alpha_n \nabla f(x_n)) \cdot \nabla f(x_n) \tag{116}$$

où θ est pris dans l'intervalle $[0, 1]$.

(116) se transforme en

$$\begin{array}{rcl} f(x_n - \alpha_n \nabla f(x_n)) = f(x_n) & - & \alpha_n (\nabla f(x_n))^2 \\ & + & \alpha_n (\nabla f(x_n) - \nabla f(x_n - \theta \alpha_n \nabla f(x_n))) \cdot \nabla f(x_n) \end{array} \tag{117}$$

et en utilisant l'inégalité de Cauchy-Schwarz

$$\begin{array}{rl} & (\nabla f(x_n) - \nabla f(x_n - \theta \alpha_n \nabla f(x_n))) \cdot \alpha_n \nabla f(x_n) \\ \leq & |\nabla f(x_n) - \nabla f(x_n - \theta \alpha_n \nabla f(x_n))| |\alpha_n \nabla f(x_n)| \end{array} \tag{118}$$

On introduit une condition de Lipschitz sur le gradient Si on suppose $(k > 0)$

$$\forall\ x, y \in E_N \quad |\nabla f(x) - \nabla f(y)| \leq k|x - y| \tag{119}$$

on obtient alors

$$(\nabla f(x_n) - \nabla f(x_n - \theta \alpha_n \nabla f(x_n))) \cdot (\alpha_n \nabla f(x_n) \leq k\theta \alpha_n^2 |\nabla f(x_n))|^2 \leq k\alpha_n^2 |\nabla f(x_n))|^2 \tag{120}$$

Et donc au total en reprenant (116)

$$f(x_{n+1}) - f(x_n) \leq (-\alpha_n + k\alpha_n^2) |\nabla f(x_n))|^2 \tag{121}$$

qui montre que, à condition que k ne soit pas infini, il existe des α_n non nul (entre 0 (exclu) et $1/k$) qui sont tels que

$$f(x_{n+1}) \leq f(x_n) \tag{122}$$

et donc tels que la suite (113) soit décroissante.

Ce qui permet de conclure :

- si le gradient est lipschitzien alors les méthodes du gradient ont la possibilité de converger.
  Ce n'est qu'une possibilité, il resterait à montrer que telle ou telle stratégie de choix de α_n entre dans ce cadre.
- cependant on n'a aucune assurance sur la vitesse de convergence.
  En effet on n'a pas effectué de majorations comme pour la méthode de Newton.

4.2.3 Commentaires sur la convergence et les méthodes

On a montré et/ou constaté que

- la méthode de Newton est dotée de ńbonne propriété de convergence locale
  i.e. proche du point minimum elle converge bien ; ailleurs on ne sait pas, ce qui pose un problème pour le choix du point initial ;
- les méthodes du gradient sont dotée, sous réserve que le gradient soit lipschitzien, de ńbonne propriétés de convergence globale
  i.e. il n'y a plus de problème pour le choix du point initial mais leur vitesse de convergence n'est pas assurée.

Il est alors tentant de fabriquer une méthode hybride qui allie les deux ńbonnes propriétés en remarquant que pour les méthodes de gradients l'analyse porte sur la longueur de pas α.

Cette méthode existe (cf. [PD77, pp : 53–61]), elle peut être appelée la méthode de Newton modifiée ou à pas variable, et elle s'écrit

$$\nabla^2 f(x_0) \cdot (x_1 - x_0) + \alpha \nabla f(x_0) = 0 \tag{123}$$

α étant une longueur de pas dont la stratégie de choix est analogue à celle de la méthode du gradient.

La méthode de Newton modifiée converge globalement sous réserve d'accepter la condition de Lipschitz non plus sur le gradient mais sur

$$\left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$$

La considération de la méthode de Newton modifiée permet d'envisager d'une classe de méthodes analogues à celle du gradient mais pour laquelle on utilise un vecteur d_0, qui s'appelle une direction de descente, à la place de $\nabla f(x_0)$, qui est la direction de descente de la plus grande pente, dans l'algorithme (96) qui devient alors

$$x_1 = x_0 - \alpha d_0 \tag{124}$$

La question est alors de trouver un procédé pour déterminer d_0 à partir de $\nabla f(x_0)$. Pour la méthode de Newton modifiée c'est

$$d_0 = \left(\nabla^2 f(x_0)\right)^{-1} \nabla f(x_0) \tag{125}$$

mais on peut envisager une forme générale

$$d_0 = H_0^{-1} \nabla f(x_0) \tag{126}$$

où H_0^{-1} est une matrice définie positive afin que d_0 soit effectivement une direction de descente.

La fabrication d'un H_0^{-1}, si on ne dispose pas d'*a priori* sur le problème traité et si, d'autre part, on ne souhaite pas faire le calcul du hessien, demande des informations complémentaires que les méthodes à un pas ne peuvent fournir.

4.3 Exercices

4.3.1 Quelque chose d'amusant

Un polynôme homogène à deux variables réelles est un polynôme de la forme

$$p(x,y) = \sum_{n=0}^{N} a_n x^n y^{N-n}$$

On considère un tel polynôme. Que ce polynôme admette ou non un minimum on lui applique une itération de Newton à partir d'un point initial (x_0, y_0) pour obtenir le point (x_1, y_1) :

a) montrer que

$$x_1 = k x_0 \quad \text{et} \quad y_1 = k y_0$$

où k ne dépend ni de x_0 ni de y_0.

b) donner l'expression exacte de k ; de quoi dépend-elle ?

4.3.2 La méthode de newton modifiée

La fonction

$$\begin{array}{rccc} f & : & E_N & \longrightarrow & R \\ & & x & \longrightarrow & f(x) \end{array}$$

est dérivable deux fois, elle admet un minimum unique sur E_N et son hessien $\nabla^2 f(x)$ est défini positif pour tout x de E_N.

Si

$$\begin{array}{rccc} k & : & R & \longrightarrow & R \\ & & u & \longrightarrow & k(u) \end{array}$$

est une fonction réelle à valeurs réelles deux fois dérivable, on définit à partir de k et de f la fonction

$$\begin{array}{rccc} F & : & E_N & \longrightarrow & R \\ & & x & \longrightarrow & F(x) = k(f(x)) \end{array}$$

a) Donner la condition nécessaire, portant sur k ou ses dérivées, pour que l'argument minimisant de f soit celui de F.

b) Donner une condition suffisante, portant sur k ou ses dérivées, pour que cet argument soit bien un argument minimisant de F.

c) Donner une condition suffisante, portant sur k ou ses dérivées, pour que le hessien $\nabla^2 F(x)$ de F soit défini positif pour x quelconque.

d) Écrire un pas de l'algorithme de Newton sur la fonction F en une formule ne contenant que k et f ou leurs dérivées (c'est à dire ni F ni ses dérivées).

e) Montrer que la formule précédente peut s'écrire ($\nabla f(x_0)$ est considéré comme un vecteur colonne)

$$x_1 = x_0 - \alpha \left(\nabla^2 f(x_0)\right)^{-1} \nabla f(x_0)$$

où α est un scalaire dont on déterminera l'expression en fonction de f et k ou de dérivées ou des inverses de ces dérivées.

On pourra utiliser que

$$\left(A + b\,{}^t b\right)^{-1} = A^{-1} - \frac{A^{-1} b\,{}^t b A^{-1}}{1 + {}^t b A^{-1} b}$$

si A est une matrice $N \times N$ inversible et symétrique, b un vecteur de dimension N et ${}^t bb$ est la matrice $N \times N$ de terme général $b_n b_m$.

f) Si

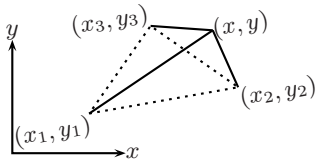
$$f(x) = \sqrt{1 + {}^t x A x}$$

où A est une matrice définie positive montrer, en utilisant un cas particulier, que l'algorithme de Newton appliquée à f directement peut ne pas converger.

g) Avec le choix précédent de f et pour $k(u) = exp(u)$ montrer alors que l'algorithme de Newton appliqué à F devient convergeant.

4.3.3 Problème de Steiner à trois points

La position de 3 points de l'espace euclidien E_2 est donnée sous forme des coordonnées cartésiennes $(x_1, y_1), (x_2, y_2), (x_3, y_3)$. On veut joindre ceux-ci par des chemins rectilignes en étoile : soit, en introduisant un point de coordonnées (x,y)



Le problème de Steiner consiste à trouver le (ou les) point(s) particulier(s) (x_∞, y_∞) qui minimise(nt) la somme des distances euclidiennes de ce point aux 3 points donnés en envisageant tous les cas que la variété des valeurs de $x_1, x_2, x_3, y_1, y_2, y_3$ laisse supposer.

Tout d'abord, par le jeux de translation, rotation et homothétie, le problème se ramène à celui pour lequel

$$\begin{array}{rcl} (x_1, y_1) & = & (0, 0) \\ (x_2, y_2) & = & (1, 0) \\ (x_3, y_3) & = & (u, v) \end{array} \tag{127}$$

Ce problème admet une solution au point de Torricelli du triangle formé si celui-ci existe, c'est à dire si $-2\pi/3 < \theta < 2\pi/3$ et au point $(0, 0)$ sinon (voir [HT84, p. 58], [ATF87, p. 21]).

L'objectif est d'utiliser une méthode numérique pour retrouver ce résultat, c'est à dire qu'on cherche (x_∞, y_∞) qui minimise

$$f(x, y) = \sqrt{x^2 + y^2} + \sqrt{(1 - x)^2 + y^2} + \sqrt{(u - x)^2 + (v - y)^2} \tag{128}$$

Point initial au loin Si (x_0, y_0) est très loin du triangle, alors les trois distances sont pratiquement confondues et on peut supposer que le comportement des méthodes numérique sera voisin du problème unidirectionnel de minimiser $|x| = \sqrt{x^2}$. La méthode de Newton ne peut alors pas être utilisée dans ce cas.

On peut cependant chercher un peu de régularité au problème en considérant que l'exemple unidirectionnel pertinent est plutôt $\sqrt{1 + x^2}$, mais on n'est pas beaucoup plus avancé puisque la méthode de Newton fournit alors l'itération

$$x_1 = -x_0^3$$

qui ne converge pas si $|x_0| > 1$, c'est à dire dans le cas qu'on s'est proposé d'étudier.

Ceci constitue le problème des fonctions ayant un comportement quasi-linéaire pour lesquelles la méthode de Newton ne converge pas. Il n'y a pas de contradiction avec l'analyse de la méthode de Newton qui prédisait la convergence si le point de départ était suffisamment proche de la solution.

Dans une telle situation, et même en revenant au cas multidimensionnel de départ, il est nécessaire de faire appel à la méthode du gradient ou encore aux méthodes à métrique variable (qui se confondent avec le gradient dans le cas unidirectionnel).

On notera que la recherche de pas optimal par la méthode de Newton unidirectionnelle est à proscrire.

Point initial proche de la solution Les défauts du gradient apparaissent, et il est préférable d'utiliser la méthode de Newton.

5 Méthodes à plus d'un pas

Si on donne une suite $x_0, x_1, \ldots, x_n$, on cherche à trouver point x_{n+1} tel que $f(x_{n+1}) \leq f(x_n)$ par un procédé utilisant la connaissance de toute la suite (ou seulement d'une partie de celle-ci) $x_0, x_1, \ldots, x_n$, le premier pas (passage de x_0 à x_1) étant réalisé par une méthode à un pas.

5.1 Méthodes de gradient conjugué

Les méthodes de gradient conjugué sont des méthodes de résolution de systèmes linéaires symétriques qui correspondent donc à des fonctions f quadratiques.

Elles sont très fortement utilisée pour résoudre les systèmes linéaires à structure très creuse qui correspondent à l'approximation numérique (éléments finis, différences finies ...) de problèmes d'équations aux dérivées partielles.

On peut en trouver des développements et extensions dans la liste non exhaustive : [BBC+91, article de J.F Maitre, pp 180–236], [Jol90, pp : 57–122], [Cia82, pp : 194–201].

5.1.1 Présentation

On considère une fonction quadratique de la forme

$$f(x) = \frac{1}{2}\ {}^t x A x - \ {}^t b x \tag{129}$$

où A est une matrice $N \times N$ symétrique définie positive.

On donne un point de départ x_0, $g_0 = Ax_0 - b$ son gradient au point x_0 (on dit aussi le résidu puisque si $g_0 = 0$ alors x_0 est la solution cherchée), d_0 une direction de descente qui n'est pas *a priori* g_0.

Le calcul du point x_1 est fait par une minimisation unidirectionnelle exacte (le problème est quadratique) dans la direction d_0, soit donc

$$x_1 = x_0 - \alpha_{opt} d_0 \tag{130}$$

avec α_{opt} l'argument minimisant de $f(x_0 - \alpha d_0)$ prise comme fonction à une variable α. On obtient

$$x_1 = x_0 - \frac{{}^t d_0 g_0}{{}^t d_0 A d_0} d_0 \tag{131}$$

le gradient en x_1 est

$$\begin{aligned} g_1 = Ax_1 - b &= g_0 - \frac{{}^t d_0 g_0}{{}^t d_0 A d_0} A d_0 \\ &= \left(I - \frac{]Ad_0, d_0[}{{}^t d_0 A d_0} \right) g_0 \\ &= P(a, d_0) g_0 \end{aligned} \tag{132}$$

C'est la projection oblique sur l'hyperplan orthogonal à d_0 dans la direction Ad_0 de g_0 (voir § 1.2.3, page 6).

Il s'agit maintenant de trouver un procédé pour déterminer d_1 à partir de g_1. On le choisit de manière que

$${}^t d_1 A d_0 = 0 \tag{133}$$

et pour cela on peut utiliser ${}^t P(A, d_0)$, la projection oblique de g_1 sur l'hyperplan orthogonal à Ad_0 dans la direction d_0 (il n'est pas nécessaire de prendre la direction d_0 pour obtenir (133) mais ce choix conduit à un algorithme efficient), et obtenir

$$d_1 = \ {}^t P(A, d_0) g_1 = \ {}^t P(A, d_0) P(a, d_0) g_0 \tag{134}$$

on dispose alors de l'algorithme

x_0 donné, $g_0 = Ax_0 - b$, $d_0 = g_0$, $P_0 = P(A, d_0)$ et

$$g_{n+1} = P_n g_n \tag{135}$$
$$d_{n+1} = {}^t P_n g_{n+1} \tag{136}$$
$$P_{n+1} = P(A, d_{n+1}) \tag{137}$$
$$x_{n+1} = x_n - \frac{{}^t d_n g_n}{{}^t d_n A d_n} d_n \tag{138}$$

qui constitue l'algorithme de la méthode du gradient conjugué et dont on va examiner les caractéristiques.

Relations d'orthogonalité Si $k < n$, on a

$${}^t d_k g_n = 0 \tag{139}$$
$${}^t g_k g_n = 0 \tag{140}$$
$${}^t d_k A d_n = 0 \tag{141}$$

Par récurrence, si (139-141) est vrai jusqu'à n alors il faut montrer que si $k \leq n$

$${}^t d_k g_{n+1} = 0 \tag{142}$$
$${}^t g_k g_{n+1} = 0 \tag{143}$$
$${}^t d_k A d_{n+1} = 0 \tag{144}$$

1. Si $k \leq n$ alors ${}^t d_k g_{n+1} = 0$ (142)
   En effet on a, à partir de (135),
   $$g_{n+1} = P_n \dots P_{k+1} P_k P_{k-1} \dots P_0 g_0 \tag{145}$$
   Or, par l'hypothèse (141) et (32) toutes les projections obliques commutent. Donc
   $$g_{n+1} = P_k P_n \dots P_{k+1} P_{k-1} \dots P_0 g_0 \tag{146}$$
   et alors, comme ${}^t P_k d_k = 0$ (cf. (31))
   $${}^t g_{n+1} d_k = \ {}^t g_0\ {}^t P_0 \dots\ {}^t P_{k-1}\ {}^t P_{k+1} \dots\ {}^t P_n\ {}^t P_k d_k = 0 \tag{147}$$
2. Si $k \leq n$ alors ${}^t g_k g_{n+1} = 0$ (143)
   En effet, à partir de (135) on peut écrire
   $$d_k = g_k - \frac{]d_{k-1}, Ad_{k-1}[}{{}^t d_{k-1} A d_{k-1}} g_k = g_k - \frac{{}^t g_k A d_{k-1}[}{{}^t d_{k-1} A d_{k-1}} d_{k-1} \tag{148}$$
   et donc remarquer que g_k appartient au plan d_k, d_{k-1}. En utilisant (142) qui vient d'être montré on obtient alors (143).

3. Si $k \leq n$ alors ${}^t d_k A d_{n+1} = 0$ (144)
   En effet, par (136)
$$d_{n+1} = {}^t P_n g_{n+1} \tag{149}$$
   donc
$${}^t d_{n+1} A d_k = {}^t g_{n+1} P_n A d_k \tag{150}$$
   Si $k = n$, on a directement ${}^t d_{n+1} A d_n = 0$ par (31) ; dans le cas contraire, en utilisant $P_n A d_k = A d_k$ (34)
$${}^t d_{n+1} A d_k = {}^t g_{n+1} P_n A d_k = {}^t g_{n+1} d_k = 0 \tag{151}$$

La récurrence des propriété est donc établie. Maintenant, à l'ordre 1, si $g_0 = d_0$,

$$g_1 = P_0 d_0 \tag{152}$$
$$d_1 = {}^t P_0 d_0 \tag{153}$$

et donc, en utilisant (31) on constate que les propriétés (139–141) sont vraies à l'ordre 1 (mais seulement si $d_0 = g_0$, sinon ${}^t g_0 g_1 = 0$ ne serait pas nécessairement vérifiée.

Fonctionnement de l'algorithme Chacune des directions g_n est orthogonale à celles qui la précédent. Donc on peut s'attendre à obtenir au bout de N' itérations (au pire en $N' = N$, la dimension du système linéaire)

$$g_{N'} = 0 \tag{154}$$

et donc, si x_∞ est le minimum de (129),

$$x_\infty = x_{N'} \tag{155}$$

L'algorithme du gradient conjugué est directe dans le sens où, si l'arithmétique utilisée était exacte, il convergerait exactement vers la solution en un nombre fini d'itérations.

De plus l'algorithme fonctionne à la condition que

$$d_{n+1} \neq 0 \tag{156}$$

Mais dans le cas contraire alors g_{n+1} appartient au noyau de ${}^t P_n$ et donc serait proportionnel à d_n ce qui oblige à ce que $g_{n+1} = 0$ puisque par constitution il appartient à l'orthogonal de d_n (voir les relations 31).

Donc l'algorithme ne présente aucune ambiguïté.

Il est intéressant de chercher dans quelle condition la convergence est obtenue rapidement ($N' << N$).

D'abord, si il arrive que $g_n = A x_n - b$ soit un vecteur propre associé à la valeur propre $\lambda > 0$ de A

$$A g_n = \lambda g_n \tag{157}$$

alors la solution x_∞ (telle que $A x_\infty = b$) est trouvée, c'est (démonstration par inspection)

$$x_\infty = x_n - \frac{1}{\lambda} g_n \tag{158}$$

Ensuite dans ce cas (157) on a, à partir de (136) puis de (139),

$$d_n = g_n - \frac{{}^t g_n A d_{n-1}}{{}^t d_{n-1} A d_{n-1}} d_{n-1} = g_n - \lambda \frac{{}^t g_n d_{n-1}}{{}^t d_{n-1} A d_{n-1}} d_{n-1} = g_n \tag{159}$$

et donc

$$P_n = I - \frac{]d_n, d_n[}{{}^t d_n dn} \tag{160}$$

c'est à dire la projection orthogonale sur l'orthogonal de $d_n = g_n$. On obtient alors

$$g_{n+1} = P_n g_n = 0 \tag{161}$$

c'est à dire la convergence.

Dans les cas pratiques, c'est ce qui arrive. Et de ce point de vue la méthode du gradient conjugué s'apparente aux méthodes de recherche de valeurs propres (et vecteurs propres) de matrices (voir [Jol90, pp : 99] pour une comparaison avec la méthode de Lanczos, les méthodes de recherche de valeurs propres, sans références particulières avec la méthode du gradient conjugué peuvent être trouvée dans [BBC+91, article de F. Chatelin, pp : 314–356]).

5.1.2 La méthode du gradient conjugué usuelle

La présentation du paragraphe précédent constitue une introduction pédagogique à la méthode du gradient conjugué, usuellement ([BBC+91, article de J.F Maitre, pp 180–236], [Jol90, pp : 57–122], [Cia82, pp : 194–201]), on le présente sous la forme équivalente (à (135–137) pratique de calcul (c'est à dire qui minimise le nombre d'opérations à faire pour le passage $n \to n+1$), x_0 étant un point initial quelconque et $g_0 = d_0 = A x_0 - b$,

$$x_{n+1} = x_n - \alpha_n d_n \tag{162}$$
$$g_n = g_{n-1} - \alpha_{n-1} A d_{n-1} \tag{163}$$
$$\alpha_n = \frac{{}^t g_n g_n}{{}^t d_n A d_n} \tag{164}$$
$$d_n = g_n + \beta_n d_{n-1} \tag{165}$$
$$\beta_n = -\frac{{}^t g_n g_n}{{}^t g_{n-1} g_{n-1}} \tag{166}$$

cet algorithme conserve évidemment les propriétés expliquées, soit

$$m < n \quad \Longrightarrow \quad {}^t d_m g_n = 0 \tag{167}$$
$$m < n \quad \Longrightarrow \quad {}^t d_m A d_n = 0 \tag{168}$$
$$m < n \quad \Longrightarrow \quad {}^t g_m g_n = 0 \tag{169}$$
$$\exists\, N' \leq N \quad \text{tel que} \quad d_{N'} = 0 \tag{170}$$

puis enfin, en introduisant le nombre de conditionnement (le rapport des valeurs propres extrêmes) de A, on établit une majoration permettant d'atteindre ce nombre N'.

5.1.3 Le préconditionnement

La vitesse de convergence de la méthode de gradient conjugué dépend fortement du nombre de conditionnement de la matrice A, et ce dernier peut être réduit par un changement de variable.

C'est pourquoi la méthode du gradient conjugué est souvent modifiée par une opération de changement de variable qui est appelée le préconditionnement et dont la variété explique celle des noms de la méthode du gradient conjugué (S.S.O.R, I.C.C.G. ...).

Le lecteur est renvoyé à [BBC+91, article de J.F Maitre, pp 212–234] et [Jol86] pour plus d'explications.

