UNIVERSITÉ DE PARIS-I PANTHÉON-SORBONNE

THÈSE DE DOCTORAT, SPÉCIALITÉ MATHÉMATIQUES APPLIQUÉES

COMMANDE OPTIMALE D'ENGINS SOUS-MARINS REMORQUES AVEC CONTRAINTES

Laurent Chauvier

Soutenue le 24 janvier 2000 devant le jury composé de

Joël Blot	*président*
Guy Chavent	*rapporteurs*
Claude Lemaréchal	
Paul Armand	*examinateurs*
Frédéric Bonnans	
Gilbert Damy	
Jean Charles Gilbert	

Directeurs de thèse : Jean Charles Gilbert & Frédéric Bonnans

Remerciements

Jean Charles Gilbert, qui a guidé mes premiers pas de jeune chercheur, est pour beaucoup dans l'achèvement de cette thèse. Je lui suis extrêmement reconnaissant de s'être montré aussi disponible et ouvert au dialogue. Qu'il sache combien j'ai apprécié de travailler sous sa direction.

Je remercie vivement Frédéric Bonnans qui, en relisant plusieurs fois certains chapitres, et en me faisant part de nombreux commentaires, m'a permis d'améliorer de façon sensible la qualité du présent manuscrit.

Claude Lemaréchal a toujours eu à mon égard une attitude bienveillante. Ses cours m'ont donné le goût de l'optimisation numérique et incité à entreprendre ce travail de recherche ; je me réjouis qu'il ait accepté de le juger. Je suis honoré que Guy Chavent ait également accepté d'en être rapporteur. Pour le temps qu'ils ont consacré à examiner mes résultats et l'intérêt qu'ils leur ont manifesté, je leur exprime à tous deux mon entière gratitude.

Je suis reconnaissant à Gilbert Damy de m'avoir fait profiter de son expérience de la modélisation des câbles sous-marins, à Joël Blot et à Paul Armand d'avoir présidé et participé aux délibérations du jury.

Enfin, je tiens à remercier les membres de l'ancien projet PROMATH et les doctorants du bâtiment 12 de l'INRIA–Rocquencourt, pour leurs conseils comme pour leur présence amicale tout au long de la préparation de cette thèse.

Table des matières

Introduction

Au fil des années, les missions d'inspection des fonds sous-marins sont devenues de plus en plus fréquentes et variées. Elles n'étaient motivées, dans un passé récent, que par des études scientifiques ou la reconnaissance d'épaves ayant un intérêt militaire ou historique. Elles sont maintenant conduites en rapport avec l'exploitation de champs pétrolifères, la pose de câbles optiques de télécommunications et l'observation de zones sismiques sous-marines actives, entre autres.

Pour l'exploration spécifique des grandes profondeurs, on utilise souvent des engins équipés d'instruments de mesure fixés à un câble et descendus ainsi à quelques dizaines de mètres du fond. Ce système a deux avantages par rapport aux véhicules autonomes, habités ou commandés depuis la surface : il accomplit des missions plus longues et il échange des données avec le navire de support à un débit nettement plus élevé. Cependant, pour de grandes longueurs de câble, c'est-à-dire plusieurs milliers de mètres, les temps de réponse aux manœuvres du navire sont très importants et le positionnement de l'engin est délicat.

C'est ce qui justifie le développement de logiciels d'aide à la navigation dont le but est de permettre un meilleur contrôle du système câble – engin. Celui de Mudie et Kastens [77] détermine une trajectoire du navire assurant que l'engin passe à distance minimale d'une cible donnée. Le problème inverse considéré par Hover [60] est le calcul d'une trajectoire du navire telle que l'engin suive une trajectoire prescrite. Nous cherchons, nous, à calculer une trajectoire du navire emmenant l'engin d'une position à une autre position données, en temps minimal. La position finale est imposée avec une certaine tolérance.

Pichon [83] avait abordé avant nous ce problème de commande optimale. La contrainte d'inégalité sur la position finale de l'engin était prise en compte au moyen de techniques de pénalisation. Mais le choix des coefficients sur lesquels repose cette approche s'est révélé épineux. Il est apparu nécessaire de mettre en œuvre un algorithme d'optimisation traitant plus efficacement les contraintes d'inégalité, sur les variables d'état en particulier.

L'étude d'un tel algorithme était notre principal objectif. Les troisième et quatrième chapitres du manuscrit lui sont consacrés. Les deux premiers chapitres traitent de la modélisation du système câble – engin et de l'intégration numérique de sa trajectoire. On applique enfin tous ces outils au contrôle du câble, et plus spécifiquement à la résolution du problème du demi-tour en temps minimal, dans le dernier chapitre.

Chapitre 1

Etude de l'équation du câble et de quelques cas particuliers

Nous nous intéressons au système composé d'un câble immergé et d'un engin fixé à son extrémité inférieure, que l'on appelle fréquemment le poisson. Ce système est remorqué par un navire, à partir duquel un treuil permet de dérouler ou d'enrouler le câble. Une équation aux dérivées partielles d'évolution, couplée à une équation algébrique, permet d'exprimer la configuration et la tension du câble au cours du temps en fonction de la position du navire et de la longueur de câble déroulée. Nous montrons l'existence d'une solution dans certains cas particuliers de cette équation.

1.1. Modélisation du système câble – engin

Précisons quelques notations. Les deux variables intervenant dans les équations sont le temps t et l'abscisse curviligne le long du câble s. Les dérivées partielles d'une fonction $x(s,t)$ sont notées

$$x'(s,t) = \frac{\partial x}{\partial s}(s,t), \quad \dot{x}(s,t) = \frac{\partial x}{\partial t}(s,t).$$

Dans ce chapitre et le chapitre suivant, sont connues, à chaque instant t, la position du navire $u(t)$ et la longueur de la partie déroulée du câble $\ell(t)$. Toute cette partie du câble est immergée. Comme indiqué sur la figure 1.1, l'abscisse curviligne s est orientée de sorte que $s = 0$ au niveau de l'engin remorqué et $s = \ell(t)$ au niveau du navire.

On cherche à déterminer la position relative du câble par rapport au navire. Celle-ci est notée $y(s,t)$ au point d'abscisse curviligne s et au temps t. Le modèle est tridimensionnel: $y(s,t) \in \mathbb{R}^3$. Cette position relative est bien évidemment nulle au point de fixation au navire, c'est-à-dire

$$y(\ell(t),t) = 0.$$

1.1.1. Hypothèse d'inextensibilité

Rappelons que l'abscisse curviligne le long du câble est définie pour une configuration de référence de celui-ci, pouvant par exemple être sa configuration "au repos" (câble déroulé et posé sur le sol). Ainsi, $y(s,t)$ désigne plus précisement la position relative à l'instant t

du point d'abscisse curviligne s dans la configuration de référence du câble. Une telle description du câble est dite lagrangienne.

On suppose que le câble est inextensible et en configuration tendue (voir Germain [50]). Ceci se traduit, pour un élément de câble de longueur ds (sous-entendu, de longueur ds dans la configuration de référence) suffisamment petit, par

$$|\, y(s+ds,t) - y(s,t)\,| = |\, ds\,|.$$

Aucune confusion n'étant à craindre, nous désignons indifféremment par $|.|$ la norme euclidienne de $\mathbb{R}^3$ et la valeur absolue sur $\mathbb{R}$. En faisant tendre ds vers 0, il en résulte que

$$|\, y'(s,t)\,| = 1.$$

Signalons qu'en pratique, le câble subit un allongement moyen inférieur à 1% de la longueur de câble immergée, ce qui semble très modeste. Les travaux de modélisation des câbles utilisent quasi-systématiquement l'hypothèse d'inextensibilité. En tout état de cause, elle n'entraîne pas d'imprécision supérieure à celle qu'induisent la méconnaissance de certains paramètres – au premier rang desquels le courant sous-marin – ou le fait de négliger des effets tels que la portance, la torsion, etc.

Le vecteur $y'(s,t)$ est tangent au câble et la condition d'inextensibilité montre qu'il est unitaire. On suppose également que le câble est parfaitement flexible. Ceci signifie (on se réfère à nouveau à [50]) que les efforts intérieurs au câble (nécessaires au maintien de la liaison d'inextensibilité) se réduisent à un vecteur tangent au câble. Au point d'abscisse s et au temps t, cette force peut donc s'écrire $T(s,t)\,y'(s,t)$. Le scalaire $T(s,t)$ est la tension dans le câble.

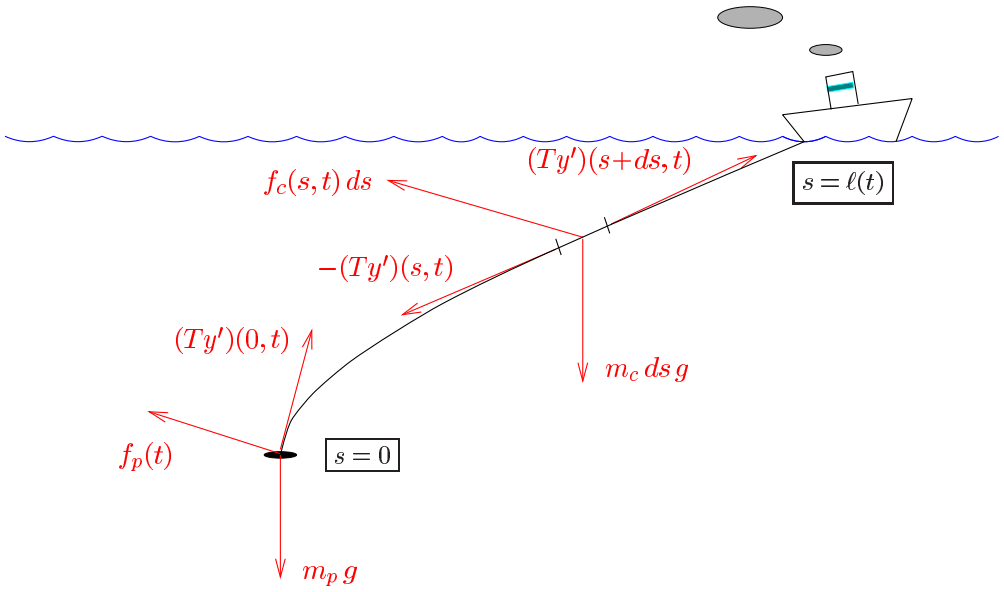


FIG. 1.1 – *forces s'appliquant à un élément de câble et au poisson.*

1.1.2. Equation d'évolution du câble

Dans le but d'appliquer le principe fondamental de la dynamique, on a représenté sur la figure 1.1 les forces prises en compte dans le modèle agissant sur un élément de câble et sur l'engin remorqué.

On note m_c la masse apparente linéique du câble – il s'agit de sa masse linéique corrigée de manière à prendre en compte la poussée d'Archimède – et g le vecteur accélération de la pesanteur. Le poids apparent d'un élément de câble de longueur ds est donc $m_c\,ds\,g$. Les forces à ses extrémités sont $(T\,y')(s+ds,t)$ et $-(T\,y')(s,t)$. Sa traînée hydrodynamique est $f_c(s,t)\,ds$. L'engin remorqué, de masse apparente m_p, est, lui, soumis à l'action de son poids apparent $m_p\,g$, de la tension à l'extrémité inférieure du câble $(T\,y')(0,t)$ et de sa traînée hydrodynamique $f_p(t)$. Toutes ces forces s'entendent comme vecteurs de $\mathbb{R}^3$ et les deux forces de traînée, détaillées dans la section suivante, sont des fonctions non linéaires de $\dot{y}$, y' et de la vitesse du navire $\dot{u}$.

Le produit de l'accélération de l'élément de câble par sa masse est égal à la somme des forces qui s'appliquent à lui. En faisant tendre sa longueur ds vers 0, il en découle que

$$m_c\,(\,\ddot{u}+\ddot{y}\,)=\big(T\,y'\big)'+f_c+m_c\,g$$

en tout point le long du câble. Cette même loi appliquée à l'engin remorqué donne

$$m_p\,(\,\ddot{u}(t)+\ddot{y}(0,t)\,)=\big(T\,y'\big)\,(0,t)+f_p(t)+m_p\,g.$$

En conclusion, si on se donne les position et vitesse relatives du câble au temps $t=0$, soient deux fonctions $y_{init}(s)$ et $v_{init}(s)$, la position du navire $u(t)$ et la longueur de câble déroulée $\ell(t)$ au cours du temps, on obtient la position relative du câble $y(s,t)\in\mathbb{R}^3$ et sa tension $T(s,t)\in\mathbb{R}$ en résolvant

$$\left\{\begin{array}{l} m_c\,\ddot{y}(s,t)=(T\,y')'\,(s,t)+f_c(s,t)+m_c\,(g-\ddot{u}(t)) \\ |\,y'(s,t)\,|=1, \end{array}\right. \qquad s\in\,]\,0,\ell(t)\,[\,,\ t\geq 0, \qquad (1.1.a)$$

en plus des conditions aux limites

$$y(\ell(t),t)=0,\quad m_p\,\ddot{y}(0,t)=(T\,y')(0,t)+f_p(t)+m_p\,(g-\ddot{u}(t)),\quad t\geq 0 \qquad (1.1.b)$$

et conditions initiales

$$y(s,0)=y_{init}(s),\quad \dot{y}(s,0)=v_{init}(s),\quad s\in[\,0,\ell(0)\,]\,. \qquad (1.1.c)$$

Ces données initiales ne peuvent être arbitrairement choisies. En particulier, il semble naturel de considérer une position relative initiale vérifiant la condition d'inextensibilité $|\,{y_{init}}'\,|=1$. On supposera ces données initiales consistantes ; nous précisons ceci pour les équations discrétisées en espace dans la section 2.2.3. Même si elle n'apparaît pas dans les équations *(1.1)*, la vitesse initiale de filage du câble $\dot{\ell}(0)$ doit également être compatible avec les autres données initiales. Par exemple, si la vitesse relative initiale v_{init} est nulle, on doit nécessairement avoir $\dot{\ell}(0)=0$.

1.1.3. Forces de traînée hydrodynamique

La force de traînée hydrodynamique du câble dépend de la vitesse du câble par rapport au courant $\mathbf{v}(s,t)$. Cette dernière s'écrit comme différence entre la vitesse absolue du câble $\dot{y}(s,t)+\dot{u}(t)$ et la vitesse du courant ω, autrement dit

$$\mathbf{v}(s,t)=\dot{y}(s,t)+\dot{u}(t)-\omega.$$

La vitesse du courant, qui peut varier selon la profondeur et le temps, nous est inconnue. Aussi avons nous supposé qu'elle était constante, et même, dans la plupart des essais numériques, nulle. On utilise une expression de la traînée hydrodynamique par unité de longueur $f_c(s,t)$ se décomposant en termes tangentiel et normal au câble : si on note respectivement

$$\mathbf{v}_t(s,t)=\Big(\mathbf{v}(s,t)^\top y'(s,t)\Big)\,y'(s,t),\quad \mathbf{v}_n(s,t)=\mathbf{v}(s,t)-\mathbf{v}_t(s,t)$$

les composantes tangentielle et normale de la vitesse du câble par rapport au courant, alors

$$f_c(s,t)=-c_t\,|\,\mathbf{v}_t(s,t)\,|\,\mathbf{v}_t(s,t)-c_n\,|\,\mathbf{v}_n(s,t)\,|\,\mathbf{v}_n(s,t).$$

Les coefficients c_t et c_n sont constants.

D'autres expressions de f_c peuvent être employées. Damy, Joannides, Le Gland, Prévosto et Rakotozafy [35] et Marichal, Dassonville et Lefort [72], négligent la composante tangentielle de la traînée $-c_t\,|\,\mathbf{v}_t\,|\,\mathbf{v}_t$ devant sa composante normale et prennent $f_c=-c_n\,|\,\mathbf{v}_n\,|\,\mathbf{v}_n$. Les raisons invoquées sont que c_t est de l'ordre de grandeur du centième de c_n – on ne s'étonnera pas que c_t soit petit devant c_n si l'on pense à la géométrie du câble – alors qu'aux vitesses de remorquage peu élevées que l'on considère dans la suite, les normes des vitesses normale $|\,\mathbf{v}_n\,|$ et tangentielle $|\,\mathbf{v}_t\,|$ sont comparables. Dans un modèle tridimensionnel, cette simplification n'entraîne cependant pas un allégement sensible des calculs, puisque la vitesse tangentielle $\mathbf{v}_t$ doit être connue pour évaluer la vitesse normale $\mathbf{v}_n$. Sanders [95] propose une formulation plus sophistiquée, $f_c=-c_t\,|\,\mathbf{v}_t\,|\,\mathbf{v}_t-c_n\,|\,\mathbf{v}_{n_1}\,|\,\mathbf{v}_{n_1}-c_n\,|\,\mathbf{v}_{n_2}\,|\,\mathbf{v}_{n_2}$ (où n_1 et n_2 sont deux directions normales au câble), mais il semble qu'elle soit peu utilisée.

En ce qui concerne la traînée hydrodynamique de l'engin remorqué, on prend

$$f_p(t)=-c_h\,|\,\mathbf{v}_h(0,t)\,|\,\mathbf{v}_h(0,t)-c_z\,|\,\mathbf{v}_z(0,t)\,|\,\mathbf{v}_z(0,t),$$

où c_h et c_z désignent des constantes et $\mathbf{v}_h$ et $\mathbf{v}_z$ sont les composantes horizontale et verticale de la vitesse de l'engin par rapport au courant $\mathbf{v}(0,t)=\dot{y}(0,t)+\dot{u}(t)-\omega$. On constate que f_p ne possède pas la même décomposition que f_c, ce qui traduit que l'engin remorqué et le câble ont des formes très différentes.

1.1.4. Quelques remarques sur la structure des équations

La complexité d'un modèle de câble peut être développée à volonté afin de décrire la réalité le mieux possible, mais, rappelons-le, de nombreux paramètres nous sont inconnus.

Le modèle auquel on aboutit semble réaliser un compromis acceptable entre inévitables hypothèses simplificatrices et précision des résultats. De nombreuses études sont basées sur ce modèle. Mentionnons entre autres les travaux de Zajac [108] sur la pose de câbles sur les fonds sous-marins (au travers des océans), ou, dans le domaine aérien, ceux de Murray [78] sur un câble tracté par un avion et remorquant un capteur.

On aura tout d'abord noté que, la variable d'espace étant l'abscisse curviligne le long du câble, l'équation *(1.1)* est en une dimension d'espace. On remarque en outre que la dynamique du poisson se traduit comme une des conditions aux limites, l'autre condition étant que la position relative du câble est nulle au point de fixation au navire. Il n'y a pas de condition initiale sur la tension. Il n'y a pas non plus de condition aux limites fixant la valeur de T. Cependant, en l'absence de poisson, auquel cas $m_p = 0$ et $f_p(t) = 0$, la condition aux limites *(1.1.b)* à l'extrémité inférieure du câble devient $T(0,t)\, y'(0,t) = 0$. L'étude de cas simplifiés de l'équation *(1.1)* confirmera le bien-fondé de ce modèle.

Nous n'avons pas repris à notre compte l'hypothèse quasi-statique qu'utilisent Pichon [83] et Robin [93]. Selon cette hypothèse, les termes d'inertie peuvent être négligés devant les autres forces mises en jeu dans *(1.1)*. Après discrétisation en temps, les équations s'en trouvent grandement simplifiées. Toutefois, si cette approximation est justifiée lors de l'étude de mouvements lents du câble, elle nous a semblé moins pertinente dans le cadre du problème de demi-tour en temps minimal que nous traitons dans le dernier chapitre.

Les équations aux dérivées partielles sont traditionnellement réparties en trois catégories (équations dites elliptiques, paraboliques et hyperboliques; voir Courant et Hilbert [32]). La première ligne de *(1.1.a)* s'apparenterait à une équation d'évolution hyperbolique si la tension était connue et, ce qui semble naturel pour un câble en traction, positive. L'exemple le plus simple de ce type d'équations est l'équation des ondes en dimension un

$$\ddot{u}(x,t) - c^2\, u''(x,t) = 0, \quad x \in \,]\,0, \ell\,[\,,\ t > 0.$$

Avec les conditions aux limites et conditions initiales

$$u(0,t) = u(\ell,t) = 0, \quad t \geq 0, \quad u(x,0) = u_0(x), \quad \dot{u}(x,0) = u_1(x), \quad x \in \,]\,0, \ell\,[\,,$$

celle-ci modélise les petites déformations $u(x,t) \in \mathbb{R}$ d'une corde vibrante de longueur ℓ, fixée à ses extrémités et qui n'est soumise à aucune force extérieure. Une forme plus générale de l'équation des ondes, étudiée par Lions et Magenes [64], est, en dimension n,

$$\ddot{u}(x,t) - \sum_{1 \leq i,j \leq n} \frac{\partial}{\partial x_j} \left(a_{i,j}\, \frac{\partial u}{\partial x_i} \right)(x,t) = f(x,t), \quad x \in \Omega \subset \mathbb{R}^n,\ t \in \,]\,0, t_{max}\,[\,, \qquad \textit{(1.2)}$$

avec conditions aux limites et conditions initiales. Les fonctions $a_{i,j}$ sont connues et on suppose notamment l'existence d'une constante $\alpha > 0$ telle que

$$\forall\, x \in \Omega,\ \forall\, t \in [\,0, t_{max}\,],\ \forall\, \xi \in \mathbb{R}^n, \quad \sum_{1 \leq i,j \leq n} a_{i,j}(x,t)\, \xi_i\, \xi_j \ \geq\ \alpha\, |\,\xi\,|^2 .$$

La première ligne de *(1.1.a)* est à comparer à *(1.2)* lorsque $n = 1$, à la différence notable que, dans la seconde équation, on se donne une fonction a minorée par une constante strictement positive, alors que, dans la première, on cherche la tension T – en s'attendant à ce que celle-ci soit strictement positive. Par rapport à l'équation des ondes, *(1.1.a)* compte donc une inconnue et une équation, la contrainte d'inextensibilité, supplémentaires.

Une autre équation modèle se rapprochant de *(1.1.a)* est l'équation de Stokes instationnaire (voir Dautray et Lions [36])

$$\begin{cases} \dot{u}(x,t) - \nu\,\Delta u(x,t) + \nabla p(x,t) = f(x,t) \\ (\text{div } u)(x,t) = 0 \end{cases} \qquad x \in \Omega \subset \mathbb{R}^n,\ t \in \left]0, t_{max}\right[$$

avec conditions aux limites et conditions initiales

$$u(x,t) = 0, \quad x \in \partial\Omega,\ t \in \left]0, t_{max}\right[, \quad u(x,0) = u_0(x), \quad x \in \Omega,$$

décrivant le comportement d'un fluide visqueux incompressible, de vitesse $u(x,t) \in \mathbb{R}^n$ et de pression $p(x,t) \in \mathbb{R}$, contenu dans un domaine Ω où il subit une force $f(x,t)$, et adhérant à la paroi $\partial\Omega$. Les dérivées en temps que les deux équations font apparaître ne portent que sur une seule de leurs variables respectives, u et y, lesquelles sont précisément sujettes aux contraintes d'incompressibilité div $u = 0$ et d'inextensibilité $|\,y'\,| = 1$. Cette analogie entre les deux équations nous a guidé dans le choix d'un schéma de discrétisation en espace du système *(1.1)*, comme nous le verrons dans le chapitre suivant.

Dans l'immédiat, on commence par étudier certains cas particuliers et dégager quelques propriétés des équations *(1.1)*.

1.2. Etude du cas statique

En l'absence de tout mouvement, c'est-à-dire lorsque le navire est immobile et que la partie déroulée du câble, de longueur constante $\ell(t) = \ell$, pend à sa verticale, la seule force s'appliquant au système est son poids. La position relative et la tension sont alors des fonctions $y(s) \in \mathbb{R}$ et $T(s) \in \mathbb{R}$ qui ne dépendent plus du temps. On considère alors le repère dont l'origine est le navire et dont le seul vecteur est unitaire ascendant. Dans ce repère, les équations *(1.1.a)* se récrivent

$$\begin{cases} (T\,y')'(s) - m_c\,|\,g\,| = 0 \\ |\,y'(s)\,| = 1, \end{cases} \qquad s \in \left]0, \ell\right[, \tag{1.3.a}$$

tandis que les conditions aux limites *(1.1.b)* deviennent

$$y(\ell) = 0, \quad T(0)\,y'(0) - m_p\,|\,g\,| = 0. \tag{1.3.b}$$

Sous la condition supplémentaire que la tension soit positive, on vérifie que ces équations admettent une unique solution. En effet, on a nécessairement

$$T(s)\,y'(s) = m_c\,|\,g\,|\,s + T(0)\,y'(0) = (m_c\,s + m_p)\,|\,g\,| > 0,$$

ce qui implique que $y'(s) > 0$, et par suite $y'(s) = 1$, $y(s) = s - \ell$ et $T(s) = (m_c\, s + m_p)\ |g|$. En revanche, si le signe de T n'est pas imposé, il n'y a pas unicité de la solution ; par exemple, si (y, T) est une solution, $(-y, -T)$ en est une autre.

Sans utiliser la connaissance explicite de solution, on va étudier, à la lumière de résultats classiques d'optimisation, plusieurs problèmes en relation avec l'équation statique (*1.3*). On en donnera une formulation variationnelle, celle-ci nous guidant ensuite dans le choix d'une formulation variationnelle pour le cas dynamique. On précisera dans quels types d'espaces fonctionnels il convient de chercher la position relative y et la tension T. Le rôle particulier joué par cette dernière sera également mis en évidence.

1.2.1. Minimisation de fonctionnelle d'énergie

Dans de fréquentes situations, la solution d'un problème physique réalise le minimum d'une certaine fonctionnelle, dite d'énergie, sur un espace de fonctions admissibles. Souvent, la fonctionnelle d'énergie est convexe coercive et l'ensemble admissible convexe fermé. L'existence de ce minimum est alors immédiate. On considère ici l'énergie potentielle de pesanteur du système câble – poisson, définie par

$$J(z) = m_c\ |g| \int_0^\ell z(s)\, ds + m_p\ |g|\ z(0),$$

sur l'ensemble des configurations admissibles

$$K = \left\{ z \in W^{1,\infty}(0,\ell) :\ z(\ell) = 0,\ \left|\, z' \,\right| = 1\ p.p. \right\}.$$

Pour tout $1 \leq p \leq +\infty$, on note $W^{1,p}(0,\ell)$ l'espace de Sobolev des fonctions $u \in L^p(0,\ell)$ admettant une dérivée au sens des distributions $u' \in L^p(0,\ell)$, c'est-à-dire une fonction telle que

$$\forall\, \varphi \in C_c^1\,(\,]\,0,\ell\,[\,)\,, \quad \int_0^\ell u(s)\, \varphi'(s)\, ds = -\int_0^\ell u'(s)\, \varphi(s)\, ds.$$

Lorsque $u \in W^{1,p}(0,\ell)$ est dérivable au sens usuel, sa dérivée et sa dérivée au sens des distributions coïncident, d'où la notation u'. On démontre (par exemple, dans Brézis [18]) que l'injection canonique $W^{1,p}(0,\ell) \subset C\,(\,[\,0,\ell\,]\,)$ est continue, c'est-à-dire que les fonctions de $W^{1,p}(0,\ell)$ sont continues et qu'il existe une constante $C > 0$ telle que

$$\forall\, z \in W^{1,p}(0,\ell), \quad ||\, z\,||_{L^\infty} \leq C\ ||\, z\,||_{W^{1,p}}\,. \tag{1.4}$$

L'ensemble $W^{1,\infty}(0,\ell)$ est donc l'espace des fonctions continues sur $[\,0,\ell\,]$ ayant une dérivée (au sens des distributions) bornée sur presque tout $]\,0,\ell\,[$. Dans la définition de K, l'écriture $z(\ell) = 0$ a bien un sens et l'égalité $|\,z'\,| = 1$ s'entend presque partout. Le recours aux espaces de Sobolev est très fréquent lors de la résolution d'équations aux dérivées partielles. On justifiera le choix spécifique de $W^{1,\infty}(0,\ell)$ dans la section 1.2.3.

On veut montrer que J atteint son minimum sur K. Mais J est linéaire donc n'est pas coercive et K n'est pas convexe. Pour montrer que ce problème a une solution, on le convexifie : on minimise J sur l'ensemble

$$K^{\#} = \left\{ z \in W^{1,\infty}(0,\ell) : \, z(\ell) = 0, \, \left| z' \right| \leq 1 \right\}$$

qui est convexe et contient K.

Proposition 1.1. *Le problème*

$$\begin{array}{c} \textit{minimiser } J(z) \\ \textit{sous la contrainte } z \in K^{\#} \end{array} \tag{1.5}$$

possède au moins une solution.

Démonstration. On fixe $p \in \left]1, +\infty\right[$ et on utilise le fait que

$$K^{\#} \subset W^{1,\infty}(0,\ell) \subset W^{1,p}(0,\ell).$$

A la différence de $W^{1,\infty}(0,\ell)$, l'espace $W^{1,p}(0,\ell)$ est réflexif. Ses sous-ensembles convexes, fermés et bornés sont donc faiblement compacts.

L'ensemble admissible $K^{\#}$ est clairement non vide, convexe, borné. Vérifions qu'il est fermé. On considère pour cela une suite convergente $(z_n) \subset K^{\#}$, de limite $z \in W^{1,p}(0,\ell)$. D'une part, du fait de l'injection continue $W^{1,p}(0,\ell) \subset C\left(\left[0,\ell\right]\right)$, les applications z_n et z sont continues et (z_n) converge donc vers z ponctuellement. Par conséquent, $z(\ell) = 0$. D'autre part, comme (z_n') converge vers (z') dans $L^p(0,\ell)$, on peut extraire une sous-suite de (z_n') qui converge presque partout vers z', ce qui implique que $|\, z' \,| \leq 1$ presque partout. On en déduit que la limite $z \in K^{\#}$ et que $K^{\#}$ est fermé dans $W^{1,p}(0,\ell)$. Il s'agit donc d'un sous-ensemble faiblement compact de $W^{1,p}(0,\ell)$.

Le critère J est linéaire et l'inégalité *(1.4)* implique que, pour toute fonction $z \in W^{1,p}(0,\ell)$,

$$|\, J(z) \,| \leq m_c \, |\, g \,| \, \ell \, \| z \|_{L^\infty} + m_p \, |\, g \,| \, \| z \|_{L^\infty} \leq C \left(m_c \, \ell + m_p \right) |\, g \,| \, \| z \|_{W^{1,p}} .$$

Ceci prouve que J est linéaire continu, donc faiblement continu sur $W^{1,p}(0,\ell)$. Il en découle que *(1.5)* a une solution, puisque toute fonction réelle continue atteint ses bornes sur un compact. □

L'existence d'au moins une solution au problème convexifié étant établie, on peut prouver que le problème non convexe consistant à minimiser J sur K a lui aussi au moins une solution.

Proposition 1.2. *Toute solution du problème (1.5) appartient à K et est donc solution du problème*

$$\begin{array}{c} \textit{minimiser } J(z) \\ \textit{sous la contrainte } z \in K. \end{array} \tag{1.6}$$

Démonstration. Soit y une solution de *(1.5)*. Puisque $y \in K^{\#}$, il nous faut démontrer que $|\,y'\,| = 1$ presque partout. Supposons qu'existe un sous-ensemble non négligeable $E \subset]\,0,\ell\,[$ sur lequel $|\,y'\,| < 1$. On va alors contredire l'optimalité de y pour *(1.5)*, en prouvant l'existence d'une fonction $y_0 \in K^{\#}$ telle que $J(y_0) < J(y)$.

On définit pour cela la fonction

$$\alpha = \frac{1}{m_c\,|\,g\,|}\left(1 - |\,y'\,|\right) \in L^\infty(0,\ell).$$

Elle est positive sur presque tout $]\,0,\ell\,[$ et strictement positive sur E. Les primitives de α étant continues, d'après le théorème des valeurs intermédiaires, on peut trouver $s_0 \in]\,0,\ell\,[$ tel que

$$\int_0^{s_0} \alpha(s)\,ds = \int_{s_0}^{\ell} \alpha(s)\,ds = \frac{1}{2}\int_0^{\ell} \alpha(s)\,ds \geq \frac{1}{2}\int_E \alpha(s)\,ds > 0.$$

Soit $\alpha_0 \in L^\infty(0,\ell)$ la fonction coïncidant avec α sur $[\,0,s_0\,[$ et $-\alpha$ sur $[\,s_0,\ell\,]$. On considère $\beta \in W^{1,\infty}(0,\ell)$ la primitive de cette fonction α_0 s'annulant en 0. β possède un certain nombre de propriétés qui nous seront utiles: $\beta(0) = \beta(\ell) = 0$; $\beta(s_0) > 0$, β est continue et positive donc

$$\int_0^{\ell} \beta(s)\,ds > 0;$$

enfin, $|\,\beta'\,| = |\,\alpha_0\,| = \alpha$.

On considère, pour conclure, la fonction $y_0 = y - m_c\,|\,g\,|\,\beta$. Il est clair que $y_0 \in W^{1,\infty}(0,\ell)$ et $y_0(\ell) = 0$. Pour presque tout $s \in [\,0,\ell\,]$, par définition de α,

$$\left|\,y_0'(s)\,\right| = \left|\,y'(s) - m_c\,|\,g\,|\,\alpha_0(s)\,\right| \leq \left|\,y'(s)\,\right| + m_c\,|\,g\,|\,\alpha = 1.$$

On a donc que $y_0 \in K^{\#}$. Par ailleurs, puisque $y_0(0) = y(0)$,

$$J(y) - J(y_0) = m_c\,|\,g\,|\int_0^{\ell} (y(s) - y_0(s))\,ds = {m_c}^2\,|\,g\,|^2 \int_0^{\ell} \beta(s)\,ds > 0.$$

On a atteint notre but, qui était de contredire l'optimalité de y pour *(1.5)*. □

Grâce à l'argument selon lequel toute solution y de *(1.5)* vérifie $|\,y'\,| = 1$ presque partout, on prouve finalement qu'il est équivalent de minimiser J sur $K^{\#}$ ou sur K.

Proposition 1.3. *Les problèmes (1.5) et (1.6) ont la même unique solution.*

Démonstration. La première étape consiste à vérifier que toute solution de *(1.6)* est solution de *(1.5)*. Considérons $y \in K$ minimisant J sur K. Si y ne minimisait pas J sur $K^{\#}$,

$$\exists\, y_1 \in K^{\#} : J(y_1) < J(y).$$

Mais alors, la proposition précédente impliquerait

$$\exists\, y_2 \in K : J(y_2) \leq J(y_1) < J(y),$$

ce qui est impossible. C'est donc que les solutions de *(1.6)* sont solutions de *(1.5)* et, pour résumer, que les deux problèmes ont les mêmes solutions.

Le problème *(1.5)* étant convexe, l'ensemble de ses solutions est également convexe. Soient $y_1, y_2 \in K^{\#}$ deux de ses solutions. Alors $\frac{1}{2}(y_1 + y_2)$ est également solution et comme l'affirme la proposition 1.2,

$$\left| y_1' \right| = \left| y_2' \right| = \left| \frac{y_1' + y_2'}{2} \right| = 1$$

sur presque tout $]0, \ell[$. Cependant, la stricte convexité de l'application $x \mapsto |x|^2$ sur $\mathbb{R}$ implique que, en chaque s où $y_1'(s) \neq y_2'(s)$,

$$\left| \frac{y_1'(s) + y_2'(s)}{2} \right|^2 < \frac{1}{2} \left(\left| y_1'(s) \right|^2 + \left| y_2'(s) \right|^2 \right) = 1.$$

On en déduit successivement que $y_1' = y_2'$ presque partout, et que $y_1 = y_2$, compte-tenu de $y_1(\ell) = y_2(\ell) = 0$; c'est-à-dire que *(1.5)* a une unique solution. □

On a vérifié qu'il existe une unique configuration du câble réalisant le minimum de la fonctionnelle d'énergie J aussi bien sur l'ensemble des configurations admissibles K que sur $K^{\#}$. On veut maintenant énoncer des conditions d'optimalité pour *(1.5)* et *(1.6)*. Dans ce but, il est nécessaire de reformuler ces deux problèmes comme suit.

On considère les cônes négatif et positif de $L^\infty(0, \ell)$

$$L^\infty(0,\ell)^- = \{z \in L^\infty(0,\ell) :\ z \leq 0 \ \ p.p.\}, \quad L^\infty(0,\ell)^+ = \{z \in L^\infty(0,\ell) :\ z \geq 0 \ \ p.p.\}$$

et la fonctionnelle f définie de $W^{1,\infty}(0,\ell)$ dans $\mathbb{R} \times L^\infty(0,\ell)$ par

$$f(z) = \Big(f_1(z), f_2(z) \Big) = \left(z(\ell), \frac{1}{2} \left(\left| z' \right|^2 - 1 \right) \right). \tag{1.7}$$

Les problèmes *(1.5)* et *(1.6)* sont respectivement équivalents à

$$\begin{array}{c} \text{minimiser } J(z) \\ \text{sous la contrainte } f(z) \in \{0\} \times L^\infty(0,\ell)^-, \end{array} \tag{1.8}$$

et minimiser $J(z)$ sous la contrainte $f(z) = 0_{\mathbb{R} \times L^\infty(0,\ell)}$.

1.2.2. Conditions d'optimalité de problèmes sous contraintes abstraites

On rappelle dans cette section quelques résultats généraux d'optimisation sous contraintes abstraites, cités notamment par Bonnans et Shapiro [14, 15]. Ces résultats seront ensuite appliqués au problème *(1.8)*.

Soient X, Y deux espaces de Banach, Y' l'espace dual de Y et K un sous-ensemble convexe fermé de Y. On se donne des applications continûment différentiables $\phi : X \to \mathbb{R}$ et $f : X \to Y$ et on considère le problème sur X

$$\begin{array}{c} \text{minimiser } \phi(x) \\ \text{sous la contrainte } f(x) \in K. \end{array} \tag{1.9}$$

Son système d'optimalité du premier ordre est

$$\phi'(x) + \langle\, \lambda\,,\, f'(x)\,\rangle = 0, \quad f(x) \in K, \quad \lambda \in N_K\,(f(x)). \tag{1.10}$$

On a noté $\langle\, .\,,\, .\,\rangle$ le crochet de dualité entre Y' et Y et $N_K\,(f(x))$ le cône normal à K au point $f(x)$, donné par

$$N_K\,(f(x)) = \{\lambda \in Y' : \forall\, z \in K, \langle\, \lambda\,,\, z - f(x)\,\rangle \leq 0\}.$$

Théorème 1.4. *Soit x une solution locale de (1.9). Sous l'hypothèse*

$$0 \in \textit{ intérieur}\,(f(x) + f'(x)\,X - K), \tag{1.11}$$

il existe au moins un multiplicateur de Lagrange λ tel que (x, λ) satisfasse le système d'optimalité du premier ordre (1.10).

L'hypothèse de qualification *(1.11)* a été formulée par Robinson [94]. Penot [81] démontre que lorsque K est d'intérieur non vide, cette condition est équivalente à

$$\exists\, h \in X : \quad f(x) + f'(x)\,h \in \text{ intérieur}\,(K).$$

Le cadre qui nous intéresse, dans lequel entre la contrainte du problème *(1.8)*, est celui de la proposition suivante.