5.2 Méthodes quasi-Newtoniennes

Les méthodes quasi-Newtoniennes, dites aussi à métrique variable, sont traitées : [BBC+91, article de J. Roux, pp 237–313], [GS80], [DM77].

Le lecteur intéressé par des formules pratiques de calcul est invité à se référer aux références précédentes. Même si certaines formules pratiques sont fournis dans ce texte, c'est surtout la compréhension globale du principe de ces méthodes qu'on cherche à susciter chez le lecteur.

Les méthodes quasi-Newtoniennes peuvent être appréhendées de deux façons complémentaires :
- elles imitent la méthode de Newton sans pour autant avoir l'inconvénient de calculer le Hessien. Elles consistent à remplacer ce dernier par une contrefaçon de Hessien.
- elles sont une manière autre que le développement de Taylor d'exprimer une approximation locale de la fonction à minimiser (un peu comme des polynômes d'approximations d'une fonction réelle à variable réelle sont une manière autre que le développement de Taylor d'approcher cette fonction par un polynôme).

Dans tous les cas elles exploitent les informations données par les pas précédent à celui qui est traité; les méthodes quasi-Newtoniennes sont donc des méthodes à plus d'un pas. Mais ńplus d'un pas va signifier seulement deux pas.

5.2.1 Principe des méthodes

Éléments de départ On dispose du point x_0, du point x_1 et on cherche le point x_2 tel que

$$x_2 = x_1 - \lambda h(x_0, x_1)^{-1} \nabla f(x_1) \tag{171}$$

où $h(x_0, x_1) = h_1$ est une matrice qu'il faut déterminer de manière à ce qu'elle ressemble le plus possible au Hessien $\nabla^2 f(x_1)$ et α est une longueur de pas comme dans la méthode du gradient.

Conditions de ressemblance sur h_1 Connaissant les gradients $\nabla f(x_0)$, $\nabla f(x_1)$ et puisque on a

$$\nabla f(x_1) - \nabla f(x_0) = \nabla^2 f(x_1)(x_1 - x_0) + o(x_1 - x_0) \tag{172}$$

on va imposer

$$\nabla f(x_1) - \nabla f(x_0) = h_1(x_1 - x_0) \tag{173}$$

qui constitue une condition (de ressemblance) sur h_1. Toujours pour des mobiles de ressemblance on peut ajouter que h_1 doit être symétrique.

Cela laisse cependant une assez grande latitude pour le choix de h_1 qui possède $N(N-1)/2$ coefficients (avec la symétrie) alors que (173) ne fournit que N équations.

Inversibilité de h_1 Si on suppose donnée $h_0 = h(x_0, x_{-1})$ une matrice dont on connaît aussi l'inverse, alors tout h_1 de la forme (voir § 1.2.3, p 6)

$$h_1 = h_0 +]v, w[\tag{174}$$

où u et v sont des vecteurs de E_N tels que $1 + {}^t u h_0^{-1} v \neq 0$, est inversible.

Ce n'est pas suffisant, il faut ajouter que h_1 doit être symétrique et alors on doit prendre $w - \xi v$, pour obtenir

$$h_1 = h_0 + \xi]v, v[\tag{175}$$

la condition de ressemblance (173) s'écrit alors

$$z = \xi({}^t p v) v \tag{176}$$

où, pour simplifier les notations,

$$y = \nabla f(x_1) - \nabla f(x_0) \tag{177}$$
$$p = x_1 - x_0 \tag{178}$$
$$z = y - h_0 p \tag{179}$$

ce qui impose de prendre

$$v = z \quad \xi = \frac{1}{{}^t zp} \tag{180}$$

et donc oblige à ce que ${}^t zp \neq 0$ ce qui n'a aucune raison de ne pas arriver (voir discussion [BBC+91, article de J. Roux, pp 256]).

Pour éviter cet inconvénient, et bien que la méthode induite par le choix (174) puisse être utilisée (elle s'appelle la méthode de Broyden ou d'actualisation de rang 1), on peut introduire un choix plus vaste que (174) (ce sera la méthode d'actualisation de rang 2) sous la forme, respectant la condition de symétrie sur h_1,

$$h_1 = h_0 + \psi]v, v[+\xi]w, w[+\tau(]v, w[+]w, v[) \tag{181}$$

où ψ, ξ et τ sont des scalaires et v et w des vecteurs à trouver pour respecter la condition de ressemblance (173). Un calcul (voir [GS80, pp : 26–27]) fait apparaître qu'il faut alors que h_1 s'écrive alors (en reprenant les notations (177-187))

$$h_1 = h_0 + \frac{\xi]z, z[-({}^t zp)]v, v[+({}^t vp)(]v, z[+]z, v[)}{\xi({}^t zp) + ({}^t vp)^2} \tag{182}$$

On remarque qu'en prenant $v = 0$ on retrouve la méthode de Broyden ou d'actualisation de rang 1 et donc que cette dernière n'est pas éliminée mais seulement plongée dans un cadre plus vaste.

Définie positivité Une condition de confort est que h_1 soit de plus définie positive : on se contentera d'admettre (c'est assez long, la démonstration figure explicitement dans [GS80, pp : 28–33]) que h_1 donné sous la forme (182) est définie positive si

- h_0 l'est (ainsi que symétrique)
- les conditions

$$\frac{({}^t vh_0^{-1}y)^2 + (\xi - ({}^t vh_0^{-1}v)({}^t zh_0^{-1}y)}{\xi({}^t zp) + ({}^t vp)^2} > 0 \tag{183}$$

$$1 + \frac{1}{2}\frac{-({}^t zp)(vh_0^{-1}v) + \xi({}^t zh_0^{-1}z) + 2({}^t vp)({}^t zh_0^{-1}v)}{\xi({}^t zp) + ({}^b p)^2} \geq 0 \tag{184}$$

sont vérifiées.

Les algorithmes de calcul Les algorithmes de calcul s'écrivent alors, en partant du point x_1, par exemple donné par un pas de gradient à partir du point x_0, et de la matrice h_0 (par exemple l'unité),

$$\begin{array}{lcl} x_{n+1} & = & x_n - \alpha_n h_n^{-1}\nabla f(x_n) \\ \alpha_n & = & \text{solution d'une minimisation unidirectionnelle} \\ h_n & = & h_{n-1} + \dfrac{\xi_n]z_n, z_n[-({}^t z_n p_n)]v_n, v_n[+({}^t v_n p_n)(]v_n, z_n[+]z_n, v_n[)}{\xi_n({}^t z_n p_n) + ({}^t v_n p_n)^2} \\ y_n & = & \nabla f(x_n) - \nabla f(x_{n-1}) \\ p_n & = & x_n - x_{n-1} \\ z_n & = & y_n - h_n p_n \\ v_n, \xi_n & = & \text{résultat d'un choix} \end{array} \tag{185}$$

Les choix des v_n, ξ_n correspondent à toute une famille de méthodes quasi-Newtoniennes ou à métrique variable avec actualisation de rang 2.

5.2.2 Quelques méthodes

P.S.B. (i.e. Powell's Symmetric Broyden) On choisit

$$\xi_n = 0 \quad v_n = p_n \tag{186}$$

et alors

$$h_n = h_{n-1} + \frac{-({}^t z_n p_n)]p_n, p_n[+p_n^2(]p_n, z_n[+]z_n, p_n[)}{p_n^2} \tag{187}$$

Ce choix ne conduit pas à des divisions par 0 parce que si $p_n = 0$ c'est que les itérations n'avancent plus et donc qu'on peut espérer (voir discussion sur la longueur de pas) avoir convergé.

La définie-positivité de h_{n-1} est transmise à h_n sous réserve des conditions (183), (184), on les vérifie pendant le calcul.

D.F.P. (i.e.Davidon–Fletcher-Powell) On choisit

$$\xi_n = 0 \quad v_n = y_n \tag{188}$$

et alors

$$h_n = h_{n-1} + \frac{-({}^t z_n p_n)]y_n, y_n[+({}^t y_n p_n)(]y_n, z_n[+]z_n, y_n[)}{({}^t y_n p_n)^2} \tag{189}$$

Il faut que ${}^t y_n p_n \neq 0$ et cette contrainte est en liaison avec la transmission du caractère de définie positivité est transmise à h_n sous réserve des conditions (183), (184), on les vérifie pendant le calcul.

B.F.G.S. (i.e. Broyden–Fletcher–Goldfarb–Shanno) On choisit

$$\xi_n = 0 \quad v_n = h_{n-1}p_n \tag{190}$$

et alors

$$h_n = h_{n-1} - \frac{]h_{n-1}p_n, h_{n-1}p_n[}{{}^t p_n h_{n-1} p_n} + \frac{]y_n, y_n[}{{}^t p_n y_n} \tag{191}$$

Les remarques de définie positivité sont les mêmes que précédemment.

La méthode B.F.G.S. est la plus pratiquée parce qu'elle est en fait une méthode d'actualisation de rang deux inverse dans le sens que ce n'est pas le hessien qui est approché par actualisation mais son inverse.

En conséquence, il existe une formule d'actualisation de h_n^{-1} à partir des mêmes éléments et de h_{n-1}^{-1} qui évite le calcul effectif de l'inverse de (191), même si celui-ci peut être fait en utilisant de proche en proche (27), et donc allège l'algorithme effectif de calcul (voir [BBC$^+$91, article de J. Roux, pp : 265–294]).

5.2.3 Convergence des méthodes

On se référera à [DM77], [BBC$^+$91, article de J. Roux, pp : 295–310], [GS80, pp : 34–66].

Les résultats sont que les méthodes quasi-Newtoniennes possèdent simultanément les propriétés globales et locales de convergence sous des hypothèses de gradient Lipschitzien.

5.3 Exercices

5.3.1 Énergie et méthode du gradient conjugué

On considère un système de courants électriques, le vecteur i, et de flux magnétiques, le vecteur φ. La liaison entre les courants et les flux est faite par la fonction énergie

$$w(\varphi) = \max_{i \in E_N} \left\{ {}^t\varphi i - \frac{1}{2}\, {}^t i M i \right\} \tag{192}$$

où M, la matrice des inductances est symétrique, définie, positive. on propose de calculer $w(\varphi)$.

On a

$$w(\varphi) = \frac{1}{2}\, {}^t\varphi M^{-1} \varphi \tag{193}$$

mais il faut alors calculer M^{-1}, et on préfère alors disposer d'un algorithme direct de résolution de (192).

On essaie la méthode du gradient conjugué (sous la forme pédagogique (136–139)). Pour simplifier, on prend $N = 2$ et

$$M = \begin{pmatrix} l & m \\ m & L \end{pmatrix} \quad {}^t\varphi = (\phi, \psi) \tag{194}$$

Au départ on prend

$$\begin{array}{rcl} i_0 & = & (0,\,0) \\ {}^t g_0 & = & (-\phi,\,-\psi) \\ {}^t d_0 & = & (-\phi,\,-\psi) \end{array} \tag{195}$$

$$P_0 = \frac{1}{\Delta} \begin{pmatrix} -(-m\,\phi - L\,\psi)\,\psi & (-l\,\phi - m\,\psi)\,\psi \\ \phi\,(-m\,\phi - L\,\psi) & -(-l\,\phi - m\,\psi)\,\phi \end{pmatrix} \tag{196}$$

avec

$$\Delta = l\,\phi^2 + 2\,\phi\,m\,\psi + L\,\psi^2 \tag{197}$$

puis on calcule

$${}^t g_1 = \frac{\Gamma}{\Delta}(\psi, -\phi) \tag{198}$$

avec

$$\Gamma = (-m\,\phi^2 - \phi\,L\,\psi + \psi\,l\,\phi + m\,\psi^2) \tag{199}$$

et

$${}^t d_1 = \frac{\Gamma}{\Delta^2}(\psi^2 + \phi^2)(m\phi + L\psi, -(l\phi + m\psi)) \tag{200}$$

puis

$$P_1 = \begin{pmatrix} \dfrac{\phi\,(l\,\phi + m\,\psi)}{\Delta} & \dfrac{(l\,\phi + m\,\psi)\,\psi}{\Delta} \\ \dfrac{(m\,\phi + L\,\psi)\,\phi}{\Delta} & \dfrac{\psi\,(m\,\phi + L\,\psi)}{\Delta} \end{pmatrix} \tag{201}$$

à partir duquel on obtient

$$\begin{array}{rcl} {}^t g_2 & = & (0,\,0) \\ {}^t d_2 & = & (0,\,0) \end{array} \tag{202}$$

L'algorithme converge en deux coups dans le cas général.

Maintenant on peut constater que la condition pour que Γ soit nulle correspond à prendre (ϕ, ψ) dans l'un ou l'autre des sous espaces de M : dans ce cas la convergence est en un coup.

6 Exemples

Pour comprendre le fonctionnement des méthodes numériques, il est intéressant d'analyser le détail de leurs itérations sur des exemples particuliers.

D'autre part, on ne voit bien qu'avec le cœur, c'est à dire ici avec ce qui prend une forme d'expression analytique.

L'ennui est que ces expressions sont très lourdes à manipuler, aussi s'est-on aidé d'un logiciel de calcul formel pour la gestion des manipulation.

6.1 Banana shapes ou longues vallées étroites

Un test assez classique des méthodes numériques est celui des longues vallées étroites. Ici on prend l'exemple le plus simple possible mais on peut trouver d'autres exemples dans la littérature. Notamment de nombreuses méthodes sont testées sur la fonction de Rosenbrock qui est étudiée plus bas. dans [Cul94, pp : 37–62]. L'article de Jean Roux dans [BBC+91, pp : 291–294] contient également des tests intéressants ainsi qu'une bibliographie de tests.

Le problème majeur de la méthode du gradient est qu'elle est assez mauvaise dans les cas de ńlongues vallées étroites comme on va le vérifier sur l'exemple : si f est une fonction de E_2 dans R de la forme

$$f(x,y) = \frac{1}{2}\left(\left(\frac{x}{a}\right)^2 + y^2\right) \tag{203}$$

où a est un paramètre qui peut devenir grand.

Le minimum de f est le point $(x_\infty, y_\infty) = (0,0)$ et on va essayer de le retrouver numériquement.

Méthode du gradient avec longueur de pas calculée par la méthode de Newton unidirectionnelle L'algorithme conduit alors à associer au point de départ (x_0, y_0) le point

$$(x_1, y_1) = \left(\frac{x_0\, {y_0}^2 a^4\left(a^2-1\right)}{{x_0}^2 + {y_0}^2 a^6}, -\frac{y_0\, {x_0}^2\left(a^2-1\right)}{{x_0}^2 + {y_0}^2 a^6}\right) \tag{204}$$

puis à l'itération suivante le point

$$(x_2, y_2) = \left(\frac{\left(a^2-1\right)^2 a^2 {y_0}^2 {x_0}^3}{\left({x_0}^2 + {y_0}^2 a^6\right)\left({y_0}^2 a^2 + {x_0}^2\right)}, \frac{\left(a^2-1\right)^2 a^2 {y_0}^3 {x_0}^2}{\left({x_0}^2 + {y_0}^2 a^6\right)\left({y_0}^2 a^2 + {x_0}^2\right)}\right) \tag{205}$$

et enfin à l'itération suivante

$$(x_3, y_3) = \left(\frac{{x_0}^3 {y_0}^4 a^6\left(a^2-1\right)^3}{\left({x_0}^2 + {y_0}^2 a^6\right)^2\left({y_0}^2 a^2 + {x_0}^2\right)}, -\frac{{x_0}^4 {y_0}^3 a^2\left(a^2-1\right)^3}{\left({x_0}^2 + {y_0}^2 a^6\right)^2\left({y_0}^2 a^2 + {x_0}^2\right)}\right) \tag{206}$$

A chacun de ces points, on associe les distances euclidiennes d_1, d_2, d_3 à l'origine : on s'attend à ce que

$$d_3 < d_2 < d_1 \tag{207}$$

et, en fixant un point de départ, on trace les courbes $d_1(a)$, $d_2(a)$, $d_3(a)$

$(x_0, y_0) = (1,1)$ $(x_0, y_0) = (1, 1/10)$

Plot: lvl1.eps **Plot: lvl2.eps**

On remarque que pour le point de départ $(x_0, y_0) = (1,1)$ les distances d_2 et d_3 sont très petites devant d_0 et cela pour tous les a, même si on observe une (petite) remontée pour a de l'ordre de 2.

Par contre le comportement du point de départ $(x_0, y_0) = (1, 1/10)$ mérite une analyse.

D'abord il faut remarquer qu'un point de départ $(x_0, y_0) = (1,0)$ aurait conduit à la solution en un pas (cf. (204). Et pourtant le point $(x_0, y_0) = (1, 1/10)$, qui lui est très voisin, semble converger difficilement puisque les distance d_2 et d_3 sont presque confondues pour les a grands, et de toute façon ne semble pas décroître très vite.

C'est une caractéristique des problèmes de longues vallées étroites : même si la longueur de pas est exacte (elle est exacte avec la méthode de Newton unidirectionnelle parce que le problème est quadratique) il y a une lenteur certaine de convergence.

Cela constitue un défaut de la méthode du gradient.

Méthode du gradient avec longueur de pas calculée par la méthode de Newton unidirectionnelle puis perturbée Si, artificiellement, on multiplie la longueur de pas par un facteur $1+\epsilon$ pour simuler les situations (réelles) où on ne peut obtenir de longueur de pas exacte et qu'on recommence l'étude on obtient :

$$(x_1, y_1) = \left(x_0 - \frac{\left({x_0}^2 + {y_0}^2 a^4\right)(1+\epsilon)\, x_0}{{x_0}^2 + {y_0}^2 a^6}, y_0 - \frac{\left({x_0}^2 + {y_0}^2 a^4\right) a^2 (1+\epsilon)\, y_0}{{x_0}^2 + {y_0}^2 a^6}\right) \tag{208}$$

les expression des points suivant étant très compliquées.

On recommence la même expérimentation que précédemment pour obtenir

$(x_0, y_0) = (1, 1)$, $\epsilon = 1/5$

Plot: lvl3.eps

$(x_0, y_0) = (1, 1/10)$, $\epsilon = 1/5$

Plot: lvl4.eps

$(x_0, y_0) = (1, 1)$, $\epsilon = -1/5$

Plot: lvl5.eps

$(x_0, y_0) = (1, 1/10)$, $\epsilon = -1/5$

Plot: lvl6.eps

Une première remarque, en faisant abstraction de la ńvaleur critique de a dans le cas $\epsilon = -1/5$, est que les distances ne diminuent plus fortement dès que $a > 2$, et ceci quelque soit le point de départ.

En fait, et même pour des valeurs de ϵ beaucoup plus petites, on peut être amené à faire des centaines d'itérations pour obtenir la convergence.

Le (bon) comportement, dans le cas $\epsilon = -1/5$, pour la ńvaleur critique de a, est une correction par l'erreur de calcul de la longueur de pas du problème inhérent à la longue vallée étroite.

Cela n'arrive que si on prend un pas plus petit que celui qu'on aurait du prendre et c'est pourquoi on a toujours intérêt à utiliser une technique de paramètre de relaxation qui revoit à la baisse la longueur de pas calculée si celle ci est faite par la méthode de Newton unidirectionnelle.

Méthode B.F.G.S. On recommence le travail des paragraphes précédents en utilisant l'algorithme B.F.G.S. : au départ on a (x_0, y_0) à partir duquel on fait un pas de gradient avec calcul de la longueur de pas par la méthode de Newton unidirectionnelle pour obtenir (comme précédemment (204))

$$(x_1, y_1) = \left(\frac{x_0 \, {y_0}^2 a^4 \left(a^2 - 1\right)}{{x_0}^2 + {y_0}^2 a^6}, -\frac{y_0 \, {x_0}^2 \left(a^2 - 1\right)}{{x_0}^2 + {y_0}^2 a^6} \right) \tag{209}$$

On choisit une matrice h_0 symétrique (pas forcement définie positive) mais à ceci près quelconque

$$h_0 := \begin{pmatrix} u_0 & t_0 \\ t_0 & u_0 \end{pmatrix} \tag{210}$$

on a

$$\begin{aligned} y_1 &= \left(-2 \frac{\left({x_0}^2 + {y_0}^2 a^4\right) x_0}{\left({x_0}^2 + {y_0}^2 a^6\right) a^2}, -2 \frac{\left({x_0}^2 + {y_0}^2 a^4\right) a^2 y_0}{{x_0}^2 + {y_0}^2 a^6} \right) \\ p_1 &= \left(-\frac{\left({x_0}^2 + {y_0}^2 a^4\right) x_0}{{x_0}^2 + {y_0}^2 a^6}, -\frac{\left({x_0}^2 + {y_0}^2 a^4\right) a^2 y_0}{{x_0}^2 + {y_0}^2 a^6} \right) \end{aligned} \tag{211}$$

$$-\frac{]h_0 p_1, h_0 p_1[}{{}^t p_1 h_0 p_1} = \begin{pmatrix} -\frac{\left(u_0 \, x_0 + t_0 \, a^2 y_0\right)^2}{u_0 \, {x_0}^2 + 2 \, x_0 \, t_0 \, a^2 y_0 + {y_0}^2 a^4 d_0} & -\frac{\left(u_0 \, x_0 + t_0 \, a^2 y_0\right)\left(t_0 \, x_0 + d_0 \, a^2 y_0\right)}{u_0 \, {x_0}^2 + 2 \, x_0 \, t_0 \, a^2 y_0 + {y_0}^2 a^4 d_0} \\ -\frac{\left(u_0 \, x_0 + t_0 \, a^2 y_0\right)\left(t_0 \, x_0 + d_0 \, a^2 y_0\right)}{u_0 \, {x_0}^2 + 2 \, x_0 \, t_0 \, a^2 y_0 + {y_0}^2 a^4 d_0} & -\frac{\left(t_0 \, x_0 + d_0 \, a^2 y_0\right)^2}{u_0 \, {x_0}^2 + 2 \, x_0 \, t_0 \, a^2 y_0 + {y_0}^2 a^4 d_0} \end{pmatrix} \tag{212}$$

$$\frac{]y_1, y_1[}{{}^t p_1 y_1} = \begin{pmatrix} 2 \frac{{x_0}^2}{a^2 \left({x_0}^2 + {y_0}^2 a^6\right)} & 2 \frac{a^2 x_0 \, y_0}{{x_0}^2 + {y_0}^2 a^6} \\ 2 \frac{a^2 x_0 \, y_0}{{x_0}^2 + {y_0}^2 a^6} & 2 \frac{{y_0}^2 a^6}{{x_0}^2 + {y_0}^2 a^6} \end{pmatrix} \tag{213}$$

Les expressions de h_1 (donnée par (191) et celle de son inverse deviennent trop importantes pour être écrites. Les expressions restant toujours aussi importantes, on calcule le point x_2 avec (185) en utilisant la méthode de Newton pour la minimisation unidirectionnelle et on trouve que, pour une matrice h_0 quelconque, $x_2 = 0$ exactement

La méthode B.F.G.S. appliquée à cet exemple, mais ce serait vrai pour toute fonction quadratique, fournit une solution exacte à l'itération qui succède directement à la première itération d'initialisation.