Proposition 1.5. *Supposons que Y soit le produit cartésien de deux espaces de Banach Y_1 et Y_2 et que $K = \{0_{Y_1}\} \times K_2$, le sous-ensemble $K_2 \subset Y_2$ étant convexe fermé et d'intérieur non vide. Dans ce cas, $f = (f_1, f_2)$ – avec pour chaque $i = 1, 2$, $f_i : X \to Y_i$. La condition (1.11) est alors équivalente, en $x \in X$, à*

$$\left\{ \begin{array}{l} {f_1}'(x) \textit{ est surjective,} \\ \exists\, h \in X : \quad {f_1}'(x)\,h = 0, \quad f_2(x) + {f_2}'(x)\,h \in \textit{intérieur}\,(K_2). \end{array} \right. \tag{1.12}$$

Démonstration. On considère x admissible, c'est-à-dire tel que $f_1(x) = 0$ et $f_2(x) \in K_2$. Soient B_{Y_1} et B_{Y_2} les boules unités des espaces Y_1 et Y_2. La condition *(1.11)* signifie qu'il existe $\varepsilon > 0$ tel que

$$\forall\, y_1 \in \varepsilon B_{Y_1}, \forall\, y_2 \in \varepsilon B_{Y_2}, \exists\, h \in X : \left\{ \begin{array}{l} y_1 = {f_1}'(x)\,h \\ y_2 \in f_2(x) + {f_2}'(x)\,h - K_2. \end{array} \right. \tag{1.13}$$

On en déduit que ${f_1}'(x)$ est surjective, et par ailleurs, en prenant $y_1 = 0$, que

$$\forall\, y_2 \in \varepsilon B_{Y_2}, \ \exists\, h \in \ker\left({f_1}'(x)\right) : \ y_2 \in f_2(x) + {f_2}'(x)\,h - K_2.$$

Il s'ensuit que $0 \in$ intérieur $\big[f_2(x) + f_2'(x)\ \ker\big(f_1'(x)\big) - K_2\big]$ et, d'après la remarque précédent la proposition, que

$$\exists\, h \in \ker\big(f_1'(x)\big) :\ f_2(x) + f_2'(x)\, h \in \text{ intérieur}\,(K_2).$$

Supposons réciproquement que la condition *(1.12)* soit satisfaite. Soient $\varepsilon > 0$, $y_1 \in \varepsilon B_{Y_1}$ et $y_2 \in \varepsilon B_{Y_2}$. Puisque $f_1'(x)$ est surjective, il existe $h_1 \in X$ tel que $y_1 = f_1'(x)\, h_1$. De plus, $f_1'(x)$ étant continue, selon le théorème de l'application ouverte, il existe une constante $C > 0$ telle que l'on puisse choisir h_1 dans $C\varepsilon B_X$. Il existe également $h_2 \in X$ tel que $0 = f_1'(x)\, h_2$ et $f_2(x) + f_2'(x)\, h_2$ soit intérieur à K_2. Remarquons que $y_2 - f_2'(x)\, h_1$ tend vers 0_{Y_2} quand ε tend vers 0; en effet, $y_2 \in \varepsilon B_{Y_2}$, $h_1 \in C\varepsilon B_X$ et $f_2'(x)$ est continue. D'autre part, $f_2(x) + f_2'(x)\, h_2 - K_2$ contient une boule δB_{Y_2}, avec $\delta > 0$, puisque 0_{Y_2} est intérieur à $f_2(x) + f_2'(x)\, h_2 - K_2$. C'est donc que pour ε assez petit,

$$y_2 - f_2'(x)\, h_1 \in f_2(x) + f_2'(x)\, h_2 - K_2.$$

En conclusion, si ε est suffisamment petit, on a trouvé $h = h_1 + h_2$ satisfaisant les conditions équivalentes *(1.13)* et *(1.11)*. □

Dans le contexte de la proposition 1.5, le cône normal $N_K\,(f(x))$ en x admissible est

$$\Big\{(\lambda_1, \lambda_2) \in {Y_1}' \times {Y_2}' :\ \forall\, z_2 \in K_2,\ \langle\, \lambda_1\,,\, 0\,\rangle_{{Y_1}',Y_1} + \langle\, \lambda_2\,,\, z_2 - f_2(x)\,\rangle_{{Y_2}',Y_2} \le 0\Big\},$$

autrement dit ${Y_1}' \times N_{K_2}\,(f_2(x))$. Les conditions d'optimalité *(1.10)* deviennent

$$\left\{\begin{array}{l} \phi'(x) + \langle\, \lambda_1\,,\, f_1'(x)\,\rangle_{{Y_1}',Y_1} + \langle\, \lambda_2\,,\, f_2'(x)\,\rangle_{{Y_2}',Y_2} = 0 \\ f_1(x) = 0, \quad f_2(x) \in K_2 \\ \lambda_1 \in {Y_1}', \quad \lambda_2 \in N_{K_2}\,(f_2(x)). \end{array}\right. \tag{1.14}$$

On peut à présent revenir au problème de la minimisation de l'énergie du câble écrit sous la forme *(1.8)*.

1.2.3. Conditions d'optimalité des problèmes de minimisation d'énergie

Le cône négatif $L^\infty(0,\ell)^-$ est un sous-ensemble convexe de $L^\infty(0,\ell)$. On vérifie aisément qu'il est fermé et d'intérieur non vide. Démontrons que la contrainte du problème *(1.8)* vérifie les conditions de régularité et de qualification nécessaires.

Lemme 1.6. *La fonctionnelle $f : W^{1,\infty}(0,\ell) \to \mathbb{R} \times L^\infty(0,\ell)$ définie par la relation (1.7) est continûment différentiable et vérifie la condition de qualification (1.12) en tout $z \in K^\#$.*

Démonstration. Nous commençons par vérifier que f est régulière. Soit $z \in W^{1,\infty}(0,\ell)$. La différentielle $f_2'(z)$ est l'application linéaire continue

$$f_2'(z)\ :\ \xi \in W^{1,\infty}(0,\ell) \mapsto z'\xi' \in L^\infty(0,\ell).$$

L'ensemble $\mathcal{L}$ des applications linéaires continues de $W^{1,\infty}(0,\ell)$ dans $L^\infty(0,\ell)$ est muni de la norme

$$\|F\|_{\mathcal{L}} = \sup_{\|\xi\|_{W^{1,\infty}}\leq 1} \|F(\xi)\|_{L^\infty}.$$

Pour tous $z_1, z_2 \in W^{1,\infty}(0,\ell)$, on a

$$\left\| f_2'(z_2) - f_2'(z_1) \right\|_{\mathcal{L}} = \sup_{\|\xi\|_{W^{1,\infty}}\leq 1} \left\| (z_2' - z_1')\xi' \right\|_{L^\infty} \leq \left\| z_2' - z_1' \right\|_{L^\infty} \leq \| z_2 - z_1 \|_{W^{1,\infty}}.$$

Ceci prouve que f_2' est continue (et même lipschitzienne) sur $W^{1,\infty}(0,\ell)$ et que f_2 est continûment différentiable. f_1 et J le sont également, car linéaires continues : $f_1'(z) = f_1$ et $J'(z) = J$.

Vérifions maintenant que la condition (*1.12*) est satisfaite. On fixe $z \in K^\#$. Il est clair que $f_1'(z) = f_1$ est surjective. D'autre part, si on prend $h = -z$, alors $f_1'(z)\,h = -z(\ell) = 0$ et

$$f_2(z) + f_2'(z)\,h = \frac{1}{2}\left|z'\right|^2 - \frac{1}{2} + z'h' = -\frac{1}{2}\left(1 + \left|z'\right|^2\right) \in \text{intérieur}\,(L^\infty(0,\ell)^-).$$

Ceci prouve que la contrainte $f(z) \in \{0\} \times L^\infty(0,\ell)^-$ est qualifiée en tout $z \in K^\#$. □

Soulignons l'importance du choix des espaces fonctionnels. On aurait pu fixer $p \in [\,2, +\infty\,[$ et considérer f comme application de $W^{1,p}(0,\ell)$ dans $\mathbb{R} \times L^{\frac{p}{2}}(0,\ell)$. Les applications f_1 et f_2 restent continûment différentiables et $f_1'(z) = f_1$ reste surjective. Mais la contrainte serait devenue

$$f_2(z) \in K_2 = \left\{ z \in L^p(0,\ell) : z' \leq 0 \ p.p. \right\}.$$

Or K_2 est d'intérieur vide**!** Le choix de $p = +\infty$ est donc primordial si l'on veut appliquer le théorème 1.4. On dispose maintenant de tous les éléments nécessaires pour écrire une condition d'optimalité pour le problème (*1.5*)**.**

Proposition 1.7. *Une fonction $y \in K$ est la solution de (1.5) si et seulement s'il existe $T \in L^\infty(0,\ell)^+$ tel que*

$$\forall\, z \in W^{1,\infty}(0,\ell) : z(\ell) = 0, \quad J(z) + \int_0^\ell T(s)\, y'(s)\, z'(s)\, ds = 0. \qquad (1.15)$$

Cette fonction T est définie de manière unique par (1.15).

Démonstration. Pour vérifier que la condition est suffisante, on admet l'existence de $T \in L^\infty(0,\ell)^+$ satisfaisant (*1.15*). Alors, pour tous $z \in K^\#$,

$$J(y) - J(z) = \int_0^\ell T(s)\, \left(y'(s)\, z'(s) - y'(s)\, y'(s)\right) ds = \int_0^\ell T(s)\, \left(y'(s)\, z'(s) - 1\right) ds \leq 0.$$

Donc y est la solution de (*1.5*). On justifie que la condition est nécessaire en plusieurs étapes.

La contrainte $f(z) \in \{0\} \times L^\infty(0,\ell)^-$ est qualifiée en y (lemme 1.6) donc le théorème 1.4 s'applique. Il existe un multiplicateur de Lagrange $(T_1, T_2) \in \mathbb{R} \times L^\infty(0,\ell)'$ solution avec y du système d'optimalité *(1.14)*, qui se récrit

$$\begin{cases} J'(y) + \langle\, T_1\,,\, f_1\,\rangle_{\mathbb{R},\mathbb{R}} + \langle\, T_2\,,\, f_2{}'(y)\,\rangle = 0 \\ T_2 \in N_{L^\infty(0,\ell)^-}(0). \end{cases}$$

Pour alléger les notations, on a noté le crochet de dualité entre $L^\infty(0,\ell)'$ et $L^\infty(0,\ell)$ simplement $\langle\, .\,,\, .\,\rangle$. On a aussi utilisé le fait que ${f_1}'(y) = f_1$ et $f_2(y) = 0$. La première ligne donne

$$\forall\, z \in W^{1,\infty}(0,\ell) :\ z(\ell) = 0, \quad J(z) + \langle\, T_2\,,\, y'z'\,\rangle = 0. \tag{1.16}$$

Dire que $T_2 \in N_{L^\infty(0,\ell)^-}(0)$ signifie que

$$\forall\, z \in L^\infty(0,\ell)^-, \quad \langle\, T_2\,,\, z\,\rangle \le 0. \tag{1.17}$$

Examinons la régularité du multiplicateur. On montre tout d'abord que $T_2 \in L^\infty(0,\ell)'$ est en fait continue sur $L^\infty(0,\ell)$ pour la norme $L^1(0,\ell)$. Fixons $\xi \in L^\infty(0,\ell)$. Soit z la primitive de $\xi\, y' \in L^\infty(0,\ell)$ s'annulant en 0. Alors, compte-tenu de $|\, y'\,| = 1$,

$$\forall\, s \in [\,0,\ell\,], \quad |\, z(s)\,| \le \int_0^s \big|\, z'(\sigma)\,\big|\, d\sigma = \int_0^s |\, \xi(\sigma)\,|\, d\sigma \le \|\, \xi\,\|_{L^1}$$

et $\|\, z\,\|_{L^1} \le \ell\, \|\, \xi\,\|_{L^1}$. L'égalité *(1.16)* à appliquée $z \in W^{1,\infty}(0,\ell)$ donne

$$|\, \langle\, T_2\,,\, \xi\,\rangle\,| = |\, J(z)\,| \le m_c\, |\, g\,|\, \|\, z\,\|_{L^1} + m_p\, |\, g\,|\, |\, z(0)\,| \le (m_c\,\ell + m_p)\, |\, g\,|\, \|\, \xi\,\|_{L^1}\,.$$

La continuité de T_2 pour la norme $L^1(0,\ell)$ en découle. Or on sait que toute application linéaire continue d'un sous-espace X_0 dense dans X vers un espace complet Y se prolonge en une unique application linéaire continue sur X. $L^\infty(0,\ell)$ est dense dans $L^1(0,\ell)$ et donc T_2 se prolonge en une application linéaire $\widetilde{T_2}$ continue sur $L^1(0,\ell)$. Ce prolongement $\widetilde{T_2} \in L^1(0,\ell)'$ s'identifie à un élément $T \in L^\infty(0,\ell)$ grâce à l'égalité

$$\langle\, T_2\,,\, z\,\rangle = \langle\, \widetilde{T_2}\,,\, z\,\rangle_{L^1(0,\ell)',L^1(0,\ell)} = \int_0^\ell T(s)\, z(s)\, ds$$

valable pour toute fonction $z \in L^\infty(0,\ell)$. La condition *(1.15)* découle de *(1.16)* et *(1.17)* devient

$$\forall\, z \in L^\infty(0,\ell)^-, \quad \int_0^\ell T(s)\, z(s)\, ds \le 0.$$

Ecrivons $T = T^+ - T^-$, où les fonctions $T^+ = \max(T,0)$ et $T^- = \min(T,0)$ appartiennent à $L^\infty(0,\ell)^+$ et sont telles que $T^+T^- = 0$. Le signe de T s'obtient en prenant $z = -T^-$ dans l'inégalité ci-dessus, ce qui donne

$$0 \ge -\int_0^\ell (T^+(s) - T^-(s))\, T^-(s)\, ds = \int_0^\ell T^-(s)^2\, ds \ge 0,$$

et par suite $T^- = 0$, soit $T \in L^\infty(0,\ell)^+$.

Il ne nous reste qu'à prouver l'unicité du multiplicateur. Considérons $T_1, T_2 \in L^\infty(0,\ell)^+$ satisfaisant la condition *(1.15)*. Alors

$$\forall\, z \in W^{1,\infty}(0,\ell) : z(\ell) = 0, \qquad \int_0^\ell (T_1 - T_2)(s)\, y'(s)\, z'(s)\, ds = 0.$$

Soit $\xi \in L^\infty(0,\ell)$; on peut appliquer cette égalité à la primitive z de $\xi\, y' \in L^\infty(0,\ell)$ s'annulant en ℓ. Comme $y'\, z' = \xi$, ceci donne

$$\forall \xi \in L^\infty(0,\ell), \qquad \int_0^\ell (T_1 - T_2)(s)\, \xi(s)\, ds = 0,$$

ce qui montre que $T_1 = T_2$ et achève cette démonstration. □

Les deux problèmes *(1.5)* et *(1.6)* ayant la même unique solution, les conclusions de la proposition 1.7 s'appliquent également à la solution de *(1.6)*. Mais on constate qu'écrire des conditions d'optimalité pour *(1.5)* apporte plus d'informations. Ce problème est convexe, donc on obtient une condition nécessaire et suffisante. De plus, prendre en compte la contrainte d'inégalité $|\, y' \,| \leq 1$ permet de préciser le signe du multiplicateur, ce qui n'est pas possible si l'on ne considère que la contrainte d'égalité $|\, y' \,| = 1$.

On a noté T le multiplicateur associé à la contrainte d'inextensibilité, tout comme la tension dans l'équation statique *(1.3)*. La proposition suivante montre qu'il s'agit bien de la même fonction.

Proposition 1.8. *Le couple $(y,T) \in K \times L^\infty(0,\ell)$ vérifie la condition d'optimalité (1.15) si et seulement si (y,T) est solution de l'équation statique (1.3).*

Démonstration. Si (y,T) est solution de *(1.3)*, alors $Ty' \in C^\infty([\,0,\ell\,])$ et

$$\forall\, z \in W^{1,\infty}(0,\ell) : z(\ell) = 0, \quad \int_0^\ell (Ty')'(s)\, z(s)\, ds - m_c\, |\, g\,| \int_0^\ell z(s)\, ds = 0;$$

on obtient *(1.15)* après intégration par parties. Réciproquement, si (y,T) satisfait *(1.15)*, il vient tout d'abord

$$\forall\, z \in C_c^1\,([\,0,\ell\,])\,, \quad \int_0^\ell T(s)\, y'(s)\, z'(s)\, ds = -m_c\, |\, g\,| \int_0^\ell z(s)\, ds.$$

Il en découle que $T\, y' \in W^{1,\infty}(0,\ell)$ et $(T\, y')' = m_c\, |\, g\,|$; on a même $Ty' \in C^\infty([\,0,\ell\,])$. Par suite, on peut intégrer *(1.15)* par parties, ce qui aboutit à

$$\forall\, z \in W^{1,\infty}(0,\ell) : z(\ell) = 0, \quad \big(\, m_p\, |\, g\,| - T(0)\, y'(0)\,\big)\; z(0) = 0.$$

Il vient $T(0)\, y'(0) = m_p\, |\, g\,|$ et on a vérifié que (y,T) est solution de *(1.3)*. □

Résumons les différents points étudiés dans cette section consacrée à l'équation *(1.3)*. Sous la condition que la tension soit positive, cette équation admet une unique solution (y, T). La proposition 1.8 montre qu'il s'agit de l'unique couple satisfaisant la condition nécessaire et suffisante d'optimalité *(1.15)*. Il découle ensuite de la proposition 1.7 que y est l'unique solution des problèmes *(1.6)* et *(1.5)*.

On a donc vérifié que la configuration du câble y obtenue en résolvant l'équation *(1.3)* réalise le minimum de l'énergie de pesanteur du système câble – engin sur l'ensemble des configurations admissibles, soumises en particulier à la contrainte d'inextensibilité $|\, z' \,| = 1$. On a également justifié que la tension T est le multiplicateur associé à cette contrainte d'inextensibilité.

On achève l'étude de l'équation statique en en donnant une formulation variationnelle. Il est raisonnable de chercher y dans

$$\mathcal{Y} = \{z \in W^{1,\infty}(0,\ell) : \, z(\ell) = 0\} \tag{1.18}$$

et T dans $L^\infty(0,\ell)$. On peut déterminer (y, T) grâce à la condition d'optimalité *(1.15)* et à la condition d'inextensibilité $f_2(y) = 0$, qui est équivalente à

$$\forall\, \tau \in L^\infty(0,\ell), \quad \frac{1}{2}\int_0^\ell f_2(y)(s)\,\tau(s)\,ds = \frac{1}{2}\int_0^\ell \Big(|\, y'(s) \,|^2 - 1 \Big)\,\tau(s)\,ds = 0.$$

On propose donc la formulation variationnelle suivante : chercher $(y, T) \in \mathcal{Y} \times L^\infty(0,\ell)$ tel que

$$\forall\, z \in \mathcal{Y}, \quad \int_0^\ell T(s)\, y'(s)\, z'(s)\, ds = -m_c\, |\, g \,| \int_0^\ell z(s)\, ds - m_p\, |\, g \,|\, z(0) \tag{1.19.a}$$

$$\forall\, \tau \in L^\infty(0,\ell), \quad \frac{1}{2}\int_0^\ell \Big(|\, y'(s) \,|^2 - 1 \Big)\,\tau(s)\,ds = 0. \tag{1.19.b}$$

D'après la proposition 1.8, les équations *(1.3)* et *(1.19)* ont la même unique solution.

1.3. Autres cas particuliers

Il est possible de justifier l'existence de solution d'autres équations se déduisant de *(1.1)*. On s'intéresse dans la suite à la situation où le câble est en équilibre dans un courant constant quelconque, puis à la situation où il se déroule verticalement sous l'action de son propre poids.

1.3.1. Le cas stationnaire

On suppose ici que la vitesse du navire est constante et non nulle (on note v_{nv} son module) et que le câble, de longueur constante $\ell(t) = \ell$, a atteint une position d'équilibre stable. La position relative et la tension du câble sont des fonctions $y(s) \in \mathbb{R}^3$ et $T(s) \in \mathbb{R}$

indépendantes du temps. Les forces à prendre en compte sont les poids du câble et de l'engin remorqué et leurs traînées hydrodynamiques $f_c(s)$ et f_p. Il s'agit de résoudre

$$\begin{cases} (T\,y')'(s) + f_c(s) + m_c\,g = 0 \\ |\,y'(s)\,| = 1 \end{cases} \qquad s \in]\,0, \ell\,[\qquad (1.20.a)$$

avec les conditions aux limites

$$y(\ell) = 0, \quad T(0)\,y'(0) + f_p + m_p\,g = 0. \qquad (1.20.b)$$

On obtient ces équations en annulant dans *(1.1)* les dérivées temporelles. On s'est placé dans un repère orthonormé inertiel, lié au navire, composé de deux vecteurs horizontaux et d'un vecteur vertical ascendant. Par simplicité, le premier des vecteurs horizontaux est choisi colinéaire et de même sens que la vitesse du navire, dont les coordonnées sont par conséquent $(v_{nv}, 0, 0)$. Celles de la vitesse du courant sous-marin sont $(\omega_1, \omega_2, \omega_3)$.

Il est connu qu'en régime stationnaire, la configuration du câble est la solution d'une équation différentielle. Ce résultat est utilisé par Marichal [71], qui cite les travaux plus anciens de Pode [86]. Nous allons brièvement montrer comment cette équation différentielle s'obtient à partir de *(1.20)*.

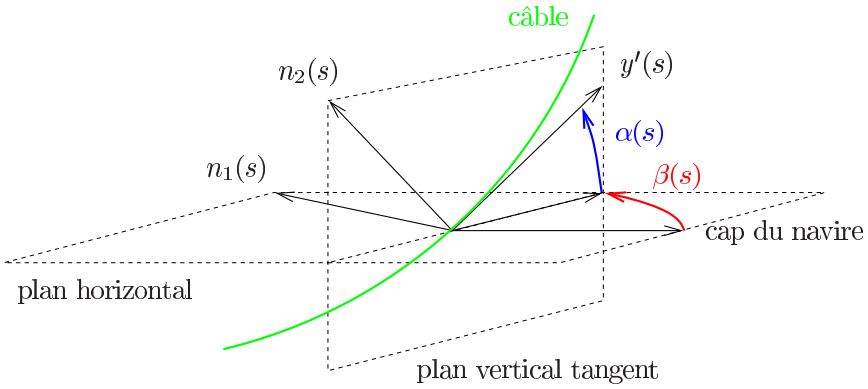


FIG. 1.2 – *angles $\alpha(s)$ et $\beta(s)$, vecteurs normaux $n_1(s)$ et $n_2(s)$.*

Ses inconnues sont, en plus de la tension $T(s)$, les angles $\alpha(s)$ entre $y'(s)$ et l'horizontale et $\beta(s)$ entre le cap du navire et le plan vertical tangent au câble. Ces angles, représentés sur la figure 1.2, sont respectivement appelés angles d'élévation et d'azimuth. Pour que le plan vertical tangent soit unique, et donc que $\beta(s)$ soit bien défini, nous supposons que le câble n'est vertical en aucun point, c'est-à-dire

$$\forall\, s \in [\,0, \ell\,], \qquad \alpha(s) \in \left]\, -\frac{\pi}{2}, \frac{\pi}{2}\,\right[.$$

Compte-tenu de la condition d'inextensibilité $|\,y'\,| = 1$, on a

$$y' = (\cos\alpha\cos\beta, \cos\alpha\sin\beta, \sin\alpha).$$

On construit une base orthonormée locale (y', n_1, n_2) en introduisant

$$n_1 = (-\sin\beta, \cos\beta, 0), \quad n_2 = (-\sin\alpha\cos\beta, -\sin\alpha\sin\beta, \cos\alpha).$$

On peut exprimer très simplement y'' dans cette base. En effet, on sait que la dérivée de la tangente unitaire est normale au câble, autrement dit que y'' est combinaison linéaire de n_1 et n_2. On trouve $y'' = \beta'\cos\alpha\, n_1 + \alpha'\, n_2$. Il s'ensuit que

$$(T\,y')' = T'\,y' + \beta'\,T\,\cos\alpha\, n_1 + \alpha'\,T\,n_2. \tag{1.21}$$

Déterminons l'expression de la force de traînée $f_c = -c_t\,|\,\mathrm{v}_t\,|\,\mathrm{v}_t - c_n\,|\,\mathrm{v}_n\,|\,\mathrm{v}_n$ dans la base (y', n_1, n_2). La vitesse du câble par rapport au courant est $\mathrm{v} = (v_{nv} - \omega_1, -\omega_2, -\omega_3)$. Ses composantes tangentielle et normale sont $\mathrm{v}_t = v_{y'}\,y'$ et $\mathrm{v}_n = v_{n_1}\,n_1 + v_{n_2}\,n_2$, avec

$$\begin{cases} v_{y'} = \mathrm{v}^\top y' = (v_{nv} - \omega_1)\,\cos\alpha\cos\beta - \omega_2\,\cos\alpha\sin\beta - \omega_3\,\sin\alpha \\ v_{n_1} = \mathrm{v}^\top n_1 = (\omega_1 - v_{nv})\,\sin\beta - \omega_2\,\cos\beta \\ v_{n_2} = \mathrm{v}^\top n_2 = (\omega_1 - v_{nv})\,\sin\alpha\cos\beta + \omega_2\,\sin\alpha\sin\beta - \omega_3\,\cos\alpha. \end{cases}$$

On obtient $f_c(s) = f_{y'}(s)\,y'(s) + f_{n_1}(s)\,n_1(s) + f_{n_2}(s)\,n_2(s)$ en posant

$$\begin{cases} f_{y'}(s) = f_c(s)^\top y'(s) = -c_t\,v_{y'}(s)\;\big|\,v_{y'}(s)\,\big| \\ f_{n_1}(s) = f_c(s)^\top n_1(s) = -c_n\,v_{n_1}(s)\,\sqrt{v_{n_1}(s)^2 + v_{n_2}(s)^2} \\ f_{n_2}(s) = f_c(s)^\top n_2(s) = -c_n\,v_{n_2}(s)\,\sqrt{v_{n_1}(s)^2 + v_{n_2}(s)^2}. \end{cases}$$

On déduit de l'expression *(1.21)* que l'équation *(1.20.a)* est équivalente à

$$\begin{cases} T'(s) = -f_{y'}(s) + m_c\,|\,g\,|\,\sin\alpha(s) \\ \alpha'(s) = \dfrac{-f_{n_2}(s) + m_c\,|\,g\,|\,\cos\alpha(s)}{T(s)} \\ \beta'(s) = \dfrac{-f_{n_1}(s)}{T(s)\,\cos\alpha(s)} \end{cases} \qquad s \in\,]\,0, \ell\,[\,. \tag{1.22}$$

Cette équation différentielle est définie en chaque point s où le câble n'est pas vertical $(\cos\alpha(s) \neq 0)$ et où sa tension $T(s)$ n'est pas nulle.

Des conditions initiales pour *(1.22)* nous sont données par *(1.20.b)*. La force de traînée hydrodynamique de l'engin remorqué a pour expression $f_p = -c_h\,|\,\mathrm{v}_h\,|\,\mathrm{v}_h - c_z\,|\,\mathrm{v}_z\,|\,\mathrm{v}_z$.

Les composantes horizontale et verticale de la vitesse de l'engin par rapport au courant sont $\mathrm{v}_h = (v_{nv} - \omega_1, -\omega_2, 0)$ et $\mathrm{v}_z = (0, 0, -\omega_3)$. On a donc

$$\begin{cases} {f_p}^\top y'(0) = c_h \mid \mathrm{v}_h \mid \cos\alpha(0)\, [(\omega_1 - v_{nv}) \cos\beta(0) + \omega_2 \sin\beta(0)] + c_z \mid \omega_3 \mid \omega_3 \sin\alpha(0) \\ {f_p}^\top n_1(0) = c_h \mid \mathrm{v}_h \mid [(v_{nv} - \omega_1) \sin\beta(0) + \omega_2 \cos\beta(0)] \\ {f_p}^\top n_2(0) = c_h \mid \mathrm{v}_h \mid \sin\alpha(0)\, [(v_{nv} - \omega_1) \cos\beta(0) - \omega_2 \sin\beta(0)] + c_z \mid \omega_3 \mid \omega_3 \cos\alpha(0). \end{cases}$$

La condition aux limites $T(0)\, y'(0) + f_p + m_p\, g = 0$ implique que

$${f_p}^\top n_1(0) = 0, \quad {f_p}^\top n_2(0) = m_p \mid g \mid \cos\alpha(0), \quad {f_p}^\top y'(0) = m_p \mid g \mid \sin\alpha(0) - T(0).$$

Examinons quelques cas particuliers. Si $\omega_1 \neq v_{nv}$, et donc $\mathrm{v}_h \neq 0$, ces relations donnent comme conditions initiales

$$\begin{cases} \beta(0) = \arctan\left(\dfrac{\omega_2}{\omega_1 - v_{nv}}\right) \\ \alpha(0) = \arctan\left(\dfrac{-c_z\, \omega_3 \mid \omega_3 \mid + m_p \mid g \mid}{c_h \mid \mathrm{v}_h \mid [(v_{nv} - \omega_1) \cos\beta(0) - \omega_2 \sin\beta(0)]}\right) \\ T(0) = c_h \mid \mathrm{v}_h \mid \cos\alpha(0)\, [(v_{nv} - \omega_1) \cos\beta(0) - \omega_2 \sin\beta(0)] \\ \qquad\quad + \sin\alpha(0)\, (-c_z\, \omega_3 \mid \omega_3 \mid + m_p \mid g \mid). \end{cases}$$

Le dénominateur de la fraction donnant $\alpha(0)$ – deuxième ligne – n'est pas nul. Il l'aurait été si $(v_{nv} - \omega_1) \cos\beta(0) - \omega_2 \sin\beta(0) = 0$, c'est-à-dire

$$\beta(0) = \arctan\left(\frac{v_{nv} - \omega_1}{\omega_2}\right).$$

Mais ceci impliquerait que $(v_{nv} - \omega_1)^2 + {\omega_2}^2 = 0$, ce qui est contradictoire avec $\mathrm{v}_h \neq 0$. Lorsque $\omega_2 \neq 0$, on a

$$\beta(0) = \operatorname{arcotan}\left(\frac{\omega_1 - v_{nv}}{\omega_2}\right)$$

et les expressions précédentes de $\alpha(0)$ et $T(0)$ restent valables. Mais si en revanche $\mathrm{v}_h = 0$, on ne dispose d'aucune valeur de $\beta(0)$. On ne peut donc pas déterminer les angles α et β et la tension T en intégrant l'équation différentielle *(1.22)*. Achevons cette discussion en remarquant qu'il ne suffit pas de disposer de conditions initiales pour pouvoir intégrer le système *(1.22)*. Le cas où la tension $T(0)$ est nulle soulève un autre problème, puisque l'équation n'est pas définie. Nous reviendrons sur ce point dans la section 2.1.5.

Lorsque l'on peut écrire des conditions initiales pour *(1.22)*, telles que $\cos\alpha(0) \neq 0$ et $T(0) \neq 0$, nous sommes assurés par le théorème de Cauchy-Lipschitz de l'existence et unicité de solutions maximales. Rappelons que dans toute cette section, la vitesse du

navire v_{nv} n'est pas nulle. Si nous supposons en outre qu'il n'y a pas de courant sous-marin, $\omega = 0$, la solution est caractérisée par $\beta = 0$ et (α, T) vérifie l'équation différentielle

$$\begin{cases} T'(s) = c_t\, {v_{nv}}^2\, \cos\alpha(s)\ |\cos\alpha(s)| + m_c\ |g|\ \sin\alpha(s) \\ \alpha'(s) = \dfrac{-c_n\, {v_{nv}}^2\, \sin\alpha(s)\ |\sin\alpha(s)| + m_c\ |g|\ \cos\alpha(s)}{T(s)} \end{cases} \qquad s \in\,]\,0, \ell\,[\qquad (1.23.a)$$

avec conditions initiales

$$\begin{cases} \alpha(0) = \arctan\left(\dfrac{m_p\ |g|}{c_h\, {v_{nv}}^2}\right) \\ T(0) = c_h\, {v_{nv}}^2\, \cos\alpha(0) + m_p\ |g|\ \sin\alpha(0). \end{cases} \qquad (1.23.b)$$

Comme c'était prévisible, la configuration du câble est plane et se trouve dans le plan vertical contenant la trajectoire du navire. On vérifie (Hetmaniuk et Paget [59]) que le système différentiel *(1.23)* possède une solution (α, T) définie sur $\mathbb{R}_+$ et a fortiori sur $[\,0, \ell\,]$. On pose alors

$$y(s) = -\left(\int_s^\ell \cos\alpha(\sigma)\, d\sigma, 0, \int_s^\ell \sin\alpha(\sigma)\, d\sigma\right), \quad s \in [\,0, \ell\,].$$

Le couple (y, T) est l'unique solution de l'équation stationnaire *(1.20)*.

1.3.2. Déroulement vertical

On suppose à nouveau que le navire est immobile et que le câble est déroulé à sa verticale. En actionnant le treuil, on impose à chaque instant la longueur $\ell(t)$ de sa partie immergée. A l'instant initial, le câble est immobile ($v_{init} = 0$ et $\dot\ell(0) = 0$) et de longueur $\ell(0) = \ell$. Les forces intervenant sont le poids du câble et celui du poisson ; on néglige les forces de traînée. Dans le repère dont l'origine est le navire et formé d'un vecteur vertical ascendant, la position relative et la tension sont des fonctions $y(s,t) \in \mathbb{R}$ et $T(s,t) \in \mathbb{R}$ satisfaisant

$$\begin{cases} m_c\, \ddot y(s,t) = (T\, y')'\,(s,t) - m_c\ |g| \\ |\,y'(s,t)\,| = 1, \end{cases} \qquad s \in\,]\,0, \ell(t)\,[\,,\ t \geq 0,$$

sous les conditions aux limites

$$y(\ell(t), t) = 0, \quad m_p\, \ddot y(0,t) = (T\, y')(0,t) - m_p\ |g|\,, \quad t \geq 0$$

et les conditions initiales $y(s,0) = s - \ell$ et $\dot y(s,0) = 0$, en tout $s \in [\,0, \ell\,]$.

Une solution de ces équations est

$$y(s,t) = s - \ell(t), \quad T(s,t) = (m_c\, s + m_p)\ \left(|g| - \ddot\ell(t)\right).$$

On retrouve la solution du cas statique $y(s) = s - \ell,\ T(s) = (m_c\, s + m_p)\,|g|$ si $\ell(t)$ est constante. On observe également que la tension $T(s,t)$ n'est nulle que lorsque l'accélération

du câble $-\ddot{\ell}(t)$ est égale à l'accélération de la pesanteur $-|\,g\,|$, autrement dit lorsqu'on laisse le câble se dérouler sous la seule action de son poids et de celui du poisson.

On ne saurait conclure ce premier chapitre sans mentionner les travaux de Reeken [91, 92], traitant de la dynamique de câbles inextensibles. Il semble que ce soit la première approche rigoureuse du sujet, dans un cadre néanmoins très restrictif : la longueur de ces câbles est infinie (donc sans poisson à leur extrémité inférieure), leur traînée hydrodynamique n'est pas prise en compte et leur position initiale est proche de la verticale. L'auteur démontre que les équations du modèle correspondant ont une solution, dans des espaces convenables.

Nous avons passé en revue plusieurs équations se déduisant de (*1.1)*, en caractérisant la ou les solutions de chacune d'entre elles. Il manque bien sûr une étude du système (*1.1)* dans toute sa généralité pour compléter ce travail. Nous avons toutefois mis en évidence, au moins dans un cas simple, le rôle de multiplicateur joué par la tension. Cette information se révèle utile dans le chapitre suivant, quant au choix d'un schéma d'intégration numérique.

Chapitre 2

Intégration numérique de l'équation du câble

On se consacre désormais à la discrétisation de l'équation du câble. Les méthodes usuelles d'approximation des problèmes d'évolution sont de deux types : ou bien on discrétise les équations simultanément en espace et en temps, grâce à un schéma de différences finies, ou bien on discrétise tout d'abord en espace et seulement ensuite en temps. Nous avons choisi cette seconde stratégie, dans laquelle la semi-discrétisation en espace est basée sur une méthode d'éléments finis. Nous l'appliquons successivement aux équations statique, stationnaire et dynamique à longueur de câble constante. Nous intégrons ensuite en temps grâce à un schéma de différences rétrogrades. Enfin, nous adaptons ce qui précède aux câbles de longueur variable.

2.1. Discrétisation de l'équation statique

Une introduction aux méthodes d'éléments finis est présentée dans les livres de Ciarlet [29] ou Raviart et Thomas [90]. Elles s'appuient toujours sur une formulation variationnelle du problème que l'on souhaite résoudre. Leur principe est d'approcher le ou les espaces fonctionnels (de dimension infinie) intervenant dans cette formulation variationnelle par des espaces fonctionnels de dimensions finies. Il convient alors de vérifier que les problèmes variationnels discrets ont une unique solution, puis que celle-ci converge vers la solution du problème original lorsque la discrétisation est de plus en plus fine.

Avant d'aborder la résolution de l'équation (*1.1*), il est instructif de revenir à l'équation statique (*1.3*). On veut s'appuyer sur sa formulation variationnelle (*1.19*). Une formulation variationnelle discrète de l'équation (*1.3*) est : chercher $(y_h, T_h) \in \mathcal{Y}_h \times \mathcal{T}_h$ tel que

$$\forall\, z_h \in \mathcal{Y}_h, \quad \int_0^\ell T_h(s)\, {y_h}'(s)^\top {z_h}'(s)\, ds = -m_c\, |\, g\, | \int_0^\ell z_h(s)\, ds - m_p\, |\, g\, |\, z_h(0) \qquad (2.1.a)$$

$$\forall\, \tau_h \in \mathcal{T}_h, \quad \frac{1}{2} \int_0^\ell \left(\left|\, {y_h}'(s)\, \right|^2 - 1 \right) \tau_h(s)\, ds = 0, \qquad (2.1.b)$$

où $\mathcal{Y}_h$ et $\mathcal{T}_h$ sont deux sous-espaces de dimension finie de $\mathcal{Y}$ et $L^\infty(0,\ell)$. Comme on va le voir, ces deux sous-espaces doivent être soigneusement choisis.

2.1.1. Une première discrétisation

Dans un premier temps, on a considéré des approximations continues affines par morceaux de y et de T, définies sur un maillage de $[0,\ell]$ que l'on peut prendre, par commodité, à pas constant : on pose, pour $m \in \mathbb{N}^*$, $h = \ell/m$ et pour $i = 0,...,m$, $s_i = i\,h$. A cette discrétisation correspondent les espaces d'éléments finis P1/P1

$$\mathcal{Y}_h = \left\{ z \in C\left([0,\ell]\right) :\ z(\ell) = 0, \quad \forall\, i = 1,...,m,\ z\big|_{[\,s_{i-1},s_i\,]} \in P_1 \right\}$$

et

$$\mathcal{T}_h = \left\{ z \in C\left([0,\ell]\right) :\ \forall\, i = 1,...,m,\ z\big|_{[\,s_{i-1},s_i\,]} \in P_1 \right\}.$$

La famille $(z_i)_{i=0,...,m}$ des fonctions de $\mathcal{T}_h$ vérifiant $z_i(s_i) = 1$ et $z_i(s_j) = 0$ si $i \neq j$ (ce sont les fonctions chapeaux habituelles) est une base de $\mathcal{T}_h$ et la famille $(z_i)_{i=0,...,m-1}$ est une base de $\mathcal{Y}_h$. Toutes les fonctions $y_h \in \mathcal{Y}_h$ et $T_h \in \mathcal{T}_h$ s'écrivent donc sous la forme

$$y_h = \sum_{i=0,...,m-1} y_i\, z_i, \quad T_h = \sum_{i=0,...,m} T_i\, z_i.$$

On a représenté sur la figure 2.1 deux fonctions de $\mathcal{Y}_h$ et $\mathcal{T}_h$. Les valeurs de la position relative y_h et de la tension T_h aux nœuds $\{s_i : i = 0,...,m-1\}$ et $\{s_i : i = 0,...,m\}$, c'est-à-dire $(y_i)_{i=0,...,m-1}$ et $(T_i)_{i=0,...,m}$, sont les $2\,m+1$ inconnues du problème discret.

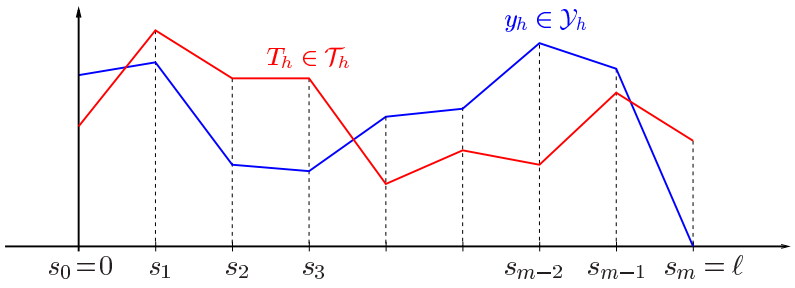


FIG. 2.1 – *position et tension pour des éléments finis P1/P1.*

Il est pratique d'introduire dans ces équations discrètes les longueurs des éléments de câble

$$\Delta y_i = y_i - y_{i-1},\ i = 1,...,m-1, \qquad \Delta y_m = -y_{m-1}.$$

De *(2.1.a)* résultent les m équations

$$\left\{ \begin{array}{l} \dfrac{(T_1 + T_0)\,\Delta y_1}{2\,h} = m_c\,|\,g\,|\,\dfrac{h}{2} + m_p\,|\,g\,| \\ \dfrac{(T_{i+1} + T_i)\,\Delta y_{i+1}}{2\,h} - \dfrac{(T_i + T_{i-1})\,\Delta y_i}{2\,h} = m_c\,|\,g\,|\,h, \qquad i = 1,...,m-1, \end{array} \right. \tag{2.2.a}$$

soit

$$\frac{(T_i + T_{i-1})\,\Delta y_i}{2\,h} = (2\,i-1)\left(m_c\,|\,g\,|\,\frac{h}{2}\right) + m_p\,|\,g\,|, \quad i = 1, ..., m.$$

De plus, *(2.1.b)* donne les $m+1$ équations

$$\begin{cases} {\Delta y_1}^2 = h^2 \\ {\Delta y_i}^2 + {\Delta y_{i+1}}^2 = 2\,h^2 \qquad i = 1, ..., m-1 \\ {\Delta y_m}^2 = h^2. \end{cases} \tag{2.2.b}$$

Celles-ci se résument à seulement m équations indépendantes ${\Delta y_i}^2 = h^2$ $(i = 1, ..., m)$. En imposant la distance entre nœuds consécutifs du câble, elles traduisent de façon très naturelle la condition d'inextensibilité.

On constate donc que, pour ce choix d'éléments finis, le système discret est sous-déterminé. Intéressons-nous néanmoins à ses solutions à tension positive. Les Δy_i sont nécessairement positifs, et par conséquent égaux à h. On obtient

$$\begin{cases} y_i = (i-m)\,h, \qquad i = 0, ..., m-1 \\ T_i = i\left(m_c\,|\,g\,|\,h\right) + m_p\,|\,g\,| + (-1)^{i+1}\left(m_p\,|\,g\,| - T_0\right), \qquad i = 1, ..., m, \end{cases}$$

où T_0 est arbitraire. En particulier, si $T_0 = m_p\,|\,g\,|$, on obtient comme solution

$$\begin{cases} y_i = (i-m)\,h, \qquad i = 0, ..., m-1 \\ T_i = i\left(m_c\,|\,g\,|\,h\right) + m_p\,|\,g\,|, \qquad i = 0, ..., m. \end{cases}$$

Rappelons que la solution à tension positive de l'équation statique est $y(s) = s - \ell$ et $T(s) = (m_c\,s + m_p)\,|\,g\,|$. Ces deux fonctions sont affines et ont les mêmes valeurs que y_h et T_h aux nœuds du maillage. On en déduit que $y_h = y$ et $T_h = T$ et que le schéma converge. En revanche, si $T_0 \neq m_p\,|\,g\,|$, on remarque que les valeurs de la tension oscillent de part et d'autre des valeurs obtenues pour $T_0 = m_p\,|\,g\,|$ et il n'y a pas convergence de y_h et T_h dans $L^\infty(0, \ell)$ lorsque h tend vers 0.

Au cours ce cette discrétisation, on perd une caractéristique de l'équation continue *(1.3)*, l'unicité de la solution à tension positive. La position y est certes correctement approchée, mais sont solutions des équations discrètes des tensions "parasites" qui ne convergent pas vers la tension T de l'équation continue. Ce phénomène est à rapprocher de celui que l'on observe si l'on tente de résoudre l'équation de Stokes à l'aide de ces mêmes éléments finis P1/P1.

2.1.2. Analogie avec l'équation de Stokes

Rappelons que l'équation de Stokes en dimension n s'écrit

$$\begin{cases} -\Delta u(x) + \nabla p(x) = f(x) \qquad x \in \Omega \subset \mathbb{R}^n \\ (\text{div } u)(x) = 0 \qquad\qquad\qquad x \in \Omega \\ u(x) = 0 \qquad\qquad\qquad\qquad x \in \partial\Omega. \end{cases}$$

En dimension $n = 2$ ou $n = 3$, les inconnues u et p sont les vitesse et pression d'un fluide incompressible en mouvement stationnaire contenu dans le domaine Ω. L'équation de Stokes est un cas simplifié de l'équation de Navier-Stokes, qui concerne elle un fluide en mouvement non stationnaire. Les ouvrages de Girault et Raviart [53], Cuvelier, Segal et van Steenhoven [34] et Pironneau [85] sont largement consacrés à l'étude de ces équations.

On suppose que $f \in H^{-1}(\Omega)^n$. D'autre part, la pression p n'étant définie que par son gradient, donc à une constante près (si Ω est connexe, ce que l'on suppose), on peut fixer cette constante en cherchant p à moyenne nulle, c'est-à-dire dans l'espace

$$L^2_0(\Omega) = \left\{ q \in L^2(\Omega) : \int_\Omega q = 0 \right\}.$$

Une formulation variationnelle de l'équation de Stokes s'obtient en multipliant la première équation par une fonction $v \in H^1_0(\Omega)^n$ et la deuxième ligne par une fonction $q \in L^2_0(\Omega)$. On intègre sur Ω et on élimine le laplacien par la formule de Green

$$\int_\Omega (\Delta u)\, v = \int_{\partial\Omega} \frac{\partial u}{\partial n}\, v - \sum_{1\le i\le n} \int_\Omega \nabla u_i \, \nabla v_i = - \sum_{1\le i\le n} \int_\Omega \nabla u_i \, \nabla v_i.$$

On a également

$$\int_\Omega (\nabla p)\; v = - \int_\Omega p \, \mathrm{div}\, v.$$

On considère donc le problème : chercher $u \in H^1_0(\Omega)^n$ et $p \in L^2_0(\Omega)$ tels que

$$\forall\, v \in H^1_0(\Omega)^n, \qquad \sum_{1\le i\le n} \int_\Omega \nabla u_i \, \nabla v_i - \int_\Omega p \, \mathrm{div}\, v = \int_\Omega f\, v \qquad (2.3.a)$$

$$\forall\, q \in L^2_0(\Omega), \qquad - \int_\Omega q \, \mathrm{div}\, u = 0, \qquad (2.3.b)$$

On démontre (Brezzi et Fortin [20], notamment) que si (u, p) est solution de ce problème variationnel, alors u est solution du problème

$$\begin{aligned} &\text{minimiser} && \frac{1}{2} \sum_{1\le i\le n} \int_\Omega |\, \nabla v_i \,|^2 - \int_\Omega f\, v \\ &\text{sous la contrainte} && v \in H^1_0(\Omega)^n,\ \mathrm{div}\ v = 0, \end{aligned}$$

et p est le multiplicateur associé à la condition d'incompressibilité $\mathrm{div}\ v = 0$. De ce point de vue, l'analogie entre l'équation statique *(1.3)* et l'équation de Stokes est évidente : les deux problèmes possèdent chacun deux inconnues, dont l'une est le multiplicateur de Lagrange associé à la contrainte portant sur l'autre inconnue.

Intéressons-nous à la discrétisation de l'équation de Stokes par des méthodes d'éléments finis s'appuyant sur la formulation variationnelle *(2.3)*. Ces méthodes, souvent qualifiées de mixtes, s'inscrivent naturellement dans un cadre abstrait ; on renvoie à [20, 53] pour une

présentation plus détaillée et la démonstration des résultats suivants. Soient deux espaces de Hilbert $\mathcal{X}$, $\mathcal{M}$, d'espaces duaux topologiques $\mathcal{X}'$, $\mathcal{M}'$, deux éléments $f \in \mathcal{X}'$, $\varphi \in \mathcal{M}'$, et deux formes bilinéaires continues

$$a : \mathcal{X}\times\mathcal{X} \to \mathbb{R}, \quad b : \mathcal{X}\times\mathcal{M} \to \mathbb{R}.$$

On peut associer à a et b les opérateurs $A : \mathcal{X} \to \mathcal{X}'$, $B : \mathcal{X} \to \mathcal{M}'$, ainsi que l'opérateur adjoint de ce dernier, $B' : \mathcal{M} \to \mathcal{X}'$, définis par

$$\begin{aligned} &\forall u \in \mathcal{X},\ \forall v \in \mathcal{X}, \quad \langle\, Au\,,\, v\,\rangle = a(u,v) \\ &\forall v \in \mathcal{X},\ \forall q \in \mathcal{M}, \quad \langle\, Bv\,,\, q\,\rangle = \langle\, v\,,\, B'q\,\rangle = b(v,q). \end{aligned}$$

Le problème que l'on s'attache à résoudre consiste à déterminer un couple $(u,p) \in \mathcal{X}\times\mathcal{M}$ tel que

$$\forall v \in \mathcal{X}, \quad a(u,v) + b(v,p) = \langle\, f\,,\, v\,\rangle \tag{2.4.a}$$

$$\forall q \in \mathcal{M}, \quad b(u,q) = \langle\, \varphi\,,\, q\,\rangle, \tag{2.4.b}$$

c'est-à-dire $Au + B'p = f$ dans $\mathcal{X}'$ et $Bu = \varphi$ dans $\mathcal{M}'$. Soit

$$V = \operatorname{Ker} B = \{v \in \mathcal{X} :\ \forall q \in \mathcal{M},\ b(v,q) = 0\}.$$

L'existence d'une solution au système *(2.4)* est assurée par la proposition ci-dessous.

Théorème 2.1. *Supposons que la forme bilinéaire a soit V-elliptique, c'est-à-dire*

$$\exists\, \alpha > 0 : \quad \forall v \in V, \quad a(v,v) \geq \alpha \| v \|_{\mathcal{X}}{}^2,$$

et que la forme bilinéaire b vérifie la condition

$$\exists\, \beta > 0 : \quad \inf_{q\in\mathcal{M}} \sup_{v\in\mathcal{X}} \frac{b(v,q)}{\| v \|_{\mathcal{X}} \| q \|_{\mathcal{M}}} \geq \beta. \tag{2.5}$$

Alors le problème (2.4) admet une unique solution $(u,p) \in \mathcal{X}\times\mathcal{M}$.

La condition inf-sup *(2.5)* est connue sous le nom de Babuška-Brezzi [5, 19]. De cette condition, on déduit que

$$\forall q \in \mathcal{M}, \quad \sup_{v\in\mathcal{X}} \frac{\langle\, v\,,\, B'q\,\rangle}{\| v \|_{\mathcal{X}}} \geq \beta \| q \|_{\mathcal{M}}.$$

Ceci implique que, pour tout $q \in \mathcal{M}$, on a $\| B'q \|_{\mathcal{X}'} \geq \beta \| q \|_{\mathcal{M}}$. On démontre, dans [18] par exemple, que cette dernière condition est équivalente à la surjectivité de B.

La formulation variationnelle *(2.3)* entre dans ce cadre abstrait : $\varphi = 0$ et les opérateurs sont

$$\begin{aligned} A : H_0^1(\Omega)^n &\to H^{-1}(\Omega)^n & B : H_0^1(\Omega)^n &\to L_0^2(\Omega) & B' : L_0^2(\Omega) &\to H_0^1(\Omega)^n \\ v &\mapsto -\Delta v, & v &\mapsto \operatorname{div} v, & q &\mapsto -\nabla q. \end{aligned}$$

Les formes a et b vérifient les hypothèses de la proposition précédente (voir Temam [99]).

On obtient une approximation interne du problème (*2.4*) en se donnant deux sous-espaces de dimensions finies $\mathcal{X}_h \subset \mathcal{X}$ et $\mathcal{M}_h \subset \mathcal{M}$. Il s'agit alors de déterminer $u_h \in \mathcal{X}_h$ et $p_h \in \mathcal{M}_h$ tels que

$$\forall\, v_h \in \mathcal{X}_h, \quad a(u_h, v_h) + b(v_h, p_h) = \langle\, f\,,\, v_h\,\rangle \tag{2.6.a}$$

$$\forall\, q_h \in \mathcal{M}_h, \quad b(u_h, q_h) = \langle\, \varphi\,,\, q_h\,\rangle. \tag{2.6.b}$$

Les dimensions des espaces $\mathcal{X}_h$ et $\mathcal{M}_h$ sont destinées à tendre vers l'infini lorsque le paramètre de discrétisation h tend vers 0. On définit l'analogue de V

$$V_h = \{v_h \in \mathcal{X}_h :\ \forall\, q_h \in \mathcal{M}_h,\ b(v_h, q_h) = 0\}.$$

Bien que $\mathcal{X}_h$ soit un sous-ensemble de $\mathcal{X}$, V_h n'est pas, en général, un sous-ensemble de V. De même, la condition inf-sup (*2.5*) n'est pas forcément satisfaite pour le couple d'espaces $(\mathcal{X}_h, \mathcal{M}_h)$. Il faut donc refaire des hypothèses pour justifier que le système (*2.6*) a une solution.

Théorème 2.2. *Supposons que la forme bilinéaire a soit V_h-elliptique (de constante α_h) et que la condition inf-sup discrète*

$$\exists\, \beta_h > 0 : \quad \inf_{q_h \in \mathcal{M}_h} \sup_{v_h \in \mathcal{X}_h} \frac{b(v_h, q_h)}{\|\, v_h \,\|_{\mathcal{X}}\, \|\, q_h \,\|_{\mathcal{M}}} \ \geq \beta_h \tag{2.7}$$

soit satisfaite. Alors le problème (2.6) admet une unique solution $(u_h, p_h) \in \mathcal{X}_h \times \mathcal{M}_h$. De plus, sous les hypothèses du théorème 2.1, il existe une constante C (dépendant de α_h, β_h et des normes de a et b) telle que

$$\|\, u - u_h \,\|_{\mathcal{X}} + \|\, p - p_h \,\|_{\mathcal{M}} \leq C \left(\inf_{v_h \in \mathcal{X}_h} \|\, u - v_h \,\|_{\mathcal{X}} + \inf_{q_h \in \mathcal{M}_h} \|\, p - q_h \,\|_{\mathcal{M}} \right).$$

La condition (*2.7*) implique que la matrice B_h, version discrète de l'opérateur B définie par

$$\forall\, v_h \in \mathcal{X}_h,\ \forall\, q_h \in \mathcal{M}_h, \quad {q_h}^{\top} B_h\, v_h = {v_h}^{\top} {B_h}^{\top} q_h = b(v_h, q_h), \tag{2.8}$$

est surjective. Il existe des critères permettant de s'affranchir de la dépendance de C en h. On dispose ainsi d'une majoration de l'erreur de dicrétisation permettant de prouver la convergence d'une méthode d'éléments finis, si les suites de sous-espaces $(\mathcal{X}_h)$ et $(\mathcal{M}_h)$ sont denses dans $\mathcal{X}$ et $\mathcal{M}$. On ne pourra cependant pas appliquer de résultat de convergence de ce type aux équations (*1.19*) et (*2.1*), puisque $W^{1,\infty}(0,\ell)$ et $L^{\infty}(0,\ell)$ ne sont pas séparables.

En ce qui concerne l'approximation de la formulation variationnelle (*2.3*), on pourra considérer

$$\mathcal{X}_h = H_h \cap H_0^1(\Omega)^n, \quad \mathcal{M}_h = L_h \cap L_0^2(\Omega),$$

où H_h et L_h sont deux sous-espaces de dimensions finies de $H^1(\Omega)^n$ et $L^2(\Omega)$, choisis en particulier de sorte que la condition inf-sup discrète (*2.7*) soit respectée. En pratique,

H_h et L_h sont des espaces d'éléments finis, H_h ayant plus de degrés de liberté que L_h. Les polynômes servant à approcher la vitesse u sont toujours de degré supérieur à ceux approchant la pression p. Deux symptômes d'un mauvais choix d'éléments finis sont une vitesse uniformément nulle – on parle alors d'un phénomène de vérouillage – et une pression déterminée à un certain nombre de modes oscillatoires parasites près. Ce dernier point est décrit par Sani, Gresho, Lee et Griffiths [96].

En plus de l'analogie entre les variables des équations statique (*1.3)* et de Stokes, il existe donc une similitude frappante entre le comportement de leurs solutions approchées pour certaines discrétisations. Bien que la formulation variationnelle (*1.19)* n'entre pas dans le cadre abstrait (*2.4)*, l'étude de ce dernier met en lumière le point suivant. Tout comme la matrice B_h dans (*2.8)*, la jacobienne de la contrainte d'inextensibilité discrétisée doit être surjective si l'on veut préserver l'unicité de la solution du problème continu (l'unicité du multiplicateur, plus précisément). Pour remplir cette condition, cette étude suggère l'emploi d'éléments finis d'ordre plus élevé pour la position relative y que pour la tension T. Nous allons présenter deux discrétisations possibles.

2.1.3. Discrétisation de l'équation statique par éléments finis P1/P0

Si l'on utilise des éléments finis P1/P0, autrement dit si l'on cherche des approximations continue affine par morceaux de y et constante par morceaux de T (figure 2.2), les espaces d'éléments finis sont

$$\mathcal{Y}_h = \left\{ z \in C\left([\,0,\ell\,]\right) :\ z(\ell) = 0, \quad \forall\, i = 1,...,m,\ z\,\big|_{[\,s_{i-1},s_i\,]} \in P_1 \right\}$$

et

$$\mathcal{T}_h = \left\{ z \in L^\infty(0,\ell) :\ \forall\, i = 1,...,m,\ z\,\big|_{]\,s_{i-1},s_i\,[} \in P_0 \right\}.$$

Les inconnues du problème discrétisé sont les valeurs $(y_i)_{i=0,...,m-1}$ de la position relative $y_h \in \mathcal{Y}_h$ aux nœuds $\{s_i : i = 0,...,m-1\}$ et les valeurs $(T_i)_{i=1,...,m}$ de la tension $T_h \in \mathcal{T}_h$ sur les éléments $\{\,]\,s_{i-1}, s_i\,[: i = 1,...,m\}$.

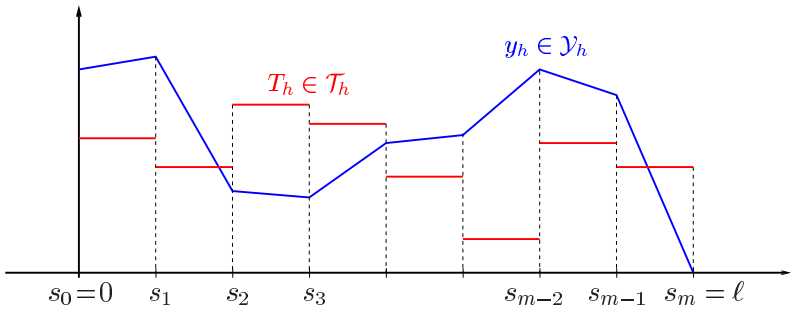


FIG. 2.2 – *position et tension pour des éléments finis P1/P0.*

Rappelons que l'on a posé $\Delta y_m = -y_{m-1}$ et $\Delta y_i = y_i - y_{i-1}$ pour tout $i = 1, ..., m-1$. Pour cette discrétisation, le système *(2.1)* donne les équations

$$\begin{cases} \dfrac{T_1\,\Delta y_1}{h} = m_c\,|\,g\,|\,\dfrac{h}{2} + m_p\,|\,g\,| \\ \dfrac{T_{i+1}\,\Delta y_{i+1}}{h} - \dfrac{T_i\,\Delta y_i}{h} = m_c\,|\,g\,|\,h, \quad i = 1, ..., m-1, \end{cases}$$

autrement dit

$$\frac{T_i\,\Delta y_i}{h} = (2\,i-1)\left(m_c\,|\,g\,|\,\frac{h}{2}\right) + m_p\,|\,g\,|, \quad i = 1, ..., m,$$

et à nouveau les équations ${\Delta y_i}^2 = h^2$, $i = 1, ..., m$. Le système discret compte donc $2\,m$ équations et autant d'inconnues. Son unique solution à tension positive est

$$\begin{cases} y_i = (i-m)\,h, \quad i = 0, ..., m-1 \\ T_i = (2\,i-1)\left(m_c\,|\,g\,|\,\dfrac{h}{2}\right) + m_p\,|\,g\,|, \quad i = 1, ..., m. \end{cases}$$

On constate que $y_h = y$, avec (y, T) solution à tension positive de l'équation statique. Par ailleurs,

$$\forall\, i = 1, ..., m, \quad T_i = T\left(s_i + \frac{h}{2}\right).$$

On en déduit la convergence de T_h vers T dans $L^\infty(0, \ell)$ lorsque h tend vers 0.

2.1.4. Discrétisation par éléments finis P1-iso-P2/P1

L'inconvénient de l'approche précédente est de ne pas fournir de tension continue. On peut remédier à cela en prenant des éléments finis d'ordres plus élevés. Nous considérons des approximations continues affines par morceaux de y et T, mais en définissant y sur un maillage deux fois plus fin, comme sur la figure 2.3. Ce type d'éléments finis est connu sous le nom de P1-iso-P2/P1 (voir [85]). Nous adopterons cette méthode de discrétisation dans le cas dynamique, après l'avoir comparée à la discrétisation par éléments finis P1/P0 dans la section suivante consacrée au cas stationnaire.

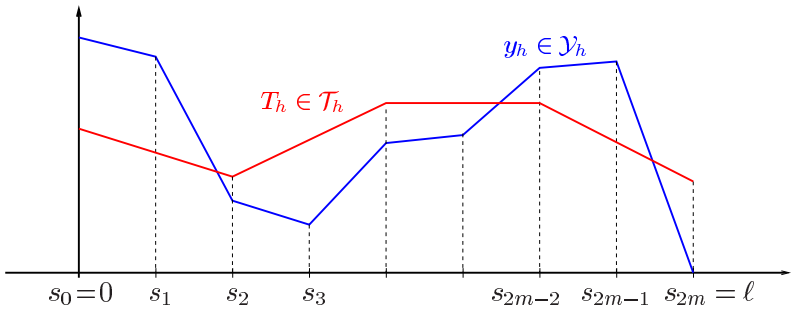


FIG. 2.3 – *position et tension pour des éléments finis P1-iso-P2/P1.*

Soient $m \in \mathbb{N}^*$, $h = \ell/(2m)$ et le maillage $(s_i) = (i\,h)_{i=0,...,2m}$. Les espaces d'éléments finis sont

$$\mathcal{Y}_h = \left\{ z \in C\left([0,\ell]\right) :\ z(\ell) = 0, \quad \forall\, i = 1,...,2\,m,\ z\,\big|_{[s_{i-1},s_i]} \in P_1 \right\}$$

et

$$\mathcal{T}_h = \left\{ z \in C\left([0,\ell]\right) :\ \forall\, j = 1,...,m,\ z\,\Big|_{\left[s_{2(j-1)},s_{2j}\right]} \in P_1 \right\}.$$

Récrivons le système *(2.1)* correspondant à cette discrétisation. Ses $3\,m+1$ inconnues sont les valeurs (y_i) de la position relative aux nœuds $\{s_i : i = 0,...,2\,m-1\}$ et les valeurs (T_j) de la tension aux nœuds $\{s_{2j} : j = 0,...,m\}$ – c'est-à-dire un nœud sur deux. Il y a maintenant $2\,m$ éléments de câble dont les longueurs sont

$$\Delta y_i = y_i - y_{i-1},\ i = 1,...,2\,m-1, \qquad \Delta y_{2m} = -y_{2m-1}.$$

Les équations *(2.1.a)* se traduisent en

$$\left\{ \begin{array}{l} \dfrac{(3\,T_0 + T_1)\,\Delta y_1}{4\,h} = m_c\,|\,g\,|\,\dfrac{h}{2} + m_p\,|\,g\,| \\[2ex] \dfrac{(T_0 + 3\,T_1)\,\Delta y_2}{4\,h} - \dfrac{(3\,T_0 + T_1)\,\Delta y_1}{4\,h} = m_c\,|\,g\,|\,h \\[2ex] \dfrac{(3\,T_1 + T_2)\,\Delta y_3}{4\,h} - \dfrac{(T_0 + 3\,T_1)\,\Delta y_2}{4\,h} = m_c\,|\,g\,|\,h \\[2ex] \quad \vdots \\[2ex] \dfrac{(3\,T_{m-1} + T_m)\,\Delta y_{2m-1}}{4\,h} - \dfrac{(T_{m-2} + 3\,T_{m-1})\,\Delta y_{2m-2}}{4\,h} = m_c\,|\,g\,|\,h \\[2ex] \dfrac{(T_{m-1} + 3\,T_m)\,\Delta y_{2m}}{4\,h} - \dfrac{(3\,T_{m-1} + T_m)\,\Delta y_{2m-1}}{4\,h} = m_c\,|\,g\,|\,h. \end{array} \right. \qquad (2.9.a)$$

Il en découle que, pour tout $i = 1,...,m$,

$$\left\{ \begin{array}{l} \dfrac{(3\,T_{i-1} + T_i)\,\Delta y_{2i-1}}{4\,h} = (4\,i-3)\left(m_c\,|\,g\,|\,\dfrac{h}{2} \right) + m_p\,|\,g\,| \\[2ex] \dfrac{(T_{i-1} + 3\,T_i)\,\Delta y_{2i}}{4\,h} = (4\,i-1)\left(m_c\,|\,g\,|\,\dfrac{h}{2} \right) + m_p\,|\,g\,|. \end{array} \right.$$

Les équations *(2.1.b)* donnent

$$\left\{ \begin{array}{l} \dfrac{3\,{\Delta y_1}^2 + {\Delta y_2}^2 - 4\,h^2}{8\,h} = 0 \\[2ex] \dfrac{{\Delta y_{2i-3}}^2 + 3\,{\Delta y_{2i-2}}^2 + 3\,{\Delta y_{2i-1}}^2 + {\Delta y_{2i}}^2 - 8\,h^2}{8\,h} = 0, \qquad i = 2,...,m, \\[2ex] \dfrac{{\Delta y_{2m-1}}^2 + 3\,{\Delta y_{2m}}^2 - 4\,h^2}{8\,h} = 0. \end{array} \right. \qquad (2.9.b)$$

Ceci représente tour à tour $2\,m$ et $m+1$ équations, soit autant que d'inconnues.

Les équations *(2.9.a)* donnent des longueurs $(\Delta y_i)_{i=1,...,2m}$ strictement positives si la tension est elle-même positive. A la différence des contraintes discrètes déjà rencontrées, la contrainte *(2.9.b)* n'impose pas, à elle seule, leur valeur. On dispose uniquement de bornes sur ces longueurs, par exemple les majorations

$$|\Delta y_1| \leq h\sqrt{\frac{4}{3}}, \qquad |\Delta y_i| \leq h\sqrt{\frac{8}{3}},\ i = 2, ..., 2m-1, \qquad |\Delta y_{2m}| \leq h\sqrt{\frac{4}{3}}.$$

Il n'est pas surprenant de retrouver la structure des équations continues *(1.19)* dans le système discret *(2.9)*, qui s'écrit sous la forme

$$\begin{cases} R(y)^\top T = D \\ C(y) = 0. \end{cases}$$

Nous notons $y = (y_i)_{i=0,...,2m-1}$ et $T = (T_i)_{i=0,...,m}$ les vecteurs des valeurs des fonctions y_h et T_h aux nœuds du maillage ; le contexte permet de ne pas les confondre avec les solutions des équations continues. La matrice $R(y)$ est la jacobienne de $C(y)$, c'est une fonction linéaire de y. Le vecteur D, constant, est le gradient $\nabla J(y_h)$ de la fonctionnelle d'énergie J (introduite page 15) définie sur $\mathcal{Y}_h$ par

$$J(z_h) = -m_c\,|g|\,h\left(\frac{z_0}{2} + z_1 + ... + z_{2m-1}\right) - m_p\,|g|\,z_0.$$

Les équations *(2.9)* sont les équations d'optimalité du problème dans $\mathcal{Y}_h$

$$\begin{array}{c} \text{minimiser } J(z_h) \\ \text{sous la contrainte } C(z) = 0. \end{array} \qquad (2.10)$$

On vérifierait aisément que le système discret de la section précédente, correspondant à la discrétisation par éléments finis P1/P0, possède cette même structure.

Nous avons remarqué qu'une condition nécessaire d'unicité de la tension discrète était la surjectivité de la jacobienne de la contrainte d'inextensibilité. Démontrons que cette condition est vérifiée.

Lemme 2.3. *La jacobienne $R(y)$ de la contrainte d'inextensibilité discrète $C(y) = 0$ est surjective en tout y satisfaisant cette contrainte.*

Démonstration. Nous allons montrer que $R(y)^\top$ est injective, c'est-à-dire que la seule solution de $R(y)^\top T = 0$ est $T = 0$. Auparavant, on remarque que si $C(y) = 0$, alors

$$\Delta y_1{}^2 > 0, \qquad \Delta y_{2i}{}^2 + \Delta y_{2i+1}{}^2 > 0, \quad i = 1, ...m-1, \qquad \Delta y_{2m}{}^2 > 0.$$

L'équation $R(y)^\top T = 0$ est donc équivalente à

$$\begin{cases} 3T_0 + T_1 = 0 \\ T_j + 3T_{j+1} = 0 \quad \text{ou} \quad 3T_{j+1} + T_{j+2} = 0, \quad j = 0, ..., m-2, \\ T_{m-1} + 3T_m = 0. \end{cases}$$

On note (S_1) le système

$$(S_1) \quad \begin{cases} 3\,T_0 + T_1 = 0 \\ T_0 + 3\,T_1 = 0 \end{cases}$$

et si $p \geq 2$,

$$(S_p) \quad \begin{cases} 3\,T_0 + T_1 = 0 \\ T_j + 3\,T_{j+1} = 0 \quad \text{ou} \quad 3\,T_{j+1} + T_{j+2} = 0, \quad j = 0, ..., p-2, \\ T_{p-1} + 3\,T_p = 0. \end{cases}$$

Il s'agit de résoudre (S_m).

Vérifions par récurrence que l'unique solution de (S_k) est $(T_j)_{j \leq k} = 0$. Ceci est vrai pour $k = 1$ et nous le supposons pour tout $k \leq p$. Les équations de (S_{p+1}) sont les suivantes : ou bien il existe un entier $k \geq 1$ tel que (S_{p+1}) soit équivalent à

$$\begin{cases} (S_k) \\ 3\,T_{j+1} + T_{j+2} = 0, \quad j = k, ..., p-1, \\ T_p + 3\,T_{p+1} = 0, \end{cases}$$

ou bien (S_{p+1}) est équivalent à

$$\begin{cases} 3\,T_j + T_{j+1} = 0, \quad j = 0, ..., p, \\ T_p + 3\,T_{p+1} = 0. \end{cases}$$

Il est clair que l'unique solution du second système est $(T_j)_{j \leq p+1} = 0$. Détaillons la solution du premier système : l'unique solution de (S_k) est $(T_j)_{j \leq k} = 0$ et les équations restantes donnent $(T_j)_{k+1 \leq j \leq p+1} = 0$. Nous parvenons donc à la conclusion voulue : l'unique solution de (S_{p+1}) est $(T_j)_{j \leq p+1} = 0$ et l'unique solution de (S_m) est $(T_j)_{j \leq m} = 0$. □

Une solution des équations *(2.9)* est

$$\begin{cases} y_i = (i - 2\,m)\,h, \quad i = 0, ..., 2\,m - 1, \\ T_j = 2\,j\,\left(m_c\,|\,g\,|\,h\right) + m_p\,|\,g\,|, \quad j = 0, ..., m. \end{cases} \tag{2.11}$$

Nous allons prouver qu'il n'existe pas d'autre solution du système *(2.9)* dont la tension soit positive. La démonstration de ce résultat est calquée sur la démarche que nous avons adoptée en dimension infinie.

Proposition 2.4. *Le système (2.9) admet une unique solution à tension positive ; cette solution est donnée par (2.11).*

Démonstration. Nous vérifions dans une première étape que toute solution de *(2.9)* à tension positive est solution du problème sur $\mathcal{Y}_h$

$$\begin{array}{c} \text{minimiser } J(z_h) \\ \text{sous la contrainte } C(z) \leq 0. \end{array} \tag{2.12}$$

En effet, ce problème est convexe. Ses conditions nécessaires et suffisantes d'optimalité sont

$$\begin{cases} R(y)^\top T = D \\ C(y) \leq 0 \\ T \geq 0 \\ C(y)^\top T = 0. \end{cases} \tag{2.13}$$

Toute solution (y, T) de *(2.9)* telle que $T \geq 0$, est également solution de *(2.13)* et *(2.12)*.

La deuxième étape consiste à prouver que toute solution de *(2.12)* est solution de *(2.10)*. Il suffit pour cela de vérifier que si y_h est solution de *(2.12)*, alors $C(y) = 0$. Supposons que pour un indice $i \geq 1$, nous ayons $C_i(y) < 0$. Au moins une des deux situations $\Delta y_{2i-3}{}^2 + 3\,\Delta y_{2i-2}{}^2 < 4\,h^2$ ou $\Delta y_{2i-1}{}^2 + 3\,\Delta y_{2i}{}^2 < 4\,h^2$ se produit. Nous allons contredire l'optimalité de y_h dans ces deux éventualités. Nous devrons distinguer selon le signe de $\Delta y_{2i-3} - \Delta y_{2i-2}$ d'une part, selon le signe de $\Delta y_{2i-1} - \Delta y_{2i}$ d'autre part, ce qui donne en fait quatre cas différents à examiner.

Supposons que $\Delta y_{2i-3}{}^2 + 3\,\Delta y_{2i-2}{}^2 < 4\,h^2$ et $\Delta y_{2i-2} \leq \Delta y_{2i-3}$. Nous modifions y_h de la manière suivante. Etant donnés des paramètres $\varepsilon > 0$ et $a, b \in \mathbb{R}$, on considère la fonction $y_{h,\varepsilon}$ définie par

$$y_{h,\varepsilon}(s) = \begin{cases} y_h(s) - h\,(a+b)\,\varepsilon & \text{si} \quad s < s_{2i-4} \\ y_h(s) - h\,b\,\varepsilon - (s_{2i-3} - s)\,a\,\varepsilon & \text{si} \quad s_{2i-4} \leq s < s_{2i-3} \\ y_h(s) - (s_{2i-2} - s)\,b\,\varepsilon & \text{si} \quad s_{2i-3} \leq s < s_{2i-2} \\ y_h(s) & \text{si} \quad s_{2i-2} \leq s. \end{cases}$$

Il est clair que $y_{h,\varepsilon}(\ell) = y_h(\ell) = 0$ et $y_{h,\varepsilon} \in \mathcal{Y}_h$. La dérivée de $y_{h,\varepsilon}$ est donnée par

$${y_{h,\varepsilon}}'(s) = \begin{cases} {y_h}'(s) & \text{si} \quad s < s_{2i-4} \\ \dfrac{\Delta y_{2i-3}}{h} + a\,\varepsilon & \text{si} \quad s_{2i-4} \leq s < s_{2i-3} \\ \dfrac{\Delta y_{2i-2}}{h} + b\,\varepsilon & \text{si} \quad s_{2i-3} \leq s < s_{2i-2} \\ {y_h}'(s) & \text{si} \quad s_{2i-2} \leq s. \end{cases}$$

Comme ${y_{h,\varepsilon}}'$ ne diffère de ${y_h}'$ que sur $[\,s_{2i-4}, s_{2i-2}\,[$, nous savons que $C_j(y_\varepsilon) \leq 0$ pour tout $j \notin \{i-1, i\}$. Puisque $C_i(y) < 0$, il est évident que $C_i(y_\varepsilon) \leq 0$ si ε est assez petit. Nous allons déterminer a et b de sorte que, pour de petites valeurs de ε, on ait $C_{i-1}(y_\varepsilon) \leq 0$ et $J(y_{h,\varepsilon}) < J(y_h)$. Cette dernière inégalité est équivalente à

$$a\left(m_c\,h\left(i - \frac{1}{2}\right) + m_p\right) + b\left(m_c\,h\left(i + \frac{1}{2}\right) + m_p\right) > 0. \tag{2.14}$$

Désignons la contrainte $C_{i-1}(y_\varepsilon)$ par

$$\phi(\varepsilon) = \frac{h}{8}\left(\left(\frac{\Delta y_{2i-5}}{h}\right)^2 + 3\left(\frac{\Delta y_{2i-4}}{h}\right)^2 + 3\left(\frac{\Delta y_{2i-3}}{h} + a\,\varepsilon\right)^2 + \left(\frac{\Delta y_{2i-2}}{h} + b\,\varepsilon\right)^2 - 8\right).$$

Si

$$\phi'(0) = \frac{3\,a\,\Delta y_{2i-3} + b\,\Delta y_{2i-2}}{4} < 0, \qquad (2.15)$$

et compte-tenu du fait que $\phi(0) = C_{i-1}(y) \leq 0$, alors $C_{i-1}(y_\varepsilon) \leq 0$ pour ε suffisamment petit. Les inégalités *(2.14)* et *(2.15)* sont satisfaites pour $a = -1$ et $b = 1$. On en déduit que, dans ce premier cas, y_h n'est pas optimal.

Supposons à présent que $3\,{\Delta y_{2i-1}}^2 + {\Delta y_{2i}}^2 < 4\,h^2$ et $\Delta y_{2i-1} \leq \Delta y_{2i}$. Nous considérons

$$y_{h,\varepsilon}(s) = \begin{cases} y_h(s) - h\,(a+b)\,\varepsilon & \text{si} \quad s < s_{2i-2} \\ y_h(s) - h\,b\,\varepsilon - (s_{2i-1} - s)\,a\,\varepsilon & \text{si} \quad s_{2i-2} \leq s < s_{2i-1} \\ y_h(s) - (s_{2i} - s)\,b\,\varepsilon & \text{si} \quad s_{2i-1} \leq s < s_{2i} \\ y_h(s) & \text{si} \quad s_{2i} \leq s. \end{cases}$$

Il s'agit de déterminer a et b tels que $J(y_{h,\varepsilon}) < J(y_h)$, c'est-à-dire

$$a\left(m_c\,h\left(i+\frac{1}{2}\right)+m_p\right) + b\left(m_c\,h\left(i+\frac{3}{2}\right)+m_p\right) > 0,$$

et tels que $a\,\Delta y_{2i-1} + 3\,b\,\Delta y_{2i} < 0$, qui nous assure que $C_{i+1}(y_\varepsilon) \leq 0$ si ε est assez petit. Ces conditions sont réalisées si $a = 1$ et

$$-\frac{m_c\,h\left(i+\frac{1}{2}\right)+m_p}{m_c\,h\left(i+\frac{3}{2}\right)+m_p} < b < -\frac{\Delta y_{2i-1}}{3\,\Delta y_{2i}}.$$

Trouver un tel b est possible car

$$-\frac{m_c\,h\left(i+\frac{1}{2}\right)+m_p}{m_c\,h\left(i+\frac{3}{2}\right)+m_p} \leq -\frac{1}{3} < -\frac{\Delta y_{2i-1}}{3\,\Delta y_{2i}}.$$

Dans cette situation encore, on contredit l'optimalité de y_h.

Supposons que ${\Delta y_{2i-3}}^2 + 3\,{\Delta y_{2i-2}}^2 < 4\,h^2$ et $\Delta y_{2i-3} < \Delta y_{2i-2}$. Nécessairement,

$${\Delta y_{2i-3}}^2 + 3\,{\Delta y_{2i-2}}^2 - (3\,{\Delta y_{2i-3}}^2 + {\Delta y_{2i-2}}^2) = 2({\Delta y_{2i-3}}^2 - {\Delta y_{2i-2}}^2) < 0,$$

donc ${\Delta y_{2i-3}}^2 + 3\,{\Delta y_{2i-2}}^2 < 4\,h^2$. Nous nous retrouvons dans la situation précédente, avec un indice i augmenté de 1. Nous pouvons montrer que y_h n'est pas optimal de façon identique.

Supposons enfin que $3\,{\Delta y_{2i-1}}^2 + {\Delta y_{2i}}^2 < 4\,h^2$ et $\Delta y_{2i} \leq \Delta y_{2i-1}$. Alors

$${\Delta y_{2i-1}}^2 + 3\,{\Delta y_{2i}}^2 < 4\,h^2$$

et on reprend les arguments de la première situation. En conclusion, nous avons prouvé que l'on ne pouvait avoir $C_i(y) < 0$ si y_h est solution du problème *(2.12)*. C'est donc

que les solutions de ce problème saturent la contrainte $C(y) \leq 0$. Par suite, ce sont des solutions de *(2.10)*.

Dans la troisième et dernière étape, nous montrons l'unicité de la solution à tension positive du système *(2.9)*. Soit $(\overline{y}, \overline{T})$ une autre solution de *(2.9)*, dont la tension $\overline{T}$ est positive. D'après la première étape de cette démonstration, y_h et $\overline{y}_h$ sont deux solutions de *(2.10)*. Ce problème étant convexe, $\frac{1}{2}(y_h + \overline{y}_h)$ en est une troisième solution. D'après la deuxième étape, y_h, $\overline{y}_h$ et $\frac{1}{2}(y_h + \overline{y}_h)$ sont solutions de *(2.12)*. Donc, pour tout $j = 1, ..., m+1$,

$$C_j(y) = C_j(\overline{y}) = C_j\left(\frac{1}{2}(y + \overline{y})\right) = 0.$$

Mais la stricte convexité de $x \mapsto |x|^2$ sur $\mathbb{R}$ implique que si ${y_h}' \neq {\overline{y}_h}'$ sur un intervalle contenu dans le support de τ_j (fonction de base de $\mathcal{T}_h$), alors

$$0 = \int_0^\ell \left(\left| \frac{{y_h}' + {\overline{y}_h}'}{2} \right|^2 - 1 \right) \tau_j < \frac{1}{2} \int_0^\ell \left(|{y_h}'|^2 - 1 \right) \tau_j + \frac{1}{2} \int_0^\ell \left(|{\overline{y}_h}'|^2 - 1 \right) \tau_j = 0,$$

ce qui est absurde. Donc ${y_h}' = {\overline{y}_h}'$ et par suite $y_h = \overline{y}_h$ puisque $y_h(\ell) = \overline{y}_h(\ell) = 0$. Les tensions T et $\overline{T}$ vérifient

$$R(y)^\top T = R(\overline{y})^\top \overline{T} = R(y)^\top \overline{T} = D.$$

On en déduit que $R(y)^\top (T - \overline{T}) = 0$ et $T = \overline{T}$ puisque $R(y)$ est surjective (lemme 2.3). Nous avons démontré l'unicité de la solution à tension positive des équations *(2.9)*. □

On achève l'étude de la discrétisation par éléments finis P1-iso-P2/P1 en remarquant que la solution discrète donnée par le système *(2.9)* coïncide avec la solution continue $y(s) = s - \ell$ et $T(s) = (m_c\, s + m_p)\, |g|$. Le schéma est donc convergent dans $L^\infty(0, \ell)$.

2.1.5. Extension au cas stationnaire

Afin de retenir une des deux méthodes de discrétisation que nous venons de présenter, nous les avons toutes deux appliquées à l'équation stationnaire *(1.20)*, sans courant sous-marin ($\omega = 0$). Une solution de référence est obtenue en résolvant l'équation différentielle *(1.23)*. On lui compare ensuite les solutions obtenues en discrétisant *(1.20)* par éléments finis P1/P0 et P1-iso-P2/P1.

Une formulation variationnelle de l'équation *(1.20)* est : chercher $y \in \mathcal{Y}^3$ (l'ensemble $\mathcal{Y}$ est défini page 24) et $T \in L^\infty(0, \ell)$ tels que

$$\forall\, z \in \mathcal{Y}^3, \quad \int_0^\ell T(s)\, y'(s)^\top z'(s)\, ds = \int_0^\ell f_c(s)^\top z(s)\, ds + f_p{}^\top z(0) + m_c \int_0^\ell g^\top z(s)\, ds + m_p\, g^\top z(0) \qquad (2.16.a)$$

$$\forall\, \tau \in L^\infty(0, \ell), \quad \frac{1}{2} \int_0^\ell \left(|y'(s)|^2 - 1 \right) \tau(s)\, ds = 0. \qquad (2.16.b)$$

Pour s'en convaincre, il suffit de reprendre la démonstration de la proposition 1.8. Il est clair que si (y,T) est solution de *(1.20)*, alors $f_c \in L^\infty(0,\ell)^3$, donc $T\,y' \in W^{1,\infty}(0,\ell)^3$ et

$$\forall\, z \in \mathcal{Y}^3, \quad \int_0^\ell (T\,y')'(s)^\top z(s)\,ds + \int_0^\ell f_c(s)^\top z(s)\,ds + m_c \int_0^\ell g^\top z(s)\,ds = 0.$$

L'égalité *(2.16.a)* résulte de l'intégration par parties du premier terme ci-dessus. Nous nous appuyons sur cette formulation variationnelle pour discrétiser l'équation *(1.20)*.