6.2 Fonction de Rosenbrock

La fonction de Rosenbrock est

$$f(x,y) = p(y-x^2)^2 + (1-x)^2 \tag{214}$$

où f est une fonction réelle des deux variables réelles x et y et p est paramètre qui peut être très grand (100 par exemple).

Il faut :

a) dessiner les lignes de niveaux qui apparaîtraient sur une carte de cette terre

b) appliquer les méthodes d'abord de la plus grande descente ensuite de Newton pour trouver un algorithme de recherche du point le plus bas de cette terre (qui est évidemment $x_\infty = y_\infty = 1$)

c) trouver des zones particulières pour le point de départ des algorithmes qui mettent en évidence :

1. l'efficacité de la méthode de Newton ;
2. puis son inefficacité ;
3. l'inefficacité de la méthode de la plus grande pente ;
4. puis son efficacité.

solution proposée La solution est $f(1,1) = 0$ puisque f est la somme de deux termes positifs qui doivent être nuls séparément, ce qui conduit à cette solution. D'autre part le gradient de f est

$$\nabla f(x,y) = (-2p(y-x^2)x - 1 + x, p(y-x^2))$$

qui n'est nul que pour cette solution, donc il n'y a pas de minimum locaux cachés.

D'abord l'efficacité de la méthode Newton. On peut chercher les points qui conduisent à la solution exacte en un nombre fini d'itération.

L'algorithme de Newton est

$$(X,Y) = (\frac{-2\,p\,x\,y + 2\,p\,x^3 + 1}{-2\,p\,y + 1 + 2\,p\,x^2},\ \frac{x\,(-2\,p\,x\,y + 2 - x + 2\,p\,x^3)}{-2\,p\,y + 1 + 2\,p\,x^2})$$

si (x,y) est le point de départ, il est amené à (X,Y) par cette formule.

La question est déjà de chercher les points (x,y) tels que $(X,Y) = (1,1)$; à la main on trouve assez facilement que ces points appartiennent à la droite

$$x = 1$$

Ensuite il faut trouver les points (x,y) tels que $(X,Y) = (1,Y)$ (profiter de l'occasion pour montrer que cette formulation est une paramétrisation) ; à la main on trouve (un peu moins facilement) que ces points appartiennent à la parabole

$$y = x^2$$

Si maintenant on cherche les points (x,y) tels que $(X,Y) = (X,X^2)$ on (maple) n'en trouve aucun.

Par contre l'inversion de l'algorithme de Newton peut être tentée (par maple), on trouve

$$\begin{cases} x &= X + \sqrt{X^2 - Y} \\ y &= \dfrac{1}{2}\,\dfrac{-2\,p\,(X + \sqrt{X^2-Y})\,Y + 4\,p\,(X + \sqrt{X^2-Y})\,X^2 - 2\,p\,Y\,X + 1 - X}{p\,\sqrt{X^2-Y}} \end{cases}$$

ou

$$\begin{cases} x &= X - \sqrt{X^2 - Y} \\ y &= -\dfrac{1}{2}\,\dfrac{-2\,p\,(X - \sqrt{X^2-Y})\,Y + 4\,p\,(X - \sqrt{X^2-Y})\,X^2 - 2\,p\,Y\,X + 1 - X}{p\,\sqrt{X^2-Y}} \end{cases}$$

ce qui porte à penser que

- les points (X,Y) tels que $Y \geq X^2$ (au dessus de la parabole) ne seront jamais atteints par l'algorithme de Newton, sauf à admettre l'existence des points à l'infini ; et donc aucun point de départ dans le plan ne conduit en trois ou plus itérations à la solution ;
- il existe (au moins) deux points (x,y) qui atteignent un même point (X,Y) ;
- toute autre remarque que l'examen de ces deux expressions peut aussi être pensée (et même aussi exprimée).

D'autre part les points tels que

$$y = x^2 + \frac{1}{2p}$$

mènent les points (X,Y) à l'infini (remarquer que cette condition est évidemment celle pour laquelle le hessien de f

$$\nabla^2 f(x,y) = \begin{pmatrix} 6px^2 - 2py + 1 & -2px \\ -2px & p \end{pmatrix}$$

comporte une valeur propre nulle et donc n'est pas inversible.), ce sont donc des points pour lesquels la méthode de Newton échoue (remarquer que pour p grand ces points sont très proches de ceux pour lesquels la méthode converge en deux coups et par conséquent que du capitole à la roche Tarpéienne il n'y a qu'un pas).

Une question intéressante est celle de savoir si un point (x, y) du plan peut être amené en (X, Y) satisfaisant $Y = X^2 + 1/2p$. Celui qui serait intéressé a toute latitude pour répondre à cette question en utilisant le canevas maple qui leur a été fourni.

Une autre question est de topographier le lieu du plan dans lesquelles le hessien est défini positif : il faut pour cela que le déterminant du hessien soit positif et aussi sa trace (pour éviter deux valeurs propres négatives) :

$$y < x^2 + \frac{1}{2p} \; ; \; y < \frac{3}{2}x^2 + \frac{1}{2} + \frac{1}{2p}$$

la première condition suffit seule.

Par considération de l'inverse, on remarque cependant que l'algorithme de Newton portera un point qui ne se trouverait pas dans ce lieu vers ce lieu.

Solution proposée pour 3 et 4 :

On peut maintenant aborder l'analyse de la méthode du gradient. L'algorithme est

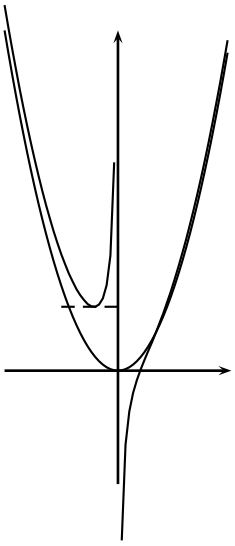
$$\begin{aligned} X &= x - \alpha(-2p(y - x^2)x - 1 + x) \\ Y &= y - \alpha(p(y - x^2)) \end{aligned}$$

où α, le pas de descente, doit être choisi au mieux (préciser ce qu'il est possible de faire pour choisir ce pas :

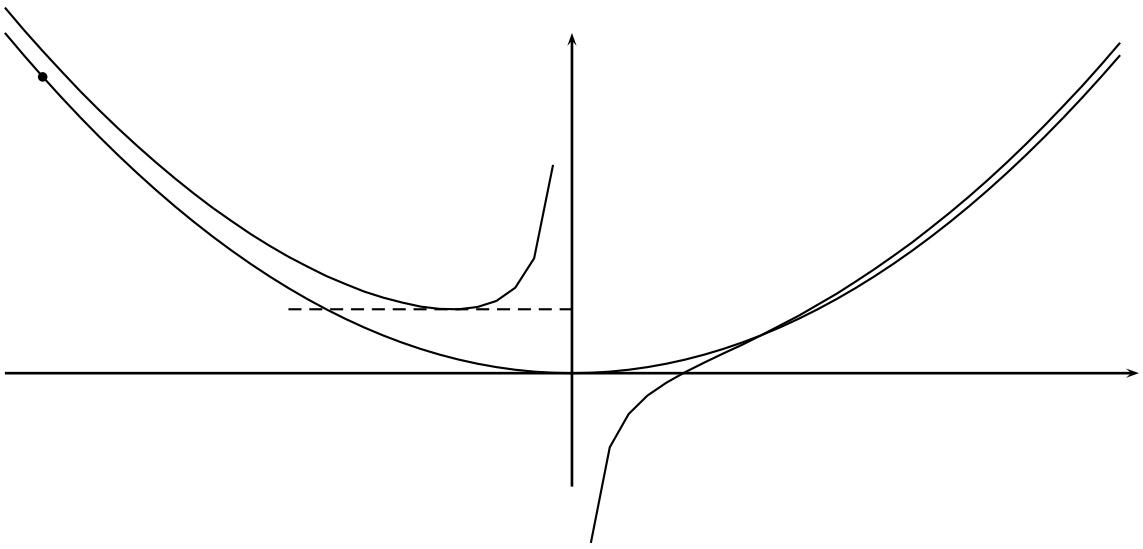
- la méthode de Newton sur le problème monodimensionel en α ;
- prendre α au hasard, le diviser ou le multiplier par 2 (par exemple) suivant les valeurs obtenue pour la fonction.
- tout autre manière qui plaira à condition de ne pas introduire des complications abstraites tirée de la condition de Zoutendijk.

)

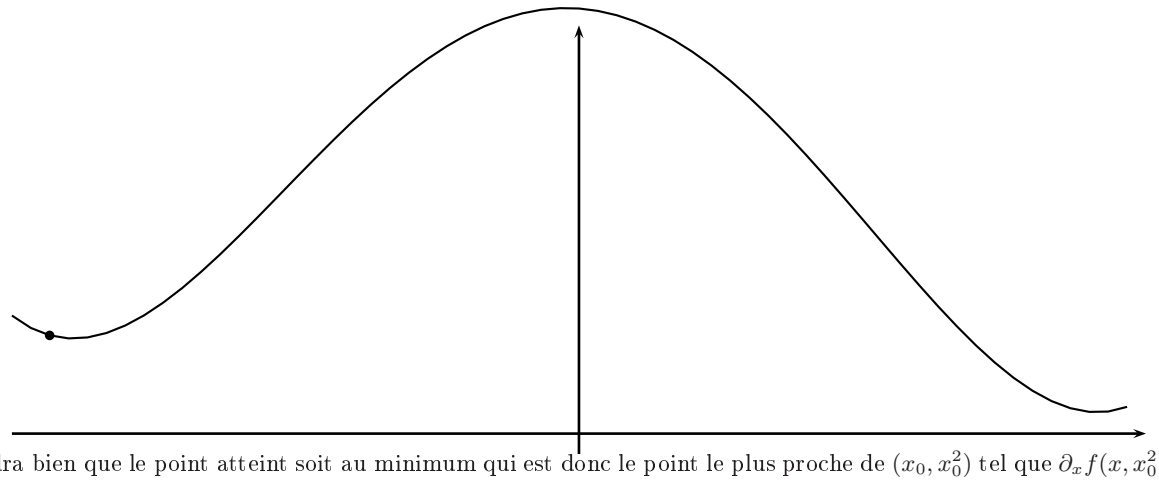
On suppose maintenant que ce pas est choisi au mieux, qu'on dispose d'une méthode inspécifiée ici qui permet de le choisir au mieux ; on fait tracer le lieux des points du plan pour lesquels $\partial_x f(x, y) = 0$; celui pour lesquels $\partial_y f(x, y) = 0$; et on part d'un point $(x_0, y_0 = x_0^2)$ avec x_0 suffisamment petit pour que le point de départ soit situé au dessus de la ligne en pointillé (le point marqué sur la parabole)



Il est un peu difficile de travailler sur cette figure aussi on change un peu l'échelle en y (remarquer qu'un tracé qualitatif est plus lisible qu'un tracé respectant les dimensions)



Si donc on part de ce point, la dérivée en y de f y est nulle et donc le déplacement sera dirigé suivant x ; comme l'application partielle $f(x, x_0^2)$ est de la forme (le point marqué est $(x0, f(x_0, x_0^2))$) (trouver que la forme ne peut être que celle-ci)



il faudra bien que le point atteint soit au minimum qui est donc le point le plus proche de (x_0, x_0^2) tel que $\partial_x f(x, x_0^2) = 0$; c'est le point le plus proche parce que l'algorithme permettant la recherche de α n'est *a priori* pas capable de franchir la montagne.

Ensuite on arrive donc en un point (x_1, x_0^2) tel que la dérivée de f par rapport à x est nulle et donc le déplacement sera dirigé suivant y et on atteint le point (x_1, x_1^2) (les tracés sont plus faciles à faire dans la direction y).

Il reste donc à continuer l'algorithme en descendant alternativement suivant x et suivant y jusqu'à ce atteindre un point (x_n, x_n^2) tel que x_n^2 soit inférieur à la ligne pointillée.

Là on arrive à l'autre branche de la courbe et on descend ensuite, toujours de façon alternée, sur le couple des deux courbes de droite.

L'inconvénient principal de la méthode du gradient est ainsi illustré : on voit bien que la descente peut être très lente. Cependant il n'y a pas de points à partir duquel on ne peut descendre.

Un phénomène intéressant est aussi illustré sur cet exemple : lors de la première descente les points successifs (x_n, y_n) sont très proches les un des autres jusqu'au changement de branche pour lequel la valeur de x change de façon spectaculaire, puis la descente reprend de façon lente.

D'une part ce phénomène est souvent rencontré (la fonction de Rosenbrock est faite pour cela) et d'autre part il s'agit d'une catastrophe au sens de Thom.

6.3 Forme d'une membrane axisymétrique

On regarde une membrane tendue sur deux cerceaux coaxiaux placés à la hauteur H l'un de l'autre.

On pense que la forme prise par cette membrane est celle qui minimise sa surface.

Il est également possible qu'on souhaite construire une telle membrane et qu'on cherche alors cette forme minimum pour dépenser le moins possible de matière dans la construction.

Ce problème admet un traitement analytique qui va être décrit explicitement. Mais l'objectif de l'exercice est de le traiter avec une méthode numérique.

6.3.1 Traitement analytique

La question centrale est de trouver la forme de la membrane qui minimise sa surface.

On décide de paramétrer le lieu des points qui appartiennent à la menbrane de la façon suivante : on choisit un plan normal à l'axe d'invariance par rotation des deux cerceaux de manière que le cerceau le plus bas soit dans ce plan ; on place l'origine O d'un repère orthonormé à l'intersection de l'axe et de ce plan et on donne deux directions par deux vecteurs du plan $\vec{i}$ et $\vec{j}$ orthogonaux ; le troisième vecteur $\vec{k}$ sera dirigé suivant l'axe.

Avec ce repère, tout point P de l'espace est défini par la donnée de ses coordonnées (x, y, z) de manière que $\vec{OP} = x\vec{i} + y\vec{j} + x\vec{k}$; et le lieu des points appartenant à la menbrane est défini par les triplets $(x, y, \aleph(x, y))$ où

$$\aleph(x, y) = \begin{cases} 0 & \text{si} \quad x^2 + y^2 > R_0^2 \\ H & \quad\quad x^2 + y^2 < R_1^1 \end{cases}$$

est une fonction (différentiable) qui détermine la forme de la menbrane.

Deux critiques de cette paramétrisations peuvent être faites :

1. on n'a pas donné d'épaisseur à la membrane et donc l'objet mathématique ne peut pas avoir de correspondance avec le monde sensible ;
2. les situations dans lesquelles la membrane comporterait une invagination

   sont exclues car alors $\aleph$ serait multivaluée.

La réponse à la première critique est que cette approche est une modélisation : on ne prétend pas qu'elle soit la réalité mais on lui demande seulement de lui ressembler ; c'est cela même qui peut être appelé la modélisation.

La réponse à la deuxième critique est qu'elle est très légitime et qu'il faudrait s'en souvenir si les calculs de l'approche conduisait à des bizarreries.

Si D est la couronne $R_1 < \sqrt{x^2+y^2} < R_0$ la surface est

$$\Sigma(\aleph) = \int_D \sqrt{1 + \left(\vec{\nabla}\aleph(x,y)\right)^2} dxdy$$

où

$$\vec{\nabla}h(x,y) = \partial_x h(x,y)\vec{e}_1 + \partial_y h(x,y)\vec{e}_2$$

et où on remarque que le mot 'surface' dénote ici une fonction à valeur réelle dont l'argument est une fonction ; ce qu'on appelle une fonctionnelle.

Si maintenant on décide d'utiliser des coordonnées cylindriques, on peut introduire $\kappa(r,\theta)$ une nouvelle fonction κ telle que

$$\kappa(r,\theta) = \aleph(r\cos\theta, r\sin\theta)$$

et donc telle que

$$\kappa(R_0,\theta) = \begin{cases} 0 & \text{si} \quad r = R_0 \\ H & \quad\quad r = R_1 \end{cases}$$

alors la nouvelle façon d'exprimer la surface sera

$$S(\kappa) = \int_{R_0}^{R_1} r dr \int_0^{2\pi} d\theta \sqrt{1 + \left(\partial_r \kappa(r,\theta)\right)^2 + \left(\frac{\partial_\theta \kappa(r,\theta)}{r}\right)^2}$$

L'analyse[5] de l'expression de la surface met en évidence que si on prend une fonction κ quelconque on peut bâtir à partir d'elle une autre fonction qui ne dépendrait plus de θ et qui conduirait à une valeur de surface plus petite (ou égale) que la première.

En effet si $\kappa(r,\theta)$ (avec $\kappa(R_0,\theta) = 0$ et $\kappa(R_1,\theta) = H$) est donnée alors en prenant

$$h(r) = \frac{1}{2\pi}\int_0^{2\pi} \kappa(r,\theta) d\theta$$

si on sait que

- $\forall f : \left(\int_a^b f(x)dx\right)^2 \leq (b-a)\int_a^b \left(f(x)\right)^2 dx$ (le carré de la moyenne est plus petit que la moyenne des carrés)
- si $\int_a^b f(x)dx \leq \int_a^b g(x)dx$ et que F est une fonction croissante alors $\int_a^b F(f(x))dx \leq \int_a^b F(g(x))dx$ [6]

on obtient alors

$$\begin{aligned} \int_{R_0}^{R_1} \left(\frac{dh}{dr}(r)\right)^2 dr &= \int_{R_0}^{R_1} \left(\frac{1}{2\pi}\int_0^{2\pi} \partial_r \kappa(r,\theta) d\theta\right)^2 dr \\ &\leq \int_{R_0}^{R_1} \left(\frac{1}{2\pi}\int_0^{2\pi} \left(\partial_r \kappa(r,\theta)\right)^2 d\theta\right) dr \end{aligned}$$

d'où

$$\Sigma(h) \leq \Sigma(\kappa)$$

Il est donc inutile de considérer le cas général d'une fonction κ qui ne serait pas à symétrie de révolution pour le problème de recherche de minimum.

Il reste alors, en appelant $h(r)$ les fonctions κ qui ne dépendent pas de θ, que l'expression de la surface est

$$S(h) = 2\pi \int_{R_0}^{R_1} r dr \sqrt{1 + \left(\frac{dh}{dr}(r)\right)^2}$$

Pour la commodité d'écriture on introduit l'espace des fonctions correspondant à des surfaces s'appuyant sur les cerceaux comme

$$\mathcal{V} = \{h \text{ telle que } h(R_0) = 0 \text{ et } h(R_1) = H\}$$

Supposons qu'il existe une fonction particulière $h_\infty \in \mathcal{V}$ telle que

$$\forall h \in \mathcal{V} : S(h_\infty) \leq S(h)$$

La question est maintenant de trouver une condition nécessaire à laquelle doit satisfaire h_∞ et de vérifier qu'elle est aussi suffisante.

Pour cela On pose $h = h_\infty + \delta\lambda h'$ (on remarquera qu'alors $h'(R_0) = h'(R_1) = 0$) et on fait un développement limité au premier ordre en $\delta\lambda$ de $S(h_\infty + \delta\lambda h')$ pour trouver cette condition.

[5] Attention l'analyse en question telle que donnée dans ce texte est fausse ! Merci à E. Plaut de l'avoir signalé. L''analyse' est pourtant laissée parce qu'elle a l'apparence du raisonnable et que le résultat auquel elle conduit est certainement vrai, mais il n'est pourtant pas démontré.

[6] C'est cette affirmation qui est fausse. Pourquoi ? Trouver un contre-exemple.

D'abord la question du calcul sordide

$$\begin{aligned}
\sqrt{1+(\dot{h}_\infty+\delta\lambda\dot{h}')^2} &= \sqrt{1+\dot{h}_\infty^2+2\delta\lambda\dot{h}'\dot{h}_\infty+\delta\lambda^2\dot{h}'^2} \\
&= \sqrt{1+\dot{h}_\infty^2}\sqrt{1+2\frac{\dot{h}'\dot{h}_\infty}{1+\dot{h}_\infty^2}\delta\lambda+\frac{\dot{h}'^2}{1+\dot{h}_\infty^2}\delta\lambda^2} \\
&= \sqrt{1+\dot{h}_\infty^2}\left(1+\left(\frac{\dot{h}'\dot{h}_\infty}{1+\dot{h}_\infty^2}\delta\lambda+\frac{\frac{1}{2}\dot{h}'^2}{1+\dot{h}_\infty^2}\delta\lambda^2\right)-\frac{1}{8}\left(\frac{\dot{h}'\dot{h}_\infty}{1+\dot{h}_\infty^2}\delta\lambda+o(\delta\lambda)\right)^2+o(\delta\lambda^)\right) \\
&= \sqrt{1+\dot{h}_\infty^2}+\frac{\dot{h}'\dot{h}_\infty}{\sqrt{1+\dot{h}_\infty^2}}\delta\lambda+\frac{1}{8}\frac{4\dot{h}'^2(1+\dot{h}_\infty^2)-(\dot{h}'\dot{h}_\infty)^2}{\left(1+\dot{h}_\infty^2\right)^{3/2}}+o(\delta\lambda^2) \\
&= \frac{\dot{h}'\dot{h}_\infty}{\sqrt{1+\dot{h}_\infty^2}}\delta\lambda+\frac{1}{8}\frac{4\dot{h}'^2+3(\dot{h}'\dot{h}_\infty)^2}{\left(1+\dot{h}_\infty^2\right)^{3/2}}+o(\delta\lambda^2)
\end{aligned}$$

Ensuite on peut écrire que si $h_\infty \in \mathcal{V}$

$$\forall h \in \mathcal{V} \ : \ S(h_\infty) \leq S(h)$$

est équivalent à

$$\forall h' \in \mathcal{V}_0; \forall \delta\lambda \ : \ S(h_\infty) \leq S(h_\infty+\delta\lambda h')$$

où

$$\mathcal{V}_0 = \left\{h' \text{ telle que } h'(R_0)=h'(R_1)=0\right\}$$

En particulier si $\delta\lambda$ devient petit alors cette deuxième condition se transforme en

$$\forall h' \in \mathcal{V}_0; \forall \delta\lambda \ : \ S(h_\infty) \leq S(h_\infty)+\delta\lambda 2\pi\int_{R_0}^{R_1}\frac{\frac{dh_\infty}{dr}\frac{dh'}{dr}}{\sqrt{1+\left(\frac{dh_\infty}{dr}\right)^2}}rdr+o(\delta\lambda)$$

soit

$$\forall h' \in \mathcal{V}_0 \ : \ 0 \leq \delta\lambda 2\pi\int_{R_0}^{R_1}\frac{\frac{dh_\infty}{dr}\frac{dh'}{dr}}{\sqrt{1+\left(\frac{dh_\infty}{dr}\right)^2}}rdr+o(\delta\lambda)$$

Si donc on peut admettre qu'il est possible de trouver un $\delta\lambda$ suffisamment petit pour que la présence de $o(\delta\lambda)$ n'influe pas sur le signe du second membre de l'inégalité, comme cette dernière condition ne présage pas du signe de $\delta\lambda$, il faudra alors que

$$\forall h' \in \mathcal{V}_0 \ : \ 2\pi\int_{R_0}^{R_1}\frac{\frac{dh_\infty}{dr}\frac{dh'}{dr}}{\sqrt{1+\left(\frac{dh_\infty}{dr}\right)^2}}rdr$$

soit simultanément positif et négatif ; donc nul.