Pour les éléments finis P1/P0, on considère à nouveau l'espace $\mathcal{Y}_h$ et sa base $(z_i)_{i=0,...,m-1}$ introduits dans la section 2.1.1. On construit une base $(\tilde{z}_i)_{i=0,...,3m-1}$ de ${\mathcal{Y}_h}^3$ en posant

$$\begin{aligned}
\tilde{z}_i &= (z_i,0,0) \ \text{ si } \ i = 0,...,m-1,\\
\tilde{z}_i &= (0,z_{i-m},0) \ \text{ si } \ i = m,...,2\,m-1,\\
\tilde{z}_i &= (0,0,z_{i-2m}) \ \text{ si } \ i = 2\,m,...,3\,m-1.
\end{aligned}$$

Les inconnues du problème sont les $3\,m$ valeurs de y_h aux nœuds $\{s_i : i = 0,...,m-1\}$ et les m valeurs de T_h sur les éléments $\{\,]\,s_{i-1},s_i\,[\,: i = 1,...,m\}$. On désigne par $\Delta\ell_i$ la longueur du $i$ème élément de câble,

$$\Delta{\ell_i}^2 = (y_i - y_{i-1})^2 + (y_{m+i} - y_{m+i-1})^2 + (y_{2m+i} - y_{2m+i-1})^2, \quad i = 1,...,m-1$$

et

$$\Delta{\ell_m}^2 = {y_{m-1}}^2 + {y_{2m-1}}^2 + {y_{3m-1}}^2.$$

La version discrète de *(2.16.b)* est

$$C_i(y) = \frac{\Delta{\ell_i}^2 - h^2}{2\,h} = 0, \qquad i = 1,...,m,$$

tandis que *(2.16.a)* devient un système de $3\,m$ équations

$$R(y)^\top T = D(y,v_{nv}),$$

avec $R(y)$ jacobienne de $C(y)$ et, pour $i = 0,...,3\,m-1$,

$$D_i(y,v_{nv}) = \int_0^\ell f_c(s)^\top \tilde{z}_i(s)\,ds + {f_p}^\top \tilde{z}_i(0) + m_c \int_0^\ell g^\top \tilde{z}_i(s)\,ds + m_p\, g^\top \tilde{z}_i(0).$$

Les intégrales des termes de traînée ${f_c}^\top \tilde{z}_i$ sont calculées grâce à la méthode des trapèzes, ce qui revient, compte-tenu de la forme des fonctions $\tilde{z}_i$, à évaluer f_c aux nœuds du maillage (voir en section 2.2.2).

Pour les éléments finis P1-iso-P2/P1, on doit définir un maillage $(s_i)_{i=0,...,2m}$. Les inconnues sont les $6\,m$ valeurs de y_h aux nœuds $\{s_i : i = 0,...,2\,m-1\}$ et les $m+1$ valeurs de T_h aux nœuds $\{s_{2j} : j = 0,...,m\}$. On pose

$$\Delta{\ell_i}^2 = (y_i - y_{i-1})^2 + (y_{2m+i} - y_{2m+i-1})^2 + (y_{4m+i} - y_{4m+i-1})^2, \quad i = 1,...,2\,m-1$$

et

$$\Delta\ell_{2m}{}^2 = y_{2m-1}{}^2 + y_{4m-1}{}^2 + y_{6m-1}{}^2.$$

Pour un maillage uniforme, la contrainte discrète généralise *(2.9.b)* et s'écrit sous la forme de $m+1$ équations $C(y)=0$, en ayant posé

$$\begin{cases} C_1(y) = \dfrac{3\,\Delta\ell_1{}^2 + \Delta\ell_2{}^2 - 4\,h^2}{8\,h} \\ C_i(y) = \dfrac{\Delta\ell_{2i-3}{}^2 + 3\,\Delta\ell_{2i-2}{}^2 + 3\,\Delta\ell_{2i-1}{}^2 + \Delta\ell_{2i}{}^2 - 8\,h^2}{8\,h}, \quad i = 2, ..., m, \\ C_{m+1}(y) = \dfrac{\Delta\ell_{2m-1}{}^2 + 3\,\Delta\ell_{2m}{}^2 - 4\,h^2}{8\,h}. \end{cases}$$

Quant à la discrétisation de *(2.16.a)*, elle conduit comme précédemment à un système $R(y)^\top T = D(y, v_{nv})$, où $R(y)$ est la jacobienne de $C(y)$. Le calcul de $D(y, v_{nv})$ est détaillé dans la section 2.2.2.

Dans les essais numériques, on a fixé la longueur de câble à $\ell = 4000$ m et la vitesse du navire à $v_{nv} = 1$ m.s^{-1}. D'une part, on résout *(1.23)* grâce à la routine Matlab `ode45`, qui est basée sur la méthode de Runge-Kutta d'ordre 4. Cette procédure permet d'imposer la précision avec laquelle on souhaite approcher la solution exacte. On a donc augmenté cette précision, jusqu'à stabiliser la position du poisson à 0,1 m près – ceci correspond à une majoration de l'erreur de 10^{-16} et un maillage de $[0, \ell]$ de 7670 points ! On obtient une position de référence du poisson que l'on note y_{ref}.

D'autre part, on résout, à l'aide de la méthode de Newton, les deux systèmes

$$\begin{cases} R(y)^\top T - D(y, v_{nv}) = 0 \\ C(y) = 0 \end{cases}$$

correspondant aux deux types d'éléments finis P1/P0 et P1-iso-P2/P1, définis sur des maillages uniformes. On note respectivement $n_{10} = 4m$ et $n_{21} = 7m + 1$ les nombres de variables de ces problèmes, et y_{10} et y_{21} les positions du poisson qu'ils donnent.

On a indiqué dans les tableaux 2.1 et 2.2 les écarts de position $|\, y_{10} - y_{ref}\,|$ et $|\, y_{21} - y_{ref}\,|$ pour différentes valeurs de m, choisies de sorte que n_{10} et n_{21} soient comparables.

	$m=5$	$m=10$	$m=20$	$m=50$	$m=100$	$m=200$
n_{10}	20	40	80	200	400	800
$\lvert\, y_{10} - y_{ref}\,\rvert$	31,4	10,8	3,0	0,5	0,2	0,1

TAB. 2.1 – *écarts (en mètres) entre les positions du poisson pour des éléments finis P1/P0.*

	$m=3$	$m=6$	$m=12$	$m=30$	$m=60$	$m=120$
n_{21}	22	43	85	211	421	841
$\lvert y_{21} - y_{ref} \rvert$	24,8	7,9	2,1	0,4	0,2	0,1

TAB. 2.2 – *écarts entre les positions du poisson pour des éléments finis P1-iso-P2/P1.*

Les deux schémas convergent, et ils fournissent des approximations très précises de la position du poisson, même pour de petites valeurs de m. Rapportés à la longueur du câble, les écarts relatifs de position ne dépassent pas 1%. On observe également que sur des maillages assez grossiers, la discrétisation par éléments finis P1-iso-P2/P1 est légérement plus précise.

La figure 2.4 montre la configuration de référence du câble, obtenue en résolvant (*1.23)*. Au vu de cette configuration, on peut penser qu'il serait souhaitable de raffiner le maillage vers $s = 0$ (extrémité inférieure du câble), où la courbure est plus importante.

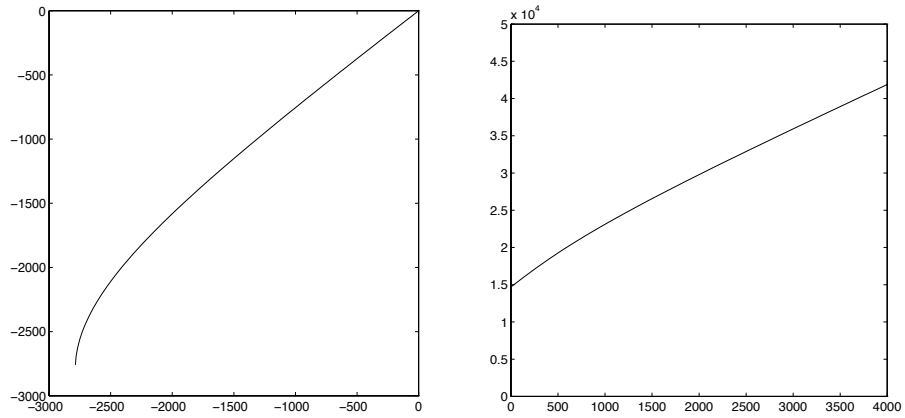


FIG. 2.4 – *configuration du câble (dans $\mathbb{R}^2$) et tension en fonction de l'abscisse curviligne.*

On a, par exemple, considéré des maillages non uniformes caractérisés par un rapport constant $\rho < 1$ entre les longueurs de mailles consécutives ; les mailles sont de plus en plus petites quand s diminue et que l'on se rapproche du poisson. Le tableau 2.3 indique les écarts de position $\lvert y_{21} - y_{ref} \rvert$ pour certaines valeurs de ρ (écarts qui doivent être comparés à ceux du tableau 2.2).

On constate que pour de petites valeurs de m, l'utilisation de ces maillages non uniformes permet de diviser l'erreur sur la position du poisson par 10. Bien entendu, on ne prétend pas ainsi minimiser l'écart $\lvert y_{21} - y_{ref} \rvert$. On montre simplement qu'il est possible d'obtenir une

approximation très précise sur un maillage extrêment grossier. Rappelons que la longueur du câble est $\ell = 4000$ m et que y_{ref} est calculée avec 7670 pas ; lorsque $m = 3$, on l'approche à 2,1 m près par y_{21} qui est calculée pour un maillage de 6 points seulement !

	$m = 3$	$m = 6$	$m = 12$	$m = 30$	$m = 60$
	$\rho = 0,6$	$\rho = 0,8$	$\rho = 0,9$	$\rho = 0,95$	$\rho = 0,95$
$\lvert y_{21} - y_{ref} \rvert$	2,1	0,6	0,2	0,1	0,1

TAB. 2.3 – *écarts (en mètres) entre les positions du poisson, en fonction de m et ρ.*

La discrétisation par éléments finis P1-iso-P2/P1 paraît avoir pour inconvénient de relaxer la condition d'inextensibilité ; en effet, la contrainte $C(y) = 0$ qui en résulte n'impose pas la longueur des éléments du câble et le taux maximal d'élongation

$$\delta\ell = \max_{i=1,\dots,2m} \left| \frac{\Delta\ell_i - (s_i - s_{i-1})}{s_i - s_{i-1}} \right|$$

n'est a priori pas nul. Le tableau 2.4 nous donne sa valeur pour des maillages uniformes ($\rho = 1$) et pour les maillages non uniformes ($\rho < 1$) du tableau 2.3. Il montre que la contrainte d'inextensibilité est numériquement bien respectée. Elle l'est d'autant plus que la finesse du maillage augmente et que celui-ci tient compte de la courbure du câble.

	$m = 3$		$m = 6$		$m = 12$		$m = 30$	
	$\rho = 1$	$\rho = 0,6$	$\rho = 1$	$\rho = 0,8$	$\rho = 1$	$\rho = 0,9$	$\rho = 1$	$\rho = 0,95$
$\delta\ell$	0,6 %	0,2 %	0,1 %	0,02 %	0,01 %	0,002 %	0,002 %	0,0002 %

TAB. 2.4 – *taux maximal d'élongation le long du câble, en fonction du maillage.*

Les tests précédents indiquent tous la convergence des deux méthodes d'éléments finis appliquées à la résolution de l'équation stationnaire. Chaque méthode fournit des solutions approchées précises même lorsqu'elles sont définies sur des maillages grossiers. Toutefois, sur ces maillages grossiers, justement, les éléments finis P1-iso-P2/P1 permettent de mieux approcher la position du poisson, tout en respectant la contrainte d'inextensibilité de façon satisfaisante.

Concluons par un dernier essai numérique, au sujet du problème que pose une tension nulle à l'extrémité inférieure du câble, et donc l'absence de poisson. Dans la section 1.3.1, nous avons remarqué que l'on ne pouvait pas intégrer l'équation différentielle *(1.23)* si la tension $T(0)$ est nulle. Pour contourner cette difficulté, certains auteurs, dont Burgess [23],

Pinto [84], Sun et Leonard [98], modifient localement les équations en introduisant des termes de rigidité en torsion ou en flexion.

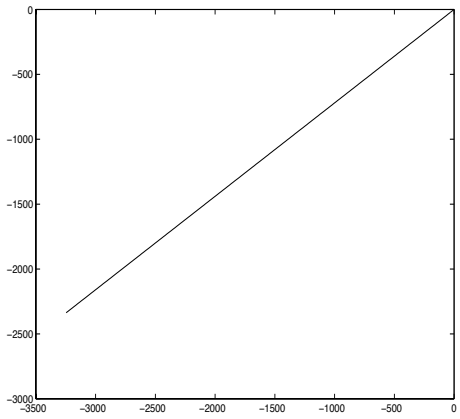


FIG. 2.5 – *configuration du câble en l'absence de poisson.*

Nous ne rencontrons pas ce problème dans notre approche par éléments finis, même pour une discrétisation très fine. La configuration du câble, calculée pour $m = 500$ et avec des éléments finis P1-iso-P2/P1, est représentée sur la figure 2.5. Cette configuration est rectiligne, ce qui est un fait bien connu des ingénieurs spécialistes des câbles. Par rapport aux méthodes basées sur l'intégration d'une équation différentielle, notre approche possède donc l'avantage de pouvoir prendre en compte une tension nulle au bout du câble.

2.2. Discrétisation de l'équation dynamique à longueur de câble fixée

L'étude du cas stationnaire suggère de discrétiser le système *(1.1)* en espace grâce à des éléments finis P1-iso-P2/P1. Les équations semi-discrètes en résultant devront ensuite être intégrées en temps. On suppose momentanément que le câble est de longueur constante ℓ. L'intérêt de cette restriction est de nous assurer la convergence du schéma d'approximation en temps que l'on a choisi d'appliquer. Dans la section 2.3, nous montrerons comment modifier les équations discrétisées pour prendre en compte les câbles de longueur variable.

2.2.1. Discrétisation en espace

Comme dans tous les cas particuliers déja étudiés, nous devons tout d'abord donner une formulation variationnelle de l'équation que l'on souhaite discrétiser. L'élément nouveau est ici la dépendance en temps des inconnues des équations *(1.1)*. La régularité de leur solution n'ayant pas été étudiée, les calculs qui suivent sont formels.

Supposons la solution (y,T) de l'équation *(1.1)* et la trajectoire du navire u régulières. Dans ce cas, la traînée f_c est également régulière, et on a pour toute fonction test $z \in \mathcal{Y}^3$,

$$m_c \int_0^\ell \ddot{y}(.,t)^\top z = \int_0^\ell (T\,y')'(.,t)^\top z + \int_0^\ell f_c(.,t)^\top z + m_c \int_0^\ell (g - \ddot{u}(t))^\top z.$$

Une intégration par parties donne, compte-tenu des conditions aux limites *(1.1.b)*,

$$\int_0^\ell (T\,y')'(.,t)^\top z = -\int_0^\ell (T\,y')(.,t)^\top z' - \Big(m_p\,\ddot{y}(0,t) - f_p(t) - m_p\,(g - \ddot{u}(t))\Big)^\top z(0).$$

Par conséquent,

$$\begin{aligned} &m_c \int_0^\ell \ddot{y}(.,t)^\top z + m_p\,\ddot{y}(0,t)^\top z(0) + \int_0^\ell (T\,y')(.,t)^\top z' = \\ &\int_0^\ell f_c(.,t)^\top z + f_p(t)^\top z(0) + m_c \int_0^\ell (g - \ddot{u}(t))^\top z + m_p\,(g - \ddot{u}(t))^\top z(0). \end{aligned}$$

On pose

$$m\,(\ddot{y}(.,t), z) = m_c \int_0^\ell \ddot{y}(.,t)^\top z + m_p\,\ddot{y}(0,t)^\top z(0)$$

et (rappelons que f_c dépend de y', $\dot{y}$ et $\dot{u}$)

$$\begin{aligned} &d\,(y'(.,t), \dot{y}(.,t), \dot{u}(t), \ddot{u}(t), z) = \\ &\int_0^\ell f_c(.,t)^\top z + f_p(t)^\top z(0) + m_c \int_0^\ell (g - \ddot{u}(t))^\top z + m_p\,(g - \ddot{u}(t))^\top z(0). \end{aligned}$$

On considère le problème variationnel suivant : pour chaque $t \geq 0$, chercher $y(.,t) \in \mathcal{Y}^3$ et $T(.,t) \in L^\infty(0,\ell)$ tels que

$$\forall\, z \in \mathcal{Y}^3, \quad m\,(\ddot{y}(.,t), z) + \int_0^\ell (T\,y')(.,t)^\top z' = d\,(y'(.,t), \dot{y}(.,t), \dot{u}(t), \ddot{u}(t), z) \qquad \textit{(2.17.a)}$$

sous la contrainte

$$\forall\, \tau \in L^\infty(0,\ell), \quad \frac{1}{2} \int_0^\ell \Big(\big|\, y'(.,t) \,\big|^2 - 1 \Big)\, \tau = 0 \qquad \textit{(2.17.b)}$$

et les conditions initiales

$$y(.,0) = y_{init}, \quad \dot{y}(.,0) = v_{init}. \qquad \textit{(2.17.c)}$$

Toute solution régulière du problème variationnel *(2.17)* est solution des équations *(1.1)* à longueur de câble constante ℓ. On le vérifie en reprenant et adaptant la démonstration de la proposition 1.8.

On peut maintenant procéder à la discrétisation en espace de *(1.1)*, en s'appuyant sur cette formulation variationnelle. On utilise les espaces d'éléments finis P1-iso-P2/P1 définis pour le cas stationnaire. Soient $(\tilde{z}_i)_{i=0,...,6m-1}$ une base de

$$\mathcal{Y}_h{}^3 = \left\{ z \in C\left([0,\ell]\right) :\ z(\ell) = 0, \quad \forall\, i = 1,...,2\,m,\ z\,\big|_{[s_{i-1},s_i]} \in P_1 \right\}^3$$

et $(\tau_i)_{i=0,...,m}$ une base de

$$\mathcal{T}_h = \left\{ z \in C\left([0,\ell]\right) :\ \forall\, j = 1,...,m,\ z\,\Big|_{[\,s_{2(j-1)},s_{2j}\,]} \in P_1 \right\}.$$

Pour tout $t \geq 0$ et toutes fonctions $y_h(.,t) \in \mathcal{Y}_h{}^3$ et $T_h(.,t) \in \mathcal{T}_h$, il existe $y(t) \in \mathbb{R}^{6m}$ et $T(t) \in \mathbb{R}^{m+1}$ tels que

$$y_h(.,t) = \sum_{i=0,...,6m-1} y_i(t)\ \tilde{z}_i, \qquad T_h(.,t) = \sum_{i=0,...,m} T_i(t)\ \tau_i.$$

On note $\Delta\ell_i(t)$ la longueur du $i$ème élément de câble ; la contrainte *(2.17.b)* aboutit à $m+1$ équations $C(y(t)) = 0$, avec

$$\left\{ \begin{array}{l} C_1(y(t)) = \dfrac{3\,\Delta\ell_1(t)^2 + \Delta\ell_2(t)^2 - 4\,h^2}{8\,h} \\[2ex] C_i(y(t)) = \dfrac{\Delta\ell_{2i-3}(t)^2 + 3\,\Delta\ell_{2i-2}(t)^2 + 3\,\Delta\ell_{2i-1}(t)^2 + \Delta\ell_{2i}(t)^2 - 8\,h^2}{8\,h}, \quad i = 2,...,m, \\[2ex] C_{m+1}(y(t)) = \dfrac{\Delta\ell_{2m-1}(t)^2 + 3\,\Delta\ell_{2m}(t)^2 - 4\,h^2}{8\,h}. \end{array} \right.$$

De *(2.17.a)* résultent $6\,m$ équations s'écrivant

$$M\,\ddot{y}(t) + R(y(t))^\top T(t) = D(y(t), \dot{y}(t), \dot{u}(t), \ddot{u}(t)),$$

avec

$$M_{i,j} = m_c \int_0^\ell z_i\, z_j + m_p\, z_i(0)\, z_j(0)$$

et (se repporter en section suivante pour le mode de calcul)

$$D_i(y(t), \dot{y}(t), \dot{u}(t), \ddot{u}(t)) = d\left(y_h(.,t), \dot{y}_h(.,t), \dot{u}(t), \ddot{u}(t), \tilde{z}_i\right).$$

La matrice M est symétrique définie positive. Comme dans chacun des cas précédents, $R(y(t)) \in \mathbb{R}^{m+1\times 6m}$ est la jacobienne de $C(y(t))$ et elle dépend linéairement de $y(t)$. Elle est surjective si la contrainte $C(y(t)) = 0$ est satisfaite ; reprendre la démonstration du lemme 2.3.

2.2.2. Calcul du vecteur D

On détaille le calcul du vecteur D pour une discrétisation par éléments finis P1-iso-P2/P1. Rappelons que l'on considère un repère orthonormé constitué de deux vecteurs horizontaux et d'un vecteur vertical ascendant. Les vecteurs des position et vitesse relatives sont y et v. La vitesse par rapport au courant du nœud i est

$$\mathrm{v} = (v_i + \dot{u}_1 - \omega_1, v_{i+2m} + \dot{u}_2 - \omega_2, v_{i+4m} + \dot{u}_3 - \omega_3).$$

Des tangentes au câble à gauche et à droite sont

$$\mathrm{t}^- = (y_i - y_{i-1}, y_{i+2m} - y_{i+2m-1}, y_{i+4m} - y_{i+4m-1}) \quad \text{(sauf au nœud 0)}$$

et

$$\begin{cases} \mathrm{t}^+ = (y_i - y_{i+1}, y_{i+2m} - y_{i+2m+1}, y_{i+4m} - y_{i+4m+1}) & \text{(au nœud } i < 2\,m-1) \\ \mathrm{t}^+ = (y_{2m-1}, y_{4m-1}, y_{6m-1}) & \text{(au nœud } 2\,m-1). \end{cases}$$

Elle permettent de définir les vitesses tangentielles à gauche et à droite du nœud i

$$\mathrm{v}_t^- = \frac{(\mathrm{v}^\top \mathrm{t}^-)\,\mathrm{t}^-}{|\,\mathrm{t}^-\,|^2} \quad \text{(sauf au nœud 0)}, \qquad \mathrm{v}_t^+ = \frac{(\mathrm{v}^\top \mathrm{t}^+)\,\mathrm{t}^+}{|\,\mathrm{t}^+\,|^2}$$

et les vitesses normales $\mathrm{v}_n^- = \mathrm{v} - \mathrm{v}_t^-$ et $\mathrm{v}_n^+ = \mathrm{v} - \mathrm{v}_t^+$ correspondantes. Enfin, pour le poisson, on a

$$\mathrm{v}_h = (v_0 + \dot{u}_1 - \omega_1, v_{2m} + \dot{u}_2 - \omega_2, 0), \qquad \mathrm{v}_z = (0, 0, v_{4m} + \dot{u}_3 - \omega_3).$$

Il s'agit de déterminer

$$D_i = \int_0^\ell {f_c}^\top \tilde{z}_i + {f_p}^\top \tilde{z}_i(0) + m_c \int_0^\ell (g - \ddot{u})^\top \tilde{z}_i + m_p\,(g - \ddot{u})^\top \tilde{z}_i(0).$$

L'intégrale du terme de traînée $f_c^\top \tilde{z}_i$ est la somme des intégrales de ce terme sur $]\,s_{i-1}, s_i\,[$ et sur $]\,s_i, s_{i+1}\,[$; on calcule ces deux intégrales grâce à la méthode des trapèzes. Tenant compte des expressions des forces de traînée (voir en section 1.1.3), on note, pour des indices $k = 1, 2, 3$ désignant les composantes des vecteurs,

$$d_{0,k} = -\frac{s_1}{2}\left(c_t \,|\, \mathrm{v}_t^+ \,|\, \mathrm{v}_{t\,k}^+ + c_n \,|\, \mathrm{v}_n^+ \,|\, \mathrm{v}_{n\,k}^+\right)$$

et si $i = 1, ..., 2m-1$,

$$d_{i,k} = -\frac{s_i - s_{i-1}}{2}\left(c_t \,|\, \mathrm{v}_t^- \,|\, \mathrm{v}_{t\,k}^- + c_n \,|\, \mathrm{v}_n^- \,|\, \mathrm{v}_{n\,k}^-\right) - \frac{s_{i+1} - s_i}{2}\left(c_t \,|\, \mathrm{v}_t^+ \,|\, \mathrm{v}_{t\,k}^+ + c_n \,|\, \mathrm{v}_n^+ \,|\, \mathrm{v}_{n\,k}^+\right).$$

On en déduit successivement

$$\begin{pmatrix} D_0 \\ D_{2m} \\ D_{4m} \end{pmatrix} = \begin{pmatrix} d_{0,1} \\ d_{0,2} \\ d_{0,3} \end{pmatrix} - \begin{pmatrix} c_h \,|\, \mathrm{v}_h \,|\, \mathrm{v}_{h1} \\ c_h \,|\, \mathrm{v}_h \,|\, \mathrm{v}_{h2} \\ c_z \,|\, \mathrm{v}_z \,|\, \mathrm{v}_{z3} \end{pmatrix} - \left(m_c\,\frac{s_1}{2} + m_p\right) \begin{pmatrix} \ddot{u}_1 \\ \ddot{u}_2 \\ |\,g\,| \end{pmatrix}$$

et si $i = 1, ..., 2m-1$,

$$\begin{pmatrix} D_i \\ D_{i+2m} \\ D_{i+4m} \end{pmatrix} = \begin{pmatrix} d_{i,1} \\ d_{i,2} \\ d_{i,3} \end{pmatrix} - m_c \left(\frac{s_{i+1} - s_{i-1}}{2} \right) \begin{pmatrix} \ddot{u}_1 \\ \ddot{u}_2 \\ |\, g \,| \end{pmatrix}.$$

La contrainte d'inextensibilité ne permet pas de garantir que t^- et t^+ ne sont pas nuls. Toutefois, cette éventualité ne s'est jamais produite dans nos essais numériques. Quoi qu'il en soit, v_t^- et v_t^+, et par suite v_n^-, v_n^+ et D, sont définis en tout point où $\mathrm{t}^- \neq 0$ et $\mathrm{t}^+ \neq 0$. En outre, en ce qui concerne sa dérivabilité par rapport à y et v, D est dérivable une fois partout où il est défini, et dérivable deux fois en chaque point où $\mathrm{v}_t^- \neq 0$, $\mathrm{v}_t^+ \neq 0$, $\mathrm{v}_n^- \neq 0$ et $\mathrm{v}_n^+ \neq 0$.

Remarquons enfin que l'expression de D dans le cas stationnaire se déduit des précédentes formules (en prenant $\ddot{u} = 0$) et que son mode de calcul est similaire pour une discrétisation par éléments finis P1/P0.

2.2.3. Discrétisation en temps

Après discrétisation en espace du système (*1.1*), on est donc ramené à la résolution du système

$$\begin{cases} M\,\ddot{y} = D(y, \dot{y}, \dot{u}, \ddot{u}) - R(y)^\top T \\ C(y) = 0, \end{cases}$$

équivalent au système

$$\begin{cases} \dot{y} = v \\ M\,\dot{v} = D(y, v, \dot{u}, \ddot{u}) - R(y)^\top T \\ C(y) = 0 \end{cases} \qquad (2.18)$$

de $13m + 1$ équations non linéaires et d'inconnues $y(t), v(t) \in \mathbb{R}^{6m}$ et $T(t) \in \mathbb{R}^{m+1}$. On le qualifie de différentiel-algébrique car il ne fait pas apparaître de dérivée de certaines de ses inconnues, en l'occurence T. Les ouvrages d'Ascher et Petzold [4], Brenan, Campbell et Petzold [16] et Hairer et Wanner [56] traitent des équations différentielles-algébriques. Des applications sont également présentées par Lubich, Nowak, Pöhle et Engstler [66]. Le système (*2.18*) est très classique et de nombreux articles sont consacrés à son étude et à sa résolution numérique. Il intervient notamment dans la conception de réseaux électriques, le contrôle de réactions chimiques, ou encore, comme c'est le cas ici, la description du mouvement de systèmes mécaniques contraints.

Le système (*2.18*) est dit d'indice 3, car au moins trois différentiations par rapport au temps sont nécessaires pour l'écrire sous forme d'une équation différentielle ordinaire. Sa deuxième ligne est équivalente à

$$\dot{v} = M^{-1} \left(D(y, v, \dot{u}, \ddot{u}) - R(y)^\top T \right);$$

une première différentiation de $C(y) = 0$ donne $R(y)\,v = 0$ et une seconde, grâce à la linéarité de R,

$$R(v)\,v + R(y)\,\dot{v} = R(v)\,v + R(y)\,M^{-1}\,D(y, v, \dot{u}, \ddot{u}) - R(y)\,M^{-1}\,R(y)^{\top}T = 0.$$

Mais $R(y)\,M^{-1}\,R(y)^{\top}$ est symétrique définie positive (car M et M^{-1} sont symétriques définies positives et $R(y)^{\top}$ est injective) donc inversible et cette dernière équation permet d'exprimer T, puis sa dérivée $\dot{T}$ après différentiation, en fonction de y, v et T.

Les deux types de méthodes les plus souvent employées pour l'intégration numérique de systèmes différentiels-algébriques sont les méthodes de différentiation rétrograde (*backward differentiation formulas*, BDF, en anglais) et les méthodes de Runge-Kutta implicites (voir Crouzeix et Mignot [33] ou Gautschi [46]). A l'origine, ces méthodes ont été conçues pour la résolution approchée d'équations différentielles

$$\begin{cases} y'(t) = f(y(t), t) \\ y(0) = y_0. \end{cases} \tag{2.19}$$

Les méthodes de Runge-Kutta permettent de calculer une valeur approchée y_{n+1} de $y(t_{n+1})$, connaissant une valeur approchée y_n de $y(t_n)$. Soit $q \in \mathbb{N}^*$. Des coefficients c_i, $a_{i,j}$ et b_j $(i, j = 1, ..., q)$ sont donnés. On pose $dt_n = t_{n+1} - t_n$ et on considère les instants intermédiaires $t_{n,i} = t_n + c_i\,dt_n$ $(i = 1, ..., q)$. On résout les q équations

$$k_{n,i} = f\left(y_n + dt_n \sum_{j=1,...,q} a_{i,j}\,k_{n,j}\ ,\ t_{n,i}\right), \quad i = 1, ..., q,$$

d'inconnues $k_{n,i}$ $(i = 1, ..., q)$. Ces équations sont non linéaires si f est non linéaire et approchent $f(y(t_{n,i}), t_{n,i})$. On pose ensuite

$$y_{n+1} = y_n + dt_n \sum_{j=1,...,q} b_j\,k_{n,j}.$$

Ceci nécessite l'évaluation de f aux q instants intermédiaires $t_{n,i}$ $(i = 1, ..., q)$. Le schéma est implicite si les coefficients $a_{i,j}$ ne sont pas nuls lorsque $j \geq i$.

En ce qui concerne les méthodes BDF, elles permettent de calculer y_{n+1} si on dispose d'approximations $y_{n-q+1}, ..., y_n$ de y aux q instants t_i $(i = n-q+1, ..., n)$. Des coefficients $a_{q,i}$ $(i = 1, ..., q)$ et b_q sont donnés. On résout alors

$$y_{n+1} = dt_n\,b_q\ f(y_{n+1}, t_{n+1}) + \sum_{i=1,...,q} a_{q,i}\ y_{n-i+1}.$$

L'évaluation de f n'est demandée qu'en t_{n+1}. Supposons que l'intervalle d'intégration de l'équation *(2.19)* soit $[\,0, t_{max}\,]$ et qu'une grille de N instants de $]\,0, t_{max}\,]$ soit fixée, grille sur laquelle on cherche à approcher y. On remarque que f doit être évaluée en $q\,N$ instants pour une méthode de Runge-Kutta et seulement N instants pour une méthode BDF.

Revenons aux équations (*2.18*). Nous ne devons pas perdre de vue qu'elles deviendront, après intégration en temps, les contraintes d'égalité du problème de demi-tour en temps minimal que l'on traite dans le chapitre 5. On utilisera un algorithme d'optimisation newtonien, qui demande à chaque itération l'évaluation non seulement de ces contraintes, mais également de leurs dérivées premières et secondes. Le calcul de $D(y, v, \dot{u}, \ddot{u})$ et de ses dérivées étant particulièrement coûteux, on préfère utiliser une méthode BDF. Signalons de plus que depuis l'article de Gear [48], l'application de cette méthode aux équations différentielles-algébriques est très courante.

Les systèmes différentiels-algébriques sont d'autant plus difficiles à intégrer numériquement que leur indice est élevé. En particulier, la résolution de (*2.18*) sous cette forme par une méthode BDF à pas dt_n variable est souvent considérée comme délicate. C'est la raison pour laquelle des techniques de réduction d'indice ont été développées (voir Gear, Leimkuhler et Gupta [49] ou Führer et Leimkuhler [44]). En revanche, lorsque le pas $dt_n = dt$ est constant, on dispose de schémas convergents (voir Brenan et Engquist [17] et Lötstedt et Petzold [65]).

Précisons quelques définitions. Si l'on veut appliquer au problème (*2.18*) une méthode BDF à q pas, on a besoin de valeurs initiales (y^k, v^k, T^k) aux temps $t_k = k\,dt$, $k = 0, ..., q-1$. On dit que ces valeurs initiales sont numériquement consistantes à l'ordre $q+1$ s'il existe une solution (y, v, T) de (*2.18*) telle que, pour tout $k = 0, ..., q-1$,

$$\left| y^k - y(t_k) \right| = O\left(dt^{q+1}\right), \quad \left| v^k - v(t_k) \right| = O\left(dt^{q+1}\right), \quad \left| C(y^k) \right| = O\left(dt^{q+2}\right). \qquad (2.20)$$

Comme dans le cas continu, il n'y a aucune condition sur les valeurs initiales de T.

Pour $k \geq q$, on note de même (y^k, v^k, T^k) les approximations de (y, v, T) aux temps $k\,dt$ $(k = q, ..., n)$. Elles sont obtenues en résolvant

$$\begin{cases} y^k = dt\, b_q\, v^k + \displaystyle\sum_{i=1,...,q} a_{q,i}\, y^{k-i} \\ v^k = dt\, b_q\, M^{-1} \left(D\left(y^k, v^k, \dot{u}(t_k), \ddot{u}(t_k)\right) - R(y^k)^\top T^k \right) + \displaystyle\sum_{i=1,...,q} a_{q,i}\, v^{k-i} \\ C(y^k) = 0. \end{cases} \qquad (2.21)$$

Etant donné un ensemble de valeurs initiales numériquement consistantes à l'ordre $q+1$, on dit que la solution $((y^k)_{k\leq n}, (v^k)_{k\leq n}, (T^k)_{k\leq n})$ de (*2.21*) converge globalement à l'ordre q vers une solution (y, v, T) de (*2.18*) si pour un indice k^* et pour tout $k = k^*, ..., n$,

$$\left| y^k - y(t_k) \right| = O\left(dt^q\right), \quad \left| v^k - v(t_k) \right| = O\left(dt^q\right), \quad \left| T^k - T(t_k) \right| = O\left(dt^q\right).$$

On trouve le résultat suivant dans [17]. Les hypothèses sont que la matrice $R(y)\, M^{-1}\, R(y)^\top$ est inversible – ce que nous avons vérifié – et que les fonctions D et C sont régulières.

Théorème 2.5. *Lorsque $q < 7$ et les valeurs initiales sont numériquement consistantes à l'ordre $q+1$, le système (2.21) possède une solution qui converge globalement à l'ordre q vers une solution de (2.18).*

Une méthode BDF à pas constant appliquée à la résolution d'une équation différentielle ordinaire est stable si son nombre de pas est strictement inférieur à 7 (voir [33]). Ceci explique la limitation $q < 7$ dans le théorème ci-dessus.

Dans nos essais numériques, à l'instant $t = 0$, le câble est en régime stationnaire. Sa vitesse relative v^0 est nulle. Sa position relative y^0 et sa tension T^0 s'obtiennent en résolvant

$$\begin{cases} R(y^0)^\top T^0 = D(y^0, 0, \dot{u}(0), 0) \\ C(y^0) = 0. \end{cases}$$

Si l'on choisit d'appliquer la méthode BDF à q pas, pour avoir un schéma d'ordre q, il est nécessaire de calculer $q-1$ autres valeurs initiales vérifiant les conditions *(2.20)*. Nous appliquons un schéma de Runge-Kutta d'ordre $q+1$, qui permet de déterminer, avec la précision voulue, (y^1, v^1, T^1) connaissant (y^0, v^0), (y^2, v^2, T^2) connaissant (y^1, v^1), et ainsi de suite jusqu'à $(y^{q-1}, v^{q-1}, T^{q-1})$.

Nous avons pris le parti de ne pas traiter les équations *(2.21)* pas de temps par pas de temps, mais sur l'ensemble des pas de temps. Cette approche paraît naturelle si l'on pense à l'utilisation de ces équations comme contraintes d'égalité dans des problèmes de contrôle du câble. En considérant le vecteur $Y = \big((y^k)_{k=1,...,n}, (v^k)_{k=1,...,n}, (T^k)_{k=1,...,n}\big)$, on écrit les différents systèmes *(2.21)* pour $k = q, ..., n$, en plus des équations permettant le calcul des valeurs initiales, sous la forme $c(Y) = 0$.

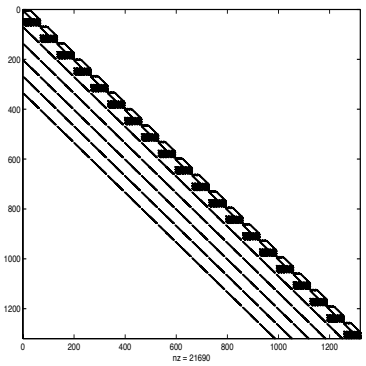


FIG. 2.6 – *éléments non nuls de $B(Y) = c'(Y)$ pour $m = 5$, $n = 20$ et $q = 5$.*

Les équations $c(Y) = 0$ dépendent aussi de la trajectoire du navire u, qui est pour l'heure donnée analytiquement. La durée de cette trajectoire étant connue, et le nombre n de pas

de temps fixé a priori, on en déduit dt. On calcule Y en résolvant $c(Y) = 0$ par la méthode de Newton. La jacobienne de c en Y, que l'on note $B(Y)$, est creuse (figure 2.6) et il est aisé de résoudre les systèmes linéaires dont elle est la matrice. L'essentiel des calculs consiste à inverser ses blocs diagonaux. Le choix d'une des deux approches (pas de temps par pas de temps ou bien ensemble des pas de temps) n'a pas de conséquence sur la quantité de calculs à effectuer.

2.2.4. Exemples de résultats numériques

La longueur du câble est toujours $\ell = 4000$ m. On suppose que le navire parcourt une ellipse centrée en l'origine et dont les axes ont pour longueurs 1000 et 4000 m. Sa loi horaire est

$$\begin{cases} u_1(t) = 1000 \cos\left(\dfrac{\pi\, t}{t_{max}}\right) \\ \\ u_2(t) = 4000 \sin\left(\dfrac{\pi\, t}{t_{max}}\right) \end{cases} \qquad t \in [\,0, t_{max}\,]\,. \tag{2.22}$$

On prend $t_{max} = 2$ h 22 min; la vitesse moyenne de parcours est alors de 1 m.s^{-1}. A l'instant $t = 0$, la configuration du câble est celle du régime stationnaire à vitesse $(0, 1)$. On fixe le nombre d'éléments de câble à 20 (donc $m = 10$). Pour juger de la façon dont la solution approchée se stabilise quand la discrétisation en temps est de plus en plus fine, on augmente le nombre n de pas de temps et on compare les positions du poisson au dernier pas de temps (tableau 2.5)

	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
$q = 3$	-491	-544	-558	-566	-574	-578	-580
	1864	2058	2117	2146	2174	2188	2196
	-3369	-3247	-3210	-3192	-3174	-3164	-3159
$q = 5$	-516	-539	-558	-566	-574	-578	-580
	1905	2057	2119	2147	2174	2188	2196
	-3339	-3247	-3210	-3192	-3174	-3164	-3159

TAB. 2.5 – *coordonnées (en mètres) de l'engin remorqué en fin de trajectoire.*

On a utilisé comme nombre de pas dans le schéma BDF successivement $q = 3$ et $q = 5$. Pour de petites valeurs de n, on obtient deux positions légèrement différentes, mais à partir de $n = 60$, les positions sont les mêmes au mètre près. L'écart entre les positions calculées pour $n = 80$ et $n = 100$ est de moins de 10 m. Rapporté à la longueur de la trajectoire du poisson, soit 8200 m environ, cet écart est de 0,12%, et rapporté aux 4000 m de longueur du câble, l'écart est de 0,25%. Au vu de ces résultats, le schéma d'approximation est convergent.

Pour la discrétisation en espace considérée ($m = 10$), le taux maximal d'allongement du câble au cours de la trajectoire est inférieur à 0,01%. La condition d'inextensibilité est bien respectée.

Les trajectoires du navire et du câble sont représentées sur la figure 2.7. La vitesse du navire à l'instant initial et sa vitesse moyenne sont de 1 m.s^{-1}. La loi horaire (*2.22)* nous indique donc que le navire accélère en début de trajectoire ; ceci explique que le poisson remonte. Le fait qu'il redescende ensuite est dû au changement de direction du navire et à son ralentissement au passage du sommet de l'ellipse.

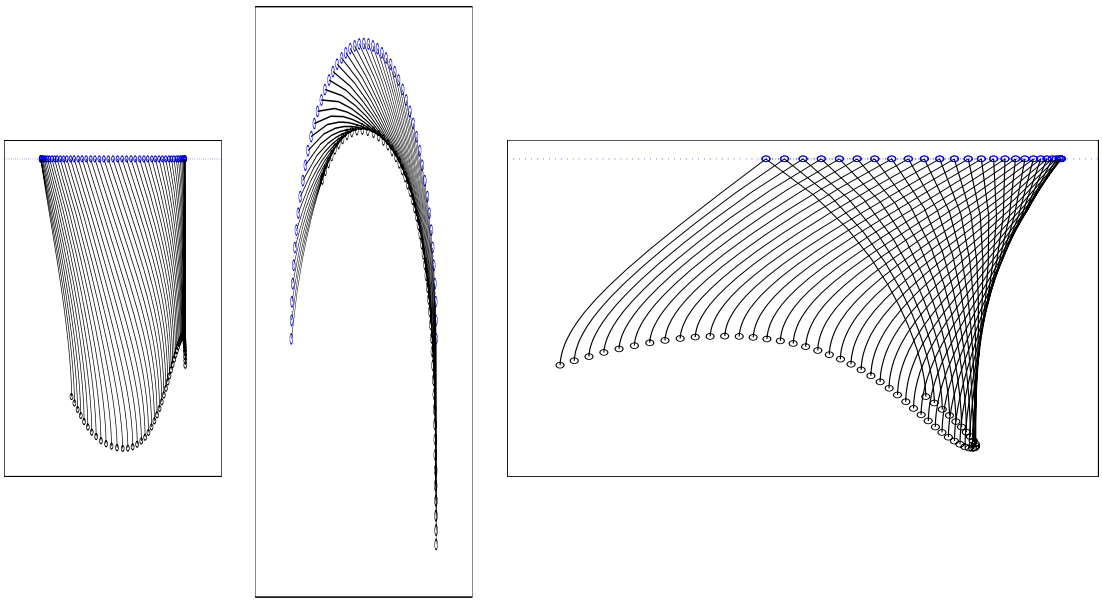


FIG. 2.7 – *au centre, vue de dessus de la trajectoire du navire et du câble, parcourue vers l'ouest ; à gauche et à droite, vues latérales dans les directions sud-nord et est-ouest.*

Ces essais numériques montrent quelles sont les limites dans la finesse de discrétisation que le traitement par ordinateur impose. Les calculs sont exécutés sous Matlab. Sur une station de travail Digital Alphastation 500 (processeur de 500 MHz), lorsque $m = 10$ et $n = 100$, une itération de Newton demande 3 minutes 10 secondes de temps CPU. Ce temps est quasi-exclusivement consacré au calcul de la jacobienne $B(Y)$ – au calcul des dérivées partielles des termes de traînée, pour être plus précis. En revanche, la résolution des systèmes linéaires diagonaux est très rapide.

2.3. Discrétisation à longueur de câble variable

Nous nous attaquons finalement à l'intégration de l'équation (*1.1)* dans sa forme la plus générale, en considérant que le câble est de longueur variable $\ell(t)$. Nous devons déterminer une reformulation du problème sur un domaine fixe.

2.3.1. Approche par changement de variable

Une manière de procéder est de ramener l'intervalle $]\,0,\ell(t)\,[$ en un intervalle fixe $]\,0,1\,[$ grâce à un changement de variable. Jusqu'à présent, nous avons utilisé l'abscisse curviligne naturelle $s \in [\,0,\ell(t)\,]$. On peut aussi introduire l'abscisse curviligne normalisée

$$\widehat{s} = \frac{s}{\ell(t)} \in [\,0,1\,].$$

Ceci n'a bien sûr de sens que si $\ell(t) > 0$. On doit alors récrire les inconnues du problème, position relative et tension, en fonction de $\widehat{s}$ et t. A toute fonction $x(s,t)$, on associe la fonction

$$\widehat{x}\,(\widehat{s},t) = x\,(\widehat{s}\,\ell(t),t) = x(s,t).$$

Les dérivées partielles de x et $\widehat{x}$ sont liées par les relations

$$\begin{cases} \dfrac{\partial x}{\partial s}\,(s,t) = \dfrac{1}{\ell(t)}\,\dfrac{\partial \widehat{x}}{\partial \widehat{s}}\,(\widehat{s},t) \\ \dfrac{\partial x}{\partial t}\,(s,t) = -s\,\dfrac{\dot{\ell}(t)}{\ell(t)^2}\,\dfrac{\partial \widehat{x}}{\partial \widehat{s}}\,(\widehat{s},t) + \dfrac{\partial \widehat{x}}{\partial t}\,(\widehat{s},t) = -\,\widehat{s}\,\dfrac{\dot{\ell}(t)}{\ell(t)}\,\dfrac{\partial \widehat{x}}{\partial \widehat{s}}\,(\widehat{s},t) + \dfrac{\partial \widehat{x}}{\partial t}\,(\widehat{s},t). \end{cases}$$

On en déduit

$$\frac{\partial}{\partial s}\left(T\,\frac{\partial y}{\partial s}\right) = \frac{1}{\ell(t)^2}\,\frac{\partial}{\partial \widehat{s}}\left(\widehat{T}\,\frac{\partial \widehat{y}}{\partial \widehat{s}}\right)$$

et

$$\begin{aligned} \frac{\partial^2 y}{\partial t^2} &= \frac{\partial}{\partial t}\left(-s\,\frac{\dot{\ell}}{\ell^2}\,\frac{\partial \widehat{y}}{\partial \widehat{s}} + \frac{\partial \widehat{y}}{\partial t}\right) \\ &= -s\left(\frac{\ddot{\ell}\,\ell - 2\,\dot{\ell}^2}{\ell^3}\right)\frac{\partial \widehat{y}}{\partial \widehat{s}} - s\,\frac{\dot{\ell}}{\ell^2}\left(-s\,\frac{\dot{\ell}}{\ell^2}\,\frac{\partial^2 \widehat{y}}{\partial \widehat{s}^2} + \frac{\partial^2 \widehat{y}}{\partial t\,\partial \widehat{s}}\right) + \left(-s\,\frac{\dot{\ell}}{\ell^2}\,\frac{\partial^2 \widehat{y}}{\partial \widehat{s}\,\partial t} + \frac{\partial^2 \widehat{y}}{\partial t^2}\right) \\ &= -\,\widehat{s}\left(\frac{\ddot{\ell}\,\ell - 2\,\dot{\ell}^2}{\ell^2}\right)\frac{\partial \widehat{y}}{\partial \widehat{s}} + \left(\widehat{s}\,\frac{\dot{\ell}}{\ell}\right)^2\frac{\partial^2 \widehat{y}}{\partial \widehat{s}^2} - 2\,\widehat{s}\,\frac{\dot{\ell}}{\ell}\,\frac{\partial^2 \widehat{y}}{\partial \widehat{s}\,\partial t} + \frac{\partial^2 \widehat{y}}{\partial t^2}. \end{aligned}$$

En particulier, l'accélération de l'engin remorqué est

$$\frac{\partial^2 y}{\partial t^2}\,(0,t) = \frac{\partial^2 \widehat{y}}{\partial t^2}\,(0,t).$$

Reformulons les équations *(1.1)*. Sont connues la position relative $y_{init}(s)$ et la vitesse relative $v_{init}(s)$ du câble à l'instant initial $t = 0$, la position du navire $u(t)$ et la longueur de la partie immergée du câble $\ell(t)$ en tout instant $t \geq 0$. On cherche $\widehat{y}(\widehat{s},t) \in \mathbb{R}^3$ et $\widehat{T}(\widehat{s},t) \in \mathbb{R}$ satisfaisant en tous $\widehat{s} \in\,]\,0,1\,[$ et $t \geq 0$

$$\begin{aligned} m_c\,\frac{\partial^2 \widehat{y}}{\partial t^2}\,(\widehat{s},t) = \frac{1}{\ell(t)^2}\,\frac{\partial}{\partial \widehat{s}}\left(\widehat{T}\,\frac{\partial \widehat{y}}{\partial \widehat{s}}\right)(\widehat{s},t) + \widehat{f}_c\,(\widehat{s},t) + m_c\left[\widehat{s}\left(\frac{\ddot{\ell}(t)\,\ell(t) - 2\,\dot{\ell}(t)^2}{\ell(t)^2}\right)\frac{\partial \widehat{y}}{\partial \widehat{s}}\,(\widehat{s},t)\right. \\ \left. - \left(\widehat{s}\,\frac{\dot{\ell}(t)}{\ell(t)}\right)^2\frac{\partial^2 \widehat{y}}{\partial \widehat{s}^2}\,(\widehat{s},t) + 2\,\widehat{s}\,\frac{\dot{\ell}(t)}{\ell(t)}\,\frac{\partial^2 \widehat{y}}{\partial \widehat{s}\,\partial t}\,(\widehat{s},t) + g - \ddot{u}(t)\right] \end{aligned}$$

et la condition d'inextensibilité

$$\left| \frac{\partial \widehat{y}}{\partial \widehat{s}} (\widehat{s}, t) \right| = \ell(t),$$

satisfaisant de plus les conditions aux limites

$$\widehat{y}(1,t) = 0, \quad m_p \frac{\partial^2 \widehat{y}}{\partial t^2} (0,t) = \frac{1}{\ell(t)} \left(\widehat{T} \frac{\partial \widehat{y}}{\partial \widehat{s}} \right) (0,t) + \widehat{f}_p(t) + m_p (g - \ddot{u}(t)), \qquad t \geq 0$$

et les conditions intiales

$$\widehat{y}(\widehat{s}, 0) = y_{init}(\widehat{s}\,\ell(0)), \quad -\widehat{s}\,\frac{\dot{\ell}(0)}{\ell(0)} \frac{\partial \widehat{y}}{\partial \widehat{s}} (\widehat{s}, 0) + \frac{\partial \widehat{y}}{\partial t} (\widehat{s}, 0) = v_{init}(\widehat{s}\,\ell(0)), \qquad \widehat{s} \in [0,1].$$

Ces équations appellent quelques commentaires. On observe que si la longueur ℓ est constante, et donc si ses dérivées sont nulles, on retrouve le système *(1.1)*. On remarque également que l'abscisse curviligne normalisée $\widehat{s}$ n'est pas définie si $\ell(t) = 0$, ce qui semble un inconvénient majeur pour la résolution du problème de changement de profil en temps minimal (chapitre 5). Rien ne garantit en effet qu'au cours de cette manœuvre, on se serait pas amené à remonter entièrement le câble, puisque ceci aurait pour effet de réduire les efforts de traînée.

Il est difficile d'interpréter la structure de ces équations ; pour le moins, cette structure est très différente de celle des équations formulées en abscisse curviligne naturelle s. Une méthode de résolution des équations à longueur de câble constante a été mise au point et nous souhaitons la modifier le moins possible pour résoudre les équations à longueur variable. Ceci ne semble pas possible en utilisant l'approche par normalisation de l'abscisse curviligne.

2.3.2. Approche par prolongement sur un intervalle fictif

Nous allons proposer une autre approche, dont l'idée sous-jacente est que si l'on prolonge l'équation *(1.1)* sur un intervalle $[0, \ell_{max}]$ contenant $[0, \ell(t)]$, quel que soit t, on ne modifie pas la configuration de la partie du câble qui nous intéresse (figure 2.8). En effet, si l'on peut déterminer $y(s,t)$ et $T(s,t)$ satisfaisant

$$\begin{cases} m_c\, \ddot{y}(s,t) = (T\, y')'(s,t) + f_c(s,t) + m_c (g - \ddot{u}(t)) \\ |\, y'(s,t)\,| = 1, \end{cases} \qquad s \in]0, \ell_{max}[,\ t \geq 0,$$

la conditions aux limites

$$m_p\, \ddot{y}(0,t) = (T\, y')(0,t) + f_p(t) + m_p (g - \ddot{u}(t)), \quad t \geq 0,$$

les conditions initiales

$$y(s,0) = y_{init}(s), \quad \dot{y}(s,0) = v_{init}(s), \quad s \in [0, \ell_{max}]$$

(on suppose que les position et vitesse initiales y_{init} et v_{init} ont été prolongées sur $[\,0, \ell_{max}\,]$) et finalement la condition

$$y(\ell(t), t) = 0,$$

alors la restriction de (y, T) à $[\,0, \ell(t)\,]$ est une solution du système *(1.1)*. On note que la condition $y(\ell(t), t) = 0$, qui joue un rôle clé dans cette reformulation de la dynamique du câble, n'est plus une condition aux limites.

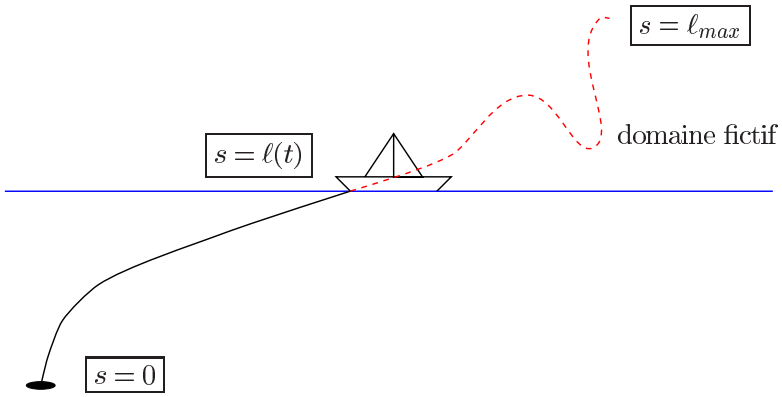


FIG. 2.8 – *introduction d'un domaine fictif.*

L'intervalle $[\,0, \ell_{max}\,]$ est souvent qualifié de fictif. Les méthodes de domaines fictifs sont d'un emploi courant. Toutes ne prennent cependant pas en compte la condition aux limites du problème initial (ici, la condition $y(\ell(t), t) = 0$) de la même manière. Certains auteurs, Glowinski, Pan et Périaux [54] par exemple, introduisent un multiplicateur (que l'on peut interpréter comme une force) pour imposer cette condition. Nous avons suivi une approche quelque peu différente.

Pour se ramener aux équations à longueur de câble constante, nous considérons comme nouvelles inconnues $\kappa(t) = y(\ell_{max}, t) \in \mathbb{R}^3$ et $w(s, t) = y(s, t) - \kappa(t)$. Le triplet (w, T, κ) vérifie, pour tous $s \in \,]\,0, \ell_{max}\,[$ et $t \geq 0$,

$$\begin{cases} m_c\, \ddot{w}(s,t) = (T\, w')'\,(s,t) + f_c(s,t) + m_c\,(g - \ddot{u}(t) - \ddot{\kappa}(t)) \\ |\,w'(s,t)\,| = 1, \end{cases} \qquad (2.23.a)$$

les conditions aux limites, en chaque $t \geq 0$,

$$w(\ell_{max}, t) = 0, \quad m_p\, \ddot{w}(0,t) = (T\, w')(0,t) + f_p(t) + m_p\,(g - \ddot{u}(t) - \ddot{\kappa}(t)) \qquad (2.23.b)$$

et les conditions initiales, en chaque $s \in [\,0, \ell_{max}\,]$,

$$w(s,0) = y_{init}(s) - \kappa(0), \quad \dot{w}(s,0) = v_{init}(s) - \dot{\kappa}(0) \qquad (2.23.c)$$

et finalement la condition, en tout $t \geq 0$,

$$w(\ell(t), t) = -\kappa(t) \qquad (2.23.d)$$

pour que $y(\ell(t), t) = 0$. Les forces de traînée f_c et f_p dépendent alors de κ, puisque la vitesse du câble par rapport au courant s'écrit maintenant $\mathbf{v}(s,t) = \dot{w}(s,t) + \dot{u}(t) + \dot{\kappa}(t) - \omega$. Notons que $\kappa(0)$ et $\dot{\kappa}(0)$ sont déterminés de manière unique par le prolongement de y_{init} et v_{init} considéré : de $\kappa(t) = y(\ell_{max}, t)$ et $\dot{\kappa}(t) = \dot{y}(\ell_{max}, t)$, on déduit

$$\kappa(0) = y_{init}(\ell_{max}, 0), \quad \dot{\kappa}(0) = v_{init}(\ell_{max}, 0).$$

Les équations *(2.23.a)* - *(2.23.b)* - *(2.23.c)* ont donc une structure similaire à celle du système *(1.1)*, mais comptent une inconnue supplémentaire, $\kappa(t) \in \mathbb{R}^3$. Celle-ci doit être déterminée grâce à l'équation *(2.23.d)*. A la différence des équations en abscisse curviligne normalisée, le système *(2.23)* a encore un sens si la longueur du câble est nulle.

Confrontons l'équation *(2.23)* à la situation décrite dans la section 1.3.2. Le navire est immobile et on déroule le câble à sa verticale. Projetées sur un axe vertical ascendant, les équations deviennent

$$\begin{cases} m_c\, \ddot{w}(s,t) = (T\, w')'\,(s,t) - m_c\,(|\, g\,| + \ddot{\kappa}(t)) \\ |\, w'(s,t)\,| = 1, \end{cases} \qquad s \in \,]\,0, \ell_{max}\,[\,,\ t \geq 0,$$

avec comme conditions aux limites

$$w(\ell_{max}, t) = 0, \quad m_p\, \ddot{w}(0,t) = (T\, w')(0,t) - m_p\,(|\, g\,| + \ddot{\kappa}(t)), \quad t \geq 0,$$

comme conditions initiales

$$w(s,0) = s - \ell_{max}, \quad \dot{w}(s,0) = 0, \quad s \in [\,0, \ell_{max}\,]$$

et finalement

$$w(\ell(t), t) = -\kappa(t), \quad t \geq 0.$$

Une solution de ce système est

$$w(s,t) = s - \ell_{max}, \quad \kappa(t) = \ell_{max} - \ell(t), \quad T(s,t) = \Big(|\, g\,| - \ddot{\ell}(t)\Big)\,(m_c\, s + m_p)$$

et on en déduit $y(s,t) = w(s,t) + \kappa(t) = s - \ell(t)$. On retrouve la solution donnée dans la section 1.3.2.

La discrétisation en espace des équations *(2.23.a)* - *(2.23.b)* - *(2.23.c)* est menée comme dans les équations à longueur de câble constante, grâce à des éléments finis P1-iso-P2/P1. Ceux-ci sont définis sur un maillage de $[\,0, \ell_{max}\,]$. Cette discrétisation conduit au système différentiel-algébrique

$$\begin{cases} M\, \ddot{w} = D(w, \dot{w}, \dot{u} + \dot{\kappa}, \ddot{u} + \ddot{\kappa}) - R(w)^\top T \\ C(w) = 0. \end{cases}$$

La condition *(2.23.d)* prend la forme $E(w,\kappa,\ell)=0$, où $E(w,\kappa,\ell)$ est le vecteur

$$w(\ell(t),t)+\kappa(t)=\begin{pmatrix} w_{i-1}(t) \\ w_{i-1+2m}(t) \\ w_{i-1+4m}(t) \end{pmatrix}+\left(\frac{\ell(t)-s_{i-1}}{s_i-s_{i-1}}\right)\begin{pmatrix} w_i(t)-w_{i-1}(t) \\ w_{i+2m}(t)-w_{i-1+2m}(t) \\ w_{i+4m}(t)-w_{i-1+4m}(t) \end{pmatrix}+\kappa(t)$$

si $\ell(t)\in[\,s_{i-1},s_i\,]$. On voit que le calcul de $w(\ell(t),t)$ est basé sur une interpolation linéaire de w, ce qui est imprécis si le nombre d'éléments de câble est réduit et si la courbure du câble est importante. Par ailleurs, la contrainte $E(w,\kappa,\ell)=0$, sous la forme présentée ici, n'est pas dérivable par rapport à ℓ aux points $\{s_i : i=0,...,2\,m\}$. Or nous aurons besoin de dériver E par rapport à ℓ : dans la suite, la longueur du câble est une inconnue. Dans notre code, nous utilisons par conséquent une version de la fonction $E(w,\kappa,\ell)$ "lissée" aux nœuds du maillage.

Finalement, nous résolvons le système différentiel-algébrique

$$\begin{cases} \dot{w}=z \\ \dot{\kappa}=\mu \\ \dot{\mu}=\nu \\ M\,\dot{z}=D(w,z,\dot{u}+\mu,\ddot{u}+\nu)-R(w)^{\top}T \\ C(w)=0 \\ E(w,\kappa,\ell)=0, \end{cases} \tag{2.24}$$

d'inconnues $w(t),z(t)\in\mathbb{R}^{6m}$, $T(t)\in\mathbb{R}^{m+1}$ et $\kappa(t),\mu(t),\nu(t)\in\mathbb{R}^3$, par la méthode des différences rétrogrades décrite précédemment. En comparaison avec le système *(2.18)*, *(2.24)* compte 3 inconnues (κ, μ et ν) et 3 équations (les deuxième, troisième et dernière lignes) supplémentaires. Une de ces inconnues supplémentaires, ν, est algébrique.

2.3.3. Résultats numériques

Voyons quels résultats donne ce modèle dans des situations simples. Dans tous les tests suivants, la longueur maximale du câble est $\ell_{max}=5000$ m et nous prenons $m=12$. La longueur totale du câble est discrétisée en 24 segments de longueur 208 m (leur longueur est de 200 m dans la section 2.2.4). A l'instant initial, la configuration du câble est toujours celle du régime stationnaire à vitesse 1 m.s^{-1} et longueur $\ell(0)=4000$ m.

Nous commençons par reprendre la trajectoire du navire *(2.22)* avec la longueur du câble fixée à $\ell=4000$ m. On compare les trajectoires du poisson obtenues en intégrant sur $n=100$ pas de temps, soit les équations *(2.18)*, soit le système *(2.24)* en supposant qu'en chaque instant t, $\ell(t)=4000$ m. L'écart maximal entre les deux trajectoires est de moins de 10 m, en dépit du facteur d'erreur que constitue la différence de longueur des éléments de câble. Il est rassurant de constater qu'en prenant une longueur de câble constante dans les équations à longueur variable, on retrouve la solution des équations à longueur constante.

Pour la même trajectoire du navire, définie par *(2.22)*, on impose, en actionnant le treuil,

$$\ell(t) = 3250 + 750 \cos\left(\frac{\pi\, t}{t_{max}}\right), \qquad t \in [\,0, t_{max}\,]. \tag{2.25}$$

Pendant toute la durée de la trajectoire, on enroule le câble, faisant passer sa longueur de 4000 à 2500 m. Comme précédemment, on augmente n et on note la position du poisson à l'instant final (tableau 2.6).

$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
-770	-794	-803	-808	-813	-817	-819
1475	1603	1649	1672	1699	1716	1730
-1915	-1809	-1768	-1746	-1722	-1707	-1696

TAB. 2.6 – *position finale du poisson.*

Cette position se stabilise, quoique peut-être légèrement moins vite que lorsque le câble est de longueur constante. Les trajectoires du navire et du câble sont représentées sur la figure 2.9.

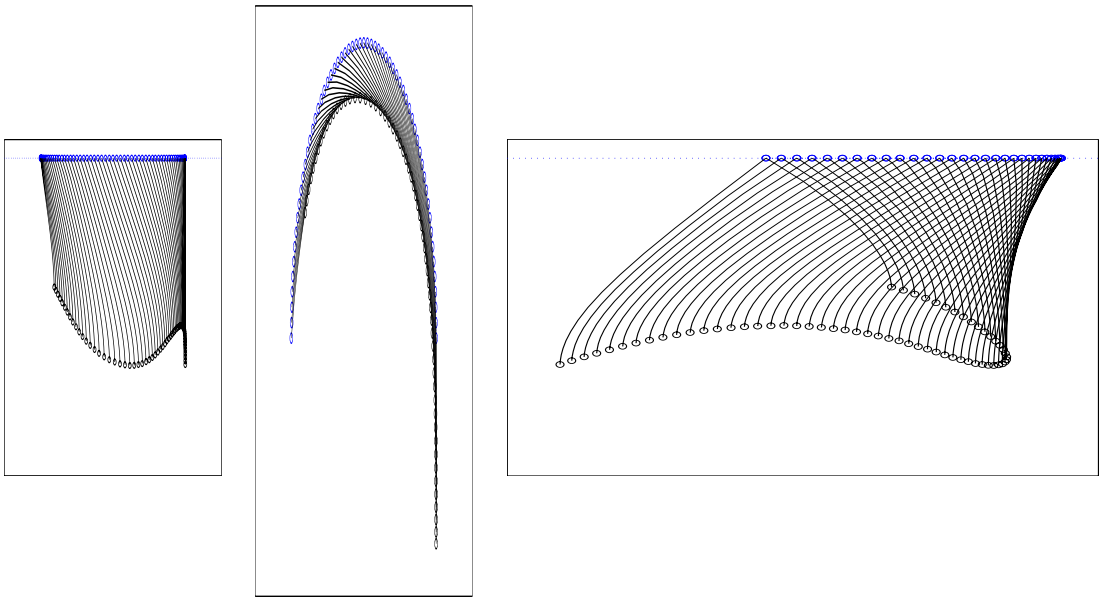


FIG. 2.9 – *vues de la trajectoire du câble lorsque sa longueur est donnée par (2.25).*

Reprenons finalement un test décrit dans [72]. La vitesse du navire est constante et de norme 1 m.s^{-1}. On prend

$$\ell(t) = 4000 + 1000 \sin\left(\frac{6\,\pi\,t}{t_{max}}\right), \qquad t \in [0, t_{max}],$$

avec $t_{max} = 2$ h 50 min. Pour cette valeur du temps final, la vitesse de filage est comprise entre 0 et 2 m.s^{-1} et la vitesse moyenne de filage est de 0,25 m.s^{-1}.

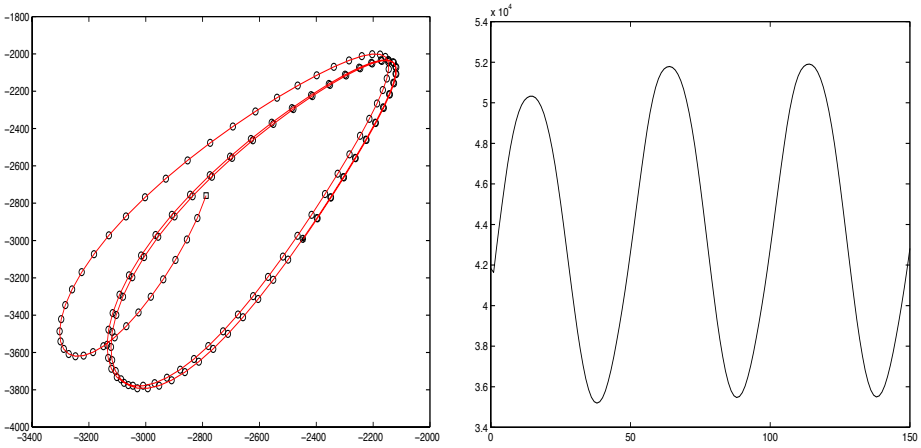


FIG. 2.10 – *position relative du poisson et tension au niveau du navire.*

La figure 2.10 montre la position relative de l'engin remorqué par rapport au navire et la tension de remorquage. Ces résultats sont similaires à ceux figurant dans [72]. La trajectoire relative du poisson par rapport au navire tend à être elliptique. Par ailleurs, comme on pouvait le prévoir, la tension oscille. On remarque que sa valeur moyenne est supérieure à la tension en régime stationnaire à longueur constante $\ell(0) = 4000$ m. Bien que cela soit difficile à distinguer sur la figure, on observe entre le temps initial et le premier pas de temps une chute de tension, due au filage du câble. Ce phénomène est d'autant plus marqué que la vitesse de filage au premier pas de temps est élevée.

Tous ces essais numériques nous semblent valider le modèle à longueur de câble variable, que nous utiliserons dans la résolution du problème du demi-tour en temps minimal. La méthode numérique choisie consiste à discrétiser la trajectoire du navire u et la longueur du câble ℓ dans les équations *(2.24)*, lesquelles deviennent des contraintes d'égalité du problème discret, puis appliquer des techniques de programmation non linéaire. Nous avons montré que l'on pouvait aisément calculer la trajectoire du câble correspondant à une trajectoire du navire en résolvant ces contraintes ; les algorithmes d'optimisation que nous proposons dans les chapitres 3 et 4 en tirent profit.

Chapitre 3

Algorithmes de Newton tronqués pour les problèmes avec contraintes d'égalité

Nous introduisons deux variantes d'une méthode de Newton tronquée pour la résolution de problèmes de programmation non linéaire avec contraintes d'égalité. Cette méthode, qui détecte les directions à courbure négative du hessien réduit au moyen d'itérations de gradient conjugué, permet de résoudre des problèmes non convexes. Sa convergence globale est assurée au moyen d'une recherche linéaire. Elle est adaptée aux problèmes de commande optimale de grande taille car le solveur de base de l'algorithme est un pas de Newton pour résoudre les équations d'état à commande fixée, opération généralement peu coûteuse dans ces problèmes.

3.1. Introduction

L'objet de ce chapitre et du chapitre suivant est de proposer un algorithme pour résoudre numériquement des problèmes de commande optimale de grande taille et de structure creuse. Nous supposons que ces problèmes peuvent s'écrire, après discrétisation, sous la forme

$$\begin{array}{rl} \text{minimiser } f(x) & \\ \text{sous les contraintes} & c_E(x) = 0 \\ & c_I(x) \leq 0. \end{array} \tag{3.1}$$

Nous notons m_E et m_I les cardinaux des ensembles d'indices E et I. La fonction objectif $f : \mathbb{R}^n \to \mathbb{R}$ et les contraintes $c_E : \mathbb{R}^n \to \mathbb{R}^{m_E}$ et $c_I : \mathbb{R}^n \to \mathbb{R}^{m_I}$ sont des fonctions non linéaires et non nécessairement convexes. Elles sont en revanche deux fois différentiables et l'algorithme utilise leurs dérivées premières et secondes. Notre méthode s'applique de manière générale à la résolution de problèmes d'optimisation non linéaires sous la forme standard *(3.1)*. Mais elle est toutefois plus particulièrement destinée aux problèmes de commande optimale, dont elle exploite la structure.

Nous définissons une partition $x = (y, u)$ des variables à optimiser. Les composantes de $y \in \mathbb{R}^{m_E}$ sont appelées variables d'état du système sur lequel porte le problème *(3.1)*. Il y en a autant que de contraintes d'égalité. Les composantes de $u \in \mathbb{R}^{n-m_E}$ sont les variables de commande. Ce sont les paramètres qui servent à modifier ou contrôler l'état

du système. On pose

$$B(x) = \frac{\partial c_E}{\partial y}(x), \quad N(x) = \frac{\partial c_E}{\partial u}(x).$$

La jacobienne des contraintes d'égalité s'écrit donc

$$c_E{}'(x) = A_E(x) = \begin{pmatrix} B(x) & N(x) \end{pmatrix}.$$

Dans les problèmes de commande optimale, il est très raisonnable de supposer que la matrice $B(x)$ est inversible. Cette propriété signifie que, grâce aux contraintes d'égalité $c_E(y,u) = 0$, appelées aussi équations d'état du système, on peut exprimer l'état y comme une fonction de la commande u. Ceci est une conséquence du théorème des fonctions implicites. Soulignons que dans *(3.1)*, les contraintes d'inégalité $c_I(y,u) \leq 0$ portent sur les deux types de variables, état et commande, de sorte que de nombreux problèmes entrent dans le formalisme que nous avons adopté.

Dans un contexte industriel, on ne décide souvent d'optimiser un système donné qu'après avoir perfectionné, au fil des années, sa modélisation. De nombreuses applications ont en commun que l'équation $F(y) = 0$ qui donne l'état du système est résolue à l'aide de la méthode de Newton. Des techniques spécifiques ont été mises au point pour résoudre le système linéaire de Newton $B\,r_y = -F$, basées par exemple sur l'exploitation de la structure creuse de la jacobienne B ou l'utilisation de parallélisme. A cet instant précis, on dispose donc de méthodes de simulation ayant prouvé leur efficacité sur ordinateur.

Puis on se demande comment ajuster des paramètres u pour minimiser ou maximiser un certain critère. Nous voulons tirer avantage des techniques numériques développées pour résoudre l'équation $F(y) = c_E(y,u) = 0$ avant que l'optimisation n'entre en jeu. L'algorithme que nous présentons entend profiter autant que possible du fait que le pas de Newton r_y est une bonne direction pour calculer l'état du système. De ce point de vue, il peut être interprété comme une méthode modifiant r_y pour atteindre l'optimalité.

C'est bien notre cheminement dans la résolution du problème du demi-tour en temps minimal. Les variables de commande du système câble – engin sont la trajectoire du navire et la longueur du câble ; elles seront discrétisées dans le chapitre 5. Les variables d'état sont la trajectoire du câble, sa tension, et d'autres inconnues auxiliaires figurant dans les équations d'état, le système différentiel-algébrique *(2.24)* intégré en temps. Nous avons constaté dans le chapitre 2 qu'il était peu coûteux de résoudre ces équations d'état à commande fixée, du fait, comme il est dit précisément plus haut, de la structure creuse de leur jacobienne par rapport aux variables d'état.

Parmi les nombreuses approches proposées pour résoudre le problème *(3.1)*, les méthodes de points intérieurs non linéaires primales-duales semblent prometteuses et sont l'objet d'une intense activité de recherche. Une abondante littérature leur est consacrée. Citons les articles de Byrd, Gilbert et Nocedal [24], El-Bakry, Tapia, Tsuchiya et Zhang [40], Forsgren et Gill [42], Gay, Overton et Wright [47] et Vanderbei et Shanno [101], parmi d'autres. Ces méthodes sont adaptées aux problèmes avec un grand nombre de contraintes d'inégalité.

En cela, elles offrent une alternative intéressante aux méthodes de gestion des contraintes actives dans le cadre de la programmation quadratique successive (*successive quadratic programming*, SQP). Leur principe est de résoudre, de façon approchée et itérative, pour des valeurs décroissantes d'un paramètre $\mu > 0$, une version perturbée des conditions d'optimalité de *(3.1)*. Pour une valeur fixée de μ, ce système d'optimalité perturbé est en fait le système d'optimalité d'un problème barrière sous contraintes d'égalité. Ceci souligne l'importance de résoudre efficacement les problèmes avec contraintes d'égalité seulement, problèmes auxquels ce chapitre est consacré.

3.2. Un premier algorithme de Newton tronqué

Nous considérons le problème de minimiser une fonction non linéaire $f : \mathbb{R}^n \to \mathbb{R}$, supposée régulière, sur une variété non linéaire définie par m contraintes d'égalité, avec $m \leq n$. Ce problème s'écrit

$$\begin{array}{c} \text{minimiser } f(x) \\ \text{sous la contrainte } c(x) = 0. \end{array} \tag{3.2}$$

La fonction $c : \mathbb{R}^n \to \mathbb{R}^m$ est également régulière. L'algorithme que nous proposons pour résoudre *(3.2)* est une extension de la méthode de Newton tronquée introduite par Dembo et Steihaug [37] pour les problèmes sans contraintes. Nous n'avons pas trouvé trace de cet algorithme dans la littérature, bien que notre approche ait des points communs avec la méthode de résolution des sous-problèmes quadratiques dans le cadre de la programmation quadratique successive sous régions de confiance (articles de Dennis, El-Alem et Maciel [39] ou Lalee, Nocedal et Plantenga [63]).

Nous désignons par $A(x) \in \mathbb{R}^{m \times n}$ la jacobienne de c en x et nous supposons qu'elle est surjective (ce pourquoi nous imposons $m \leq n$). Alors, pour toute solution x_* de *(3.2)*, il existe un unique multiplicateur de Lagrange $\lambda_* \in \mathbb{R}^m$ tel que (voir Fletcher [41] ou Luenberger [67])

$$\left\{ \begin{array}{l} \nabla f(x_*) + A(x_*)^\top \lambda_* = 0 \\ c(x_*) = 0. \end{array} \right. \tag{3.3}$$

Le membre de gauche de la première équation est le gradient par rapport à x du lagrangien $\ell(x, \lambda) = f(x) + \lambda^\top c(x)$. Si un couple (x_*, λ_*) satisfait ces conditions, x_* est appelé point stationnaire du problème *(3.2)*.

La version la plus simple de l'algorithme SQP pour déterminer une solution de *(3.2)* n'est autre que la méthode de Newton appliquée à la résolution de *(3.3)*; voir par exemple Boggs et Tolle [11] ou Bonnans, Gilbert, Lemaréchal et Sagastizábal [12]. Proche d'une solution, une itération en (x, λ) consiste à résoudre le système linéaire

$$\left(\begin{array}{cc} L(x,\lambda) & A(x)^\top \\ A(x) & 0 \end{array} \right) \left(\begin{array}{c} \Delta x \\ \lambda_{\text{PQ}} \end{array} \right) = - \left(\begin{array}{c} \nabla f(x) \\ c(x) \end{array} \right), \tag{3.4}$$

où $L(x, \lambda) = \nabla^2_{xx}\, \ell(x, \lambda)$ est le hessien du lagrangien, puis mettre à jour $x^+ = x + \Delta x$ et $\lambda^+ = \lambda_{\text{PQ}}$. La notation λ_{PQ} rappelle que ce multiplicateur est associé à la contrainte du

problème quadratique

$$\begin{array}{c}\text{minimiser } \frac{1}{2}\,\Delta x^\top L(x,\lambda)\,\Delta x + \nabla f(x)^\top \Delta x \\ \text{sous la contrainte } A(x)\,\Delta x = -c(x),\end{array}$$

dont les équations *(3.4)* constituent le système d'optimalité. On dit qu'une solution de *(3.3)* vérifie les conditions suffisantes d'optimalité du second ordre du problème *(3.2)* si $L(x_*, \lambda_*)$ est défini positif sur le noyau de $A(x_*)$. Si l'algorithme est initialisé suffisamment proche d'une telle solution, il est bien défini (puisque le système *(3.4)* est inversible) et il converge quadratiquement.

3.2.1. Réduction du système de Newton

Pour résoudre le système de Newton *(3.4)*, on procède en général de la manière suivante. Comme nous supposons que $A(x)$ est surjective, la première équation de *(3.4)* permet de déterminer λ_{PQ} dès que Δx est connu. Nous nous concentrons donc sur le calcul de Δx, que nous écrivons sous la forme

$$\Delta x = r + t\,;$$

r est une solution particulière de la contrainte linéarisée $A(x)\,r = -c(x)$ et t est un élément de $N(A(x))$, le noyau de $A(x)$. On dit que r est un pas de restauration et que t est un pas tangent, puisqu'en effet tangent à la variété $\{\bar{x} : c(\bar{x}) = c(x)\}$.

Comme nous l'avons déjà souligné dans l'introduction, dans les problèmes de commande optimale, une partition de la jacobienne $A(x)$ est

$$A(x) = \begin{pmatrix} B(x) & N(x) \end{pmatrix} = \begin{pmatrix} \dfrac{\partial c_E}{\partial y}(x) & \dfrac{\partial c_E}{\partial u}(x) \end{pmatrix},$$

avec $B(x)$ inversible. On peut alors choisir

$$r = \begin{pmatrix} r_y \\ r_u \end{pmatrix} = - \begin{pmatrix} B(x)^{-1} \\ 0 \end{pmatrix} c(x). \qquad (3.5)$$

Le vecteur $r_y = -B(x)^{-1}\,c(x)$ est le pas de Newton pour résoudre l'équation $c(.,u) = 0$ à commande u fixée. Calculer r de cette manière est fréquemment peu coûteux. C'est le cas lorsque la contrainte provient de la discrétisation d'une équation différentielle exprimant la dynamique du système. La matrice $B(x)$ est alors souvent triangulaire inférieure par blocs et creuse. On note que l'on peut prendre le pas tangent dans l'espace image de la matrice

$$\begin{pmatrix} -B(x)^{-1}\,N(x) \\ I \end{pmatrix}, \qquad (3.6)$$

dont les colonnes forment une base de $N(A(x))$.