La condition nécessaire cherché est

$$\forall h' \in \mathcal{V}_0 \ : \ 2\pi\int_{R_0}^{R_1}\frac{\frac{dh_\infty}{dr}\frac{dh'}{dr}}{\sqrt{1+\left(\frac{dh_\infty}{dr}\right)^2}}rdr=0$$

Si cette condition est remplie alors l'inégalité devient ($\dot{f}=\frac{df}{dr}$)

$$\forall h' \in \mathcal{V}_0 \ : \ 0 \leq \int_{R_0}^{R_1}\frac{1}{8}\frac{4\dot{h}'^2+3(\dot{h}'\dot{h}_\infty)^2}{\left(1+\dot{h}_\infty^2\right)^{3/2}}rdr+o(\delta\lambda)$$

ce qui semble vrai en reconduisant l'hypothèse portant sur $o(\delta\lambda^2)$. Mais évidemment comme on n'a pas précisé l'espace dans lequel devaient être les h, il est possible de trouver des fonctions h' non nulle sur des portions négligeable de $[R_0,R_1]$ et pour lesquels cette quantité serait nulle et donc pour lesquels il faudrait pousser le développement à l'ordre suivant ; ce qui d'ailleurs est la raison pour laquelle il est intéressant d'introduire les espaces fonctionnels.

On n'entre pas dans ces considérations et on se contente de dire que la positivité avérée de l'ordre deux est la condition suffisante.

Il reste maintenant à trouver une équation différentielle à laquelle doit satisfaire h_∞ et vérifier que des conditions aux limites en quantité suffisante sont disponibles pour qu'une solution unique puisse exister.

On a

$$\int_{R_0}^{R_1}\frac{\frac{dh_\infty}{dr}\frac{dh'}{dr}}{\sqrt{1+\left(\frac{dh_\infty}{dr}\right)^2}}rdr=\left[h'\frac{r\frac{dh_\infty}{dr}}{\sqrt{1+\left(\frac{dh_\infty}{dr}\right)^2}}\right]_{R_0}^{R_1}-\int_{R_0}^{R_1}h'\frac{d}{dr}\left(\frac{r\frac{dh_\infty}{dr}}{\sqrt{1+\left(\frac{dh_\infty}{dr}\right)^2}}\right)dr$$

et donc en utilisant les conditions $h' \in \mathcal{V}_0$, la condition nécessaire devient

$$\forall h' \in \mathcal{V}_0 \ : \ \int_{R_0}^{R_1} h' \frac{d}{dr} \left(\frac{r \frac{dh_\infty}{dr}}{\sqrt{1 + \left(\frac{dh_\infty}{dr}\right)^2}} \right) dr = 0$$

Si maintenant on croit que c'est équivalent (le seul obstacle serait des espaces fonctionnels dans lesquels ce ne serait pas vrai) à

$$\frac{d}{dr} \left(\frac{r \frac{dh_\infty}{dr}}{\sqrt{1 + \left(\frac{dh_\infty}{dr}\right)^2}} \right) = 0$$

on a trouvé l'équation différentielle cherché et on possède évidemment les conditions aux limites correspondant à la condition $h_\infty \in \mathcal{V}$ soit

$$h_\infty(R_0) = 0 \text{ et } h_\infty(R_1) = H$$

Il suffit alors de résoudre cette équation différentielle et d'exprimer la solution en fonction de R_0, R_1, H quand c'est possible

On trouve déjà

$$r \frac{dh_\infty}{dr} = k \sqrt{1 + \left(\frac{dh_\infty}{dr}\right)^2}$$

où k est une constante. Si $R_0 < R_1$ cette constante est positive puisque le changement de signe de $\frac{dh_\infty}{dr}$ est rendu impossible par cette première résolution et négative dans le cas contraire.

On obtient par résolution algébrique

$$(r^2 - k^2) \left(\frac{dh_\infty}{dr}\right)^2 = k^2$$

qui n'admet de solutions que si

$$k^2 < R_0 (< R_1)$$

dans ce cas on a aussi

$$\frac{dh_\infty}{dr} \geq 0$$

et donc

$$k > 0$$

on peut alors écrire

$$\frac{dh_\infty}{dr} = \frac{k}{\sqrt{r^2 - k^2}}$$

dont la solution est

$$h_\infty(r) = k \log \left(r + \sqrt{x^2 - k^2} \right) + k'$$

où k' est une autre constante. La condition à la limite R_0 permet d'écrire

$$h_\infty(r) = k \log \left(\frac{r + \sqrt{r^2 - k^2}}{R_0 + \sqrt{R_0^2 - k^2}} \right)$$

et la condition à la limite R_1 fournit le calcul de k sous la forme

$$H = k \log \left(\frac{R_1 + \sqrt{R_1^2 - k^2}}{R_0 + \sqrt{R_0^2 - k^2}} \right)$$

Donc si la résolution de cette équation en k est possible (et si il y a une solution unique) alors on a résolu le problème entier.

La membrane sera une surface de révolution dont le profil sera la fonction $h_\infty(r)$ trouvée.

Hélas un cas d'impossibilité apparait. Il n'est en effet pas certain du tout que la résolution de l'équation précédente soit possible.

Examinons cette question. En posant

$$x = \tfrac{k}{H} \quad x_1 = \tfrac{R_1}{H} \quad x_0 = \tfrac{R_0}{H}$$

l'équation devient

$$x \log \frac{x_1 + \sqrt{x_1^2 - x^2}}{x_0 + \sqrt{x_0^2 - x^2}} = 1$$

et il faut trouver une solution telle que

$$0 < x < x_0 < x_1$$

d'autre part on sait que $\frac{dh_\infty}{dr} > 0$ et donc que cette fonction de x est également croissante.

On a donc

$$x \log \frac{x_1 + \sqrt{x_1^2 - x^2}}{x_0 + \sqrt{x_0^2 - x^2}} \leq x_0 \log \frac{x_1 + \sqrt{x_1^2 - x_0^2}}{x_0}$$

et il y aura une solution et elle sera unique si

$$x_0 \log \frac{x_1 + \sqrt{x_1^2 - x_0^2}}{x_0} \geq 1$$

Il n'y aura donc pas de solutions possibles si, R_0 et R_1 étant fixés, H est trop grand.

Que dire alors ?

À la limite de la condition

$$x = x_0$$

donc

$$k = R_0$$

et

$$h_\infty(r) = R_0 \log \left(\frac{r}{R_0} + \sqrt{\left(\frac{r}{R_0}\right)^2 - 1} \right)$$

et surtout

$$\frac{dh_\infty}{dr}(r) = \frac{R_0}{\sqrt{r^2 - R_0^2}}$$

la pente du profil de la membrane est infinie en R_0 !

On peut alors se dire que la paramétrisation qui supposait que h soit une fonction atteint ses limites : dans ce cas il serait préférable d'utiliser plutôt une paramétrisation dont le cœur serait une fonction dont l'argument est z et la valeur r (soit h^{-1}).

Le faire présenterait cependant l'inconvénient majeur de faire apparaître une équation différentielle qui ne s'intégre pas si facilement que celle qu'on vient de traiter et donc le numérique devrait venir au secours de l'analytique.

Dans ce dernier cas il devient préférable de traiter le problème comme cela sera proposé au paragraphe suivant.

Avant d'y venir, et parce que cela sera utile dans ce paragraphe, on peut cependant s'interroger sur une question subsidiaire.

Comment calculer la dérivée première de $S(h_\infty)$ par rapport à R_0, R_1, H ?

On peut évidemement calculer

$$\begin{aligned} S(h_\infty) &= \int_{R_0}^{R_1} \sqrt{1 + \left(\frac{dh_\infty}{dr}\right)^2} r dr \\ &= \int_{R_0}^{R_1} \frac{r}{r^2 - k^2} r dr \\ &= \left[r + \frac{k}{2} \log \frac{k - r}{k + r} \right]_{R_0}^{R_1} = (R_1 - R_0) + \frac{k}{2} \log \frac{(R_1 - k)(k + R_0)}{(R_0 - k)(k + R_1)} \end{aligned}$$

(attention la formule n'est pas valable dans le cas limite) et calculer les dérivées directement.

Mais on peut également utiliser les calculs de variation déjà fait ce qui fait apparaître des variations 'horizontales' pour les variations δR_0 et δR_1 et une variation verticale pour δH.

Le fait de faire des variations des paramétres du problème entraîne une variation de premier ordre δh_∞ et donc

$$S(h_\infty + \delta h_\infty) = 2\pi \int_{R_0 + \delta R_0}^{R_1 + \delta R_1} r dr \sqrt{1 + \left(\frac{d(h_\infty + \delta h_\infty)}{dr}(r) \right)^2}$$

compte tenu de la condition nécessaire permettant le calcul de h_∞ et également compte tenu que $\delta h_\infty(R_1) = \delta H \neq 0$, le développement à l'ordre 1 de cette expression conduit à

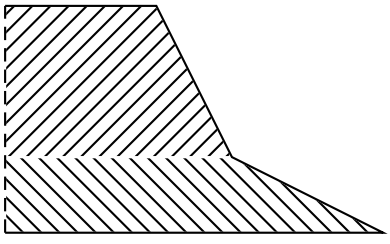
$$\delta S / 2\pi = \delta R_1 \sqrt{1 + \left(\frac{dh_\infty}{dr}(R_1)\right)^2} - \delta R_0 \sqrt{1 + \left(\frac{dh_\infty}{dr}(R_0)\right)^2} + \delta H \frac{R_1 \frac{dh_\infty}{dr}(R_1)}{\sqrt{1 + \left(\frac{dh_\infty}{dr}(R_1)\right)^2}}$$

d'où on tire les dérivées demandées.

6.3.2 Traitement numérique

On a développé longuement le traitement analytique du problème de lampadaire pour que tous les aspects en soient clairs afin d'utiliser ce dernier à des fins de vérification d'algorithmes numériques.

Deux tronçons de cône Si on paramétrise la surface par deux tronçons de cônes ; cela permet alors de ramener le problème à l'étude d'une fonction à deux variables dont il faut chercher le minimum en espérant que l'abscisse du point libre ne devienne pas négative.



Il faut le faire.

N tronçons de cônes ? Si maintenant on la paramétrise par un nombre quelconque N de section de cônes ; cela permet alors de ramener le problème à l'étude d'une fonction à $2N$ variables dont il faut chercher le minimum

Il constater cette méthode prise brutalement n'a aucune chance de réussir puisque si par exemple trois points succcessifs des N points sont alignés alors on peut déplacer le point milieu sans changer la valeur de la surface.

N tronçons de cônes : le support est l'ordonnée Aussi on peut décider de fixer les ordonnées de ces N points et de classer ces derniers par ordonnées croissantes (équiréparties par exemple), du cercle inférieur au cercle supérieur. La surface devient alors une fonction des abscisses des N points, soit une fonction à N variables

Comme monsieur Jourdain qui faisait de la prose sans le savoir, on fait ainsi une approximation du problème de départ par des éléments finis monodimensionnels de Lagrange P1.

N tronçons de cônes : le support est l'abscisse Plutôt que de fixer les ordonnées on peut décider de fixer les abscisses.

Troisième partie

Minimisation en dimension finie de fonctions deux fois différentiables sous contraintes égalité et inégalité

Si une fonction F d'argument le vecteur X élément de l'espace euclidien E_N, différentiable 2 fois, est donnée, il s'agit de trouver le ou les X_∞ élément(s) d'un sous-ensemble D de E_N qui minimise(nt) cette fonction, c'est à dire :

$$\begin{array}{l} \forall X'' \in D \quad F(X_\infty) \leq F(X'') \\ X_\infty \in D \end{array} \tag{215}$$

Plusieurs possibilités apparaissent :

1. *D est donné par paramétrisation*

$$D = \{X \in E_N \text{ tel que } \exists x \in E_{N-P} \text{ avec } X = L(x)\} \tag{216}$$

   où E_{N-P} est un espace de dimension inférieures à E_N et L est une application deux fois différentiables de E_{N-P} dans E_N qui est connue ;

2. *D est donné par les équations*

$$D = \{X \in E_N \text{ tel que } G(X) = 0\} \tag{217}$$

   où G est une application deux fois différentiable de E_N dans E_P (E_P étant un espace de dimension P) ;

3. *D est donné par les inéquations*

$$D = \{X \in E_N \text{ tel que } G(X) \leq 0\} \tag{218}$$

   où G est l'application précédente et le signe $\leq$ appliqué à deux vecteurs signifie qu'il agit sur chacune des composantes des vecteurs.

4. *Il y a encore d'autres possibilités ; le panachage des trois premières ; le renoncement à la différentiabilité de G ou L ; ...*

   mais on les écarte tout de suite le traitement de cette diversité des objectifs de cette partie dans laquelle seul le simple (et donc le beau ?) a droit de cité.

Le premier cas est favorable, il suffit d'introduire la fonction f d'argument $x \in E_{N-P}$ telle que

$$f(x) = F(L(x)) \tag{219}$$

pour ramener le problème (215) à un problème de minimisation sans contrainte qu'on suppose ici parfaitement dominé.

Le second cas est plus gênant ; sauf s'il se ramène facilement à celui de la paramétrisation (216).

Le troisième cas apparaît comme encore plus gênant et effectivement il nécessite une discussion préalable sur la nature géométrique des ensembles définis par (218).

Donc avant de tenir un discours sur la ou les façons générales de résoudre (216) il n'est pas inutile de visiter, si ce n'est pas déjà fait le paragraphe hypersurface page 9.

7 Une seule contrainte

La seule volonté libre peut engager,
Et jamais la contrainte.
LENOBLE, le Cerf et la Brebis.

7.1 La contrainte égalité

Il s'agit de résoudre le problème de trouver X_∞ tel que, une fonction G soumise aux conditions

$$\exists X \text{tel que } G(X) = 0 \text{ et de plus si } G(X) = 0 \text{ alors } \nabla G(X) \neq 0 \tag{220}$$

étant donnée, on a

$$\begin{array}{l} G(X_\infty) = 0 \\ \forall X \text{ tels que } G(X) = 0 : F(X_\infty) \leq F(X) \end{array} \tag{221}$$

Et, pour éviter de permanentes précautions oratoires, on suppose dans ce paragraphe que les F et G sont suffisamment bien choisis pour qu'il existe une solution unique au problème.

Bien sûr, ce qui va être raconté peut être appliqué aux cas pour lesquels il n'y a pas nécessairement de solution unique ; il faudra simplement noter qu'on obtiendra des solutions locales.

7.1.1 Le multiplicateur de Lagrange

Les conditions données sur G permettent d'introduire une paramétrisation L telle que

$$\forall x \in E_{N-1} : G(L(x)) = 0 \tag{222}$$

et, même si on sait que dans de nombreux cas il ne sera pas possible de trouver effectivement un expression de L en fonction de l'expression G, de chercher à résoudre le problème sans contrainte qui consiste à trouver x_∞ tel que si

$$f(x) = F(L(x)) \tag{223}$$

alors

$$\forall x \in E_{N-1} : f(x_\infty) \leq f(x) \tag{224}$$

On dispose d'une condition nécessaire qui est que le gradient de f au point x_∞ doit être nul et d'une condition suffisante qui est que la matrice hessienne de f au point x_∞ doit être définie positive.

La recopie de ces conditions, compte tenu de (223) passe par le développement de Taylor de

$$F(L(x_\infty + \delta\lambda x')) \tag{225}$$

et donc tout d'abord de celui de (cf. (65))

$$L(x_\infty + \delta\lambda x') = L(x_\infty) + \delta\lambda \nabla L(x_\infty)x' + \frac{1}{2}\delta\lambda^2[\nabla^2 L(x_\infty), x', x'] + \&c \tag{226}$$

où $\nabla L(x_\infty)$ est la matrice jacobienne de L au point x (dimensions $N \times N-1$) (cf. (66)) ; $[\nabla^2 L(x_\infty), x', x']$ est une notation pour désigner le terme quadratique en x' issu du développement de Taylor (cf. (67)) ; &c contient tous les termes de puissance supérieure à 2 en $\delta\lambda$.

Il vient alors

$$\begin{array}{rcl} F(L(x_\infty + \delta\lambda x')) & = & F(L(x_\infty)) \\ & + & \delta\lambda \nabla F(L(x_\infty))\nabla L(x_\infty)x' \\ & + & \frac{1}{2}\delta\lambda^2 \left({}^t x'\, {}^t\nabla L(x_\infty)\nabla^2 F(L(x_\infty))\nabla L(x_\infty)x' + \nabla F(L(x_\infty))[\nabla^2 L(x_\infty), x', x'] \right) \\ & + & \&c \end{array} \tag{227}$$

d'où il sort que

- si x_∞ minimise f (cf. (224)) alors nécessairement

$$\nabla F(L(x_\infty))\nabla L(x_\infty) = 0 \tag{228}$$

- dans les conditions de (228) il est alors suffisant que

$$\begin{array}{l} \forall x' \in E_{N-1} : \text{si } x' \neq 0 \text{ alors} \\ {}^t x'\, {}^t\nabla L(x_\infty)\nabla^2 F(L(x_\infty))\nabla L(x_\infty)x' + \nabla F(L(x_\infty))[\nabla^2 L(x_\infty), x', x'] > 0 \end{array} \tag{229}$$

Ces conditions ne sont pas très intéressantes puisqu'elle font intervenir L et ses dérivées qu'on sait ne pas toujours pouvoir connaître. Aussi il faut maintenant faire disparaître L et ses dérivées.

Pour cela on dispose de (222) qui peut tout aussi bien s'écrire

$$\forall x' \in E_{N-1} \wedge \forall \delta\lambda > 0 : G(L(x_\infty + \delta\lambda x')) = 0 \tag{230}$$

et donc conduire (cf. (66)) à

$$\begin{array}{rcl} G(L(x_\infty + \delta\lambda x')) & = & G(L(x_\infty)) \\ & + & \delta\lambda \nabla G(x_\infty)\nabla L(x_\infty)x' \\ & + & \frac{1}{2}\delta\lambda^2 \left({}^t x'\, {}^t\nabla L(x_\infty)\nabla^2 G(L(x_\infty))\nabla L(x_\infty)x' + \nabla G(x_\infty)[\nabla^2 L(x_\infty), x', x'] \right) \\ & + & \&c \\ & = & 0 \end{array} \tag{231}$$

qui ne peut être réalisée pour tout $\delta\lambda$ positif que si

$$G(L(x_\infty)) = 0 \tag{232}$$

puis

$$\nabla G(x_\infty)\nabla L(x_\infty) = 0 \tag{233}$$

et encore

$$\begin{array}{l} \forall x' \in E_{N-1} : \text{si } x' \neq 0 \text{ alors} \\ {}^t x'\, {}^t\nabla L(x_\infty)\nabla^2 G(L(x_\infty))\nabla L(x_\infty)x' + \nabla G(x_\infty)[\nabla^2 L(x_\infty), x', x'] = 0 \end{array} \tag{234}$$

ainsi d'ailleurs que les conditions contenues dans le &c et qui ne seront pas exploitées ici.

En posant

$$X_\infty = L(x_\infty) \tag{235}$$

on obtient avec (232)

$$G(X_\infty) = 0 \tag{236}$$

puis en remarquant que (228) et (233) signifient que les vecteurs de dimension N ${}^t\nabla F(X_\infty)$ et ${}^t\nabla G(X_\infty)$ sont tous deux orthogonaux aux $N-1$ vecteurs que compose la matrice jacobienne $\nabla L(x_\infty)$ il vient que nécessairement les vecteurs ${}^t\nabla F(X_\infty)$ et ${}^t\nabla G(X_\infty)$ sont alignés, soit

$$\nabla F(X_\infty) = y_\infty \nabla G(X_\infty) \tag{237}$$

où y_∞ est le coefficient de proportionnalité ; il est mis comme multiplicateur de $\nabla G(X_\infty)$, et non pas de $\nabla F(X_\infty)$, parce que, par hypothèse sur G, $\nabla G(X_\infty) \neq 0$, alors qu'on n'a pas cette assurance sur $\nabla F(X_\infty)$.

Ensuite, en tenant compte de (235) et (237), les conditions (229) et (234) peuvent être soustraites l'une de l'autre, après avoir multiplié (234) par y_∞, pour donner

$$\begin{array}{l} \forall x' \in E_{N-1} : \text{si } x' \neq 0 \text{ alors} \\ {}^t x' \, {}^t\nabla L(x_\infty)(\nabla^2 F(X_\infty) - y_\infty \nabla^2 G(X_\infty))\nabla L(x_\infty) x' > 0 \end{array} \tag{238}$$

et les dernières traces de L peuvent enfin être effacées dans (238) en la reformulant comme

$$\begin{array}{l} \forall X' \in E_N \text{ tel que } \nabla G(x_\infty) X' = 0 : \text{si } X' \neq 0 \text{ alors} \\ {}^t X'(\nabla^2 F(X_\infty) - y_\infty \nabla^2 G(X_\infty)) X' > 0 \end{array} \tag{239}$$

Cette dernière condition est alors suffisante pour que X_∞ soit un argument minimisant de (221).