Cette approche s'inscrit dans un cadre plus général que nous détaillons à présent. Puisque la matrice $A(x)$ est surjective, elle admet un inverse à droite $A^-(x) \in \mathbb{R}^{n\times m}$. Il s'agit d'une matrice nécessairement injective vérifiant $A(x)\,A^-(x) = I_m$. Une solution particulière de

la contrainte linéarisée est alors $r = -A^-(x)\,c(x)$. Soit par ailleurs $Z^-(x) \in \mathbb{R}^{n\times(n-m)}$ une matrice dont les vecteurs colonnes forment une base du noyau $N(A(x))$, telle que la matrice donnée par *(3.6)*. On a clairement $A(x)\,Z^-(x) = O_{m\times(n-m)}$. On peut mettre Δx sous la forme

$$\Delta x = r + t = -A^-(x)\,c(x) + Z^-(x)\,\tau,$$

où $\tau \in \mathbb{R}^{n-m}$ doit encore être spécifié. En substituant cette expression de Δx dans le système *(3.4)*, on obtient

$$L(x,\lambda)\,Z^-(x)\,\tau - L(x,\lambda)\,A^-(x)\,c(x) + A(x)^\top\lambda_{\mathrm{PQ}} = -\nabla f(x),$$

puis en multipliant à gauche par $Z^-(x)^\top$,

$$H(x,\lambda)\,\tau = -g(x) + Z^-(x)^\top L(x,\lambda)\,A^-(x)\,c(x). \qquad (3.7)$$

La matrice $H(x,\lambda) = Z^-(x)^\top L(x,\lambda) Z^-(x)$ et le vecteur $g(x) = Z^-(x)^\top \nabla f(x)$ sont appelés hessien réduit du lagrangien et gradient réduit. Le système réduit *(3.7)* permet de déterminer τ et par suite Δx.

On observe en premier lieu que $H(x,\lambda)$ est une matrice carrée d'ordre $n-m$, ce qui, dans les problèmes de commande optimale, correspond au nombre de variables de commande. La résolution de *(3.7)* s'en trouvant facilitée, cette réduction du système de Newton est d'autant plus avantageuse que la dimension de la commande est petite par rapport à celle de l'état.

On peut comparer le système *(3.7)* au système de Newton en optimisation sans contrainte,

$$\nabla^2 f(x)\,\Delta x = -\nabla f(x),$$

pour la résolution duquel Dembo et Steihaug ont proposé la méthode de Newton tronquée. N'exploitant que la partie définie-positive de $\nabla^2 f(x)$, cette méthode permet de traiter des problèmes non convexes. Elle détermine une direction de descente de f même en des points éloignés d'un minimum de f. Ce ne serait pas nécessairement le cas si on résolvait le système de Newton ci-dessus de manière exacte. D'autres méthodes sont basées sur une factorisation de Cholesky modifiée du hessien ou du hessien réduit ; voir par exemple le livre de Gill, Murray et Wright [52] ou l'article plus récent de Cheng et Higham [28].

L'algorithme ne requiert pas un calcul explicite des matrices $A^-(x)$, $Z^-(x)$ et $H(x,\lambda)$. En effet, seuls interviennent leurs produits par des vecteurs. Dans le cadre des problèmes de commande optimale, $-A^-(x)\,c(x)$ est déterminé grâce à *(3.5)*, qui consiste à résoudre le système linéaire de matrice $B(x)$ et de second membre $c(x)$. De même, on calcule $Z^-(x)\,\tau$ à l'aide de *(3.6)*, en résolvant le système linéaire de matrice $B(x)$ et second membre $N(x)\,\tau$. Toujours parce que seul le produit de $H(x,\lambda)$ par un vecteur est nécessaire, notons qu'au lieu des dérivées secondes du critère et des contraintes, on pourrait se contenter des dérivées directionnelles de leurs gradients.

Nous allons appliquer la méthode de Newton tronquée à la résolution du système *(3.7)*. Mais il nous a semblé également intéressant d'appliquer cette méthode à la résolution de $H(x,\lambda)\,\tau = -g(x)$. Cette variante est détaillée dans la section 3.4.

3.2.2. Itérations de gradient conjugué tronquées

Soit à résoudre le système linéaire

$$H\tau = v, \tag{3.8}$$

de matrice H symétrique mais pas nécessairement définie positive. Nous en calculons une solution par itérations de gradient conjugué, itérations que nous interrompons si une direction à courbure négative est rencontrée.

Par abus de langage, nous appelons direction à courbure négative toute direction conjuguée le long de laquelle H n'a pas un quotient de Rayleigh suffisamment positif, c'est-à-dire toute direction δ telle que

$$\delta^\top H\,\delta \leq \varepsilon\,|\,\delta\,|^2\,,$$

où ε est une constante strictement positive. On désigne par $|\,\delta\,|$ la norme euclidienne de δ. L'algorithme peut être interrompu à tout moment avant de détecter une de ces directions. Si on l'applique à la résolution de *(3.7)*, on souhaitera s'arrêter d'autant plus vite que l'on est éloigné d'une solution de *(3.3)*, autrement dit en des points où les dérivées secondes n'apportent pas d'information utile. Il est délicat de définir un tel test d'arrêt. On remarquera cependant que ce qui suit s'étend facilement au cas où ε dépend de l'itération externe, pourvu que la suite de ses valeurs soit minorée par une constante strictement positive.

Soit $\rho(\tau) = H\tau - v$ le résidu de l'équation *(3.8)*. On fixe une constante $\varepsilon > 0$. L'algorithme est le suivant.

Itération 0.
a. Poser $i = 0$, $\tau_0 = 0$ et $\delta_0 = -\rho(0)$.
b. Si ${\delta_0}^\top H\,\delta_0 \leq \varepsilon\,|\,\delta_0\,|^2$, terminer avec $\tau = \delta_0$.
c. Calculer $\tau_1 = -\dfrac{\rho(0)^\top \delta_0}{{\delta_0}^\top H\,\delta_0}\,\delta_0$ et passer à l'itération 1.

Itération i, $i \geq 1$.
a. Si un test de convergence sur $\rho(\tau_i)$ est satisfait, terminer avec $\tau = \tau_i$.
b. Calculer $\delta_i = -\rho(\tau_i) + \dfrac{|\,\rho(\tau_i)\,|^2}{|\,\rho(\tau_{i-1})\,|^2}\,\delta_{i-1}$.
c. Si ${\delta_i}^\top H\,\delta_i \leq \varepsilon\,|\,\delta_i\,|^2$, terminer avec $\tau = \tau_i$.
d. Calculer $\tau_{i+1} = \tau_i - \dfrac{\rho(\tau_i)^\top \delta_i}{{\delta_i}^\top H\,\delta_i}\,\delta_i$ et passer à l'itération $i+1$.

Théoriquement, cette procédure doit converger en un nombre fini d'itérations, qui est au plus la dimension de H. Mais il est connu que la méthode du gradient conjugué est sensible aux erreurs d'arrondi et ce nombre maximal d'itérations peut être largement dépassé. C'est pourquoi nous limitons en pratique le nombre d'itérations. Un seul produit

matrice – vecteur est nécessaire à l'itération i (le produit $H\,\delta_i$) si l'on utilise la formule de mise à jour du résidu

$$\rho(\tau_{i+1}) = \rho(\tau_i) - \frac{\rho(\tau_i)^\top \delta_i}{\delta_i{}^\top H \delta_i} H\delta_i.$$

L'algorithme construit des itérés $(\tau_i)_{i=0,\dots,p}$ en partant de $\tau_0 = 0$. La première direction explorée est l'opposée du résidu initial $\delta_0 = -\rho(0) = v$. En optimisation sans contrainte, il s'agit de la direction opposée au gradient, autrement dit la direction de descente de plus forte pente. L'algorithme s'arrête sur cette direction si elle est à courbure négative. La solution du système *(3.8)* est alors approchée par $\tau = v$. Dans le cas contraire, un ensemble de directions conjuguées $(\delta_i)_{i=0,\dots,p}$ est généré. Celles-ci vérifient

$$\forall\, i = 0, \dots, p, \quad \rho(\tau_i)^\top \delta_i = (H\,\tau_i - v)^\top \delta_i = -v^\top \delta_i,$$

puisque τ_i est dans le sous-espace engendré par les $(\delta_j)_{j=0,\dots,i-1}$ et $\delta_j{}^\top H\,\delta_i = 0$ si $j \neq i$. On approche la solution de *(3.8)* par

$$\tau = \sum_{i=0,\dots,p} \frac{\delta_i{}^\top v}{\delta_i{}^\top H\,\delta_i}\,\delta_i = \left(\sum_{i=0,\dots,p} \frac{\delta_i\,\delta_i{}^\top}{\delta_i{}^\top H\,\delta_i} \right) v.$$

Pour résumer l'alternative précédente, on considère comme approximation de la solution du système $H\,\tau = v$ le vecteur $\tau = J\,v$, où la matrice J est donnée par

$$J = \begin{cases} I & \text{si } v^\top H\,v \leq \varepsilon\,|\,v\,|^2 \\ \displaystyle\sum_{i=0,\dots,p} \frac{\delta_i\,\delta_i{}^\top}{\delta_i{}^\top H\,\delta_i} & \text{sinon.} \end{cases} \qquad (3.9)$$

Insistons sur le fait que la matrice semi-définie positive J n'est pas calculée explicitement. Nous ne l'introduisons ici que pour faciliter l'analyse de l'algorithme. Elle peut être vue comme une approximation de l'inverse de H ; on a même $J = H^{-1}$ si H est définie positive et si toutes les directions de l'espace ont été explorées. Mais si l'on s'intéresse à la mise en œuvre de l'algorithme, on ne verra en $J\,v$ qu'une notation pour désigner la solution approchée du système *(3.8)* calculée par la procédure d'itérations de gradient conjugué tronquées que nous venons de décrire.

Revenons à la résolution du système réduit *(3.7)*. Si l'on omet les arguments x et λ, son second membre s'écrit $-g + Z^{-\top} L\,A^- c$. Nous obtenons, en appliquant l'algorithme précédent, une approximation $J\left(-g + Z^{-\top} L\,A^- c\right)$ de sa solution. Par suite, la direction de Newton tronquée

$$\Delta x = -A^- c + Z^- J\left(-g + Z^{-\top} L\,A^- c\right) = -A^-_{\text{NT}}\,c - Z^- J\,g \qquad (3.10)$$

est une solution approchée de *(3.4)*. Nous avons ainsi défini un nouvel inverse à droite de A, la matrice

$$A^-_{\text{NT}} = \left(I - Z^- J\,Z^{-\top} L\right) A^-.$$

On peut considérer le pas de restauration $-\bar{A}_{\text{NT}}\, c$ comme la somme du pas de restauration $-A^{-}c$ "fourni par l'utilisateur" (le pas de Newton à commande fixée, dans les problèmes de commande optimale) et d'une correction $Z^{-}J\, Z^{-\top}L\, A^{-}c$ de ce pas dans l'espace tangent. C'est ce qu'illustre la figure 3.1.

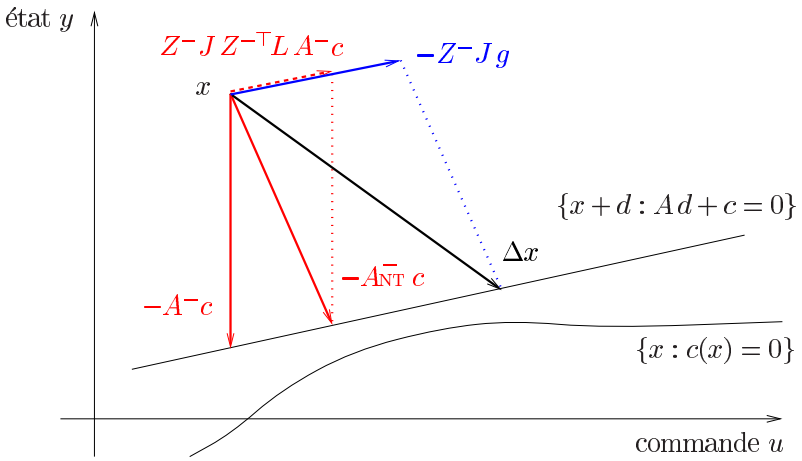


FIG. 3.1 – *pas de restauration* $-\bar{A}_{\text{NT}}\, c$ *et pas tangent* $-Z^{-}J\, g$.

La direction *(3.10)* est une solution exacte de la contrainte linéarisée et une solution approchée de la condition d'optimalité linéarisée dans le système *(3.4)*. Si toutefois le hessien réduit H est défini positif (ce qui se produit au voisinage d'un point stationnaire de *(3.2)* vérifiant les conditions suffisantes d'optimalité du second ordre) et si ε est assez petit, les itérations de gradient conjugué peuvent être menées jusqu'à résolution exacte du système réduit. Dans ce cas, la direction de Newton tronquée Δx est la direction de Newton exacte et

$$\lambda_{\text{PQ}} = -A^{-\top}(\nabla f + L\, \Delta x).$$

Détaillons une itération de l'algorithme de Newton tronqué. On calcule, par itérations de gradient conjugué tronquées, une solution approchée du système réduit *(3.7)*. On en déduit une solution approchée $(\Delta x, \lambda^{+})$ du système de Newton *(3.4)*; Δx est donné par la formule *(3.10)* et $\lambda^{+} = -A^{-\top}(\nabla f + L\, \Delta x)$. Enfin, on met à jour $x^{+} = x + \Delta x$. Cette méthode souffre bien sûr du même inconvénient que la méthode de Newton classique: elle n'est pas globalement convergente. Pour remédier à cela, on peut recourir à deux classes de techniques différentes, celles des régions de confiance et celles des recherches linéaires.

3.3. Globalisation par recherche linéaire

Nous allons tout d'abord expliquer pourquoi globaliser la méthode par recherche linéaire nous semble mieux approprié, malgré la réputation de robustesse que se sont forgée les méthodes à régions de confiance.

A ce stade de leur développement, ces dernières ne peuvent pas employer avec profit le pas de restauration à commande fixée. Byrd, Hribar et Nocedal [25], Lalee, Nocedal et Plantenga [63] ou Omojokun [80] utilisent un pas de restauration orthogonal à la contrainte linéarisée, du type $-A^\top \left(A\,A^\top\right)^{-1} c$; l'extrémité de ce déplacement est le centre de la boule $B_\delta = \{x+d : A\,d+c=0,\, |\,d\,| \leq \delta\}$, intersection entre la contrainte linéarisée et la région de confiance (de rayon δ ajusté de sorte que B_δ ne soit pas vide). Ils déterminent, à partir de ce point, un pas tangent par itérations de gradient conjugué tronquées.

Dans les problèmes de commande optimale de grande taille, le pas à commande fixée $-A^- c$ est bien moins coûteux qu'un pas orthogonal. Dans certaines situations, comme celle qu'illustre la figure 3.2, $-A^- c$ pointe vers le bord de B_δ, de même que $-A_{\rm NT}^- c$ si, par exemple, peu d'itérations de gradient conjugué ont été effectuées. Or les techniques actuelles ne permettent pas de calculer un pas tangent à partir d'un point excentré de B_δ.

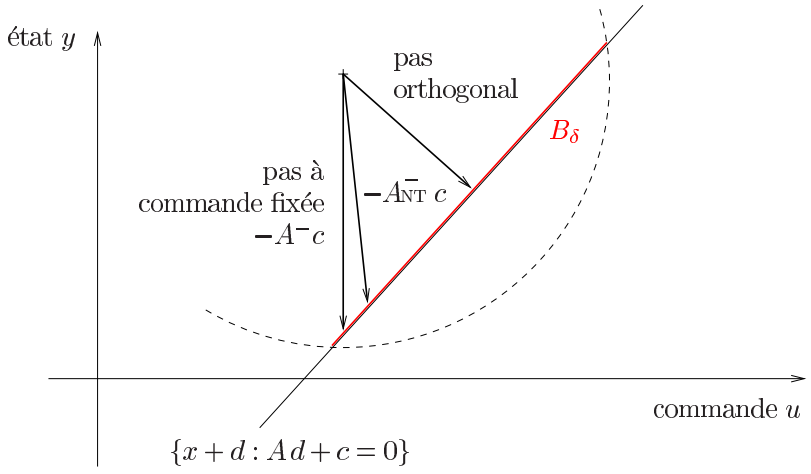


FIG. 3.2 – *pas de restauration orthogonal et à commande fixée.*

Pour ces raisons, l'utilisation du pas de restauration à commande fixée, avantageuse dans le cadre qui nous intéresse, exclut une globalisation de la méthode de Newton tronquée au moyen de régions de confiance.

3.3.1. Choix d'une fonction de mérite

Dans la globalisation par recherche linéaire, on remplace la formule de mise à jour locale $x^+ = x + \Delta x$ par la formule $x^+ = x + \alpha\,\Delta x$. Le pas $\alpha > 0$ sera choisi de manière à faire suffisamment décroître une certaine fonction de mérite. Il s'agit d'une fonction prenant en compte les deux objectifs du problème *(3.2)* : minimiser le critère f et satisfaire la contrainte $c(x) = 0$. Nous proposons d'utiliser à cet effet la fonction de pénalisation exacte

$$\phi_\sigma(x) = f(x) + \sigma\,||\,c(x)\,||\,. \qquad (3.11)$$

La norme $||\,.\,||$ est une norme arbitraire sur $\mathbb{R}^m$, dont nous notons $||\,.\,||_d$ la norme duale :

$$||\,y\,||_d = \max_{||\,x\,||=1} y^\top x.$$

On dit d'une fonction de pénalisation qu'elle est exacte lorsque tout minimum local de *(3.2)* est un minimum local de cette fonction. Han et Mangasarian [58] ont demontré que si x_* est un minimum local de *(3.2)* satisfaisant, avec un multiplicateur λ_*, les conditions suffisantes du second ordre de ce problème, alors x_* est un minimum local strict de ϕ_σ dès que $\sigma > ||\,\lambda_*\,||_d$.

Une autre fonction de pénalisation exacte est le lagrangien augmenté

$$\mathcal{L}(x,\lambda,\sigma) = f(x) + \lambda^\top c(x) + \frac{\sigma}{2}\,||\,c(x)\,||^2\,.$$

Certains auteurs, dont Conn, Gould et Toint [30] ou Gill, Murray, Saunders et Wright [51], l'utilisent comme fonction de mérite. D'autres, Han [57], Powell [88] ou Pshenichnyi et Danilin [89], emploient une fonction de type ϕ_σ, le plus souvent avec la norme $\ell 1$.

Proposition 3.1. *Supposons que x ne soit pas un point stationnaire de (3.2). Posons*

$$\lambda_{\text{NT}} = -\bar{A_{\text{NT}}}^\top \nabla f = -A^{-\top}\left(\nabla f - L\,Z^-\,J\,g\right). \tag{3.12}$$

La direction de Newton tronquée Δx donnée par (3.10) est une direction de descente de la fonction ϕ_σ en x si $\sigma > ||\,\lambda_{\text{NT}}\,||_d$.

Démonstration. La fonction ϕ_σ n'est pas différentiable en un point x satisfaisant la contrainte $c(x) = 0$. Elle admet toutefois des dérivées directionnelles en tout point. On calcule facilement sa dérivée directionnelle dans une direction satisfaisant la contrainte linéarisée $A(x)\,\Delta x = -c(x)$. Pour un tel Δx, et pour $t > 0$,

$$c(x + t\,\Delta x) = c(x) + t\,A(x)\,\Delta x + o(t) = (1-t)\,c(x) + o(t).$$

Il s'ensuit que

$$\lim_{t\to 0} \frac{||\,c(x+t\,\Delta x)\,|| - ||\,c(x)\,||}{t} = -\,||\,c(x)\,||$$

et donc $\phi_\sigma{}'(x,\Delta x) = \nabla f(x)^\top \Delta x - \sigma\,||\,c(x)\,||$. Si de plus Δx est calculé selon *(3.10)*, on obtient

$$\begin{aligned}\phi_\sigma{}'(x,\Delta x) &= -g^\top J\,g - \nabla f^\top \bar{A_{\text{NT}}}\,c - \sigma\,||\,c\,|| = -g^\top J\,g + \lambda_{\text{NT}}{}^\top c - \sigma\,||\,c\,|| \\ &\le -g^\top J\,g + (||\,\lambda_{\text{NT}}\,||_d - \sigma)\,||\,c\,||\,. \end{aligned} \tag{3.13}$$

Supposons que $\sigma > ||\,\lambda_{\text{NT}}\,||_d$ et $\phi_\sigma{}'(x,\Delta x) = 0$. Alors $g^\top J\,g = 0$ et $c = 0$. La première direction interne est alors $\delta_0 = -g$; si elle est à courbure positive, l'inégalité

$$g^\top J\,g = \sum_{i=0,\ldots,p} \frac{\left(\delta_i{}^\top g\right)^2}{\delta_i{}^\top H\,\delta_i} \ge \frac{\left(\delta_0{}^\top g\right)^2}{\delta_0{}^\top H\,\delta_0} = \frac{|\,g\,|^4}{g^\top H\,g}$$

implique que $g = 0$. Dans le cas contraire, $J = I$ et on a de même $g = 0$. On en déduit que $\nabla f(x) \in N(Z^-(x)^\top)$, qui est l'espace image de $A(x)^\top$, et par conséquent qu'il existe un multiplicateur λ tel que $\nabla f(x) + A(x)^\top \lambda = 0$. Par suite, (x, λ) est une solution de *(3.3)*. En conclusion, si $\sigma > \| \lambda_{\text{NT}} \|_d$ et si x n'est pas un point stationnaire de *(3.2)*, alors $\phi_\sigma{}'(x, \Delta x) < 0$, c'est-à-dire que Δx est une direction de descente de ϕ_σ. □

On reconnait en λ_{NT} l'unique solution de l'équation $A^\top \lambda_{\text{NT}} = -\nabla f + L\, Z^- J\, g$. On a donc

$$\begin{cases} -L\, Z^- J\, g + A^\top \lambda_{\text{NT}} = -\nabla f \\ -A\, Z^- J\, g = 0, \end{cases}$$

c'est-à-dire que $(-Z^- J\, g, \lambda_{\text{NT}})$ est un point stationnaire du problème

$$\begin{array}{c} \text{minimiser } \frac{1}{2}\, \Delta x^\top L(x, \lambda)\, \Delta x + \nabla f(x)^\top \Delta x \\ \text{sous la contrainte } A(x)\, \Delta x = 0. \end{array}$$

Ceci montre que λ_{NT} est un multiplicateur de Lagrange. Nous observons par ailleurs que $\lambda_{\text{NT}} - \lambda_{\text{PQ}} = -A^{-\top} L\, A_{\text{NT}}^{-} c$. Les deux multiplicateurs sont égaux lorsque l'itéré courant x satisfait la contrainte $c(x) = 0$. Si l'algorithme converge vers une solution (x_*, λ_*) du système d'optimalité *(3.3)*, λ_{PQ} et λ_{NT} ont même limite λ_*.

Précisons le mode de calcul de λ_{NT}. Il est nécessaire d'évaluer $-J\, g$, qui est égal à

$$\begin{cases} -g & \text{si} \quad {\delta_0}^\top H\, \delta_0 \leq \varepsilon\, |\, \delta_0\, |^2 \\ -\displaystyle\sum_{i=0,\ldots,p} \frac{\delta_i\, {\delta_i}^\top}{{\delta_i}^\top H\, \delta_i}\, g = -\sum_{i=0,\ldots,p} \frac{{\delta_i}^\top g}{{\delta_i}^\top H\, \delta_i}\, \delta_i & \text{sinon.} \end{cases}$$

On rappelle que les directions conjuguées $(\delta_i)_{i=0,\ldots,p}$ sont générées par des itérations de gradient conjugué sur le système réduit $H\, \tau = -g + Z^{-\top} L\, A^- c$. Calculer $-J\, g$ n'est rien de plus qu'appliquer la méthode des directions conjuguées au système $H\, \tau = -g$, en partant de $\tau_0 = 0$ et en utilisant les directions $(\delta_i)_{i=0,\ldots,p}$. Par conséquent, on mène la résolution des deux systèmes

$$\begin{cases} H\, \tau = -g + Z^{-\top} L\, A^- c \\ H\, \tau = -g \end{cases}$$

de front, en utilisant les mêmes directions conjuguées pour chacun d'eux. Cette procédure ne demande toujours qu'un seul produit matrice – vecteur par itération (car la mise à jour des résidus du second système linéaire et le calcul du pas se font avec les mêmes produits matrice – vecteur). Le calcul de $-J\, g$ n'entraîne donc aucun surcoût par rapport à la résolution du système réduit. Il ne reste ensuite qu'à déterminer $-A^{-\top}(\nabla f - L\, Z^- J\, g)$, ce qui, pour un problème de commande optimale ayant la structure décrite en section 3.2.1, revient à résoudre un système linéaire de matrice $B^\top$.

A l'itération k, nous devons choisir un paramètre de pénalisation $\sigma_k \geq \| \lambda_{\text{NT}\,k} \|_d$ qui nous garantit que la direction de Newton tronquée *(3.10)* est une direction de descente de ϕ_{σ_k}.

Mais on ne doit pas se contenter de cette inégalité. On peut fixer une constante $\overline{\sigma} > 0$ et demander que

- pour chaque k, $\sigma_k \geq || \lambda_{\text{NT}\,k} ||_d + \overline{\sigma}$, *(3.14.a)*
- il existe k_0 tel que, si $k \geq k_0$ et $\sigma_{k-1} \geq || \lambda_{\text{NT}\,k} ||_d + \overline{\sigma}$, alors $\sigma_k = \sigma_{k-1}$, *(3.14.b)*
- si (σ_k) est bornée, (σ_k) prenne un nombre fini de valeurs. *(3.14.c)*

La condition (b) entraîne qu'après k_0 itérations, σ_k ne sera plus modifié si la condition (a) est remplie, et implique que $(\sigma_k)_{k \geq k_0}$ est croissante. Le point (c) concerne le cas favorable où la fonction de mérite est asymptotiquement indépendante de l'itération, ce qui est essentiel dans la preuve de convergence. Soit σ son paramètre constant ; la condition (a) assure que $\sigma > || \lambda_* ||_d$ si la suite $(\lambda_{\text{NT}\,k})$ converge vers λ_*, c'est-à-dire que la fonction ϕ_σ est exacte.

Les trois conditions *(3.14)* définissent un cadre général pour le choix du paramètre σ. Elles ne sont cependant pas respectées par la règle de mise à jour suivante, que nous avons mise en pratique. Nous prenons

$$\sigma_k = \begin{cases} \dfrac{1}{2}\Big(\sigma_{k-1} + || \lambda_{\text{NT}\,k} ||_d + \overline{\sigma}\Big) & \text{si} \quad \sigma_{k-1} \geq c_1\Big(|| \lambda_{\text{NT}\,k} ||_d + \overline{\sigma}\Big) \\ \max\Big(c_2\,\sigma_{k-1}, || \lambda_{\text{NT}\,k} ||_d + \overline{\sigma}\Big) & \text{si} \quad \sigma_{k-1} < || \lambda_{\text{NT}\,k} ||_d + \overline{\sigma} \\ \sigma_{k-1} & \text{sinon.} \end{cases} \qquad (3.15)$$

Les constantes c_1 et c_2 sont strictement supérieures à 1, par exemple $c_1 = 1,1$ et $c_2 = 1,5$. On remarque que $\sigma_k \leq \sigma_{k-1}$ si $\sigma_{k-1} \geq c_1\,(|| \lambda_{\text{NT}\,k} ||_d + \overline{\sigma})$, alors que les conditions *(3.14)* ne nous autorisent pas à réduire σ. Cela peut pourtant être avantageux lorsque, par exemple, les premiers termes de la suite $(\lambda_{\text{NT}\,k})$ sont de beaucoup trop grandes estimations de λ_*. On évite ainsi que σ_k soit trop grand par rapport à λ_* et le mauvais conditionnement de ϕ_{σ_k} que cela pourrait occasionner.

Le pas α est déterminé comme suit. Dans la globalisation des méthodes newtoniennes, on impose généralement que soit satisfait le critère d'Armijo [3]

$$\phi_{\sigma_k}(x_k + \alpha\,\Delta x_k) \leq \phi_{\sigma_k}(x_k) + \omega\,\alpha\,\phi_{\sigma_k}{}'(x_k, \Delta x_k), \qquad (3.16)$$

où $\omega \in \,]\,0,1\,[$ est un coefficient constant. Trouver un pas α vérifiant l'inégalité *(3.16)* est toujours possible, car $\phi_{\sigma_k}{}'(x_k, \Delta x_k) < 0$. Nous avons effectivement

$$\lim_{\alpha \to 0} \frac{\phi_{\sigma_k}(x_k + \alpha\,\Delta x_k) - \phi_{\sigma_k}(x_k)}{\alpha} = \phi_{\sigma_k}{}'(x_k, \Delta x_k) < \omega\,\phi_{\sigma_k}{}'(x_k, \Delta x_k),$$

inégalité qui prouve que *(3.16)* est vérifiée pour $\alpha > 0$ suffisamment petit. Dans le même temps, nous tenons compte du fait que $\alpha = 1$ est une valeur idéale, car permettant d'avoir la convergence quadratique de la méthode de Newton. Nous calculons donc α en

appliquant la procédure de rebroussement suivante, où β est une constante choisie dans l'intervalle $]\,0;0,5\,]$.

a. Initialiser : $\alpha = 1$.
b. Tant que α ne satisfait pas le critère d'Armijo (*3.16*), choisir un nouveau pas α dans $[\,\beta\,\alpha, (1-\beta)\,\alpha\,]$.
c. Poser $\alpha_k = \alpha$.

La boucle *b.* se termine toujours en un nombre fini d'étapes. On s'offre la possibilité d'utiliser des formules d'interpolation de ϕ_{σ_k} en prenant un nouvel α dans $[\,\beta\,\alpha, (1-\beta)\,\alpha\,]$. Comme α_k est le premier (donc le plus grand) des pas satisfaisant l'inégalité (*3.16*), celui-ci ne sera pas trop petit ; nous montrerons que la suite (α_k) a un minorant strictement positif.

Il est bien connu que le pas unité peut ne pas être asymptotiquement accepté dans (*3.16*). On perd alors la convergence quadratique locale de l'algorithme. On peut remédier à cela en ajoutant au pas de Newton, lorsque nécessaire, un pas appelé correction du second ordre. Cette technique, que nous détaillerons dans la section 4.2.2, ne modifie pas la structure des algorithmes de Newton ou, en ce qui nous concerne, de Newton tronqué. C'est la raison pour laquelle nous n'insistons pas, dans l'immédiat, sur l'utilisation de cette correction du second ordre, malgré son importance du point de vue pratique.

3.3.2. Convergence globale de l'algorithme

Formulons notre algorithme pour résoudre le problème avec contraintes d'égalité (*3.2*).

Initialisation.
Poser $k = 0$. Choisir x_0. Passer à l'itération 0.

Itération k, $k \geq 0$.
a. Calculer le multiplicateur de moindres carrés $\lambda_k = -A_k^{-\top} \nabla f_k$.
b. Si un test de convergence portant sur la norme des conditions d'optimalité (*3.3*) est satisfait, terminer.
c. Calculer
- la direction de Newton tronquée Δx_k donnée par (*3.10*),
- le multiplicateur $\lambda_{\text{NT}\,k}$ donné par (*3.12*).
d. Mettre à jour σ_k en vérifiant les hypothèses (*3.14*).
e. Déterminer α_k par rebroussement à partir de 1, jusqu'à satisfaire (*3.16*).
f. Mettre à jour $x_{k+1} = x_k + \alpha_k\,\Delta x_k$ puis passer à l'itération $k+1$.

Nous restons dans le cadre fixé par les conditions (*3.14*). Le multiplicateur courant, qui n'intervient que dans le hessien du lagrangien, est mis à jour en début d'itération grâce à

$$\lambda_{\text{MC}} = -A^{-\top} \nabla f,$$

et non plus, comme dans les algorithmes locaux que nous avons mentionnés jusqu'à présent, grâce à $\lambda_{\text{PQ}} = -A^{-\top}(\nabla f + L\,\Delta x)$, en fin d'itération. Il ne dépend que de x ; l'algorithme est

purement primal. La désignation λ_{MC} vient du fait que, si nous avions considéré comme inverse à droite $A^- = A^\top(AA^\top)^{-1}$, ce multiplicateur serait la solution du problème de moindres carrés

$$\text{minimiser} \quad \tfrac{1}{2} \left| A^\top\lambda + \nabla f \right|^2 .$$

Il s'agit de la troisième estimation du multiplicateur optimal λ_* que nous rencontrons, après λ_{PQ} et λ_{NT}. Nous allons prouver la convergence de l'algorithme en supposant que les suites issues du simulateur sont bornées, et en particulier que (A_k^-) et (∇f_k) sont bornées. Nous sommes alors certains que $(\lambda_{MC\,k})$ est bornée. Rien ne permet de l'affirmer concernant $(\lambda_{PQ\,k})$, puisque λ_{PQ} dépend du hessien du lagrangien, qui lui-même dépend du multiplicateur !

Est également bornée, par construction, la suite (J_k). Effectivement, à l'iteration k, si des directions internes à courbure positive, vérifiant

$$\forall\, i = 0, ..., p_k, \qquad {\delta_i}^\top H_k\, \delta_i \geq \varepsilon \left| \delta_i \right|^2 ,$$

ont été générées, la plus grande valeur propre de J_k est telle que

$$\lambda_{max}(J_k) \leq \sum_{i=0,...,p_k} \frac{\lambda_{max}\left(\delta_i\, {\delta_i}^\top\right)}{\varepsilon \left| \delta_i \right|^2} \leq \frac{n-m}{\varepsilon}. \tag{3.17}$$

D'un autre côté, $J = I$ si la première direction interne est à courbure négative ; nous avons donc $\lambda_{max}(J_k) \leq \max(1, (n-m)/\varepsilon)$. Il en découle que les suites (J_k) et $(J_k^{1/2})$ sont bornées.

Proposition 3.2. *Supposons que f et c soient deux fois continûment différentiables et que les suites (∇f_k), (A_k^-), (Z_k^-), $(\nabla^2 f_k)$ et $(\nabla^2 c_k)$ soient bornées. Alors la suite des paramètres de pénalisation (σ_k) est constante à partir d'un certain rang, de valeur σ. Si, de plus, la suite $(\phi_\sigma(x_k))$ est bornée inférieurement, les suites des contraintes (c_k) et des gradients réduits (g_k) convergent vers 0.*

Démonstration. On vérifie facilement, compte-tenu de nos hypothèses, que $(\lambda_{MC\,k})$, (L_k), $(A_{NT\,k}^-)$ et $(\lambda_{NT\,k})$ sont bornées. Il résulte du point *(3.14.b)* que, pour k suffisamment grand, $\sigma_k \neq \sigma_{k-1}$ entraîne que $\sigma_{k-1} < \| \lambda_{NT\,k} \|_d + \overline{\sigma}$. Si (σ_k) n'était pas bornée, étant croissante, elle tendrait vers $+\infty$. La précédente inégalité impliquerait que $(\lambda_{NT\,k})$ n'est pas non plus bornée, ce que nous savons être faux. Donc (σ_k) est bornée et d'après le point *(3.14.c)*, il existe k_1 et σ tels que $\sigma_k = \sigma$ pour tout $k \geq k_1$.

Puisque le critère d'Armijo *(3.16)* est satisfait pour chaque k, le fait que $(\phi_\sigma(x_k))$ soit minorée se traduit par

$$\lim\, \alpha_k\, {\phi_\sigma}'(x_k, \Delta x_k) = 0.$$

Il découle de *(3.13)* que, pour $k \geq k_1$,

$$ {\phi_\sigma}'(x_k, \Delta x_k) \leq -{g_k}^\top J_k\, g_k - \overline{\sigma} \| c_k \| . \tag{3.18}$$

Par suite, $\lim\, \alpha_k \| c_k \| = 0$ et $\lim\, \alpha_k\, {g_k}^\top J_k\, g_k = 0$. Montrons que (α_k) admet un minorant strictement positif.

Si $k \geq k_1$ et $\alpha_k < 1$, le pas α_k a été choisi après un essai infructueux dans la recherche linéaire. Il existe donc $\overline{\alpha}_k$ tel que $\alpha_k \in [\,\beta\,\overline{\alpha}_k, (1-\beta)\,\overline{\alpha}_k\,]$ et

$$\phi_\sigma(x_k + \overline{\alpha}_k\,\Delta x_k) > \phi_\sigma(x_k) + \omega\,\overline{\alpha}_k\,\phi_\sigma{}'(x_k, \Delta x_k).$$

Nous avons supposé que f et c sont régulières et que $(\nabla^2 f_k)$ et $(\nabla^2 c_k)$ sont bornées. Nous avons donc

$$f(x_k + \overline{\alpha}_k\,\Delta x_k) = f(x_k) + \overline{\alpha}_k\,\nabla f_k^{\top}\Delta x_k + O\left(\overline{\alpha}_k{}^2\,|\,\Delta x_k\,|^2\right)$$

et

$$c(x_k + \overline{\alpha}_k\,\Delta x_k) = c(x_k) + \overline{\alpha}_k\,A_k\,\Delta x_k + O\left(\overline{\alpha}_k{}^2\,|\,\Delta x_k\,|^2\right) = (1-\overline{\alpha}_k)\,c(x_k) + O\left(\overline{\alpha}_k{}^2\,|\,\Delta x_k\,|^2\right).$$

Par conséquent, $\|\,c(x_k + \overline{\alpha}_k\,\Delta x_k)\,\| \leq (1-\overline{\alpha}_k)\,\|\,c(x_k)\,\| + C\,\overline{\alpha}_k{}^2\,|\,\Delta x_k\,|^2$, où C est une constante positive. On peut donc écrire

$$\phi_\sigma(x_k + \overline{\alpha}_k\,\Delta x_k) \leq \phi_\sigma(x_k) + \overline{\alpha}_k\,\phi_\sigma{}'(x_k, \Delta x_k) + C_1\,\overline{\alpha}_k{}^2\,|\,\Delta x_k\,|^2,$$

avec $C_1 > 0$. Il en découle que $(\omega - 1)\,\phi_\sigma{}'(x_k, \Delta x_k) < C_1\,\overline{\alpha}_k\,|\,\Delta x_k\,|^2$. Grâce à *(3.18)*,

$$\|\,c_k\,\| + \left|\,J_k^{1/2}\,g_k\,\right|^2 < C_2\,\overline{\alpha}_k\,|\,\Delta x_k\,|^2, \quad C_2 = \frac{C_1}{(1-\omega)\,\min(1,\overline{\sigma})}.$$

Il résulte de *(3.10)* et de ce que $(A_{\overline{\text{NT}}\,k})$, (Z_k^-) et $(J_k^{1/2})$ sont bornées, qu'il existe une constante $C_3 > 0$ telle que

$$|\,\Delta x_k\,| \leq C_3\,\left(\|\,c_k\,\| + \left|\,J_k^{1/2}\,g_k\,\right|\right).$$

Par conséquent,

$$\|\,c_k\,\| + \left|\,J_k^{1/2}\,g_k\,\right|^2 < C_4\,\overline{\alpha}_k\,\left(\|\,c_k\,\|^2 + \left|\,J_k^{1/2}\,g_k\,\right|^2\right), \quad C_4 = 2\,C_2\,{C_3}^2,$$

ou de manière équivalente

$$(1 - C_4\,\overline{\alpha}_k)\,\left|\,J_k^{1/2}\,g_k\,\right|^2 + \|\,c_k\,\| < C_4\,\overline{\alpha}_k\,\|\,c_k\,\|^2.$$

Supposons maintenant que 0 soit un point d'adhérence de $(\overline{\alpha}_k)$, autrement dit qu'existe une sous-suite $(\overline{\alpha}_{\varphi(k)})$ de limite 0. Nous aurions, pour k assez grand, $1 - C_4\,\overline{\alpha}_{\varphi(k)} > 0$. Par suite,

$$\|\,c_{\varphi(k)}\,\| < C_4\,\overline{\alpha}_{\varphi(k)}\,\|\,c_{\varphi(k)}\,\|^2,$$

et (puisque $c_{\varphi(k)} \neq 0$ d'après l'inégalité ci-dessus, et $\beta\,\overline{\alpha}_k \leq \alpha_k$)

$$1 < C_4\,\overline{\alpha}_{\varphi(k)}\,\|\,c_{\varphi(k)}\,\| \leq \frac{C_4}{\beta}\,\alpha_{\varphi(k)}\,\|\,c_{\varphi(k)}\,\|.$$

Ceci est impossible puisque $\lim\,\alpha_k\,c_k = 0$. On en déduit que $(\overline{\alpha}_k)$ et (α_k) sont minorées par des constantes strictement positives, et que $\lim\,c_k = 0$ et $\lim\,{g_k}^{\top} J_k\,g_k = 0$.

Il nous reste à justifier que $\lim\ g_k = 0$. Nous savons que ${g_k}^\top J_k\, g_k = |\, g_k\,|^2$ si la première direction interne $\delta_0 = -g_k + {Z_k^-}^\top L_k\, A_k^- c_k$ est à courbure négative. Dans le cas contraire,

$$ {g_k}^\top J_k\, g_k = \sum_{i=0,\ldots,p} \frac{\left({\delta_i}^\top g_k\right)^2}{{\delta_i}^\top H_k\, \delta_i} \geq \frac{\left({\delta_0}^\top g_k\right)^2}{{\delta_0}^\top H_k\, \delta_0}. \qquad (3.19) $$

Notons $\mathcal{K} \subset \mathbb{N}$ l'ensemble des indices k pour lesquels se produit cette seconde éventualité. Si $\mathcal{K}$ est fini, il est clair que $\lim\ g_k = 0$. Supposons donc que $\mathcal{K}$ est infini. Les suites (g_k) et (H_k) sont bornées (car (∇f_k), (Z_k^-) et (L_k) le sont). De plus, il existe une constante $C_5 > 0$ telle que $|\,\delta_0\,| \leq C_5(|\,g_k\,| + ||\,c_k\,||)$. On en déduit que la suite des dénominateurs $({\delta_0}^\top H_k\, \delta_0)$ est bornée et que

$$ \lim_{\substack{k\in\mathcal{K} \\ k\to+\infty}} {\delta_0}^\top g_k = 0. $$

Mais $|\,g_k\,|^2 \leq -{\delta_0}^\top g_k + C_6\,|\,g_k\,|\,||\,c_k\,||$, avec $C_6 > 0$. Or (g_k) est bornée et (c_k) tend vers 0, la sous-suite $(g_k)_{k\in\mathcal{K}}$ est donc de limite nulle. Comme $(g_k)_{k\in\mathbb{N}\setminus\mathcal{K}}$ converge également vers 0, nous avons prouvé que $\lim\ g_k = 0$. □

Rappelons que proche d'une solution vérifiant les conditions suffisantes d'optimalité du second ordre, la direction de Newton tronquée *(3.10)* coïncide avec la direction de Newton exacte si le paramètre ε de troncature des itérations de gradient conjugué est suffisamment petit et si celles-ci ont été poussées jusqu'à convergence. Si nous supposons que le pas unité est asymptotiquement accepté par la recherche linéaire, notre algorithme ne diffère alors de la méthode de Newton locale que dans le choix de multiplicateur utilisé pour évaluer le hessien du lagrangien. Comme λ est fonction de x, notre algorithme est purement primal et la convergence quadratique des itérés (x_k) est préservée. Nous démontrons ce résultat (proposition 4.6) pour les problèmes avec contraintes d'égalité et d'inégalité.