Ainsi une solution locale[7] de (220), (221) est obtenue par les conditions nécessaires

$$\begin{array}{l} G(X_\infty) = 0 \\ \nabla F(X_\infty) = y_\infty \nabla G(X_\infty) \end{array} \tag{240}$$

et il est suffisant que $(\nabla^2 F(X_\infty) - y_\infty \nabla^2 G(X_\infty))$ soit définie positive sur l'orthogonal de ${}^t\nabla G(X_\infty)$ pour que X_∞ soit un argument minimisant du problème (221).

On a raisonné sur une paramétrisation de l'hypersurface définie par (220) ; on a ensuite supprimé toute trace de cette paramétrisation au prix de l'introduction d'un coefficient y_∞, appelé un multiplicateur de Lagrange associé à la contrainte (220) ; et on a obtenu les conditions pour qu'un point X_∞ soit un argument minimisant du problème (221).

7.1.2 Le lagrangien

Les résultats du paragraphe précédent conduisent à des équations dont la résolution permet de calculer X_∞ et y_∞ mais comme on ne sait pas, en général résoudre les équations on ne dispose pas encore de méthode effective de calcul.

Toutefois les éléments pour trouver de telles méthodes sont en place : il y a un multiplicateur de Lagrange qui doit être calculé en plus de X_∞ ; et donc il ne s'agit pas d'opérer dans E_N mais plutôt dans $E_N \times R$.

Cela mène à chercher à fabriquer une fonction dont les arguments sont dans $E_N \times R$ et dont on espère qu'elle puisse mener, par une opération de minimisation au résultat escompté.

Cette fonction peut être

$$\mathcal{L}(X, y) = F(X) - yG(X) \tag{241}$$

qui est une primitive de la deuxième équation de (240) et elle peut être (elle l'est) appelée le lagrangien du problème (221).

L'analyse de ce lagrangien doit maintenant être faite.

Généralités

Si accepte la restriction par rapport aux résultats obtenus au paragraphe précédent que pour toute valeur de y il existe des points $X_\infty(y)$ qui minimisent $\mathcal{L}(X, y)$ par rapport à X, alors ces points satisfont à

$$\nabla F(X_\infty(y)) - y\nabla G(X_\infty(y)) = 0 \tag{242}$$

et donc aussi, en dérivant par rapport à y, à

$$(\nabla^2 F(X_\infty(y)) - y\nabla^2 G(X_\infty(y)))\frac{dX_\infty}{dy}(y) = {}^t\nabla G(X_\infty(y)) \tag{243}$$

Ces points sont bien des arguments minimisant de (241) si

$$\nabla^2 F(X_\infty(y)) - y\nabla^2 G(X_\infty(y)) \tag{244}$$

est une matrice définie positive.

On peut alors s'intéresser à la fonction d'une variable y

$$\mathcal{L}(X_\infty(y), y) = F(X_\infty(y)) - yG(X_\infty(y)) \tag{245}$$

La dérivée par rapport à y de cette fonction est

$$\frac{d}{dy}\mathcal{L}(X_\infty(y), y) = -G(X_\infty(y)) \tag{246}$$

puisque le terme $\nabla F(X_\infty(y)) - y\nabla G(X_\infty(y))$ qui serait multiplié par $dX_\infty/dy(y)$ est nul par définition (cf. (242)) [8].

La dérivée seconde est alors

$$\frac{d^2}{dy^2}\mathcal{L}(X_\infty(y), y) = -\nabla G(X_\infty(y))\frac{dX_\infty}{dy}(y) \tag{247}$$

et on dispose des éléments permettant d'écrire de développement de Taylor

$$\mathcal{L}(X_\infty(y + \delta y), y + \delta y) \quad = \quad \mathcal{L}(X_\infty(y), y) - \delta y G(X_\infty(y)) - \frac{\delta y^2}{2}\nabla G(X_\infty(y))\frac{dX_\infty}{dy}(y) + \&c \tag{248}$$

[7] et non pas globale ; pour les mêmes raisons que dans le cas où il n'y a pas de contraintes.

[8] Cette dérivée est celle qu'on aurait obtenue sans prendre en compte la dépendance de $X(y)$ par rapport à y : pour bien comprendre cela il peut être utile de considérer le problème de minimiser $f(x, g(x))$ par rapport à x où $g(x)$ est l'argument minimisant de $f(x, y)$ par rapport à y (f étant une fonction à deux arguments réels et g une fonction à un argument réel).

Compte tenu de (243), on peut aussi écrire

$$\begin{array}{rcl} \mathcal{L}(X_\infty(y+\delta y), y+\delta y) & = & \mathcal{L}(X_\infty(y), y) \\ & - & \delta y G(X_\infty(y)) \\ & - & \dfrac{\delta y^2}{2} \nabla G(X_\infty(y))(\nabla^2 F(X_\infty(y)) - y\nabla^2 G(X_\infty(y)))^{-1}\, {}^t\nabla G(X_\infty(y)) \\ & + & \&c \end{array} \tag{249}$$

et constater que puisque la matrice (244) a été supposée définie positive pour que $X_\infty(y)$ existe alors son inverse est aussi définie positive et donc une valeur particulière y_∞ qui satisfaisait à

$$G(X(y_\infty)) = 0 \tag{250}$$

serait un argument maximisant (et pas minimisant, voir les signes de (249)) de $\mathcal{L}(X_\infty(y), y)$.

Comme (250) et (242) sont des conditions identiques à (240), que de plus on suppose la définie positivité de (244) (ce qui, dans la suffisance, est une condition plus forte que celle qui était demandée pour que $X_\infty = X_\infty(y_\infty)$ et y_∞ soient les arguments minimisants de (221)), il est ainsi possible de trouver des méthodes de calcul à partir de ces remarques.

Ces méthodes de calcul sont appelée méthodes lagrangiennes et aussi méthodes 'max/min' parce qu'elles s'écrivent

$$\mathcal{L}(X_\infty, y_\infty) = \max_y \min_X \mathcal{L}(X, Y) \tag{251}$$

Leur intérêt pratique par rapport à une simple tentative de résolution des équations non linéaires

$$\begin{array}{rcl} \nabla F(X_\infty) & = & y_\infty \nabla G(X_\infty) \\ G(X_\infty) & = & 0 \end{array} \tag{252}$$

est qu'elles sont efficaces

- quand cette résolution directe est possible ;
- et même quand elle ne l'est plus.

Mais voyons maintenant le détail des opérations à faire.

Méthode lagrangienne générique

Un telle méthode procède en deux temps : on devine un y_0 d'où on déduit un $X_\infty(y_0)$ par un procédé de minimisation par rapport à X, $y = y_0$ étant fixé ; on trouve un nouveau y_1 à partir d'un procédé de maximisation par rapport à y, $X = X_\infty(y_0)$ étant fixé.

Il faut cependant décider des méthodes avec lesquelles seront faites ces maxi/mini-misations élémentaires et la diversité de choix de ces méthodes explique le pluriel de «méthodes lagrangiennes».

Par exemple, si on aime utiliser la méthode de Newton, on peut agir comme suit :

1. on considère le lagrangien (241)
$$\mathcal{L}(X, y) = F(X) - yG(X) \tag{253}$$
2. on devine des valeurs initiales y_0 de y et X_0 de X
3. on minimise (253) par rapport à X, au voisinage de X_0 en maintenant y fixé à y_0, avec la méthode de Newton, soit chercher X_1 solution de
$$(\nabla^2 F(X_0) - y_0\nabla^2 G(X_0))(X_1 - X_0) = -\,{}^t\nabla F(X_0) + y_0\,{}^t\nabla G(X_0) \tag{254}$$
le point X_1 n'est pas encore le minimum $X_\infty(y0)$ demandé, et cette étape peut être éventuellement réitérée plusieurs fois, mais pour ne pas compliquer les notations cette complication ne va pas être introduite ici et on fera comme si $X_1 = X_\infty(y0)$.
4. on calcule ensuite la variation de $\delta y\dot{X}_1$ de X_1 pour une variation δy de y_0, pour cela on peut utiliser (243)
$$(\nabla^2 F(X_1) - y_0\nabla^2 G(X_1))\dot{X}_1 = {}^t\nabla G(X_1) \tag{255}$$
5. on utilise enfin la méthode de Newton pour calculer la valeur y_1 succédant à y_0 sur la fonction (245), soit en utilisant (248)
$$(y_1 - y_0)\nabla G(X_1)\dot{X}_1 + G(X_1) = 0 \tag{256}$$
mais là on ne peut pas légitimement réitérer plusieurs fois l'algorithme de Newton, sauf à calculer les termes d'ordre supérieurs à 2 dans le développement (248).
6. ainsi on dispose du nouveau point X_1 et de y_1 avec lesquels le processus peut être réitéré.

Pour donner un nom propre à l'une des méthodes de lagrangien, on peut aussi utiliser la méthode d'Uzawa[9] pour laquelle :

1. on considère le lagrangien (241)
$$\mathcal{L}(X, y) = F(X) - yG(X) \tag{257}$$
2. on devine des valeurs initiales y_0 de y et X_0 de X
3. on minimise (257) par rapport à X, au voisinage de X_0 en maintenant y fixé à y_0, avec une méthode quelconque poussée jusqu'au bout de manière à obtenir une bonne estimation de $X_\infty(y0)$

[9] Attention pour ce nom : l'algorithme d'Uzawa de [Cia82, p. 226] et celle de [Cul94, p. 130] ne semblent pas coïncider. Il faudrait trouver la source directe pour en avoir le cœur net ; ce sera fait dans une version ultérieure du texte.

4. on calcule alors y_1 par une méthode de remontée (pour ne pas dire descente puisqu'il faut maximiser et non pas minimiser) à pas fixe, soit
$$y_1 = y_0 + \rho G(X_1) \tag{258}$$
où ρ est un coefficient positif (le pas fixe).
5. ainsi on dispose du nouveau point X_1 et de y_1 avec lesquels le processus peut être réitéré.

Cette dernière méthode a l'avantage de ne pas nécessiter l'étape 4 de la première méthode proposée. Mais surtout les deux méthodes, ainsi d'ailleurs que toutes les autres qui s'inspirent des considérations faites dans le paragraphe 'généralité', ont le désavantage de faire l'hypothèse que (244) est définie positive, ce qui peut ne pas être le cas.

Un Exemple :

Voyons maintenant le résultat de ce type d'algorithme sur un exemple simple. Si
$$X = \begin{pmatrix} u \\ v \end{pmatrix} \quad F(X) = \frac{1}{2}(u^2 + v^2) \quad G(X) = u + v - 1 \tag{259}$$
l'étape 1 revient à considérer la fonction
$$\mathcal{L}(X, y) = \frac{1}{2}(u^2 + v^2) - y(u + v - 1) \tag{260}$$
l'étape 2 à choisir u_0, v_0 et y_0 ; l'étape 3 à minimiser (260) par rapport à u et v en maintenant $y = y_0$ par la méthode de Newton, ce qui conduit à
$$\begin{pmatrix} u_1 \\ v_1 \end{pmatrix} = \begin{pmatrix} y \\ y \end{pmatrix} \tag{261}$$
l'étape 4 conduit à
$$\begin{pmatrix} \dot{u}_1 \\ \dot{v}_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{262}$$
et enfin l'étape 5 à
$$y_1 = \frac{1}{2} \tag{263}$$
Si on recommence avec ces nouvelles valeurs en indiçant 1 ce qui était indicé 0 et 2 ce qui était indicé 1, l'étape 3 donne
$$\begin{pmatrix} u_2 \\ v_2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{264}$$
l'étape 4 (262) où les indices 2 sont substitués à 1 et finalement l'étape 5 à
$$y_2 = \frac{1}{2} \tag{265}$$
de laquelle on déduit que le processus a convergé et on peut vérifier que les conditions (240) sont vérifiées pour les valeurs trouvées.

La méthode d'Uzawa sur ce même exemple serait plus lente si le pas ρ était mal choisi, c'est l'inconvénient intrinsèque aux méthodes de descente signalé dans 'minimisation des fonctions deux fois différentiables'.

Un autre exemple :

Il ne faudrait cependant pas déduire de cet exemple que les choses sont toujours aussi faciles : par exemple si on choisit pour l'expression de G '$uv - 1$' plutôt que '$u + v - 1$', on est directement conduit à un des inconvénients de la méthode parce qu'alors (244) ne sera pas nécessairement définie positive.

On trouvera en effet que le hessien de
$$\mathcal{L}(X, y) = \frac{1}{2}(u^2 + v^2) - y(uv - 1) \tag{266}$$
est
$$\begin{pmatrix} 1 & -y \\ -y & 1 \end{pmatrix} \tag{267}$$
et donc que la méthode ne peut fonctionner que si
$$|y| < 1 \tag{268}$$
Il faudra ainsi deviner au départ un y_0 satisfaisant à (268) ; ce qui peut introduire le paragraphe suivant.

7.1.3 La pénalisation

D'autorité, les méthodes suggérées par l'introduction du lagrangien ont l'inconvénient qu'elles supposent que le point de départ des algorithmes correspondants ne soit pas situé trop loin de l'argument minimisant cherché.

Comme il est certain qu'en général on n'a aucune idée de l'endroit où se situe ce point, il faut compléter les méthodes de lagrangien par une méthode capable de trouver des points de départ pour ces méthodes.

On peut pour cela utiliser la méthode de pénalisation dont la première étape est de fabriquer une fonction
$$F(X) + \frac{1}{\epsilon}U(G(X)) \tag{269}$$

où U est une fonction à argument réel, à valeurs toujours positive, dont la borne inférieure, 0, est atteinte seulement quand son argument est nul ; et où ϵ est un paramètre positif destiné à devenir très petit.

Si on appelle $X_\infty(\epsilon)$ la solution du problème de minimisation appliqué à (269), il semble certain que le point X_∞ cherché est

$$X_\infty = \lim_{\epsilon \longrightarrow 0} X_\infty(\epsilon) \tag{270}$$

En effet si ce n'était pas le cas alors on n'aurait pas $G(X_\infty) = 0$, et donc la fonction (269) prendrait une valeur infinie, ce qui est difficilement compatible avec sa vocation à être un minimum.

On ne peut pas pratiquement positionner ϵ à 0 mais on peut le prendre très petit. Aussi le point $X_\infty(\epsilon)$ est peut-être un bon candidat comme point initial d'une méthode de lagrangien.

Il est même possible d'utiliser l'extrapolation de Richardson pour améliorer le calcul de la limite.

On rappelle en effet que si $X_\infty(\epsilon)$ est différentiable par rapport à ϵ alors

$$X_\infty(\epsilon) = X_\infty(0) + \epsilon \dot{X}_\infty(0) + \frac{\epsilon^2}{2} \ddot{X}_\infty(0) + \&c \tag{271}$$

mais aussi

$$X_\infty(k\epsilon) = X_\infty(0) + k\epsilon \dot{X}_\infty(0) + k^2 \frac{\epsilon^2}{2} \ddot{X}_\infty(0) + \&c \tag{272}$$

d'où en soustrayant (272) de (271) préalablement multiplié par k

$$X_\infty(k\epsilon) - kX_\infty(\epsilon) = (1-k)X_\infty(0) - k(1-k)\frac{\epsilon^2}{2} \ddot{X}_\infty(0) + \&c \tag{273}$$

d'où on déduit que

$$\frac{X_\infty(k\epsilon) - kX_\infty(\epsilon)}{1-k} \tag{274}$$

approche $X_\infty(0)$ à un ordre 2 en ϵ près.

Cependant si on peut ainsi obtenir une bonne estimation en X du point de départ, il est également indispensable d'avoir une bonne estimation en y.

Cela peut également être fait, au cas par cas en précisant le choix de la fonction U. Si

$$U(u) = u^2 \tag{275}$$

il est facile de rapprocher (269) de (241) et de se demander si une bonne estimation de y ne serait pas

$$-\frac{G(X_\infty(\epsilon))}{\epsilon} \tag{276}$$

C'est effectivement presque le cas. Pour le voir il suffit d'écrire la condition nécessaire pour que $X_\infty(\epsilon)$ soit un minimum,

$$\nabla F(X_\infty(\epsilon)) + 2\frac{G(X_\infty(\epsilon))}{\epsilon} \nabla G(X_\infty(\epsilon)) = 0 \tag{277}$$

et de comparer cette condition (277) à la condition nécessaire (240) pour trouver que

$$y_\infty = \lim_{\epsilon \longrightarrow 0} -2\frac{G(X_\infty(\epsilon))}{\epsilon} \tag{278}$$

et donc que

$$-2\frac{G(X_\infty(\epsilon))}{\epsilon} \tag{279}$$

et non pas (276) peut servir d'estimation à y. Et on peut même utiliser l'extrapolation de Richardson pour préférer

$$-2\frac{G(X_\infty(k\epsilon)) - kG(X_\infty(\epsilon))}{k(1-k)\epsilon} \tag{280}$$

Ainsi avec la pénalisation on dispose d'un outil rustique qui permet de palier les inconvénients des délicates méthodes de lagrangien : et l'alternance de ces deux méthodes permet d'obtenir des résultats qu'aucune d'entre elles n'eût obtenu seule[10][11].

[10] On notera que cette remarque est identique à celle qui a été fait dans 'Minimisation de fonction deux fois différentiables' à propos de la méthode de la plus grande descente et de la méthode de Newton : la rustique méthode de la plus grande descente permet de palier les inconvénients de la délicate méthode de Newton.

Le fait de rencontrer cette fois encore ce type de résultat porterait à inférer que l'accomplissement d'une chose doit être réalisé non pas de façon unique mais avec deux techniques différentes : l'une peu précise mais robuste, l'autre précise mais peu robuste.

Les exemples de la vie civile ne manquent d'ailleurs pas pour illustrer cette dernière remarque...

[11] En langage un peu technocratique, c'est la synergie : synergie n. f. 1778 ; gr. sunergia "coopération" Didact. 1- Action coordonnée de plusieurs organes, association de plusieurs facteurs qui concourent à une action, à un effet unique. ...Dictionnaire Le petit Robert.

Un exemple :

En reprenant l'exemple de la minimisation de $\frac{1}{2}(u^2+v^2)$ sous la contrainte $uv=1$, on fabrique alors

$$F_\epsilon(u,v)=\frac{1}{2}(u^2+v^2)+\frac{1}{\epsilon}(uv-1)^2 \tag{281}$$

la résolution directe (mais bien sûr en général on est obligé d'utiliser une méthode numérique pour trouver le minimum de F_ϵ) de

$$\nabla F_\epsilon=0 \text{ soit } \left\{\begin{array}{l} u+\frac{2v}{\epsilon}(uv-1)=0 \\ v+\frac{2u}{\epsilon}(uv-1)=0 \end{array}\right. \tag{282}$$

procure les solutions possibles

$$(u_\infty,v_\infty)=\{(0,0),\pm\sqrt{1-\frac{\epsilon}{2}}(1,1)\} \tag{283}$$

Le hessien de F_ϵ est

$$\left(\begin{array}{cc} 1+2\frac{v^2}{\epsilon} & 2\frac{2uv-1}{\epsilon} \\ 2\frac{2uv-1}{\epsilon} & 1+2\frac{u^2}{\epsilon} \end{array}\right) \tag{284}$$

Comme la trace est positive (ϵ est positif), le déterminant doit être positif pour que cette matrice 2×2 soit définie positive ; ce déterminant est

$$\frac{-12u^2v^2+16uv-4+\epsilon^2+2\epsilon(u^2+v^2)}{\epsilon^2} \tag{285}$$

Au point $(0,0)$ ce déterminant vaut donc $1-4/\epsilon^2$ et il est négatif si ϵ est petit : ce point n'est pas un argument minimisant de F_ϵ ;

Aux points $\sqrt{1-\frac{\epsilon}{2}}(1,1)$ et $-\sqrt{1-\frac{\epsilon}{2}}(1,1)$ ce déterminant vaut $8/\epsilon-4$; il est positif pour ϵ suffisamment petit et donc on ne retient comme solutions que

$$(u_\infty,v_\infty)=\pm\sqrt{1-\frac{\epsilon}{2}}(1,1) \tag{286}$$

On peut vérifier que la limite $\epsilon=0$ donne bien les solutions du problème de départ ; mais ce n'est pas le propos ici.

Plutôt puisque $\sqrt{1-\frac{\epsilon}{2}}(1,1)$, par exemple, est candidat à être un point de départ pour une méthode lagrangienne, on peut chercher une valeur de départ pour le multiplicateur de Lagrange avec (276). C'est $1/2$ qui est bien inférieur à 1, comme demandé pour que le hessien de (266) soit défini positif.

Si on veut, dans cet exemple, améliorer le calcul du point de départ par l'extrapolation de Richardson on obtient que

$$(u_\infty,v_\infty)=\pm\frac{\sqrt{1-\frac{k\epsilon}{2}}-k\sqrt{1-\frac{\epsilon}{2}}}{1-k}(1,1) \tag{287}$$

(qui est bien en ϵ^2) et aucune amélioration sur le multiplicateur de Lagrange puisqu'il vaut la valeur constante $1/2$ qui est d'ailleurs la valeur exacte.

7.2 La contrainte inégalité

Il s'agit de résoudre le problème de trouver X_∞ tel que, une fonction G, soumises aux conditions

$$\exists X \text{tel que } G(X)=0 \text{ et de plus si } G(X)=0 \text{ alors } \nabla G(X)\neq 0 \tag{288}$$

étant donnée, on a

$$\begin{array}{l} G(X_\infty)\leq 0 \\ \forall X \text{ tels que } G(X)\leq 0 : f(X_\infty)\leq f(X) \end{array} \tag{289}$$

Et, pour éviter de permanentes précautions oratoires, on suppose dans ce paragraphe que les F et G sont tels qu'il existe une solution unique au problème.

7.2.1 Problématique

Il y a deux cas possibles :

1. le problème sans contrainte d'inégalité admet une solution et cette solution X_∞ satisfait à $G(X_\infty)\leq 0$; il est alors inutile de résoudre le problème avec contrainte d'inégalité ;
2. le problème sans contrainte n'admet pas de solution ou il en admet une mais celle-ci, X_∞, ne satisfait pas à $G(X_\infty)\leq 0$; le problème avec contrainte d'inégalité devient alors très proche du problème de contrainte égalité (221).

Le premier cas ne demande pas de développement supplémentaires, le second peut être traité.

En ne s'intéressant d'abord qu'au premier ordre on suppose avoir trouvé X_∞ et y_∞ tels que

$$\begin{array}{l} G(X_\infty)=0 \\ \nabla F(X_\infty)=y_\infty\nabla G(X_\infty) \end{array} \tag{290}$$

Trois cas sont possibles :

1. Si
$$y_\infty > 0 \tag{291}$$
alors compte tenu qu'un déplacement δX, à partir de X_∞ dans la direction et le sens de $\nabla G(X_\infty)$
$$\text{pour } \delta\lambda > 0 : \delta X = \delta\lambda\, {}^t\nabla G(X_\infty) \tag{292}$$
conduit à augmenter G et donc que
$$G(X_\infty + \delta\lambda X) = \delta\lambda \nabla G(X_\infty)\, {}^t\nabla G(X_\infty) > 0 \tag{293}$$
pour $\delta\lambda$ suffisamment petit, on voit qu'il est possible de diminuer F en se déplaçant dans cette même direction et dans le sens opposé.
Le point trouvé ne peut pas correspondre au minimum cherché ; il correspond peut-être à un maximum.
2. Si
$$y_\infty < 0 \tag{294}$$
au contraire, et pour les mêmes raisons, le point X_∞ correspond au minimum cherché.
3. Si
$$y_\infty = 0 \tag{295}$$
alors c'est que le problème de recherche de minimum sous contrainte (égalité comme inégalité) n'avait à être traité que comme un problème de recherche de minimum sans contrainte du tout.

On voit donc comme le cas de la contrainte inégalité est finalement plus simple que celui de l'égalité ; il ne suppose pas de développement au second ordre pour conclure.

Toutefois il conduit à une situation nouvelle pour lesquelles il s'agit de disposer d'algorithmes capables de séparer les différents cas et de traiter chacun d'une façon appropriée.

Un exemple :

Ces cas peuvent être illustrés par le simple problème de minimiser
$$X = (u)\,;\; F(X) = u^2/2\,;\; G(X) = (u-a)(u-b)/2 \text{ avec } a < b \tag{296}$$
Le cas pour lequel la contrainte d'inégalité est superflue correspond à
$$a < 0 < b \tag{297}$$
Si
$$0 < a < b \tag{298}$$
la solution de (289) est
$$u_\infty = a \tag{299}$$
et on a
$$y_\infty = \frac{-2a}{b-a} < 0 \tag{300}$$
Si on avait cru que la solution étaient
$$u_\infty = a \tag{301}$$
on aurait alors trouvé
$$y_\infty = \frac{2b}{b-a} > 0 \tag{302}$$
et on aurait compris que b ne pouvait pas être solution.
Si maintenant
$$a = 0 < b \tag{303}$$
on obtient
$$u_\infty = 0 \text{ et } y_\infty = 0 \tag{304}$$
ce qui correspond au troisième cas indiqué.

7.2.2 La pénalisation

Dans le cas des inégalités, il n'est pas impossible que la solution cherchée à (289) ne nécessite pas de prendre en compte l'inégalité parce que celle-ci est avérée pour cette solution.

Aussi il est nécessaire de modifier la méthode de pénalisation introduite dans le cas de la contrainte égalité pour tenir compte de ce fait.

Si on introduit la fonction de pénalisation
$$F(X) + U(G(X)) \tag{305}$$
où
$$U(u) = \frac{1}{\epsilon} \begin{cases} 0 & \text{si} \quad u \le 0 \\ U_+(u) > 0 & \text{sinon} \end{cases} \tag{306}$$
et qu'on cherche le minimum de (305) alors on utilise une méthode pénalisation extérieure ; le mot 'extérieur' indiquant que U a vocation à devenir infinie quand ϵ tend vers 0 seulement à l'extérieur du domaine de E_N dans lequel on veut que soit le minimum cherché ; et donc cette pénalisation sera neutre (ne jouera aucun rôle) s'il était de la nature du problème que ce minimum y fût.

Si on introduit la fonction de pénalisation

$$U(u) = \begin{cases} U_-(u) \approx 0 & \text{si} \quad u < 0 \\ +\infty & \text{si} \quad u = 0 \\ U_+(u) \approx +\infty & \text{sinon} \end{cases} \tag{307}$$

où $U_+(u)$ prend des valeurs proches de 0 quand $u < 0$ et où U_+ n'a pas nécessairement à être défini. Si on cherche le minimum de (305) en prenant garde que le point initial soit tel que $G(X_0) < 0$ alors les points explorés lors du processus de recherche resteront toujours dans ce domaine puisque s'ils approchaient de la frontière $G(X) = 0$ ils en seraient rejeté par le fait que U y est infinie. Cette pénalisation porte le nom de pénalisation intérieure.

Les choix possibles de fonctions de pénalisation intérieur dépendent fortement du problème traité, par contre pour la pénalisation intérieur on peut prendre

$$U(u) = \frac{1}{\epsilon} W(u) \tag{308}$$

où W est la fonction de Whitney définie par (43). On a alors une fonction de pénalisation indéfiniment différentiable.

Plus simplement la fonction (306) pour laquelle

$$U_+(u) = u^2 \tag{309}$$

n'est que deux fois différentiable mais cela suffit largement.

Un exemple :

Si on veut minimiser $1/2(u^2 + v^2)$ sous la contrainte $(1 - uv) \leq 0$ on peut introduire

$$F_\epsilon(u, v) = \frac{1}{2}(u^2 + v^2) + \frac{1}{\epsilon} W_2(1 - uv) \tag{310}$$

où W_2 est une version affaiblie (différentiable seulement jusqu'à l'ordre 2) de la fonction de Whitney

$$W_2(t) = \begin{cases} t^2 & \text{si} \quad t > 0 \\ 0 & \text{sinon} \end{cases} \tag{311}$$

Si on prend un point initial (u_0, v_0) tel que $u_0 v_0 > 1$ la fonction F_ϵ devient $1/2(u^2 + v^2)$ et avec la méthode de Newton on trouve un point (u_1, v_1) en $(0, 0)$; ensuite le problème devient l'analogue du problème décrit en exemple dans le paragraphe sur la méthode de pénalisation sous contrainte d'égalité à laquelle il faut se reporter pour conclure cet exemple (puisque le point trouvé sans contrainte n'est pas admissible pour la contrainte c'est qu'il fallait prendre la contrainte comme un contrainte d'égalité).

On ajoutera cependant que comme finalement le multiplicateur de Lagrange trouvé est négatif (c'est -1/2) le point trouvé est bien un argument minimisant du problème.

7.2.3 Proposition d'algorithme

Avec les paragraphes précédents on dispose de la base nécessaire pour construire des algorithmes effectifs de calcul. Par exemple on peut :

1. chercher un point X_0 par la méthode de pénalisation intérieur
2. dans le cas où ce point satisfaisait la condition $G(X_0) \leq 0$ alors le point X_∞ serait trouvé si la matrice $\nabla^2 F(X_0)$ est définie positive.
3. sinon c'est que le point cherché est tel que $G(X_\infty = 0$ auquel cas il suffit de résoudre le problème avec contrainte d'égalité correspondant ; et avec une méthode de lagrangien puisque on doit être assez proche de X_∞ pour que l'une de ces méthodes soit efficace.

7.3 Exercices

Pour récompenser l'éventuel lecteur d'avoir bien voulu arriver à ce point de l'exposé, on propose maintenant de le détendre par l'exposé de quelques problèmes de minimisation sous contrainte historiques et/ou d'usage pratiques.

7.3.1 Le problème de Kepler

L'histoire est rapporté par Alekseev [ATF87, p. 6]

> *En 1615, Kepler publia son livre «Nouvelle stéréométrie des tonneaux de vin.»Kepler commence son livre par les mots suivants : "L'année où je me suis marié, les vendanges donnèrent une bonne récolte et le vin était bon marché ; étant un bon maître de maison, il me fallait faire des réserves de vin. J'en achetais plusieurs tonneaux. Après un certain temps, le commerçant qui me les vendit vint pour mesurer la capacité des tonneaux et, sans aucune forme de calcul, nommait immédiatement le contenu en vin de ce tonneau."*
>
> *Kepler fut très surpris. Il lui semblait étrange qu'il fût possible d'effectuer, grâce à une seule mesure, le calcul de tonneaux de capacité différentes....*

Effectivement : si un tonneau est assimilé à un tronçon de cylindre à section circulaire de hauteur h et de rayon r, il semblerait qu'il faut deux mesures : l'une pour h, l'autre pour r afin de trouver que le volume est $\pi r^2 h$. Le marchand n'en fait qu'une : en gros il plonge une règle en diagonale dans le tonneaux et donc ne peut déduire que $\sqrt{h^2 + 4r^2}$.

Alors comment peut-il déduire cette valeur $\pi r^2 h$?

La réponse est la suivante : les tonneaux représentent un certain investissement ; si on néglige le coût de main d'œuvre de leur fabrication il reste la matière avec laquelle ils sont fabriqués dont le coût est proportionnel à la surface du tonneau soit $2\pi r^2 + 2\pi rh$ en comptant les couvercles ; d'autre part la fonction d'un tonneau est de contenir et donc son efficacité est proportionnelle au volume du tonneau soit $\pi r^2 h$.

Si on ajoute à ces données que l'art de fabriquer des tonneaux est pluri-séculaire, même à l'époque de Kepler ; on comprend alors que tous les fabriquants de tonneaux qui ne fournissaient pas des tonneaux d'efficacité maximum à coût donné ont été éliminés par leurs concurrents[12].

Donc tous les tonneaux fabriqués à l'époque de Kepler étaient optimum au sens précédent ; c'est à dire que r et h étaient les arguments maximisant $\pi r^2 h$ sous la contrainte $2\pi r^2 + 2\pi rh = s_0$; et la solution de ce problème, laissée en exercice, conduit à une relation entre r et h.

Le reste se déduit facilement : avec une mesure le marchand trouvait l'autre relation nécessaire pour déterminer r et h puis en déduisait le volume $\pi r^2 h$[13].

Kepler passa une dizaine d'année à déduire cela.

Et il déduisit d'ailleurs quelque chose de plus : au voisinage d'un optimum la valeur de la fonction optimisée varie très peu (le gradient de la forme paramétrisée du problème est nul) ; donc les écarts par rapport à cet optimum peuvent être comptés pour rien.

7.3.2 Le problème de Didon

C'est un autre problème célèbre, raconté aussi par Alekseev, mais dont il est possible de trouver la source directe :

> ...
> *His commota fugam Dido sociosque parabat :*
> *conveniunt, quibus aut odium crudele tyranni*
> *aut metus acer erat ; navis, quae forte paratae,*
> *corripiunt, onerantque auro : portantur avari*
> *Pygmalionis opes pelago ; dux femina facti.*
> *Devenere locos, ubi nunc ingentia cernis*
> *moenia surgentemque novae Karthaginis arcem,*
> *mercatique solum, facti de nomine Byrsam,*
> *taurino quantum possent circumdare tergo.*
> ...
> http ://www.promo.net/pg (chercher 'Vergil' ou 'Aeneid')
> On peut trouver une version Française : Virgile, "L'énéide", Garnier-Flammarion, 1965 (première édition 19 av. J.C.), p. 40.

La question est alors[14] sur ce : sachant que le roi Iarbas veut bien donner à Didon et à ses compagnons un territoire que peut embrasser la peau d'un taureau ; que Didon se dit qu'elle peut découper la peau du taureau en une fine lanière qui servira à délimiter ce territoire ; comment pourra-t-elle procéder pour disposer du plus grand territoire possible et fonder ainsi Carthage (Byrsa) ?

C'est le problème isopérimétrique dont la solution est un cercle si il n'y a pas de contraintes supplémentaires.

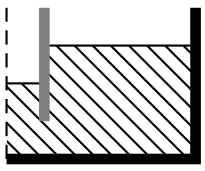
Si maintenant on donne une contrainte que le territoire doit posséder un accès à la mer ; que la côte à une forme déterminée ; et qu'au lieu de chercher une courbe différentiable on cherche au contraire une courbe polygonale ; alors le problème peut devenir plus compliqué.

Par exemple la terre est plate et on y repère la position par des coordonnées cartésiennes (x, y) ; la mer occupe le domaine $y < 0 \quad \wedge \quad y - x < 0$; la courbe polygonale comporte deux tronçons contigus ; le premier tronçon part du point $(1, 0)$ et arrive au point X, Y qui est le point de départ du second tronçon, lequel arrive au point $(-t^2, -t^2)$; comment trouver (X, Y, t) pour que la somme des longueurs des tronçons $\sqrt{(X-1)^1 + Y^2} + \sqrt{(X+t^2)^2 + (Y+t^2)} = L$, une longueur fixée et que la surface délimitée par les deux tronçons complétés par ces deux autres qui constituent la côte soit maximale ?

7.3.3 Capillarité

> *La physique utilisée est due à Laplace, mais c'est Thomson (Lord Kelvin) qui a l'appliquée au phénomène de capilarité dans les arbres si on croit Maxwell [?, pp. 359–375].*
>
> *Noter que la lecture du traité de la chaleur de Maxwell (le même que celui des équations de Maxwell) est très intéressante : notamment pour les pages citées on y apprend comment éliminer une tache de graisse sur un drap, et aussi on a une interprétation d'un passage obscur des Proverbes sur la force du vin...*

Un vase en verre axisymétrique (en traits noirs épais) est rempli d'un liquide (zone hachurée). On place dans ce vase un tube axisymétrique en verre (en traits gris épais) de manière que les axes de ces deux objets coïncident.



[12] Ce pourrait être l'occasion d'introduire des méthodes d'optimisation autres que les méthodes déterministes dont on s'occupe ici : les méthodes génétiques. Mais ce ne sera pas fait

[13] Évidemment, le marchand utilisait une règle graduée empiriquement.

[14] Inutile de chercher la question dans l'énéide, le seul passage consacré au problème de Didon est celui-ci ; et Virgile passe beaucoup plus de temps à décrire des scènes de conflits qu'à analyser ce problème.
Au contraire des Grecs, les Romains n'ont jamais eu le sens des valeurs.

On observe que le niveaux du liquide dans le tube n'est pas le même qu'à l'extérieur.

C'est dû au phénomène de capillarité.

On introduit les valeurs des trois surfaces et les trois tensions superficielles

S_{lv}	σ_{lv}	entre le liquide et le verre
S_{va}	σ_{va}	entre le verre et l'air (à extérieur au liquide)
S_{al}	σ_{al}	entre l'air et le liquide

telles que les énergies de surface soient

$$\mathcal{E}_s = \sigma_{lv} S_{lv} + \sigma_{va} S_{va} + \sigma_{al} S_{al}$$

Puis on introduit la masse volumique ρ du liquide (celle de l'air est négligée) à partir de laquelle on peut trouver l'énergie potentielle $\mathcal{E}_p$

Sachant de plus que le volume du liquide doit rester constant, il faut calculer la position des deux niveaux en fonction de σ_{lv}, σ_{va}, σ_{al}, ρ, le rayon intérieur du vase, les rayons intérieurs et extérieur du tube et le volume de liquide.

8 Plusieurs contraintes

La situation du paragraphe précédent était idéale : une seule hypersurface était introduite ; on supposait qu'elle n'était pas réduite à l'ensemble vide ; et bien qu'on ne l'ait pas signalé il ne s'agissait que de trouver un système de coordonnées curvilignes adapté pour analyser les problèmes de contraintes.

Le cas de plusieurs contraintes amène une certaine complexité liée à la diversité des situations possibles :

- les contraintes peuvent être hétérogènes (égalités et inégalité à la fois) ;
- deux contraintes peuvent impossible a réaliser simultanément) ;
- deux contraintes peuvent se réduire à une seule (par exemple $v^2 - u = 0$ $u \leq 0$ s'il s'agit de minimiser $u^2 + v^2 + w^2$ sous ces contraintes).

Il ne semble pas exister de taxonomie satisfaisante actuellement pour gérer cette complexité.

Aussi va-t-on procéder de façon empirique : c'est à dire suivre une stratégie consistant à ignorer *a priori* les problèmes qui peuvent se poser jusqu'à ce qu'ils se posent effectivement.

8.1 Plusieurs contraintes égalité

Dans le cas de plusieurs contraintes égalité la forme de l'expression du problème (221)

$$\begin{array}{l} G(X_\infty) = 0 \\ \forall X \text{ tels que } G(X) = 0 : f(X_\infty) \leq f(X) \end{array} \tag{312}$$

n'a pas à être modifiée ; il suffit d'ajouter que G est une fonction de E_N dans E_P si il y a P égalités, soit

$$G(X) = \begin{pmatrix} G_1(X) \\ \vdots \\ G_P(X) \end{pmatrix} \tag{313}$$

où les G_p sont des fonctions à valeur dans R.

Par contre la condition (220) sur G doit être remaniée. On ne s'intéresse encore qu'aux hypersurfaces, aussi on veut que

$$\forall p = 1 \ldots : \exists X \text{tel que } G_p(X) = 0 \text{ et de plus si } G_p(x) = 0 \text{ alors } \nabla G_p(X) \neq 0 \tag{314}$$

qui peut s'écrire comme (220)

$$\exists X \text{tel que } G(X) = 0 \text{ et de plus si } G(X) = 0 \text{ alors } \nabla G(X) \neq 0 \tag{315}$$

où $\nabla G(X)$ est la matrice dont les lignes sont formées des composantes des gradients des $G_p(X)$ et le signe $=$ s'applique aux lignes entières et non pas à chacune des composantes de la matrice.

Moyennant ces modifications l'intégralité du paragraphe 'La contrainte égalité' peut être ré-écrite avec de légères modifications.

8.1.1 Les multiplicateurs de Lagrange

Les conditions données sur G permettent d'introduire une paramétrisation L telle que

$$\forall x \in E_{N-P} : G(L(x)) = 0 \tag{316}$$

et, même si on sait que dans de nombreux cas il ne sera pas possible de trouver effectivement un expression de L en fonction de l'expression G, de chercher à résoudre le problème sans contrainte de trouver x_∞ tel que si

$$f(x) = F(L(x)) \tag{317}$$

alors

$$\forall x \in E_{N-P} : f(x_\infty) \leq f(x) \tag{318}$$

On dispose d'une condition nécessaire qui est que le gradient de g au point x_∞ doit être nul et d'une condition suffisante qui est que la matrice hessienne de g au au point x_∞ doit être définie positive.

La recopie de ces conditions, compte tenu de (317) passe par le développement de Taylor de

$$F(L(x + \delta\lambda x')) \tag{319}$$

et donc tout d'abord de celui de

$$L(x+\delta\lambda x') = L(x) + \delta\lambda \nabla L(x) x' + \frac{1}{2}\delta\lambda^2 [\nabla^2 L(x), x', x'] + \&c \tag{320}$$

où $\nabla L(x)$ est la matrice jacobienne de L au point x (dimensions $N \times N-1$) ; $[\nabla^2 L(x), x', x']$ est une notation pour désigner le terme quadratique en x' issue du développement de Taylor ; &c contient tous les termes de puissance supérieure à 2 en $\delta\lambda$.

Il vient alors

$$\begin{array}{rcl} F(L(x+\delta\lambda x')) & = & F(L(x)) \\ & + & \delta\lambda \nabla F(L(x)) \nabla L(x) x' \\ & + & \frac{1}{2}\delta\lambda^2 \left({}^t x' \, {}^t\nabla L(x) \nabla^2 F(L(x)) \nabla L(x) x' + \nabla F(L(x))[\nabla^2 L(x), x', x'] \right) \\ & + & \&c \end{array} \tag{321}$$

d'où on tire que

– si x_∞ minimise f ((318)) alors nécessairement

$$\nabla F(L(x_\infty)) \nabla L(x_\infty) = 0 \tag{322}$$

– dans les conditions de (322) il est alors suffisant que

$$\begin{array}{l} \forall x' \in E_{N-1} : \text{si } x' \neq 0 \text{ alors} \\ {}^t x' \, {}^t\nabla L(x_\infty) \nabla^2 F(L(x_\infty)) \nabla L(x_\infty) x' + \nabla F(L(x_\infty))[\nabla^2 L(x_\infty), x', x'] > 0 \end{array} \tag{323}$$

Ces conditions ne sont pas très intéressantes puisqu'elle font intervenir L et ses dérivées qu'on sait ne pas pouvoir connaître toujours. Aussi il faut maintenant faire disparaître L et ses dérivées.

Pour cela on dispose de (88) soit

$$\begin{array}{rcl} G(L(x+\delta\lambda x')) & = & G(L(x)) \\ & + & \delta\lambda \nabla G(L(x)) \nabla L(x) x' \\ & + & \frac{1}{2}\delta\lambda^2 \left([\nabla^2 G(L(x)), \nabla L(x) x', \nabla L(x) x'] + \nabla G(L(x))[\nabla^2 L(x), x', x'] \right) \\ & + & \&c \\ & = & 0 \end{array} \tag{324}$$

de laquelle on déduit d'abord que nécessairement (cf. (89))

$$\nabla G(L(x_\infty)) \nabla L(x_\infty) = 0 \tag{325}$$

qui, rappelons le, traduit l'orthogonalité de chacun des $N-P$ vecteurs colonnes $\partial_p L(x_\infty)$ (de dimensions N) avec chacun des P vecteurs lignes $\nabla G_p(L(x_\infty))$ (de dimensions N).

Le rapprochement de (322) et (325) met en évidence que le vecteur de dimension N ${}^t\nabla F(L(x_\infty))$ et les P vecteurs de dimension N ${}^t\nabla G(L(x_\infty))$ sont orthogonaux aux $N-P$ vecteurs que compose la matrice jacobienne $\nabla L(x_\infty)$; il existe donc un vecteur y_∞

$$y = \begin{pmatrix} y_\infty^1 \\ \vdots \\ y_\infty^P \end{pmatrix} \tag{326}$$

de dimension P tel que

$${}^t\nabla F(L(x)) = {}^t\nabla G(L(x)) y_\infty \left(= \sum_{p=1}^{P} y_\infty^p \, {}^t\nabla G_p(L(x)) \right) \tag{327}$$

Soit en posant

$$X_\infty = L(x_\infty) \tag{328}$$

on obtient avec (324)

$$G(X_\infty) = 0 \tag{329}$$

puis

$${}^t\nabla F(X_\infty) \quad = \quad {}^t\nabla G_p(X_\infty) y_\infty \tag{330}$$

Il y a cependant quelque chose à ajouter. Si les P vecteurs de dimension N $\nabla G_p(X_\infty)$ ne forment pas une famille libre (si au moins l'un d'entre eux est linéairement dépendant des autres) alors on a introduit trop de coefficients y_p. Dans ce cas il faut reformuler (330) en introduisant une famille libre construite sur cet ensemble de vecteurs. On suppose ici que cette situation n'arrive pas.