3.4. Un autre algorithme de Newton tronqué

Nous nous intéressons dans cette section à une variante de la méthode de Newton tronquée que nous venons de présenter. On résout, par itérations de gradient conjugué, non plus le système réduit $H\,\tau = -g + {Z^-}^\top L\, A^- c$, mais le système

$$ H\,\tau = -g. $$

Sa solution est approchée par $-J\,g$, où la matrice J est définie, comme précédemment, à l'aide des directions conjuguées internes. Soulignons le fait que cette matrice est différente de celle de la section 3.2.2, puisque la première direction $-g + {Z^-}^\top L\, A^- c$ est ici remplacée par $-g$. Il s'agit donc d'une seconde approximation de l'inverse de H. Nous allons étudier l'algorithme utilisant cette nouvelle matrice J dans les calculs de la direction de Newton tronquée *(3.10)* et du multiplicateur *(3.12)*.

Revenons, pour motiver cette approche, au système de Newton *(3.4)*. Sa solution Δx est la somme du pas de restauration Δx_r et du pas tangent Δx_t solutions de

$$ \begin{cases} L(x,\lambda)\,\Delta x_r + A(x)^\top \lambda_r = 0 \\ A(x)\,\Delta x_r = -c(x), \end{cases} \qquad \begin{cases} L(x,\lambda)\,\Delta x_t + A(x)^\top \lambda_t = -\nabla f(x) \\ A(x)\,\Delta x_t = 0. \end{cases} $$

Les pas Δx_r et Δx_t sont respectivement points stationnaires des problèmes

$$\begin{array}{c}\text{minimiser } \frac{1}{2}\,\Delta x^\top L(x,\lambda)\,\Delta x \\ \text{sous la contrainte } A(x)\,\Delta x = -c(x)\end{array}$$

et

$$\begin{array}{c}\text{minimiser } \frac{1}{2}\,\Delta x^\top L(x,\lambda)\,\Delta x + \nabla f(x)^\top \Delta x \\ \text{sous la contrainte } A(x)\,\Delta x = 0.\end{array}$$

Il existe donc une décomposition très naturelle de la direction de Newton dans laquelle restauration de la contrainte et progression vers l'optimalité sont nettement séparées. Le pas tangent est en particulier indépendant de la valeur de la contrainte au point courant. C'est de cette propriété que nous nous inspirons en résolvant $H\,\tau = -g$ au lieu de *(3.7)*. Dans cette nouvelle approche, la contrainte n'intervient plus dans la définition de J ni par conséquent dans celle du pas tangent $-Z^- J\,g$.

Reprenons l'algorithme de Newton tronqué défini page 79. On ne modifie que le mode de calcul de la direction de Newton tronquée, qui garde cependant l'expression *(3.10)*. On détermine $-J\,g$ par itérations de gradient conjugué et, en utilisant les directions ainsi générées, $J\,(-g + Z^{-\top} L\,A^- c)$ grâce à la méthode des directions conjuguées. En d'autres termes, on résout simultanément, de manière approchée, les deux systèmes

$$\left\{\begin{array}{l} H\,\tau = -g \\ H\,\tau = -g + Z^{-\top} L\,A^- c, \end{array}\right.$$

en appliquant cette fois-ci la méthode du gradient conjugué tronquée au système $H\,\tau = -g$. On en déduit Δx et λ_{NT}.

Proposition 3.3. *Sous les mêmes hypothèses, les conclusions des propositions 3.1 et 3.2 restent valables si l'algorithme utilise la nouvelle matrice J.*

Démonstration. Montrer que la direction de Newton tronquée *(3.10)* est encore une direction de descente de ϕ_σ ne présente pas de difficulté. Il suffit d'adapter la preuve de la proposition 3.2 pour montrer que l'algorithme converge. Les suites $(\lambda_{\text{MC}\,k})$, (L_k), (J_k), $(A\bar{}_{\text{NT}\,k})$, $(\lambda_{\text{NT}\,k})$ sont bornées. On vérifie de façon similaire que (σ_k) est constante à partir d'un certain rang, que (α_k) est minorée par une constante strictement positive, puis que $\lim\, c_k = 0$ et $\lim\, {g_k}^\top J_k\, g_k = 0$.

Or ${g_k}^\top J_k\, g_k = |\,g_k\,|^2$ si la première direction interne $\delta_0 = -g_k$ est à courbure négative, ou sinon, d'après *(3.19)*,

$${g_k}^\top J_k\, g_k \geq \frac{|\,g_k\,|^4}{{g_k}^\top H_k\, g_k}.$$

La suite des dénominateurs $({g_k}^\top H_k\, g_k)$ est bornée. Par suite, $\lim\, g_k = 0$. □

Sous certaines hypothèses, la méthode de Newton tronquée que nous venons d'introduire possède deux caractéristiques primordiales pour la résolution de problèmes industriels: sa convergence est globale (au sens où elle converge même si elle est initialisée loin d'une solution) et sa vitesse de convergence est asymptotiquement quadratique, ce qui en fait une méthode rapide. Nous comparerons les résultats donnés par ses deux variantes dans le chapitre 5.

Il est néanmoins restrictif de supposer que la suite des inverses à droite (A_k^-) est bornée. Cette hypothèse est en défaut si au cours des itérations, (x_k) converge vers un point $\widetilde{x}$ qui soit un minimum local de $|c|^2$ sans être un zéro de c. On a simultanément $A(\widetilde{x})^\top c(\widetilde{x}) = 0$ et $c(\widetilde{x}) \neq 0$, ce qui montre que $A(\widetilde{x})$ n'est pas surjective. Donc A n'a pas d'inverse à droite en $\widetilde{x}$ et $(A^-(x_k))$ explose. C'est précisément une faiblesse de notre algorithme par rapport aux méthodes à régions de confiance, pour lesquelles la surjectivité de A n'est pas nécessaire. Chercher à s'affranchir de cette restriction semble donc important. Une autre question ouverte est: comment exploiter les directions à courbure négative du hessien réduit du lagrangien? Leur compatibilité avec les directions à courbure négative de la fonction de mérite ϕ_σ reste à étudier.

Chapitre 4

Algorithmes de Newton tronqués et de points intérieurs

Nous nous intéressons désormais à la résolution de problèmes avec contraintes d'égalité et d'inégalité, non convexes. L'algorithme proposé est une extension de la méthode de Newton tronquée présentée dans le chapitre précédent, basée sur une approche par points intérieurs. Nous examinons particulièrement deux de ses aspects : sa convergence globale et sa convergence quadratique avec admission asymptotique du pas unité.

4.1. Une méthode primale-duale de points intérieurs

Récrivons le problème sous contraintes d'égalité et d'inégalité auquel ce chapitre est consacré,

$$\begin{array}{rl} \text{minimiser } f(x) & \\ \text{sous les contraintes} & c_E(x) = 0 \\ & c_I(x) \leq 0. \end{array} \tag{3.1}$$

Les fonctions $f : \mathbb{R}^n \to \mathbb{R}$, $c_E : \mathbb{R}^n \to \mathbb{R}^{m_E}$ et $c_I : \mathbb{R}^n \to \mathbb{R}^{m_I}$ sont régulières. Le lagrangien de ce problème est $\ell(x, \lambda_E, \lambda_I) = f(x) + {\lambda_E}^\top c_E(x) + {\lambda_I}^\top c_I(x)$ et son système d'optimalité du premier ordre s'écrit

$$\left\{ \begin{array}{l} \nabla_x \ell(x_*, \lambda_{E*}, \lambda_{I*}) = \nabla f(x_*) + A_E(x_*)^\top \lambda_{E*} + A_I(x_*)^\top \lambda_{I*} = 0 \\ c_E(x_*) = 0 \\ c_I(x_*) \leq 0, \quad {\lambda_{I*}}^\top c_I(x_*) = 0, \quad \lambda_{I*} \geq 0. \end{array} \right. \tag{4.1}$$

Les notations $c_I(x) \leq 0$ et $\lambda_I \geq 0$ signifient que les composantes de $c_I(x)$ sont négatives et celles de λ_I positives. Les matrices $A_E(x) \in \mathbb{R}^{m_E \times n}$ et $A_I(x) \in \mathbb{R}^{m_I \times n}$ sont les jacobiennes de c_E et c_I en x.

Une hypothèse commune à l'ensemble du chapitre est que $A_E(x)$ soit surjective. Pour cela, il est nécessaire que $m_E \leq n$. La surjectivité de $A_E(x)$ implique l'existence d'un inverse à droite $A_E^-(x)$, par exemple l'inverse à droite à commande fixée dans les problèmes de commande optimale.

4.1.1. Conditions d'optimalité

Les équations *(4.1)* ont une solution $(x_*, \lambda_{E*}, \lambda_{I*})$ si x_* est un minimum local de *(3.1)* et si les gradients des contraintes actives en x_* sont linéairement indépendants (ou sous

d'autres hypothèses de qualification des contraintes plus faibles, comme la condition de Mangasarian et Fromovitz [69]). Rappelons une condition suffisante d'optimalité du second ordre. Le cône des directions critiques en x_* est l'ensemble

$$C(x_*) = \left\{ \Delta x \in N(A_E(x_*)) : \nabla f(x_*)^\top \Delta x \leq 0,\ \nabla c_i(x_*)^\top \Delta x \leq 0 \ \text{ si } \ i \in I \text{ et } c_i(x_*) = 0 \right\}.$$

Si $(x_*, \lambda_{E*}, \lambda_{I*})$ est une solution des équations d'optimalité *(4.1)* telle que le hessien du lagrangien $\nabla^2_{xx}\, \ell(x_*, \lambda_{E*}, \lambda_{I*}) = L(x_*, \lambda_{E*}, \lambda_{I*})$ soit défini positif sur $C(x_*)$, au sens où

$$\forall\ \Delta x \in C(x_*)\ :\ \Delta x \neq 0, \qquad \Delta x^\top L(x_*, \lambda_{E*}, \lambda_{I*})\, \Delta x > 0,$$

alors x_* est un minimum local strict de *(3.1)*.

L'algorithme de référence pour déterminer une solution de *(3.1)* est SQP (se référer à nouveau à [11, 12, 41, 67]). Une itération de la méthode locale s'organise comme suit. La linéarisation des conditions d'optimalité conduit à résoudre le sous-problème quadratique

$$\begin{array}{rl} \text{minimiser } \frac{1}{2}\, \Delta x^\top L(x, \lambda_E, \lambda_I)\, \Delta x + \nabla f(x)^\top \Delta x & \\ \text{sous les contraintes} \quad c_E(x) + A_E(x)\, \Delta x = 0 & \\ \qquad\qquad\qquad\quad c_I(x) + A_I(x)\, \Delta x \leq 0. & \end{array} \tag{4.2}$$

Au voisinage d'une solution de *(4.1)* en laquelle les gradients des contraintes actives sont linéairement indépendants, et vérifiant les conditions suffisantes du second ordre citées auparavant, le problème *(4.2)* a au moins une solution primale-duale, que nous désignons par $(\Delta x, \lambda_{\text{PQ}\,E}, \lambda_{\text{PQ}\,I})$. On met ensuite à jour $x^+ = x + \Delta x$, ${\lambda_E}^+ = \lambda_{\text{PQ}\,E}$ et ${\lambda_I}^+ = \lambda_{\text{PQ}\,I}$.

Toute la difficulté de la méthode est donc concentrée dans la résolution du problème quadratique *(4.2)*. On a souvent recours à des méthodes de gestion des contraintes actives (voir [41, 52]) pour le traitement de la contrainte d'inégalité linéarisée $c_I(x) + A_I(x)\, \Delta x \leq 0$. Ces méthodes ont un aspect combinatoire qui les rend d'autant plus coûteuses que le nombre de contraintes d'inégalité est élevé, et que l'on ne dispose pas d'information sur les contraintes actives en la solution. Il est également possible d'appliquer des techniques de points intérieurs pour résoudre *(4.2)*.

4.1.2. Perturbation des conditions d'optimalité

Les algorithmes de points intérieurs ont d'abord été mis au point pour la résolution de problèmes linéaires ou quadratiques; voir l'article pionnier de Karmarkar [61], ou bien den Hertog [38], Terlaky, Vial et Roos [100] ou Wright [107]. Notre démarche vise à les généraliser au problème non linéaire non convexe *(3.1)*. Une interprétation de ces méthodes est de transformer un problème avec contraintes d'inégalité en une suite de problèmes avec contraintes d'égalité seulement. Par ce biais, nous allons étendre aux problèmes avec contraintes d'inégalité la méthode de Newton tronquée que nous avons présentée dans le chapitre précédent.

Il est pratique de récrire le système d'optimalité *(4.1)* sous la forme équivalente

$$\begin{cases} \nabla_x \ell(x_*, \lambda_{E*}, \lambda_{I*}) = \nabla f(x_*) + A_E(x_*)^\top \lambda_{E*} + A_I(x_*)^\top \lambda_{I*} = 0 \\ S_* \, \lambda_{I*} = 0 \\ c_E(x_*) = 0 \\ c_I(x_*) + s_* = 0 \\ s_* \geq 0, \;\; \lambda_{I*} \geq 0. \end{cases} \tag{4.3}$$

On appelle les composantes du vecteur $s_* \in \mathbb{R}^{m_I}$ variables d'écart ; on note $S_* = \mathrm{diag}\,(s_*)$ la matrice diagonale dont les éléments sont les composantes de s_*. De même, on notera $\Lambda_I = \mathrm{diag}\,(\lambda_I)$.

Le principe de base des méthodes de points intérieurs est de remplacer la condition de complémentarité $S_* \, \lambda_{I*} = 0$ par $S \, \lambda_I = \mu \, e$, où μ est un paramètre strictement positif et $e \in \mathbb{R}^{m_I}$ est le vecteur dont toutes les composantes valent 1. On résout, de façon répétée, le système d'optimalité perturbé qui en résulte,

$$\begin{cases} \nabla f(x) + A_E(x)^\top \lambda_E + A_I(x)^\top \lambda_I = 0 \\ S \, \lambda_I = \mu \, e \\ c_E(x) = 0 \\ c_I(x) + s = 0 \\ s, \lambda_I > 0, \end{cases} \tag{4.4}$$

pour des valeurs décroissantes de μ tendant vers 0. La recherche linéaire permettant de maintenir les composantes de s et λ_I strictement positives, les contraintes $s, \lambda_I > 0$ sont implicites dans *(4.4)*. On a donc remplacé le système *(4.3)*, difficile à résoudre du fait des contraintes d'inégalité $s, \lambda_I \geq 0$, par une suite de systèmes d'équations non linéaires.

Expliquons pourquoi *(4.4)* est plus aisé à résoudre que *(4.3)*. Les itérés successifs, ainsi que la solution du problème *(4.4)*, vérifient $s, \lambda_I > 0$. Les déplacements de Newton ou de Newton tronqué $(\Delta s, \Delta \lambda_I)$ ne risquent donc pas, au moins localement, de venir buter sur les contraintes de positivité $s, \lambda_I \geq 0$. Tel ne serait pas le cas si l'on appliquait l'algorithme à *(4.3)*, dont la solution se trouve justement sur le bord de l'orthant $s, \lambda_I \geq 0$. Il est par conséquent plus avantageux de résoudre *(4.4)* car, le déplacement le long de la direction de Newton étant potentiellement plus grand, la convergence doit être plus rapide.

On a coutume d'appeler trajectoire centrale l'ensemble des solutions de *(4.4)*, obtenu en faisant varier μ. Nous supposons dans les sections suivantes que le paramètre μ est fixé et nous résolvons le système d'optimalité perturbé *(4.4)* par une méthode de Newton tronquée. Cependant, comme notre objectif est avant tout de résoudre *(4.3)*, on conçoit qu'il n'est pas utile de déterminer une solution exacte de *(4.4)* pour chaque valeur de μ générée par l'algorithme. Une solution approchée devrait en effet devenir un itéré initial satisfaisant pour la résolution de ce même système après réduction de μ. Nous examinons, dans la section 4.3, les problèmes liés au choix d'un tel critère d'arrêt, ainsi qu'à la mise à jour de μ.

L'introduction des variables d'écart s mérite un commentaire. Au lieu de *(4.4)*, nous aurions pu considérer

$$\begin{cases} \nabla f(x) + A_E(x)^\top \lambda_E + A_I(x)^\top \lambda_I = 0 \\ \Lambda_I \, c_I(x) = -\mu \, e \\ c_E(x) = 0 \\ c_I(x) < 0, \quad \lambda_I > 0. \end{cases}$$

A chaque itération, on devrait vérifier que $c_I(x) < 0$ et $\lambda_I > 0$ au lieu de $s, \lambda_I > 0$. Mais déterminer un itéré initial x tel que $c_I(x) < 0$ est parfois difficile. Une manière de contourner cette difficulté est d'introduire, où cela est nécessaire, des variables d'écart directement dans le problème *(3.1)* – voir [47]. En ce qui concerne la formulation *(4.4)*, il s'agit tout simplement de choisir $s > 0$. Un autre avantage des variables d'écart est qu'il est plus facile de les maintenir positives le long d'une direction donnée que de maintenir les contraintes non linéaires c_I négatives. C'est d'ailleurs en ce sens que l'algorithme est intérieur : on garde $s > 0$ au cours des itérations, au lieu de $c_I(x) < 0$.

Une interprétation du système *(4.4)* est basée sur le problème barrière

$$\begin{array}{rl} \text{minimiser } f(x) - \mu \displaystyle\sum_{i \in I} \ln s_i & \\ \text{sous les contraintes} & c_E(x) = 0 \\ & c_I(x) + s = 0. \end{array} \qquad (4.5)$$

Ses conditions d'optimalité sont

$$\begin{cases} \nabla f(x) + A_E(x)^\top \lambda_E + A_I(x)^\top \lambda_I = 0 \\ \lambda_I = \mu \, S^{-1} \, e \\ c_E(x) = 0 \\ c_I(x) + s = 0 \\ s, \lambda_I > 0. \end{cases}$$

Elles s'obtiennent à partir de *(4.4)* en substituant à la condition de complémentarité perturbée $S \, \lambda_I = \mu \, e$ la condition équivalente $\lambda_I = \mu \, S^{-1} \, e$.

4.1.3. Résolution des conditions d'optimalité perturbées

La linéarisation des conditions *(4.4)* donne le système

$$\begin{pmatrix} L(x, \lambda_E, \lambda_I) & 0 & A_E(x)^\top & A_I(x)^\top \\ 0 & \Lambda_I & 0 & S \\ A_E(x) & 0 & 0 & 0 \\ A_I(x) & I & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta s \\ \Delta \lambda_E \\ \Delta \lambda_I \end{pmatrix} = - \begin{pmatrix} \nabla \ell(x, \lambda_E, \lambda_I) \\ S \, \lambda_I - \mu \, e \\ c_E(x) \\ c_I(x) + s \end{pmatrix}. \qquad (4.6)$$

En multipliant par S^{-1} sa deuxième équation, et en posant $\lambda_{\text{PQ}} = \Delta\lambda + \lambda$, on obtient le système primal-dual symétrique

$$\begin{pmatrix} L(x, \lambda_E, \lambda_I) & 0 & A_E(x)^\top & A_I(x)^\top \\ 0 & S^{-1} \Lambda_I & 0 & I \\ A_E(x) & 0 & 0 & 0 \\ A_I(x) & I & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta s \\ \lambda_{\text{PQ}\,E} \\ \lambda_{\text{PQ}\,I} \end{pmatrix} = - \begin{pmatrix} \nabla f(x) \\ -\mu \, S^{-1} \, e \\ c_E(x) \\ c_I(x) + s \end{pmatrix}. \qquad (4.7)$$

Nous aurions obtenu un système purement primal si nous avions appliqué la méthode de Newton à la résolution du système d'optimalité du problème *(4.5)*. Dans la matrice précédente, le terme primal-dual $S^{-1}\,\Lambda_I$ serait remplacé par le terme primal $\mu\,S^{-2}$.

Les méthodes primales-duales se sont révélées très efficaces en programmation linéaire. De plus, Conn, Gould, et Toint [31] et Wright [105] ont montré que, dans les itérations qui suivent une réduction du paramètre μ, des composantes du pas de Newton primal peuvent être trop grandes, et ceci indépendamment du mauvais conditionnement de la matrice lorsque μ tend vers 0. Villalobos, Tapia et Zhang [103] étudient la convergence des deux types de méthodes, et concluent que pour de petites valeurs de μ, la résolution de *(4.4)* demande moins d'itérations primales-duales que purement primales. Pour ces raisons, les recherches en programmation linéaire et non linéaire s'orientent principalement vers les méthodes primales-duales.

Définissons, pour alléger les notations,

$$z = \begin{pmatrix} x \\ s \end{pmatrix}, \quad c(z) = \begin{pmatrix} c_E(x) \\ c_I(x) + s \end{pmatrix}, \quad f_\mu(z) = f(x) - \mu \sum_{i \in I} \ln s_i,$$

$$A(z) = \begin{pmatrix} A_E(x) & 0 \\ A_I(x) & I \end{pmatrix}, \quad \lambda = \begin{pmatrix} \lambda_E \\ \lambda_I \end{pmatrix}, \quad M(z,\lambda) = \begin{pmatrix} L(x,\lambda_E,\lambda_I) & 0 \\ 0 & S^{-1}\Lambda_I \end{pmatrix}.$$

Le problème *(4.5)* consiste à minimiser $f_\mu(z)$ sous la contrainte $c(z) = 0$. La matrice $A(z)$ est la jacobienne de c en z et $M(z,\lambda)$ est une modification primale-duale du hessien du lagrangien de *(4.5)* – on remplace $S^{-1}\Lambda_I$ par $\mu\,S^{-2}$. Le système de Newton *(4.7)* se récrit

$$\begin{pmatrix} M(z,\lambda) & A(z)^\top \\ A(z) & 0 \end{pmatrix} \begin{pmatrix} \Delta z \\ \lambda_{\text{PQ}} \end{pmatrix} = - \begin{pmatrix} \nabla f_\mu(z) \\ c(z) \end{pmatrix}.$$

Il s'agit de l'analogue de *(3.4)* pour les problèmes avec contraintes d'égalité et d'inégalité. Comme dans la section 3.2, on résout une forme réduite de *(4.7)* au moyen d'itérations de gradient conjugué et on force la convergence globale de l'algorithme par recherche linéaire.

Nous supposons que la jacobienne des contraintes d'égalité $A_E(x)$ est surjective. Alors $A(z)$ est également surjective, donc admet un inverse à droite $A^-(z)$. Plusieurs calculs du pas de restauration $-A^-(z)\,c(z)$, définissant différentes solutions particulières de la contrainte linéarisée $A(z)\,\Delta z = -c(z)$, sont examinés dans la section suivante. Les bases de $N(A(z))$ ont elles une structure bien spécifique.

Lemme 4.1. *Les colonnes de $Z^-(z)$ forment une base du noyau $N(A(z))$ si et seulement s'il existe une matrice $Z_E^-(x)$, dont les colonnes forment une base de $N(A_E(x))$, telle que*

$$Z^-(z) = \begin{pmatrix} I \\ -A_I(x) \end{pmatrix} Z_E^-(x). \tag{4.8}$$

Démonstration. Il s'agit de prouver que toute base de $N(A)$ est du type *(4.8)*. Soit une matrice injective $Z^- \in \mathbb{R}^{(n+m_I)\times(n-m_E)}$ telle que $A\,Z^- = O_{(m_E+m_I)\times(n-m_E)}$. Alors il existe nécessairement $z_E \in \mathbb{R}^{n\times(n-m_E)}$ vérifiant $A_E\,z_E = O_{m_E\times(n-m_E)}$ et

$$Z^- = \begin{pmatrix} I \\ -A_I \end{pmatrix} z_E.$$

Supposons que, pour $\tau \in \mathbb{R}^{n-m_E}$, on ait $z_E\,\tau = 0$. Alors $Z^-\tau = 0$. L'injectivité de Z^- implique que $\tau = 0$. Donc z_E est également injective. Par suite, ses colonnes forment une base de $N(A_E)$. □

Un pas Δz satisfait la contrainte linéarisée si et seulement s'il existe $\tau \in \mathbb{R}^{n-m_E}$ tel que $\Delta z = -A^-(z)\,c(z) + Z^-(z)\,\tau$. On détermine τ en reportant cette expression de Δz dans le système de Newton ci-dessus, ce qui donne

$$H(z,\lambda)\,\tau = -g(z) + Z^-(z)^\top M(z,\lambda)\,A^-(z)\,c(z). \qquad (4.9)$$

Le vecteur

$$g(z) = Z^-(z)^\top \nabla f_\mu(z) = Z_E^-(x)^\top \Big(\nabla f(x) + \mu\,A_I(x)^\top S^{-1}\,e\Big)$$

est le gradient réduit de f_μ en z. Quant à la matrice

$$H(z,\lambda) = Z^-(z)^\top M(z,\lambda)\,Z^-(z) = Z_E^-(x)^\top \Big(L(x,\lambda_E,\lambda_I) + A_I(x)^\top S^{-1}\Lambda_I\,A_I(x)\Big)\,Z_E^-(x),$$

il s'agit d'une approximation primale-duale du hessien réduit du lagrangien de *(4.5)*.

Que l'on résolve, par itérations de gradient conjugué tronquées, ou bien le système *(4.9)*, ou bien, comme en section 3.4, le système $H(z,\lambda)\,\tau = -g(z)$, la direction de Newton tronquée Δz se met sous la forme

$$\Delta z = -A^-c + Z^-J\,\Big(-g + Z^{-\top}M\,A^-c\Big) = -A_{\text{NT}}^-\,c - Z^-J\,g. \qquad (4.10)$$

La matrice J est la matrice des directions conjuguées internes définie par la formule *(3.9)* et

$$A_{\text{NT}}^- = \Big(I - Z^-J\,Z^{-\top}M\Big)\,A^-.$$

En un point stationnaire du problème *(4.5)*, c'est-à-dire en une solution du système *(4.4)*, nous avons $\lambda_I = \mu\,S^{-1}\,e$, donc $S^{-1}\,\Lambda_I = \mu\,S^{-2}$. Les matrices M et H sont alors exactement le hessien du lagrangien de *(4.5)* et son hessien réduit. Nous en déduisons qu'au voisinage d'un tel point, satisfaisant de surcroît les conditions suffisantes d'optimalité du second ordre pour *(4.5)*, H est définie positive et *(4.7)* admet une unique solution. Les itérations de gradient conjugué peuvent se poursuivre jusqu'à résolution complète du système *(4.9)*, auquel cas Δz est la direction de Newton exacte. Nous en déduisons

$$\lambda_{\text{PQ}\,I} = \mu\,S^{-1}\,e - S^{-1}\,\Lambda_I\,\Delta s$$

et finalement

$$\lambda_{\text{PQ}\,E} = -A_E^{-\top}\Big(\nabla f + L\,\Delta x + A_I{}^\top\lambda_{\text{PQ}\,I}\Big),$$

où A_E^- est un inverse à droite de A_E.

4.1.4. Calcul du pas de restauration

Nous supposons connus un inverse à droite A_E^- de A_E et une matrice Z_E^- dont les colonnes forment une base du noyau $N(A_E)$. Le lemme 4.1 indique comment en déduire une base de $N(A)$, sous forme d'une matrice Z^-. On dispose de plus de latitude dans le choix d'un inverse à droite A^- de A, lequel n'a pas de structure aussi bien déterminée que Z^-.

Une manière simple de résoudre la contrainte linéarisée $A\,\Delta z = -c$ est d'utiliser l'inverse à droite de A

$$A^- = \begin{pmatrix} A_E^- & 0 \\ -A_I\,A_E^- & I \end{pmatrix}. \qquad (4.11)$$

En l'absence de contraintes d'égalité (ou lorsque celles-ci sont satisfaites), cet inverse à droite définit comme pas de restauration $\Delta z = -A^- c = (0\,, -(c_I + s))$.

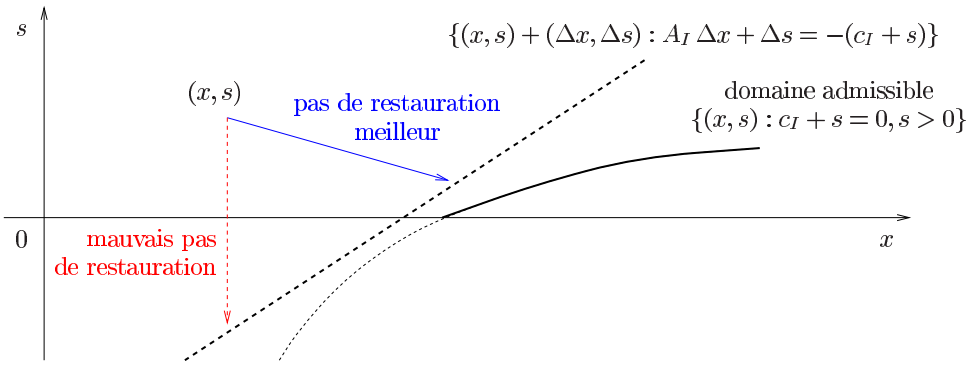


FIG. 4.1 – *prise en compte de la contrainte $s > 0$ dans le calcul du pas de restauration.*

La figure 4.1 suggère que ce pas Δz pourrait être fortement réduit pour garder $s > 0$. En outre, si les itérations de gradient conjugué sont rapidement interrompues, il ne sera que légèrement corrigé dans l'espace tangent, si bien que le pas $-A_{\text{NT}}^- c$ sera lui aussi fortement réduit. Il paraît donc important de considérer la contrainte $s > 0$ dans le calcul du pas de restauration $-A^- c$.

On pourrait penser à prendre Δz solution du problème de moindres carrés

$$\begin{aligned} &\text{minimiser } \tfrac{1}{2}\left| S^{-1}\,\Delta s \right|^2 \\ \text{sous les contraintes } \quad & c_E + A_E\,\Delta x = 0 \\ & c_I + s + A_I\,\Delta x + \Delta s = 0. \end{aligned}$$

De cette manière, on pénalise d'autant plus tout déplacement d'une composante de s vers le bord $s = 0$ du domaine admissible (mais aussi, vers l'intérieur du domaine admissible) que cette composante est déja très petite. Toutefois, nous avons constaté numériquement que cette approche peut donner un déplacement Δx trop grand.

Il nous semble préférable de minimiser simultanément $|\Delta x|$ et $|\Delta s|$. Une première idée est de résoudre le problème

$$\begin{array}{rl} \text{minimiser } \frac{1}{2}\left(\Delta x^\top P_x\,\Delta x + \Delta s^\top P_s\,\Delta s\right) & \\ \text{sous les contraintes} & c_E + A_E\,\Delta x = 0 \\ & c_I + s + A_I\,\Delta x + \Delta s = 0. \end{array} \tag{4.12}$$

Si les matrices symétriques P_x et P_s sont définies positives, ce problème est quadratique strictement convexe et admet une unique solution. Son système d'optimalité se met sous la forme

$$\begin{pmatrix} P_x & 0 & {A_E}^\top & {A_I}^\top \\ 0 & P_s & 0 & I \\ A_E & 0 & 0 & 0 \\ A_I & I & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta s \\ \lambda_{\text{RST}E} \\ \lambda_{\text{RST}I} \end{pmatrix} = - \begin{pmatrix} 0 \\ 0 \\ c_E \\ c_I + s \end{pmatrix}. \tag{4.13}$$

On retrouve la matrice du système (*4.7*) lorsque $P_x = L(x, \lambda_E, \lambda_I)$ et $P_s = S^{-1}\Lambda_I$.

Il existe un lien intéressant entre (*4.13*) et le problème de centre analytique

$$\begin{array}{rl} \text{minimiser } -\mu \displaystyle\sum_{i\in I} \ln s_i & \\ \text{sous les contraintes} & c_E(x) = 0 \\ & c_I(x) + s = 0. \end{array} \tag{4.14}$$

Soient $\lambda_{\text{CA}E}$ et $\lambda_{\text{CA}I}$ les multiplicateurs associés à ses contraintes $c_E(x) = 0$ et $c_I(x)+s = 0$, $R(x, \lambda_{\text{CA}E}, \lambda_{\text{CA}I})$ le hessien par rapport à x de son lagrangien. Le système de Newton associé à (*4.14*) est (annuler ∇f dans (*4.7*) et remplacer $S^{-1}\,\lambda_I$ par $\mu\,S^{-2}$)

$$\begin{pmatrix} R & 0 & {A_E}^\top & {A_I}^\top \\ 0 & \mu\,S^{-2} & 0 & I \\ A_E & 0 & 0 & 0 \\ A_I & I & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta s \\ {\lambda_{\text{CA}E}}^+ \\ {\lambda_{\text{CA}I}}^+ \end{pmatrix} = - \begin{pmatrix} 0 \\ -\mu\,S^{-1}\,e \\ c_E \\ c_I + s \end{pmatrix}.$$

On retrouve le système (*4.13*) en prenant $P_x = R$ et $P_s = \mu\,S^{-2}$ dans la matrice ci-dessus et en annulant le terme $\mu\,S^{-1}\,e$ dans le second membre. Autrement dit, pour ce choix de P_x et P_s, on utilise la courbure du problème de centre analytique, sans toutefois chercher à résoudre ce problème (puisqu'on annule $-\mu\,S^{-1}\,e$).

Revenons à la résolution de (*4.13*). Après élimination de $\lambda_{\text{RST}I}$ et Δs, on obtient

$$\left(P_x + {A_I}^\top P_s\,A_I\right)\Delta x + {A_E}^\top \lambda_{\text{RST}E} + {A_I}^\top P_s\,(c_I + s) = 0.$$

On définit $P = P_x + {A_I}^\top P_s\,A_I$. Puisque Δx satisfait la contrainte d'égalité linéarisée $c_E + A_E\,\Delta x = 0$, il existe τ tel que $\Delta x = -A_E^-\,c_E + Z_E^-\,\tau$. Il vient

$${Z_E^-}^\top P\,Z_E^-\,\tau = {Z_E^-}^\top \left(P\,A_E^-\,c_E - {A_I}^\top P_s\,(c_I + s)\right). \tag{4.15}$$

Les deux matrices P_x que nous avons mentionnées correspondent à des choix naturels mais ne sont pas définies positives. L'existence et l'unicité d'une solution du problème *(4.12)* ne sont alors pas garanties et rien n'assure que la matrice $Z_E^{-\top} P\, Z_E^-$ soit inversible, ni a fortiori définie positive. Mais notre objectif premier n'étant pas la résolution exacte de *(4.12)* ou de *(4.15)*, on peut se contenter d'en déterminer des solutions approchées. On considère donc $\tau = K\, Z_E^{-\top} \left(P\, A_E^- \, c_E - A_I{}^\top P_s \, (c_I + s) \right)$; la matrice K est générée lors de la résolution (par itérations de gradient conjugué tronquées, factorisation de Cholesky modifiée, ...) du système *(4.15)*. On en déduit Δx, puis Δs et enfin

$$A^- = \begin{pmatrix} \left(I - Z_E^-\, K\, Z_E^{-\top} P \right) A_E^- & Z_E^-\, K\, Z_E^{-\top} A_I{}^\top P_s \\ -A_I \left(I - Z_E^-\, K\, Z_E^{-\top} P \right) A_E^- & I - A_I\, Z_E^-\, K\, Z_E^{-\top} A_I{}^\top P_s \end{pmatrix}. \qquad (4.16)$$

Puisque la résolution du problème *(4.12)* ou de ses conditions d'optimalité *(4.13)* est basée sur l'utilisation de la formule $\Delta x = -A_E^-\, c_E + Z_E^-\, \tau$, une deuxième idée est de résoudre le problème de moindres carrés en $(\tau, \Delta s)$

$$\begin{gathered} \text{minimiser } \tfrac{1}{2} \left(\tau^\top P_\tau\, \tau + \Delta s^\top P_s\, \Delta s \right) \\ \text{sous la contrainte } A_I\, Z_E^-\, \tau + \Delta s = -c_I - s + A_I\, A_E^-\, c_E. \end{gathered}$$

Des conditions d'optimalité de ce problème,

$$\begin{cases} P_\tau\, \tau + Z_E^{-\top} A_I{}^\top \lambda_{\text{RST}} = 0 \\ P_s\, \Delta s + \lambda_{\text{RST}} = 0 \\ A_I\, Z_E^-\, \tau + \Delta s = -c_I - s + A_I\, A_E^-\, c_E, \end{cases}$$

nous déduisons

$$\left(P_\tau + Z_E^{-\top} A_I{}^\top P_s\, A_I\, Z_E^- \right) \tau = Z_E^{-\top} A_I{}^\top P_s \left(-c_I - s + A_I\, A_E^-\, c_E \right). \qquad (4.17)$$

On note à nouveau K la matrice construite lors de la résolution, éventuellement approchée, de ce système. Ceci conduit à $\tau = K\, Z_E^{-\top} A_I{}^\top P_s \left(-c_I - s + A_I\, A_E^-\, c_E \right)$. Il en résulte que

$$A^- = \begin{pmatrix} \left(I - Z_E^-\, K\, Z_E^{-\top} A_I{}^\top P_s\, A_I \right) A_E^- & Z_E^-\, K\, Z_E^{-\top} A_I{}^\top P_s \\ -A_I \left(I - Z_E^-\, K\, Z_E^{-\top} A_I{}^\top P_s\, A_I \right) A_E^- & I - A_I\, Z_E^-\, K\, Z_E^{-\top} A_I{}^\top P_s \end{pmatrix}. \qquad (4.18)$$

Concluons cette discussion du calcul du pas de restauration par quelques observations. Il n'est pas nécessaire de former explicitement A^-. Mais, comme nous le soulignerons dans la suite, l'enchaînement des diverses opérations de l'algorithme impose de stocker K, ou du moins les directions conjuguées internes à partir desquelles K est construite. Ceci est pénalisant si le nombre de variables de commande (égal à la dimension de K) est important.

On retrouve la matrice *(4.11)* à partir des matrices *(4.16)* ou *(4.18)* en prenant, selon l'approche considérée, $P_x = 0$, $P_\tau = 0$ ou $P_s = 0$. Le pas de restauration obtenu à partir de *(4.16)* ou *(4.18)* est la somme du pas de restauration associé à *(4.11)* et d'une correction dans l'espace tangent, censée traduite la contrainte $s > 0$. De toute évidence, le choix des matrices P conditionne l'efficacité des opérateurs de restauration ; ceci a été confirmé par nos tests numériques.

4.1.5. Globalisation par recherche linéaire

La méthode de Newton tronquée pour résoudre *(4.4)* est la suivante. On détermine un pas de restauration $-A^{-}c$ à l'aide d'une des approches que nous venons de décrire. Puis on calcule la direction de Newton tronquée $\Delta z = (\Delta x, \Delta s)$ donnée par *(4.10)*. On en déduit

$$\left\{ \begin{array}{l} {\lambda_I}^+ = \mu\, S^{-1}\, e - S^{-1}\, \Lambda_I\, \Delta s \\ \Delta\lambda_I = {\lambda_I}^+ - \lambda_I, \end{array} \right. \tag{4.19}$$

qui satisfait la condition de complémentarité perturbée linéarisée (ce qui est capital dans la suite). On met ensuite à jour $x^+ = x + \Delta x$ et $s^+ = s + \Delta s$. L'influence de λ_E étant moins importante, on peut utiliser comme nouvel itéré le multiplicateur du sous-problème quadratique $\lambda_{\mathrm{PQ}\,E}$

$$ {\lambda_E}^+ = -A_E^{-\top} \left(\nabla f + L\, \Delta x + {A_I}^\top {\lambda_I}^+ \right) $$

ou le multiplicateur de moindres carrés $\lambda_{\mathrm{MC}\,E}$

$$ {\lambda_E}^+ = -A_E \left(x^+\right)^{-\top} \left(\nabla f \left(x^+\right) + A_I \left(x^+\right)^\top {\lambda_I}^+ \right). $$

Si Δz est la direction de Newton et si ${\lambda_E}^+ = \lambda_{\mathrm{PQ}\,E}$, on a résolu le système *(4.7)* de manière exacte, donc la convergence de la méthode est quadratique. Nous montrerons que ceci reste vrai si ${\lambda_E}^+ = \lambda_{\mathrm{MC}\,E}$. Retenons que $(\Delta z, \lambda^+)$ est dans tous les cas une solution exacte des trois dernières équations de *(4.7)*, mais peut n'être qu'une solution approchée de sa première équation.