On a déjà éliminé la paramétrisation L pour le terme de premier ordre en $\delta\lambda$, il reste à faire de même pour le terme de second ordre.

la multiplication à gauche par ${}^t y_\infty$ du vecteur nul de dimension P (cf (90))

$$[\nabla^2 G(L(x_\infty)), \nabla L(x_\infty) x', \nabla L(x_\infty) x'] + \nabla G(L(x_\infty))[\nabla^2 L(x), x', x'] = 0 \tag{331}$$

est le scalaire nul

$${}^t y_\infty \left([\nabla^2 G(L(x_\infty)), \nabla L(x_\infty) x', \nabla L(x_\infty) x'] + \nabla G(L(x_\infty))[\nabla^2 L(x_\infty), x', x'] \right) \quad = \quad 0 \tag{332}$$

soit

$${}^t x' \, {}^t\nabla L(x_\infty) ({}^t y_\infty \nabla^2 G(L(x_\infty))), \nabla L(x_\infty) x' + ({}^t y_\infty \nabla G(L(x_\infty)))[\nabla^2 L(x_\infty), x', x'] \quad = \quad 0 \tag{333}$$

et donc compte tenu de (330)

$$ {}^t x' \, {}^t\nabla L(x_\infty)({}^t y_\infty \nabla^2 G(L(x_\infty))), \nabla L(x_\infty)x' + (\nabla F(L(x_\infty)))[\nabla^2 L(x_\infty), x', x'] \quad = \quad 0 \tag{334} $$

Il reste à soustraire (334) au terme de second ordre de (321) pour trouver que ce dernier est

$$ {}^t x' \, {}^t\nabla L(x_\infty)(\nabla^2 F(L(x_\infty)) - {}^t y_\infty \nabla^2 G(L(x_\infty)))\nabla L(x_\infty)x' \tag{335} $$

et après avoir utilisé (328) les dernières traces de L sont alors effacées en formulant la condition suffisante pour que X_∞ soit un argument minimisant F sous la contrainte que $G(X) = 0$ comme

$$ \begin{array}{l} \forall X' \in E_N \text{ tel que } \nabla G(L(x_\infty))X' = 0 : si X' \neq 0 \text{ alors} \\ {}^t X'(\nabla^2 F(X_\infty) - y_\infty \nabla^2 G(X_\infty))X' > 0 \end{array} \tag{336} $$

Et donc on obtient qu'une solution locale[15] de (312), (313) est obtenue par les conditions nécessaires

$$ \begin{array}{l} G(X_\infty) = 0 \\ {}^t\nabla F(X_\infty) = {}^t\nabla G(X_\infty) y_\infty \end{array} \tag{337} $$

qui, rappelons le, signifie

$$ \begin{array}{ll} \forall p = 1 \ldots P & G_p(X_\infty) = 0 \\ & \nabla F(X_\infty) = \sum_{p=1}^P y_\infty^p \nabla G_p(X_\infty) \end{array} \tag{338} $$

et qu'il est suffisant que $\nabla^2 F(L(x_\infty)) - y_\infty \nabla^2 G(L(x_\infty))$ c'est à dire

$$ \nabla^2 F(L(x_\infty)) - \sum_{p=1}^P y_\infty^p \nabla^2 G_p(x_\infty) $$

soit définie positive sur l'orthogonal de ${}^t\nabla G(X_\infty)$, c'est à dire le sous espace vectoriel formé généré par les vecteurs

$$ {}^t\nabla G_1(X_\infty), \ldots, {}^t\nabla G_P(X_\infty) $$

qui impérativement doit être de dimension P (voir la remarque faite supra), que X_∞ soit un minimum.

On a raisonné sur une paramétrisation de l'hypersurface définie par (312) ; on a ensuite supprimé toute trace de cette paramétrisation au prix de l'introduction d'un coefficient y_∞, appelé un multiplicateur de Lagrange associé à la contrainte (312) ; et on a obtenu les conditions pour qu'un point X_∞ soit un argument minimisant du problème (313).

8.1.2 Le lagrangien

Il faut maintenant encore recopier, mutatis mutandi, ce qui a déjà été dit à ce propos dans cas le d'une seule contrainte à propos du lagrangien.

Allons seulement un peu plus vite ; il s'agit de trouver une décomposition algorithmique au problème de 'min/max' (251).

On peut alors procéder de la façon suivante

1. on considère le lagrangien
$$ \mathcal{L}(X, y) = F(X) - {}^t y G(X) \tag{339} $$
où y est un vecteur de dimension P ;
2. on devine des valeurs initiales y_0 de y et X_0 de X
3. on minimise (339) par rapport à X, au voisinage de X_0 en maintenant y fixé à y_0, avec la méthode de Newton, soit chercher X_1 solution de
$$ (\nabla^2 F(X_0) - \sum_{p=1}^P y_0^p \nabla^2 G_p(X_0))(X_1 - X_0) = - {}^t\nabla F(X_0) + {}^t y_0 \, {}^t\nabla G(X_0) \tag{340} $$
le point X_1 n'est pas encore le minimum $X_\infty(y0)$ demandé, et cette étape peut être éventuellement réitérée plusieurs fois, mais pour ne pas compliquer les notations cette complication ne va pas être introduite ici et on fera comme si $X_1 = X_\infty(y0)$.
4. on calcule ensuite la variation de $\dot{X}_1 \delta y$ de X_1 pour une variation δy de y_0, pour cela on peut partir de la condition nécessaire pour que X_1 soit un argument minimisant de $\mathcal{L}(X, y_0)$ soit
$$ \nabla F(X_1) = {}^t y_0 \nabla G(X_1) \tag{341} $$
et écrire qu'alors si y_0 varie de δy alors la variation à l'ordre 1 $\dot{X}_1 \delta y$ de X_1 sera solution de
$$ (\nabla^2 F(X_1) - \sum_{p=1}^P y_0^p \nabla^2 G_p(X_1))\dot{X}_1 \delta y = {}^t\nabla G(X_1)\delta y \tag{342} $$
on obtient un moyen de calcul de $\dot{X}_1$ qui est une matrice de dimension $N \times P$ après avoir éliminé δy de (342) (qui doit être valable pour tout δy)

[15] et non pas globale ; pour les mêmes raisons que dans le cas où il n'y a pas de contraintes.

5. on utilise enfin la méthode de Newton pour calculer la valeur y_1 succédant à y_0 sur la fonction $\mathcal{L}(X_\infty(y), y)$ (la même que (245)) pour le cas où y est un vecteur), soit

$$\nabla G(X_1)\dot{X}_1(y_1 - y_0) + G(X_1) = 0 \tag{343}$$

mais là on ne peut pas légitimement réitérer plusieurs fois l'algorithme de Newton, sauf à calculer les termes d'ordre supérieurs à 2 dans le développement de Taylor des fonctions G.

6. ainsi on dispose du nouveau point X_1 et de y_1 avec lesquels le processus peut être réitéré.

De la même façon encore que dans le cas où il n'y avait qu'une seule contrainte on peut également réécrire la méthode d'Uzawa, ce qui devrait être direct.

Il est maintenant intéressant de traiter un exemple.

Un exemple :

Si on donne la fonction

$$F(u, v, w) = \alpha u + \beta v + \gamma w \tag{344}$$

et la fonction

$$G(u, v, w) = \begin{pmatrix} G(u, v, w) \\ H(u, v, w) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} u^2 + v^2 + w^2 - 1 \\ u^2 + v^2 - r^2 \end{pmatrix} \tag{345}$$

On peut introduire le lagrangien

$$\mathcal{L}(u, v, w, y, z) = F(u, v, w) - yG(u, v, w) - zH(u, v, w) \tag{346}$$

donner des valeurs initiales u_0, v_0, w_0, y_0, z_0 à u, v, w, y, z et d'abord minimiser (346) par rapport à u, v, w en maintenant $y = y_0$ et $z = z_0$ fixes.

Ça n'est possible que si

$$y_0 < 0 \quad \text{et} \quad y_0 + z_0 < 0 \tag{347}$$

ce qu'on suppose et on trouve alors que les valeurs de u, v, w qui minimisent le lagrangien sont

$$\begin{pmatrix} u_1 \\ v_1 \\ w_1 \end{pmatrix} = \begin{pmatrix} \alpha/(y_0 + z_0) \\ \beta/(y_0 + z_0) \\ \gamma/y_0 \end{pmatrix} \tag{348}$$

ce qui est cohérent avec (347).

En entrant ces valeurs directement dans (346) on obtient

$$\begin{aligned} 2\mathcal{L}(u_1, v_1, w_1, y_0, z_0) \quad = \quad & y_0 + r^2 z_0 + \frac{(\alpha^2 + \beta^2 + \gamma^2)y_0 + \gamma^2 z_0}{y_0(y_0 + z_0)} \\ & \left((1 - r^2)y_0 + \frac{\gamma^2}{y_0}\right) + \left(r^2(y_0 + z_0) + \frac{\alpha^2 + \beta^2}{y_0 + z_0}\right) \end{aligned} \tag{349}$$

qui d'abord, comme on peut le constater, n'est plus une fonction linéaire en y_0 et z_0 et donc peut être un objet susceptible de subir une maximisation par rapport à ces variables ; et ensuite peut effectivement être maximisée en considérant l'expression de la deuxième ligne qui est la somme de deux fonctions indépendantes : la première d'argument y_0, la seconde d'argument $y_0 + z_0$.

Si $r^2 > 1$ on voit qu'il n'y aura pas de maximum pour la première fonction (c'est la somme de deux fonctions décroissantes) et c'est normal puisque dans ce cas il n'y a pas de solutions à $G(u, v, w) = 0$.

Dans le cas contraire l'argument maximisant cette première fonction est $y_1 = -\sqrt{\gamma^2/(1 - r^2)}$ (le signe moins à cause de (347) mais aussi, comme on peut le voir en traçant le graphe de cette fonction parce que le signe + correspondrait à un argument minimisant) ; de même, mutatis mutandi, l'argument minimisant la deuxième fonction est $y_1 + z_1 = -\sqrt{(\alpha^2 + \beta^2)/r^2}$.

Il vient alors que l'argument minimisant du problème est (avec $r > 0$)

$$\begin{pmatrix} u_2 \\ v_2 \\ w_2 \end{pmatrix} = \begin{pmatrix} -\alpha r/\sqrt{\alpha^2 + \beta^2} \\ \\ -\beta r/\sqrt{\alpha^2 + \beta^2} \\ -\gamma\sqrt{1 - r^2}/\sqrt{\gamma^2} \end{pmatrix} \tag{350}$$

comme on aurait pu le trouver directement en extrayant de $G(u, v, w) = 0$ les deux cercles solutions puis en voyant dans la forme de F la fonction linéaire qui varie uniquement suivant la direction (α, β, γ).

Mais ce n'est pas ainsi que l'algorithme a été décrit : on a profité de la forme particulière du problème pour éviter les étapes de calcul des variations.

Il faut donc plutôt, conformément à l'étape 4 de l'algorithme, calculer

$$\begin{pmatrix} \partial_y u_1 & \partial_z u_1 \\ \partial_y u_1 & \partial_z u_1 \\ \partial_y u_1 & \partial_z u_1 \end{pmatrix} \tag{351}$$

solution de

$$\left(-y_0\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - z_0\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}\right)\begin{pmatrix} \partial_y u_1 & \partial_z u_1 \\ \partial_y v_1 & \partial_z v_1 \\ \partial_y w_1 & \partial_z w_1 \end{pmatrix} = \begin{pmatrix} u_1 & u_1 \\ v_1 & v_1 \\ w_1 & 0 \end{pmatrix}\begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \tag{352}$$

et enfin calculer (y_1, z_1) comme solution de

$$\begin{pmatrix} u_1 & v_1 & w_1 \\ u_1 & v_1 & 0 \end{pmatrix} \begin{pmatrix} \partial_y u_1 & \partial_z u_1 \\ \partial_y v_1 & \partial_z v_1 \\ \partial_y w_1 & \partial_z w_1 \end{pmatrix} \begin{pmatrix} y_1 - y_0 \\ z_1 - z_0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} u_1^2 + v_1^2 + w_1^2 - 1 \\ u_1^2 + v_1^2 - r^2 \end{pmatrix} = 0 \tag{353}$$

à partir desquels on peut recommencer les opérations jusqu'à convergence vers le résultat qui a été obtenu par analyse directe du problème.

Celui/lle qui penserait que ces calculs sont un peu lourds ne pourrait pas être mis au pilori avec un écriteau autour du cou proclamant qu'il/elle est paresseux/se. Ils le sont effectivement ; et c'est pourquoi on les donne à effectuer par des ordinateurs/rices.

La méthode du lagrangien dans le cas de plusieurs contraintes souffre du même inconvénient que dans le cas d'une seule contrainte ; et c'est pourquoi on utilise une méthode de pénalisation au préalable.

8.1.3 La pénalisation

Il suffit de pénaliser chacune des contraintes : c'est à dire d'introduire la fonction

$$F(X) + \sum_{p=1}^{P} \frac{1}{\epsilon_p} U_p(G_p(X)) \tag{354}$$

où les ϵ_p, positifs, ont vocation à devenir très petits ; où les fonctions U_p sont soumises aux mêmes conditions que U de (269).

Une question pourrait être de savoir s'il est préférable de plus pénaliser une des contraintes qu'une autre (c'est à dire de prendre l'ϵ correspondant plus petit que qu'un autre ϵ).

Mais, tant que F et les G_p sont indéfinis, cette question n'est pas sans rappeler celle que faisait poser Rabelais à de savants docteurs : «La chimère bruissant dans le vide peut-elle dévorer la seconde intention ? [Ils discutèrent les dix à douze semaines suivantes de la méthodologie à mener pour formaliser correctement cet important problème ...] »Aussi sera-t-elle abandonnée.

8.2 Plusieurs contraintes d'inégalités (caetera desunt)

Le cas de plusieurs contraintes inégalités ne sera pas traité dans le module.

La raison est qu'il faut introduire les problèmes de programmation linéaire qui

> *[sont] plus difficile à traiter que certain problèmes de programmation non linéaires, notamment quadratiques.*
> Ciarlet [Cia82, p. 231]

et que donc à chaque semestre suffit sa peine.

8.3 Exercices

8.3.1 Tracer une route

Sur une terre plate la position est repérée par les coordonnées cartésiennes (x, y) ; le relief est donné par une fonction $h(x, y)$ (la hauteur par rapport au niveau de la mer).

Il faut donner le tracé d'une route joignant deux villes, l'une à la position $(0, 0)$, l'autre à la position (X, Y) avec la contrainte qu'en aucun point de la route la pente ne soit supérieure à une limite définie p.

Comme on ne veut pas traiter des problèmes en dimension infinie, il faut plutôt aborder en aborder une version discrète. Celle-ci peut être de chercher la suite de coordonnées (x_0, y_0), $\ldots X_{N+1}, Y_{N+1}$ telle que

$$\begin{array}{lllllll} x_0 & = & 0 & ; & y_0 & = & 0 \\ x_{N+1} & = & X & ; & y_{N+1} & = & Y \end{array} \tag{355}$$

et pour $n = 0 \ldots N$ quelque chose comme

$$\max_{t \in [0,1]} \sqrt{(\partial h_x(x_n + (x_{n+1} - x_n)t, y_n + (y_{n+1} - y_n)t)^2 + (\partial h_y(x_n + (x_{n+1} - x_n)t, y_n + (y_{n+1} - y_n)t))^2} \leq p \tag{356}$$

Si on décide que le relief n'est pas suffisamment torturé pour jouer significativement dans la longueur de la route alors la longueur est celle qu'elle aura sur une carte, soit

$$\sum_{n=0}^{N} \sqrt{(x_{n+1} - x_n)^2 + (y_{n+1} - y_n)^2} \tag{357}$$

qu'il faut minimiser par rapport aux $2N$ variables x_n et y_n sous les $N + 1$ contraintes (356).

Voilà donc un problème pratique mettant en jeu un nombre important de contraintes de type inégalité.

N'essayez cependant pas de le résoudre sous cette forme : il est plus facile de le poser sous forme continue et de seulement le discrétiser après.

Le problème n'est donné que pour montrer l'adéquation entre la question de l'optimisation sous contrainte et la vie pratique.

D'ailleurs on pourrait le rendre plus proche des nécessités réelles en introduisant des lacs qu'il faudrait contourner ; en donnant aux villes des dimensions, et donc les points extrémités n'auraient plus de positions fixes mais seraient assujetties à passer par les contours des villes ; ...

8.3.2 Problème de Lagrange

Lagrange est l'inventeur des multiplicateurs de Lagrange [Lag65, t. 1,p. 69] ; en trahissant un peu le texte, on peut considérer qu'il traite le problème de statique qui consiste à considérer N masses ponctuelles pesantes.

La n^emasse a un poids M_n et ses coordonnées sont (x_n, y_n, z_n) ; la pesanteur est orientée suivant l'axe z, vers les z décroissants ; il s'agit alors de trouver l'équilibre de l'ensemble avec les contraintes :

- que certaines positions des masses sont fixées ; ou qu'elles appartiennent à une surface ou une ligne de l'espace ;
- que les distances entre certains couples de masses sont imposées (liaison rigides) ;
- que certains couples de masses sont reliés par des ressorts

La mise en équation de ce problème conduit à trouver l'énergie du système de masses et à minimiser celui-ci (on est en statique, donc l'énergie cinétique engendrées par les déplacements d'une configuration à l'autre n'est pas à prendre en compte) sous les contraintes précédemment indiquées.

Rien de plus intéressant que de traiter ce problème : on s'aperçoit notamment que les multiplicateurs de Lagrange associés aux contraintes sont les forces de réaction qui maintiennent l'équilibre[16] ; que la pénalisation est équivalente à remplacer une liaison rigide par une liaison avec un ressort dont la raideur est très grande ; et d'autres grandes et belles choses.

[16] Historiquement d'ailleurs les forces de réaction sont définies par Lagrange comme ce qu'on appelle maintenant des multiplicateurs de Lagrange

Quatrième partie

Logiciels utilisés et éléments de programmation

Il y a deux écoles antagonistes en matière de logiciels : la première préconise d'utiliser des 'produits standards' et commerciaux afin de ne pas réinventer la roue ; la seconde préconise d'utiliser des langages un peu souples parce que

I do not want to re-invent the wheel. The language must have hash-tables, lists, arrays, conditions &c standard.

Sam Steingold, The right tool for the job,

`http://www.podval.org/~sds/tool.html`

(Cet article pointe sur de nombreux liens utiles pour le calcul scientifique en général.)

On voit donc que l'argument est le même dans les deux camps : il ne faut pas réinventer la wheel. Les affidés de l'une des écoles croient qu'en payant avec de l'argent des logiciels ils ne commettront pas cette faute et les autres croient pouvoir éviter la faute en payant avec du temps.

9 Logiciels libres

Comme ceux qui préconisent de passer du temps plutôt que de dépenser de l'argent affirment que quand on dépense de l'argent il faudra en plus dépenser du temps pour que le logiciel acheté devienne vraiment utilisable et que cet argument semble correspondre à une certaine réalité, il parait raisonnable de choisir le camp des adorateurs de langages un peu souples.

Évidemment cela mène au LISP éventuellement un peu assisté de FORTRAN et de de C [17] pour la rapidité d'exécution.

LISP, ou tout autre membre de sa famille, est un langage attachant qui a l'avantage supplémentaire d'être disponible immédiatement dès lors qu'on utilise 'Emacs'.

Le but de ces recommandations n'est évidemment pas de faire de la manipulation sectaire. Aussi la recommandation finale sera d'utiliser

SCILAB

téléchargeable à partir de

http ://www-rocq.inria.fr/scilab/

pour de nombreuses plates-formes, UNIX/LINUX bien sûr, mais aussi des systèmes plus exotiques comme WINDOWS.

Celui qui voudrait chercher à télécharger SCILAB, sous UNIX ou LINUX devrait vérifier au préalable si son système ne le contient pas déjà.

Pour cela, taper 'scilab' ; si cela ne donne rien, tenter 'whereis scilab' ; et, si cela ne donne encore rien, 'find / -name scilab -print' (mais cette commande est un peu longue).

Dans le cas ou 'scilab' existe mais que l'environemment de l'utilisateur ne lui permette pas de l'exécuter : il faut ajouter dans la variable PATH le chemin d'accès de 'scilab' (quelque chose comme (setenv PATH ($PATH <le chemin d'accès>)) ; il faut aussi positionner une nouvelle variable d'environnement SCI au chemin d'accès du répertoire principal de scilab (quelque chose comme (setenv SCI <le répertoire principal qui doit être de la forme '/(une suite de nom)/scilab-2.5')).

10 Logiciels non libres

Ce ne sont pas les rois, mais bien les courtisans, Qui de la liberté redoutent les accen[t]s.

A. NAUDET, l'Assemblée des Animaux.

Ils sont donc liés ? Mais en fait c'est plutôt l'utilisateur qui est lié. Que dire de plus ?

Peut-être raconter les tribulations d'un pauvre chercheur. Des années 80 à 90 celui-ci a sucessivement utilisé : la norme graphique du Pascal UCSD sur Apple II_e ; celle du turbo Pascal sur carte Z80 adaptée sur un Apple II_e ; les normes graphiques tektro (au moins 3 d'entre elles emboîtées) sous DPS8 avec émulation de terminal sur un Victor S3 et un modem 1200 Bauds ; l'interface *UIS* sous VMS avec un microVax ; *GKS* sous ce même système ; puis *GKS* sous un SUN3 avec Sunview ; et enfin X11 directement sous Unix.

Depuis X11, le chercheur n'a plus eu de problèmes graphiques et en plus ses programmes sont vraiment portables.

11 Logiciel de calcul formel

Il existe des logiciels libres de calcul formel, par exemple CALC écrit entièrement en ELISP.

Et de la même manière qu'il est possible de fabriquer ses propres programmes d'analyse numérique ordinaire, il est également possible de fabriquer sans grands frais ses propres programmes de calcul formel.

Mais pour des raisons (probablement mauvaises) on va donner quelques éléments d'utilisation d'un logiciel non libre : MAPLE. La raison de ce choix est que ce logiciel est relativement facile à trouver et nulle autre : il n'est ni pire ni meilleur que ses frères.

Voici tout d'abord deux procédures : la première 'newton' retourne l'itération de Newton et la seconde l'itération du gradient dont la longueur de pas est calculée avec la méthode de Newton unidirectionnelle.

[17] La trilogie FORTRAN, LISP et C est très attachante. Le FORTRAN est naît dans les années 50, le LISP dans les années 60 et le C dans les années 70. Le premier (formula translator) répond aux besoins de calculs bruts de trajectoires de fusées pour aller sur la lune ; le second (list processing) est une tentative pour faire de l'intelligence artificielle, et donc des fusées sans pilote ; le troisième est le langage d'UNIX. Il est sans doute injuste de ne pas parler de COBOL, de PASCAL, ADA et JAVA, des différentes variantes de BASIC, de FORTH et de bien d'autres, mais il faut bien faire un choix.

```
newton:= proc(expression,variables)
local gradient, hessien,formule;
description 'Sort l'iteration de Newton.
            Fonctionne avec linalg (with(linalg))';
gradient:= grad(expression,variables):
hessien:= hessian(expression,variables):
formule:= convert(linsolve(hessien,evalm(-1*gradient)),list)
          +convert(variables,list):
map(simplify,formule)
end:

g_newton:= proc(expression,variables)
local gradient,hessien,alpha,formule;
description 'Sort l'itération du gradient avec pas calculé par Newton.
     Fonctionne avec linalg (with(linalg))';
gradient:= grad(expression,variables):
hessien:= hessian(expression,variables):
alpha:= -1 * innerprod(gradient,gradient)
           / innerprod(gradient,hessien,gradient):
formule:= convert(evalm(alpha * gradient),list)
         + convert(variables,list):
map(simplify,formule);
end:
```

Puis des exemples d'utilisation de ces deux procédures :

```
with(linalg):
newton:=... end:<- écrire ici la définition de newton donnée plus haut

newton(x^2/a^2+y^2,[x,y]);
```

ce qui renvoie évidemment le vecteur $[0, 0]$.

```
with(linalg):
g_newton:=... end:<- écrire ici la définition de g_newton donnée

iteration:=g_newton(x^2/a^2+y^2,[x,y]);
```

ce qui renvoie le vecteur

$$iteration := \left[\frac{xy^2a^4\left(a^2-1\right)}{x^2+y^2a^6}, -\frac{yx^2\left(a^2-1\right)}{x^2+y^2a^6}\right]$$

On peut essayer d'estimer l'effet de l'itération en calculant

```
facteur:=innerprod(iteration,iteration)/(x^2+y^2);
```

qui renvoie

$$facteur := \frac{x^2 * y^2 * (a^2-1)^2 * (y^2 * a^8 + x^2)}{(x^2+y^2*a^6)^2(x^2+y^2)}$$

Pourquoi ne pas tracer les lignes de niveaux de $facteur$ dans le cas où $a = 3$?

```
with(plots);
f3:=subs(a=3,facteur);
plot3d(f3,x=0..4,y=0..4);
```

on obtient un dessin comme (d'un point de vue technique on n'a rien fait pour en améliorer la présentation et donc le résultat est un peu minable)

qui peut être interprété si on voit correctement dans l'espace.

On peut aussi tenter quelques manipulation comme

```
fc:=simplify(subs(a=tan(alpha),x=r*cos(theta),y=r*sin(theta),facteur));
```

qui renvoie une expression un peu décourageante

```
fc :=
-cos(theta)^2*(-1-41*cos(alpha)^8+44*cos(alpha)^6-40*cos(theta)^4*cos(alpha)^8
+16*cos(theta)^4*cos(alpha)^10+8*cos(alpha)^2-26*cos(alpha)^4+20*cos(alpha)^10
-4*cos(alpha)^12-16*cos(theta)^2*cos(alpha)^2+52*cos(theta)^2*cos(alpha)^4
-88*cos(theta)^2*cos(alpha)^6+81*cos(theta)^2*cos(alpha)^8
-36*cos(theta)^2*cos(alpha)^10+4*cos(theta)^2*cos(alpha)^12-cos(theta)^4
+2*cos(theta)^2+8*cos(theta)^4*cos(alpha)^2-26*cos(theta)^4*cos(alpha)^4
+44*cos(theta)^4*cos(alpha)^6)/(1-2*cos(theta)^2-36*cos(theta)^2*cos(alpha)^8
+42*cos(theta)^2*cos(alpha)^6-6*cos(alpha)^2+12*cos(theta)^2*cos(alpha)^2
-30*cos(theta)^2*cos(alpha)^4+18*cos(theta)^2*cos(alpha)^10
-4*cos(theta)^2*cos(alpha)^12-6*cos(theta)^4*cos(alpha)^2
+15*cos(theta)^4*cos(alpha)^4-22*cos(theta)^4*cos(alpha)^6
+21*cos(theta)^4*cos(alpha)^8-12*cos(theta)^4*cos(alpha)^10
```

```
+4*cos(theta)^4*cos(alpha)^12+cos(theta)^4+15*cos(alpha)^4-20*cos(alpha)^6
+15*cos(alpha)^8-6*cos(alpha)^10+cos(alpha)^12)
```

qui ne dépend cependant que de α et θ ; ce qui permet le tracé

```
with(plots);
f3:=subs(a=3,facteur);
plot3d(fc,alpha=0..2*Pi,theta=0..2*Pi);
```

qui renvoie

un paysage de montagne qui peut être interprété.

Si maintenant on désire effectuer les itérations dans un langage plus rapide que celui de MAPLE, on peut utiliser

```
map(fortran,iteration);
```

qui renvoie

```
t0 = x*y**2*a**4*(a**2-1)/(x**2+y**2*a**6)
t0 = -y*x**2*(a**2-1)/(x**2+y**2*a**6)
```

c'est à dire des lignes directement utilisables dans un programme FORTRAN.

On l'aura compris, l'apport logistique d'un programme de calcul formel est loin d'être négligeable.

Cinquième partie

Éléments bibliographiques

12 Mathématiques

12.1 Mathématiques de base

L'étudiant en science appliquées sentira toute sa vie qu'il est ignorant en mathématiques. Ce Sisyphe cherchera alors à combler ce manque en collectionnant des manuels. Voici, parmi les livres qui se présentent comme des manuels, une liste choisie en essayant précisément d'éliminer ceux qui ne sont que des manuels :

- Saint-Guilhem [SG89] *«s'adresse [...] à des adultes désireux d'apprendre et prêts à un certain effort, mais sans soucis d'examens associés à un programme officiel» ;*
- *Rudin [Rud95b] réalise un manuel dont le contenu est, avec quelques extensions, celui des classes préparatoires.*

De plus on peut trouver sur internet des textes très intéressants pour les mathématiques de bases parmi des textes plus spécialisés.

Par exemple voici un site qui contient de nombreux cours de mathématiques

```
http://spoirier.citeweb.net//liensmaths.html
```

dans ce site on peut trouver un cours de mathématiques de classes préparatoires

```
clicker 'Gérard Lavau'
```

mais aussi on peut trouver des cours de niveau supérieur

```
clicker 'Cours de second cycle'
```

dans lequel on trouve

```
clicker 'Mathématiques pour l'agrégation (version pas à jour, ça viendra)'
```

qui est un gros fichier contenant des théorèmes et des démonstrations, puis

```
clicker 'Mathématiques'
```

Un autre site qui pointe d'ailleurs sur le précédent est

```
http://homer.span.ch/~spaw2581/maths.htm
```

qui contient également quelques cours de mathématiques.

12.2 Mathématiques par catégories

12.2.1 Géométrie et calcul différentiel

S'il faut comprendre ce qu'est la géométrie analytique on pourra d'abord lire Descartes [Des91].

S'il faut faire de la géométrie différentielle, les leçons de géométries de Postnikov sont très utiles [Pos81b], [Pos81a], [Pos90a], [Pos90b] ; notamment le second tome et le début du troisième.

Si on s'intéresse au calcul différentiel, qui est d'ailleurs fortement relié à la géométrie différentielle, on pourra lire Dieudonné [Die] qui a voulu réaliser un ouvrage didactique et aussi Cartan [Car77].

On pourra aussi s'intéresser aux publication de l'IREM [IRE99] qui inévitablement conduiront à s'intéresser à Archimede, Leibnitz et Newton plus directement. À cet égard quelques pages de Maxwell sont également très instructives [Max54].

12.2.2 Algèbre

L'algèbre actuel s'intéresse surtout aux structures algébriques et est d'un abord un peu délicat. Toutefois, si on ne souhaite pas devenir un véritable mathématicien on peut trouver un intérêt à lire Vuillemin [Vui62] qui recense et explique les travaux des fondateurs de l'algèbre moderne.

Deux manuels de niveau Deug contiennent une information disponible sans difficulté [CC95b] [CC95a] et on trouve de l'algèbre dans [SG89].

12.2.3 Analyse

De l'analyse peut être trouvé dans la partie 'géométrie et calcul différentiel', mais plus spécifiquement, il y a aussi : Une analyse assez didactique de Kolmogorof [KF94] ; une analyse fonctionnelle de Sobolev [LS89] ; celle de Rudin [Rud95a] qui est assez abstraite ; et finalement les 'Dautrey et Lions' [DLC$^+$87], [DLA$^+$87g], [DLB$^+$87], [DLA$^+$87a], [DLA$^+$87c], [DLA$^+$87b], [DLA$^+$87e], [DLA$^+$87f], [DLA$^+$87d] dont le volume est impressionnant mais ou on trouve aussi des explications très claires, notamment celle de l'article de Cessenat, dans le volume 5, sur la décomposition de Helmholtz des champs de vecteurs.

Mais on peut encore conseiller Mawhin [Maw97] qui donne une base de référence historique très intéressante.

Des références mathématiques des équations différentielles ordinaires et des équations aux dérivées partielles sont données dans la partie 'Analyse numérique'.

13 Analyse numérique

Est il vraiment nécessaire de séparer l'analyse numérique des mathématiques ? Pour paraphraser Clausewitch a propos de choses plus sympathiques : l'analyse numérique ne serait elle pas la continuation de mathématiques par d'autre moyens ?

La question reste ouverte quand bien même cela vient d'être fait ici.

13.1 Analyse numérique de base

L'analyse numérique est avant la mise en œuvre d'algorithmes, il est intéressant d'avoir un point de vue sur ce qu'est un algorithme et pour cela la lecture d'«Histoire d'algorithmes»[Ca94] donne des informations très intéressantes.

Des ressources en ligne importantes existent

`http://www.netlib.org/index.html`

on trouve des algorithmes de calcul et des liens intéressants comme

`http://www.nr.com/`

où le 'numerical recipes' peut être lu et même un peu téléchargé. Des cours interactifs existent comme

`http://dmawww.epfl.ch/rappaz.mosaic/Support/support/support.html`

qui contient à peu près ce que doivent savoir des étudiants sortant d'une école d'ingénieur et est une démarche publicitaire intelligente pour que lassé de l'écran on achète le livre correspondant. Il y a aussi

`http://lumimath.univ-mrs.fr/~jlm/cours/analnum/`

ou on trouve un aide mémoire intéressant.

Plus classiquement il existe aussi le livre de Sibony [SM84] qui traite d'un peu tout ; mais surtout le couple

- «*Introduction à l'analyse numérique de Baranger* »*[Bar77]*
- *complété de* «Analyse numérique»[BBC$^+$91]

dans lesquels on trouve tout ce qui ce qui peut être utile.

Par contre la lecture d'ouvrages didactiques comme [Nou87], [Euv90] est toujours décevante ; on y trouve des formules mais peu d'explications sur ces formules. Si on aime les formules le mieux est encore le Numerical Recipes.

13.2 Analyse numérique par catégories

13.2.1 Optimisation

L'optimisation est une matière importante de l'analyse numérique, et en plus de nombreux ouvrages sont disponibles :

- d'abord la référence de Luenberger [Lue84] ;
- puis une autre référence [Cia82] et encore une troisième [GS80] qu'on peut compléter par un article fondateur [DM77] ;
- également un livre récent [Cul94] ;
- enfin des livres un peu vieillis mais qui gardent néanmoins un intérêt certain comme ceux de Minoux [Min83], de Karmanov [Kar77], de Céa [Céa71] qu'on aurait tort de négliger si on ne veut pas réinventer ce qui y est, par exemple l'artifice de Valentine pour le dernier.

Il est également intéressant de lire des livres de vulgarisation comme celui de S. Hildbrandt et al. [HT84] dans lesquels les exemples célèbres de problèmes d'optimisation sont évoqués.

13.2.2 Équations différentielles ordinaires

En plus de l'ouvrage collectif de Baranger [BBC$^+$91] on peut lire le livre très complet de Crouzeix et Mignot [CM89] ; ou le livre plus didactique de Demailly [Dem96] ; et surtout le livre de Hubbard [HW99] dans lequel la problématique du calcul numérique des équations différentielles est replacé dans son contexte.

Comme on n'a rien cité de spécifique aux équations différentielles dans la partie 'mathématiques' de cette liste, il est maintenant utile de proposer : Arnold [Arn74] et Pontriaguine [Pon75] pour les edos en général ; puis [ATF87] et [PBGM74] pour le contrôle optimal sur les edos, qui d'ailleurs peuvent également émarger à la partie 'optimisation'.

13.2.3 Équations aux dérivées partielles

Les 'Dautrey et Lions' [DLC$^+$87] à [DLA$^+$87d] traitent presque exclusivement des équations aux dérivées partielles, du point de vue des mathématiques appliquées vues par les mathématiciens.

Mais il y a aussi les nombreux livres dont le titre contient l'un des groupes nominaux 'éléments finis', 'calcul scientifique' . . . :

- d'abord la référence de Zinckiewicz [Zin71] dans laquelle presque tous les aspects de la méthode des éléments finis sont abordés ;
- puis le très intéressant livre de Ciarlet [Cia78] qui est très didactique ;
- et aussi le livre didactique de Lucquin et Pironneau [LP96]
- et, finalement, il ne faut pas oublier [Jol90]

Cette liste est évidemment loin d'être complète.

Références

[Arn74] V. Arnold. *Équations différentielles ordinaires.* Mir, 1974.

[ATF87] A. Alekseev, V. M. Tikhomirov, and S. V. Fomin. *Optimal control.* Plenum, 1987. existe en francais collection Mir.

[AVGZ86] V. Arnold, A. Varchanko, and S. Goussein-Zadé. *Singularité des applications différentiables.* Mir, 1986. 2 tomes.

[Bar77] J. Baranger. *Introduction à l'analyse numérique.* Hermann, 1993 (premier tirage 1977).

[BBC$^+$91] J. Baranger, C. Brezinsky, C. Carasso, J.M. Chassery, F. Chatelin, J.F. Maitre, J. Roux, and G. Wanner. *Analyse numérique.* Hermann, 1991.

[Ca94] J.-L. Chabert and al. *Histoire d'algorithmes.* Belin, collection Regards sur la science, 1994.

[Car77] E. Cartan. *Cours de calcul différentiel.* Hermann, 1977.

[CC95a] A. Calvo and B. Calvo. *Algèbre générale.* Masson, 1995.

[CC95b] A. Calvo and B. Calvo. *Algèbre linéaire.* Masson, 1995.

[Céa71] J. Céa. *Optimisation : théorie et algorithmes.* Dunod, 1971.

[Cia78] P. G. Ciarlet. *The finite element method for elliptic problems.* North-Holland, 1978.

[Cia82] P. G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation.* Masson, 1982.

[CM89] M. Crouzeix and A.L. Mignot. *Analyse numérique des équations différentielles.* Masson, 1989.

[Cul94] J.C. Culioli. *Introduction à l'optimisation.* Ellipse, 1994.

[Dem96] J.-P. Demailly. *Analyse numérique et équations différentielles.* Presses Universitaires de Grenoble, 1996.

[Des91] R. Descartes. *La géométrie.* Jean Gabay (d'après l'édition de Hermann en MDCCLXXXVI), 1991. On le trouve aussi à la B.N.F. 'http ://gallica.bnf.fr'.

[Die] J. Dieudonne. *Calcul infinitésimal.*

[DLA+87a] R. Dautrey, J. L. Lions, M. Artola, M. Authier, M. Cessenat, J. M. Combes, B. Mercier, and C. Wild. *Analyse mathématique et calcul numérique*, volume 4. Masson, 1987. Tome 1, chapitre 6, 7. Méthodes variationnelles.

[DLA+87b] R. Dautrey, J. L. Lions, M. Artola, P. Bénilan, M. Bernadou, M. Cessenat, J. C. Nedelec, and J. Planchard. *Analyse mathématique et calcul numérique*, volume 6. Masson, 1987. Tome 2, chapitre 11, 12, 13. Méthodes intégrales et numériques.

[DLA+87c] R. Dautrey, J. L. Lions, M. Artola, M. Cessenat, J. M. Combes, and B. Scheurer. *Analyse mathématique et calcul numérique*, volume 5. Masson, 1987. Tome 2, chapitre 8, 9, 10. Spectre des opérateurs.

[DLA+87d] R. Dautrey, J. L. Lions, M. Artolat, C. Bardos, M. Cessenat, A. Kavenoly, P. Lascaux, B. Mercier, O. Pironneau, and R. Sentis. *Analyse mathématique et calcul numérique*, volume 9. Masson, 1987. Tome 3, chapitre 20, 21. Évolution : numérique, transport.

[DLA+87e] R. Dautrey, J. L. Lions, M. Artolat, M. Cessenat, and H. Lanchon. *Analyse mathématique et calcul numérique*, volume 7. Masson, 1987. Tome 3, chapitre 14, 15, 16. Évolution : Fourier, Laplace.

[DLA+87f] R. Dautrey, J. L. Lions, M. Artolat, M. Cessenat, and B. Scheurer. *Analyse mathématique et calcul numérique*, volume 8. Masson, 1987. Tome 3, chapitre 17, 18, 19. Évolution : semi-groupe, variationnel.

[DLA+87g] R. Dautrey, J. L. Lions, M. Authier, P. Bénilan, and M. Cessenat. *Analyse mathématique et calcul numérique*, volume 2. Masson, 1987. Tome 1, chapitre 2. L'opérateur de Laplace.

[DLB+87] R. Dautrey, J. L. Lions, P. Bénilan, M. Cessenat, B. Mercier, and C. Zuily. *Analyse mathématique et calcul numérique*, volume 3. Masson, 1987. Tome 1, chapitre 3, 4, 5. Transformation, Sobolev, opérateurs.

[DLC+87] R. Dautrey, J. L. Lions, M. Cessenat, A. Gervat, and H. Lanchon. *Analyse mathématique et calcul numérique*, volume 1. Masson, 1987. Tome 1, chapitre 1. Modèles physiques.

[DM77] J. E. Dennis and J. J. Moré. Quasi-newton methods, motivation and theory. *SIAM review*, 19(19) :pp 46–89, 1977.

[Euv90] D. Euvrard. *Résolution numérique des équations aux dérivées partielles.* Masson, 2° edition, 1990.

[Fey95] R. Feynmann. *Électromagnétisme.* InterEdition, nouveau tirage de 1975 edition, 1995. 2 tomes.

[GS80] W.A. Gruver and E. Sach. *Algorithmic method in optimal control.* Pitman Advanced, 1980.

[HT84] S. Hildbrandt and A. Tromba. *Mathématiques et formes optimales.* Bibliothèque de la revue pour la science, 1984.

[HW99] J. Hubbard and B. West. *Équations différentielles et systèmes dynamiques.* Cassini, 1999.

[IRE99] IREM. *Aux origines du calcul infinitésimal.* Ellipse, 1999.

[Jol86] P. Joly. Présentation de synthèse des méthodes de gradient conjugué. *MAN – Mathematical Modelling and Numerical analysis – Modélisation máthématique et analyse numérique*, 20(4) :pp 639–665, 1986.

[Jol90] P. Joly. *Mise en oeuvre de la méthode des éléments finis.* Mathématique & application. Ellipse, 1990.

[Kar77] V. Karmanov. *Programmation mathématique.* Mir, 1977.

[KF94] A. Kolmogorov and S. Fomine. *Élements de la théorie des fonctions et de l'analyse fonctionnelle.* Mir-Ellipse, 3° edition, 1994.

[Lag65] J.-L. Lagrange. *Mécanique analytique.* Librairie scientifique Albert Blanchard, Paris, 1965. 2 tomes. Édition originale 1788.

[LP96] B. Lucquin and O. Pironneau. *Introduction au calcul scientifique.* Masson, 1996.

[LS89] L. Lusternik and V. Sobolev. *Précis d'analyse fonctionnelle.* Mir, 1989.

[Lue84] D. G. Luenberger. *Linear and nonlinear programming.* Addison-Wesley Publishing company, 1984.

[Maw97] J. Mawhin. *Analyse.* De Boeck université, 1997.

[Max54] J. C. Maxwell. *A treatise on electricity & magnetism*, volume 1 & 2. Dover, New York, 1954. from the third edition Clarendon press, 1891.

[Min83] M. Minoux. *Programmation mathématique : théorie et algorithmes.* Dunod, 1983.

[Nou87] J.P. Nougier. *Méthodes de calcul numériques.* Masson, third edition, 1987.

[PBGM74] L. Pontriaguine, Y. Boltianski, R. Gamkrélidzé, and E. Michtchenko. *Théorie mathématique des processus optimaux.* Mir, 1974.

[PD77] B. Pchénitchny and Y. Daniline. *Méthode numérique dans les problèmes d'extrémum.* Mir, 1977.

[Pol89] G. Polya. *Comment poser et résoudre un problème.* Jean Gabay (d'après l'édition Dunod de 1965), 1989.

[Pon75] L. Pontriaguine. *Équations différentielles ordinaires.* Mir, 1975.

[Pos81a] M. Postnikov. *Leçons de géométrie : algébre linéaire et géométrie différentielle.* Mir, 1981.

[Pos81b] M. Postnikov. *Leçons de géométrie : géométrie analytique.* Mir, 1981.

[Pos90a] M. Postnikov. *Leçons de géométrie : géométrie différentielle.* Mir, 1990.

[Pos90b] M. Postnikov. *Leçons de géométrie : variétés différentiables.* Mir, 1990.

[Rai93] D. Raikov. *Analyse mathématique multidimensionnelle.* Mir, 1993.

[Rud95a] W. Rudin. *Analyse Fonctionnelle.* Ediscience, 2° edition, 1995. traduit de "Functional analysis".

[Rud95b] W. Rudin. *Principes d'analyse mathématique.* Ediscience, 3° edition, 1995. traduit de "Principles of mathématical analysis".

[Sch81] L. Schwartz. *Cours d'analyse.* Hermann, 1981.

[SG89] R. Saint-Guilhem. *Notions fondamentales de mathématiques modernes.* Ellipse, 1989. 2 tomes.

[SM84] M. Sibony and J. C. Mardon. *Analyse numérique.* Hermann, 1984. 3 tomes.

[Vui62] J. Vuillemin. *La philosophie de l'algèbre.* Presses Universitaires de France, collection Épiméthée, 1993 (première édition 1962).

[Zin71] O.C. Zinckiewicz. *The finite element method in engineering science.* Mc Graw Hill, 1971.