La convergence de la méthode n'étant cependant que locale, on la globalise par recherche linéaire, en adaptant les idées de la section 3.3. Nous considérons comme fonction de mérite

$$ \psi_\sigma(z, \lambda_I) = f_\mu(z) + \sigma \,||\, c(z) \,|| + \gamma \left(s^\top \lambda_I - \mu \sum_{i \in I} \ln s_i\, \lambda_i. \right) \tag{4.20} $$

La norme $||\,.\,||$ est une norme arbitraire sur $\mathbb{R}^{m_E + m_I}$, σ et γ sont des paramètres strictement positifs. La fonction ψ_σ est la somme de la fonction de pénalisation exacte $f_\mu + \sigma \,||\, c \,||$, qui n'est autre que la fonction de mérite *(3.11)* associée au problème barrière *(4.5)*, et du produit de γ par le terme primal-dual

$$ v(s, \lambda_I) = s^\top \lambda_I - \mu \sum_{i \in I} \ln s_i\, \lambda_i\,, $$

également utilisé par Vial [1, 102] et Armand, Gilbert et Jan-Jégou [2]. On note d'ores et déjà une inégalité de traitement entre les multiplicateurs λ_E et λ_I, justifiée par le rôle essentiel que λ_I joue dans le système de Newton.

L'intérêt du terme $v(s, \lambda_I)$ dans $\psi_\sigma(z, \lambda_I)$ est de permettre de contrôler les déplacements en λ_I, comme nous le montrerons dans la suite. La fonction v possède de nombreuses propriétés avantageuses. Elle est strictement convexe par rapport à s et λ_I séparément

(mais pas globalement). Son minimum, $v(s, \mu\, S^{-1}\, e) = m_I\, \mu\, (1 - \ln \mu)$, est atteint sur la trajectoire centrale. Sa valeur $v(s, \lambda_I)$ tend vers $+\infty$ dès qu'au moins une des composantes de $S\, \lambda_I$ tend vers 0 ou bien vers $+\infty$. Nous allons vérifier que v ne remet pas en question l'exactitude de ψ_σ.

Proposition 4.2. *Supposons que $z_* = (x_*, s_*)$ soit un minimum local de (4.5) et qu'il existe un multiplicateur $\lambda_* = (\lambda_{E\,*}, \lambda_{I\,*})$ tel que (z_*, λ_*) satisfasse les conditions suffisantes d'optimalité du second ordre de ce problème. Alors $(z_*, \lambda_{I\,*})$ est un minimum local strict de ψ_σ si $\sigma > \| \lambda_* \|_d$.*

Démonstration. Le résultat standard d'optimisation sous contraintes d'égalité déjà cité page 76 implique que z_* est un minimum local strict de la fonction $f_\mu + \sigma \| c \|$ si $\sigma > \| \lambda_* \|_d$. En d'autres termes, il existe une boule ouverte $\mathcal{B}(z_*)$ centrée en z_* telle que

$$\forall\, z \in \mathcal{B}(z_*) \,:\, z \neq z_*, \qquad f_\mu(z_*) + \sigma \| c(z_*) \| < f_\mu(z) + \sigma \| c(z) \|\,.$$

Par ailleurs, (z_*, λ_*) vérifiant les conditions d'optimalité du premier ordre de *(4.5)*, que nous savons être équivalentes à *(4.4)*, on a $S_*\, \lambda_{I\,*} = \mu\, e$. Par conséquent,

$$\forall\, s, \lambda_I > 0, \qquad v(s_*, \lambda_{I\,*}) \leq v(s, \lambda_I).$$

En sommant les deux inégalités précédentes, on prouve que $(z_*, \lambda_{I\,*})$ est un minimum local strict de ψ_σ. □

La proposition suivante est l'analogue de la proposition 3.1. Nous utilisons le fait que la seconde équation de *(4.7)* – la condition de complémentarité perturbée linéarisée – est résolue exactement.

Proposition 4.3. *Supposons que z ne soit pas un point stationnaire de (4.5). Posons*

$$\lambda_{\text{NT}} = -\bar{A_{\text{NT}}}^{\top} \nabla f_\mu = -A^{-\top} \left(\nabla f_\mu - M\, Z^-\, J\, g \right). \tag{4.21}$$

La direction de Newton tronquée Δz donnée par (4.10) et le déplacement $\Delta \lambda_I$ donné par (4.19) forment une direction de descente de la fonction ψ_σ en (z, λ) si $\sigma > \| \lambda_{\text{NT}} \|_d$.

Démonstration. La fonction v est différentiable en tout point intérieur, c'est-à-dire tel que $(s, \lambda_I) > 0$. Calculons sa dérivée directionnelle dans une direction $(\Delta s, \Delta \lambda_I)$ vérifiant l'équation *(4.19)*. Nous avons tout d'abord, pour $t > 0$,

$$\begin{aligned} v(s + t\,\Delta s, \lambda_I + t\,\Delta \lambda_I) &= (s + t\,\Delta s)^{\top} (\lambda_I + t\,\Delta \lambda_I) - \mu \sum_{i \in I} \ln(s_i + t\,\Delta s_i)(\lambda_i + t\,\Delta \lambda_i) \\ &= v(s, \lambda_I) + t \sum_{i \in I} \left(s_i\,\Delta \lambda_i + \lambda_i\,\Delta s_i - \mu\, \frac{s_i\,\Delta \lambda_i + \lambda_i\,\Delta s_i}{s_i\,\lambda_i} \right) + o(t). \end{aligned}$$

Par ailleurs, d'après *(4.19)*, $s_i\,\Delta \lambda_i + \lambda_i\,\Delta s_i = \mu - s_i\,\lambda_i$ pour tout $i \in I$. Il s'ensuit que la dérivée directionnelle de v est

$$v'(s, \lambda_I; \Delta s, \Delta \lambda_I) = \sum_{i \in I} (\mu - s_i\,\lambda_i) \left(1 - \frac{\mu}{s_i\,\lambda_i} \right) = -\sum_{i \in I} \frac{(\mu - s_i\,\lambda_i)^2}{s_i\,\lambda_i}.$$

La dérivée directionnelle de ψ_σ dans la direction $(\Delta z, \Delta\lambda_I)$ donnée par *(4.10)* et *(4.19)* vaut donc

$$\psi_\sigma{}'(z,\lambda_I;\Delta z,\Delta\lambda_I) = \nabla f_\mu{}^\top \Delta z - \sigma \,||\, c \,||\, + \gamma\, v'(s,\lambda_I;\Delta s,\Delta\lambda_I) \qquad (4.22)$$

$$= -g^\top J\, g - \nabla f_\mu{}^\top A^-_{\rm NT}\, c - \sigma \,||\, c \,||\, - \gamma \sum_{i\in I} \frac{(\mu - s_i\,\lambda_i)^2}{s_i\,\lambda_i}$$

$$\leq -g^\top J\, g + (||\, \lambda_{\rm NT} \,||_d - \sigma) \,||\, c \,||\, - \gamma \sum_{i\in I} \frac{(\mu - s_i\,\lambda_i)^2}{s_i\,\lambda_i}. \qquad (4.23)$$

Il en résulte que $\psi_\sigma{}'(z,\lambda_I;\Delta z,\Delta\lambda_I) \leq 0$ si $\sigma \geq ||\, \lambda_{\rm NT} \,||_d$.

Supposons maintenant que $\psi_\sigma{}'(z,\lambda_I;\Delta z,\Delta\lambda_I) = 0$ et $\sigma > ||\, \lambda_{\rm NT} \,||_d$. On déduit de *(4.23)* que $g^\top J\, g = 0$, $c_E = 0$, $c_I + s = 0$ et $S\,\lambda_I = \mu\, e$. Comme dans le chapitre 3, on démontre que, quelle que soit la matrice des directions conjuguées J (construite en résolvant $H\,\tau = -g$ ou $H\,\tau = -g + Z^{-\top} M\, A^-\, c$), on a $g = 0$. Par suite, $\nabla f + \mu\, A_I{}^\top S^{-1} e = \nabla f + A_I{}^\top \lambda_I \in N\left(Z_E^{-\top}\right)$, qui n'est autre que l'espace image de $A_E{}^\top$. Nous avons démontré l'existence de λ_E tel que $\nabla f + A_E{}^\top \lambda_E + A_I{}^\top \lambda_I = 0$. Autrement dit, $(x, s, \lambda_E, \lambda_I)$ est une solution de *(4.4)*, donc $z = (x, s)$ est un point stationnaire de *(4.5)*. Le résultat annoncé en découle. □

On détermine les différentes expressions du multiplicateur $\lambda_{\rm NT} = (\lambda_{{\rm NT}\,E}, \lambda_{{\rm NT}\,I})$ après quelques calculs sans difficulté. On pose

$$\xi = \nabla f + \mu\, A_I{}^\top S^{-1}\, e - \left(L + A_I{}^\top S^{-1}\, \Lambda_I\, A_I\right)\, Z_E^-\, J\, g.$$

Si l'on utilise les inverses à droite *(4.16)* ou *(4.18)*, on obtient respectivement

$$\lambda_{{\rm NT}\,E} = -A_E^{-\top}\left(I - P\, Z_E^-\, K\, Z_E^{-\top}\right)\xi, \qquad \lambda_{{\rm NT}\,E} = -A_E^{-\top}\left(I - A_I{}^\top P_s\, A_I\, Z_E^-\, K\, Z_E^{-\top}\right)\xi,$$

mais à chaque fois, puisque les deux matrices *(4.16)* et *(4.18)* ont les mêmes blocs à droite,

$$\lambda_{{\rm NT}\,I} = \mu\, S^{-1}\, e - S^{-1}\, \Lambda_I\, A_I\, Z_E^-\, J\, g - P_s\, A_I\, Z_E^-\, K\, Z_E^{-\top}\, \xi.$$

Dans le cas le plus simple, qui correspond à l'inverse à droite *(4.11)*,

$$\begin{cases} \lambda_{{\rm NT}\,E} = -A_E^{-\top}\xi \\ \lambda_{{\rm NT}\,I} = \mu\, S^{-1}\, e - S^{-1}\, \Lambda_I\, A_I\, Z_E^-\, J\, g. \end{cases}$$

On retrouve les multiplicateurs précédents en prenant des matrices P nulles (ou, ce qui revient au même, $K = 0$).

Lorsque l'algorithme emploie les inverses à droite *(4.16)* ou *(4.18)*, les systèmes linéaires devant être résolus sont les suivants. Tout d'abord, déterminer le pas de restauration $-A^-c$ demande de résoudre un des systèmes *(4.15)* ou *(4.17)*, que nous notons sous la forme $Q\,\tau = Q_E\, c_E + Q_I\, (c_I + s)$. La matrice K est générée au cours de cette opération. Cette étape

achevée, le calcul du pas tangent $Z^{-\top}\tau$ se fonde sur la résolution de $H\,\tau = -g + Z^{-\top} M\, A^{-} c$. Vient ensuite le calcul du multiplicateur λ_{NT}, qui nécessite de résoudre successivement $H\,\tau = -g$, ce qui donne $-J\,g$ et par suite ξ, puis $Q\,\tau = Z_E^{-\top}\xi$, dont résulte $K\,Z_E^{-\top}\xi$.

Il en ressort que les seules opérations pouvant être parallélisées sont les résolutions des deuxième et troisième systèmes ci-dessus,

$$\begin{cases} H\,\tau = -g + Z^{-\top} M\, A^{-} c \\ H\,\tau = -g. \end{cases}$$

En particulier, quand bien même $Q = H$ dans *(4.15)*, il n'est pas possible de résoudre simultanément les deux premiers systèmes, car la solution du premier figure dans le membre de droite du second. Rappelons, du reste, que les calculs des pas de restauration et pas tangent sont bien indépendants dans l'algorithme de la section 3.2. Une conséquence de la séquentialisation des résolutions de systèmes linéaires est donc de nous obliger à garder en mémoire la matrice K.

Nous adaptons maintenant l'algorithme de la page 79 à la résolution du système *(4.4)*. On doit prendre en compte la contrainte $s, \lambda_I > 0$, propre aux problèmes avec contraintes d'inégalité. Nous calculons les pas maximaux α_s et α_λ autorisant, à partir de (s, λ_I), un déplacement le long des directions Δs et $\Delta\lambda_I$ sans violer les contraintes de positivité. Ce sont

$$\alpha_s = \min_{\{i \in I\,:\,\Delta s_i < 0\}} -\frac{s_i}{\Delta s_i}, \qquad \alpha_\lambda = \min_{\{i \in I\,:\,\Delta\lambda_i < 0\}} -\frac{\lambda_i}{\Delta\lambda_i}$$

(avec la convention que ces minima valent $+\infty$ si les ensembles d'indices sont vides). On utilise la même technique de rebroussement pour déterminer α, en ne modifiant que l'initialisation de cette procédure. On prend comme pas initial

$$\alpha_{max} = \min\{1\,;\ \zeta\,\alpha_s\,;\ \zeta\,\alpha_\lambda\},$$

avec une constante $\zeta \in \,]0,1[\,$ (par exemple $\zeta = 0,99$). Il s'agit toujours de satisfaire le critère d'Armijo

$$\psi_\sigma(z + \alpha\,\Delta z, \lambda_I + \alpha\,\Delta\lambda_I) \leq \psi_\sigma(z, \lambda_I) + \omega\,\alpha\,\psi_\sigma{}'(z, \lambda_I; \Delta z, \Delta\lambda_I). \qquad (4.24)$$

Voyons comment sont gérés les différents multiplicateurs intervenant dans l'algorithme. La mise à jour du coefficient de pénalisation σ de la fonction de mérite ψ_σ est effectuée comme précédemment et s'appuie sur le multiplicateur *(4.21)*. La direction $\Delta\lambda_I$ dans laquelle on fait évoluer λ_I permet, avec la direction de Newton tronquée *(4.10)*, de résoudre exactement la condition de complémentarité perturbée linéarisée, et par conséquent est une direction de descente de ψ_σ. On dispose de plus de liberté pour mettre à jour λ_E et par analogie avec l'algorithme précédent, on utilise le multiplicateur de moindres carrés

$$\lambda_{\text{MC}\,E} = -A_E^{-\top}\left(\nabla f + A_I{}^{\top}\lambda_I\right).$$

4.1.6. Preuve de convergence

Résumons notre algorithme de Newton tronqué pour résoudre le système d'optimalité perturbé (*4.4*). Nous montrerons dans la section 4.2.2 comment y inclure une correction du second ordre.

Initialisation.
Poser $k = 0$. Choisir x_0 et $s_0, \lambda_{I\,0} > 0$. Passer à l'itération 0.

Itération k, $k \geq 0$.
a. Calculer le multiplicateur de moindres carrés $\lambda_{E\,k} = \lambda_{\text{MC}\,E\,k}$.
b. Si un test de convergence portant sur la norme des conditions d'optimalité (*4.4*) est satisfait, terminer.
c. Calculer
- le pas de restauration $-A_k^- c_k$, grâce à (*4.11*), (*4.16*) ou (*4.18*),
- la direction de Newton tronquée $\Delta z_k = (\Delta x_k, \Delta s_k)$ donnée par (*4.10*),
- la direction $\Delta\lambda_{I\,k}$ donnée par (*4.19*),
- le multiplicateur $\lambda_{\text{NT}\,k}$ donné par (*4.21*).

d. Mettre à jour σ_k en vérifiant les hypothèses (*3.14*).
e. Déterminer α_k par rebroussement à partir de $\alpha_{max\ k}$, jusqu'à satisfaire (*4.24*).
f. Mettre à jour $z_{k+1} = z_k + \alpha_k \Delta z_k$ et $\lambda_{I\,k+1} = \lambda_{I\,k} + \alpha_k \Delta\lambda_{I\,k}$ puis passer à l'itération $k+1$.

Une itération est basée sur la connaissance de x, s et λ_I. Il s'agit donc d'une extension primale-duale de l'algorithme de la section 3.3. Nous allons montrer sa convergence globale sous l'hypothèse, forte, que la suite des paramètres (σ_k) est bornée, donc stationnaire à partir d'un certain rang, d'après les conditions (*3.14*). Dans le chapitre 3, que (σ_k) soit bornée résultait de (*3.14*) et de ce que $(\lambda_{\text{NT}\,k})$ était bornée par construction. C'est ce dernier point que l'on ne peut plus garantir. Nous étudions tout d'abord le comportement de s et λ_I.

Lemme 4.4. *Supposons que (f_k) soit bornée inférieurement, que $(c_{E\,k})$ et $(c_{I\,k})$ soient bornées et que (σ_k) soit stationnaire à partir d'un certain rang, de valeur σ. Alors les composantes de (s_k) et $(\lambda_{I\,k})$ et leurs inverses sont bornées, autrement dit il existe des constantes strictement positives s_{min}, s_{max}, $\lambda_{I\,min}$, $\lambda_{I\,max}$ telles que*

$$\forall\ k, \quad \forall\ i \in I, \qquad 0 < s_{min} \leq s_{i\,k} \leq s_{max}, \quad 0 < \lambda_{I\,min} \leq \lambda_{i\,k} \leq \lambda_{I\,max}.$$

Démonstration. Le test d'Armijo étant satisfait à chaque itération, $(\psi_\sigma(z_k, \lambda_{I\,k}))$ est décroissante à partir d'un certain rang, et majorée. Comme f et v sont minorées, nous en déduisons qu'il existe une constante C_1 telle que

$$C_1 \geq -\mu \sum_{i \in I} \ln s_{i\,k} + \sigma \,\|\, c_k \,\|\,.$$

Supposons qu'existe une sous-suite $(s_{\varphi(k)})$ de limite $+\infty$. Comme $(c_{E\,k})$ et $(c_{I\,k})$ sont bornées, $\left(\left|\left|\, c_{\varphi(k)} \,\right|\right|\right)$ serait équivalente à $\left(\left|\, s_{\varphi(k)} \,\right|\right)$, suite devant laquelle

$$\left(\sum_{i \in I} \ln s_{i\,\varphi(k)} \right)$$

est négligeable. Dans l'inégalité précédente, le terme de droite tendrait vers $+\infty$, ce qui est impossible. On en déduit l'existence de s_{max}.

Tenant compte de ce que la fonction exponentielle est croissante, il découle de l'inégalité

$$C_1 \geq -\mu \sum_{i \in I} \ln s_{i\,k}$$

que le produit des composantes de s_k est minoré par une constante $C_2 = \exp(-C_1/\mu)$. Ceci implique que, pour chaque k,

$$\forall\, i \in I, \qquad s_{i\,k} \geq C_2 \prod_{j \in I,\, j \neq i} {s_{j\,k}}^{-1} \geq C_2\ {s_{max}}^{1-m_I} > 0,$$

et l'existence de s_{min}.

Supposant que f est minorée, et ayant prouvé que (s_k) et $({S_k}^{-1}\, e)$ sont bornées, nous avons justifié que $(f_\mu(z_k))$ est minorée. Puisque $(\psi_\sigma(z_k, \lambda_{I\,k}))$ est majorée, il existe donc une constante C_3 majorant $v(s_k, \lambda_{I\,k})$, quel que soit k. Nous avons déjà remarqué que $v(s, \lambda_I)$ tend vers $+\infty$ si au moins une composante de $S\,\lambda_I$ tend vers 0 ou $+\infty$. Par conséquent, il existe des constantes $C_3, C_4 > 0$ telles que $C_3 \geq S_k\,\lambda_{I\,k} \geq C_4$, d'où

$$\forall\, k, \quad \forall\, i \in I, \qquad \frac{C_3}{s_{max}} \geq \lambda_{i\,k} \geq \frac{C_4}{s_{min}} > 0.$$

On en conclut l'existence de $\lambda_{I\,min}$ et $\lambda_{I\,max}$. □

Le rôle du terme primal-dual $v(s, \lambda_I)$ apparaît clairement dans la dernière partie de cette démonstration. Il nous assure qu'au cours des itérations, chaque composante de λ_I ne devient ni trop petite, ni trop grande. Venons-en à la convergence globale de l'algorithme.

Proposition 4.5. *Supposons que f, c_E et c_I soient deux fois continûment différentiables, que (f_k) soit bornée inférieurement, que $(c_{E\,k})$, $(c_{I\,k})$, (∇f_k), $(A^-_{E\,k})$, $(Z^-_{E\,k})$, $(A_{I\,k})$, $(\nabla^2 f_k)$, $(\nabla^2 c_{E\,k})$ et $(\nabla^2 c_{I\,k})$ soient bornées, que les matrices K, P_x, P_s ou P_τ intervenant dans le calcul du pas de restauration soient également bornées, et enfin que (σ_k) soit stationnaire à partir d'un certain rang, de valeur σ. Alors $(c_{E\,k})$, $(c_{I\,k} + s_k)$ et (g_k) convergent vers 0 et $(S_k\,\lambda_{I\,k})$ vers $\mu\, e$, ce qui veut dire que tout point d'adhérence de la suite $(x_k, s_k, \lambda_{E\,k}, \lambda_{I\,k})$ vérifie les conditions d'optimalité perturbées (4.4).*

Démonstration. Compte-tenu de nos hypothèses, les expressions *(4.8)*, *(4.11)*, *(4.16)* et *(4.18)* montrent que (Z_k^-) et (A_k^-) sont bornées. La fonction de mérite ψ_σ est minorée et le critère d'Armijo est vérifié à chaque itération. Par suite,

$$\lim\ \alpha_k\, \psi_\sigma{}'(z_k, \lambda_{I\,k}; \Delta z_k, \Delta\lambda_{I\,k}) = 0.$$

L'inégalité *(4.23)* et les majorations données par le lemme 4.4 impliquent

$$\psi_\sigma{}'(z_k, \lambda_{I\,k}; \Delta z_k, \Delta\lambda_{I\,k}) \leq -\overline{\sigma}\,\|\, c_k \,\| - {g_k}^\top J_k\, g_k - \gamma\, \frac{|\, S_k\, \lambda_{I\,k} - \mu\, e\,|^2}{s_{max}\, \lambda_{I\,max}} \leq 0. \qquad (4.25)$$

Donc $\lim\ \alpha_k\ \|\, c_k \,\| = 0$, $\lim\ \alpha_k\, {g_k}^\top J_k\, g_k = 0$ et $\lim\ \alpha_k\ |\, S_k\, \lambda_{I\,k} - \mu\, e\,|^2 = 0$. Montrons que (α_k) est minorée par une constante strictement positive.

Considérons, pour k suffisamment grand, un pas $\alpha_k < 1$. Le pas unité ne satisfait pas ou bien le critère d'Armijo, ou bien la contrainte de positivité sur (s, λ_I). Dans le premier cas, il existe $\overline{\alpha}_k \leq 1$ tel que $\alpha_k \in [\,\beta\,\overline{\alpha}_k, (1-\beta)\,\overline{\alpha}_k\,]$ et

$$\psi_\sigma(z_k + \overline{\alpha}_k\, \Delta z_k, \lambda_{I\,k} + \overline{\alpha}_k\, \Delta\lambda_{I\,k}) > \psi_\sigma(z_k, \lambda_{I\,k}) + \omega\, \overline{\alpha}_k\, \psi_\sigma{}'(z_k, \lambda_{I\,k}; \Delta z_k, \Delta\lambda_{I\,k}). \qquad (4.26)$$

Dans le second cas, d'après la définition du paramètre ζ (page 97), au moins une des composantes de

$$(s_k, \lambda_{I\,k}) + \frac{\alpha_k}{\zeta}\, (\Delta s_k, \Delta\lambda_{I\,k})$$

est nulle. Ceci entraîne que

$$\lim_{\alpha \to \zeta^{-1}\, \alpha_k} \psi_\sigma(z_k + \alpha\, \Delta z_k, \lambda_{I\,k} + \alpha\, \Delta\lambda_{I\,k}) = +\infty.$$

Par conséquent, il existe $\overline{\alpha}_k \in \left]\, \alpha_k, \zeta^{-1}\, \alpha_k \,\right[$ satisfaisant *(4.26)*. Pour résumer les deux cas, il existe toujours $\overline{\alpha}_k \leq 1$ satisfaisant *(4.26)* et tel que $\alpha_k \in [\,\gamma\,\overline{\alpha}_k, \overline{\alpha}_k\,]$, avec $\gamma = \min(\beta, \zeta)$.

Compte-tenu de la régularité de f et du fait que $(\nabla^2 f_k)$ et (S_k^{-2}) sont bornées,

$$f_\mu(z_k + \overline{\alpha}_k\, \Delta z_k) = f_\mu(z_k) + \overline{\alpha}_k\, \nabla f_\mu(z_k)^\top \Delta z_k + O\left(\overline{\alpha}_k{}^2\ |\, \Delta z_k\,|^2\right).$$

La fonction c est également régulière et ses dérivées secondes $(\nabla^2 c_{E\,k})$, $(\nabla^2 c_{I\,k})$ bornées, donc il existe une constante $C_1 > 0$ telle que

$$\|\, c(z_k + \overline{\alpha}_k\, \Delta z_k) \,\| \leq (1 - \overline{\alpha}_k)\ \|\, c(z_k) \,\| + C_1\, \overline{\alpha}_k{}^2\ |\, \Delta z_k\,|^2.$$

Enfin, la fonction v est régulière, $\left(S_k^{-2}\right)$ et $\left(\Lambda_{I\,k}^{-2}\right)$ sont bornées, donc

$$\begin{aligned} v(s_k + \overline{\alpha}_k\, \Delta s_k, \lambda_{I\,k} + \overline{\alpha}_k\, \Delta\lambda_{I\,k}) = {} & v(s_k, \lambda_{I\,k}) + \overline{\alpha}_k\, v'(s_k, \lambda_{I\,k}; \Delta s_k, \Delta\lambda_{I\,k}) \\ & + O\left(\overline{\alpha}_k{}^2\ \left(|\, \Delta s_k\,|^2 + |\, \Delta\lambda_{I\,k}\,|^2\right)\right). \end{aligned}$$

On en déduit qu'il existe une constante $C_2 > 0$ telle que

$$\psi_\sigma(z_k + \overline{\alpha}_k \Delta z_k, \lambda_{I\,k} + \overline{\alpha}_k \Delta\lambda_{I\,k}) \leq \psi_\sigma(z_k, \lambda_{I\,k}) + \overline{\alpha}_k \psi_\sigma{}'(z_k, \lambda_{I\,k}; \Delta z_k, \Delta\lambda_{I\,k}) + C_2 \overline{\alpha}_k{}^2 \left(|\Delta z_k|^2 + |\Delta\lambda_{I\,k}|^2 \right).$$

Il résulte de (*4.26)* que

$$(\omega - 1)\psi_\sigma{}'(z_k, \lambda_{I\,k}; \Delta z_k, \Delta\lambda_{I\,k}) < C_2 \overline{\alpha}_k \left(|\Delta z_k|^2 + |\Delta\lambda_{I\,k}|^2 \right)$$

et de (*4.25)* que

$$g_k{}^\top J_k\, g_k + \|c_k\| + |S_k \lambda_{I\,k} - \mu e|^2 < C_3\, \overline{\alpha}_k \left(|\Delta z_k|^2 + |\Delta\lambda_{I\,k}|^2 \right).$$

Grâce au fait que (Z_k^-), (A_k^-), (J_k) et (M_k) sont bornées, $(A_{\mathrm{NT}\,k}^-)$ est bornée. L'existence d'une constante $C_4 > 0$ telle que

$$|\Delta z_k|^2 = \left| -A_k^- c_k + Z_k^- J_k v_k \right|^2 \leq C_4 \left(\|c_k\|^2 + \left| J_k^{1/2} g_k \right|^2 \right)$$

découle de (*4.10)*. Par ailleurs, d'après (*4.19)*, il existe $C_5 > 0$ telle que

$$|\Delta\lambda_{I\,k}|^2 = \left| S_k{}^{-1}(\mu e - S_k \lambda_{I\,k}) - S_k{}^{-1} \Lambda_{I\,k} \Delta s_k \right|^2 \leq C_5 \left(|S_k \lambda_{I\,k} - \mu e|^2 + |\Delta z_k|^2 \right).$$

Les deux dernières inégalités se traduisent par

$$\|c_k\| + \left| J_k^{1/2} g_k \right|^2 + |S_k \lambda_{I\,k} - \mu e|^2 < C_6\, \overline{\alpha}_k \left(\|c_k\|^2 + \left| J_k^{1/2} g_k \right|^2 + |S_k \lambda_{I\,k} - \mu e|^2 \right),$$

qui implique que $(\overline{\alpha}_k)$ et (α_k) ont des minorants strictement positifs, et que $\lim \|c_k\| = 0$, $\lim g_k{}^\top J_k\, g_k = 0$ et $\lim |S_k \lambda_{I\,k} - \mu e|^2 = 0$.

Les suites (g_k) d'une part, $(\lambda_{E\,k})$, (L_k) et (H_k) d'autre part, sont bornées. Selon que l'on calcule le pas tangent au moyen d'itérations de gradient conjugué sur le système *(4.9)* ou $H\tau = -g$, on conclut que $\lim g_k = 0$ comme dans la démonstration de la proposition 3.2 ou de la proposition 3.3. □

4.2. Comportement asymptotique de l'algorithme

Poursuivons l'étude de l'algorithme en montrant, dans un premier temps, la convergence quadratique en (x, s, λ_I) de sa version locale – c'est-à-dire en supposant que le pas unité est asymptotiquement accepté par la recherche linéaire. Mais nous savons que cette hypothèse n'est pas toujours satisfaite. En revanche, si on ajoute au pas de Newton une correction dite du second ordre, nous sommes alors certains que le pas unité est accepté ; c'est ce que nous montrerons ensuite. Comme cette correction ne remet pas en cause la convergence quadratique, nous aurons donc un algorithme tout à la fois globalement convergent et à vitesse de convergence localement quadratique.

4.2.1. Convergence quadratique locale

Nous avons déjà observé que, si le paramètre de troncature des itérations de gradient conjugué ε est assez petit, proche d'une solution primale-duale (z_*, λ_*) du problème *(4.5)* satisfaisant ses conditions suffisantes d'optimalité du second ordre, Δz est la direction de Newton exacte. Si nous n'effectuons pas de recherche linéaire, une itération de notre algorithme ne diffère donc d'une itération de l'algorithme de Newton local que dans la mise à jour de λ_E (qui pour nous est une quantité auxiliaire).

Proposition 4.6. *Supposons que f, c_E et c_I soient deux fois continûment différentiables, de dérivées secondes lipschitziennes, et que A_E^- soit également lipschitzien. Considérons la version locale de l'algorithme dans laquelle Δz_k est la direction de Newton exacte et $\alpha_k = 1$. Soit un point (z_*, λ_*) vérifiant les conditions suffisantes d'optimalité du second ordre du problème (4.5). Si $(z_0, \lambda_{I\,0})$ est assez proche de $(z_*, \lambda_{I\,*})$, la suite de terme général $(z_k, \lambda_{I\,k})$ converge vers $(z_*, \lambda_{I\,*})$ et son taux de convergence est quadratique.*

Démonstration. Notons les conditions d'optimalité perturbées *(4.4)*, que l'on souhaite résoudre, sous la forme d'un vecteur

$$F(z,\lambda) = F(x,s,\lambda_E,\lambda_I) = \begin{pmatrix} \nabla_x \ell(x,\lambda_E,\lambda_I) \\ S\,\lambda_I - \mu\,e \\ c_E(x) \\ c_I(x) + s \end{pmatrix};$$

soit ∇F la jacobienne de F. Il est commode de définir $v_{k+1} = (x_{k+1}, s_{k+1}, \lambda_{\text{PQ}\,E\,k}, \lambda_{I\,k+1})$, $w_k = (x_k, s_k, \lambda_{E\,*}, \lambda_{I\,k})$ et $v_* = (z_*, \lambda_*) = (x_*, s_*, \lambda_{E\,*}, \lambda_{I\,*})$.

L'accroissement $(\Delta x_k, \Delta s_k, \Delta\lambda_I)$ généré par l'algorithme local est solution du système de Newton *(4.6)*, dans lequel le multiplicateur λ_E intervenant dans le hessien du lagrangien $L(x, \lambda_E, \lambda_I)$ est de moindres carrés ; ce système est équivalent à

$$\nabla F(z_k,\lambda_k) \begin{pmatrix} \Delta x_k \\ \Delta s_k \\ \lambda_{\text{PQ}\,E\,k} - \lambda_{E\,*} \\ \Delta\lambda_{I\,k} \end{pmatrix} = -\begin{pmatrix} \nabla_x \ell(x_k,\lambda_{E\,*},\lambda_{I\,k}) \\ S_k\,\lambda_{I\,k} - \mu\,e \\ c_E(x_k) \\ c_I(x_k) + s_k \end{pmatrix},$$

c'est-à-dire $\nabla F(z_k,\lambda_k)\,(v_{k+1} - w_k) = -F(w_k)$ de façon condensée. Comme $\nabla F(z_k,\lambda_k)$ est inversible (et d'inverse borné) dans un voisinage de v_*, nous avons

$$\begin{aligned} v_{k+1} - v_* = w_k - v_* + v_{k+1} - w_k &= w_k - v_* - \nabla F(z_k,\lambda_k)^{-1}\,F(w_k) \\ &= \nabla F(z_k,\lambda_k)^{-1}\Big(\nabla F(z_k,\lambda_k)\,(w_k - v_*) - F(w_k)\Big). \end{aligned}$$

Nos hypothèses impliquent que F est continûment différentiable, d'où

$$F(w_k) = F(w_k) - F(v_*) = \int_0^1 \nabla F(v_* + t\,(w_k - v_*))\,(w_k - v_*)\,dt.$$

On en déduit que

$$v_{k+1} - v_* = \nabla F(z_k, \lambda_k)^{-1} \int_0^1 \Big(\nabla F(z_k, \lambda_k) - \nabla F(v_* + t\,(w_k - v_*)) \Big)\,(w_k - v_*)\,dt.$$

Par conséquent, tenant compte du fait que ∇F est d'inverse borné, nous avons l'existence d'une constante $C_1 > 0$ telle que, proche de v_*,

$$|\, v_{k+1} - v_* \,| \le C_1\, |\, w_k - v_* \,| \int_0^1 |\, \nabla F(z_k, \lambda_k) - \nabla F(v_* + t\,(w_k - v_*)) \,|\, dt. \qquad (4.27)$$

Comme l'inverse à droite A_E^- est lipschitzien, il existe une constante $C_2 > 0$ telle que

$$|\, \lambda_{E\,k} - \lambda_{E\,*} \,| \le C_2 \left| \begin{pmatrix} x_k \\ \lambda_{I\,k} \end{pmatrix} - \begin{pmatrix} x_* \\ \lambda_{I\,*} \end{pmatrix} \right|.$$

Par suite,

$$\begin{pmatrix} z_k \\ \lambda_k \end{pmatrix} - v_* = O\left(|\, w_k - v_* \,|\right).$$

Sachant que la jacobienne ∇F est lipschitzienne, l'inégalité *(4.27)* nous donne finalement

$$|\, w_{k+1} - v_* \,| \le |\, v_{k+1} - v_* \,| = O\left(|\, w_k - v_* \,|^2 \right),$$

d'où la convergence quadratique de la suite $(x_k, s_k, \lambda_{I\,k})$. □

L'admission asymptotique du pas unité est garantie en ajoutant à la direction de Newton une correction du second ordre. Nous allons voir que la convergence quadratique du nouvel algorithme, incluant cette correction, découle de la proposition précédente.

4.2.2. Correction du second ordre

Replaçons-nous pendant quelques instants dans le cadre des problèmes *(3.2)* n'ayant que des contraintes d'égalité. La fonction de mérite *(3.11)* ne permet pas toujours d'avoir admission asymptotique du pas unité. Des situations ont été identifiées dans lesquelles $\alpha = 1$ ne satisfait pas le test d'Armijo *(3.16)*, même lorsque l'itéré courant x_k est arbitrairement proche d'un minimum. Ce phénomène est connu sous le nom d'effet Maratos [70] (voir aussi l'exemple de Chamberlain, Lemaréchal, Pedersen et Powell [26]).

Pour l'éviter, Gabay [45] et Mayne et Polak [73] ont proposé la technique de correction du second ordre suivante (toujours pour des problèmes avec contraintes d'égalité seulement). Soit une suite (x_k) définie par la relation de récurrence $x_{k+1} = x_k + \Delta x_k$ et convergeant quadratiquement vers une solution x_*. On considère

$$e_k = -A^-(x_k)\, c\,(x_k + \Delta x_k),$$

où A^- est un inverse à droite de A. Remarquons que

$$c\,(x_k + \Delta x_k) = c\,(x_k + \Delta x_k) - c(x_*) = O\left(|\, x_k + \Delta x_k - x_* \,| \right) = O\left(|\, x_k - x_* \,|^2 \right)$$

si la contrainte c est lipschitzienne. Par suite, si $(A^-(x_k))$ est également bornée,

$$e_k = O\left(|\, x_k - x_* \,|^2\right),$$

d'où le nom de correction du second ordre, et

$$x_k + \Delta x_k + e_k - x_* = (x_k + \Delta x_k - x_*) + e_k = O\left(|\, x_k - x_* \,|^2\right).$$

Ceci montre que la convergence quadratique de (x_k) vers x_* est préservée si l'on remplace la formule de mise à jour $x_{k+1} = x_k + \Delta x_k$ par $x_{k+1} = x_k + \Delta x_k + e_k$.

L'utilisation de la correction du second ordre exige de modifier la procédure de recherche linéaire et notamment son critère d'arrêt *(3.16)*. Puisque nous ne pouvons pas affirmer que $\Delta x_k + e_k$ est une direction de descente de la fonction de mérite ϕ_{σ_k}, il n'y a aucun sens à effectuer une recherche linéaire dans cette direction en visant à faire décroître ϕ_{σ_k}. On cherche en fait un pas α le long de l'arc $\alpha \mapsto x_k + \alpha\,\Delta x_k + \alpha^2\, e_k$. Comme ce dernier est tangent à la direction Δx_k au point x_k, et comme Δx_k est une direction de descente de ϕ_{σ_k}, on sait que pour $\alpha > 0$ petit,

$$\phi_{\sigma_k}\left(x_k + \alpha\,\Delta x_k + \alpha^2\, e_k\right) \leq \phi_{\sigma_k}(x_k) + \omega\,\alpha\,\phi_{\sigma_k}{}'(x_k\,;\Delta x_k). \qquad (4.28)$$

Comme dans la recherche linéaire, on détermine α par rebroussement, en partant de 1 et jusqu'à satisfaire le critère *(4.28)*. Sous certaines hypothèses énoncées dans [12, 45, 73], notamment en fixant $\omega \in\,]\,0;0,5\,[$, le pas unité vérifie *(4.28)* si x_k est assez proche d'une solution de *(3.2)*.

Sous les hypothèses de la proposition 3.2, l'algorithme de Newton tronqué avec correction du second ordre et recherche curviligne est globalement convergent. Pour s'en convaincre, il suffit de reprendre une partie de la démonstration de cette proposition (page 80). Si $\alpha_k < 1$, il existe $\overline{\alpha}_k$ tel que $\alpha_k \in [\,\beta\,\overline{\alpha}_k, (1-\beta)\,\overline{\alpha}_k\,]$ et

$$\phi_\sigma(x_k + \overline{\alpha}_k\,\Delta x_k + \overline{\alpha}_k{}^2\, e_k) > \phi_\sigma(x_k) + \omega\,\overline{\alpha}_k\,\phi_\sigma{}'(x_k, \Delta x_k).$$

Puisque

$$e_k = -A^-(x_k)\left(c(x_k) + A(x_k)\,\Delta x_k + O\left(|\,\Delta x_k\,|^2\right)\right) = O\left(|\,\Delta x_k\,|^2\right),$$

les développements en $O(|\,\Delta x_k\,|^2)$ des différents termes de $\phi_\sigma(x_k + \overline{\alpha}_k\,\Delta x_k + \overline{\alpha}_k{}^2\, e_k)$ restent valables. On a donc, comme pour $\phi_\sigma(x_k + \overline{\alpha}_k\,\Delta x_k)$,

$$\phi_\sigma(x_k + \overline{\alpha}_k\,\Delta x_k + \overline{\alpha}_k{}^2\, e_k) \leq \phi_\sigma(x_k) + \overline{\alpha}_k\,\phi_\sigma{}'(x_k, \Delta x_k) + C_1\,\overline{\alpha}_k{}^2\;|\,\Delta x_k\,|^2,$$

avec $C_1 > 0$. Le reste de la preuve est inchangé. Toutefois, l'intérêt de cet algorithme est de garantir une convergence rapide au voisinage d'une solution. La correction du second ordre n'apporte aucun avantage dans les premières itérations, durant lesquelles on peut se contenter d'effectuer une recherche linéaire dans la direction Δx_k.

Revenons aux problèmes avec contraintes d'égalité et d'inégalité. Nous proposons d'utiliser comme correction du second ordre pour la résolution des problèmes barrières *(4.5)*

$$e_{z\,k} = \begin{pmatrix} e_{x\,k} \\ e_{s\,k} \end{pmatrix} = -A^{-}(z_k)\, c\,(z_k + \Delta z_k)\,. \qquad (4.29)$$

Elle ne porte que sur $z = (x, s)$. Il n'est pas utile de considérer de correction en λ_I car cette variable n'apparaît que dans v. Or la fonction v a la même courbure que le modèle quadratique du problème *(4.5)* utilisé pour calculer $(\Delta z, \Delta\lambda_I)$ – modèle quadratique dont est issu le système de Newton *(4.7)*.

Dans un cadre plus général que celui que nous avons étudié jusqu'à présent, nous allons montrer qu'ajoutée à Δz, la correction e_z force l'admissibilité du pas unité au voisinage d'une solution. Nous nous intéressons à une version quasi-newtonienne de l'algorithme, dans laquelle le hessien du lagrangien $L(x_k, \lambda_{E\,k}, \lambda_{I\,k})$ est approché par une matrice B_k – pouvant en pratique être mise à jour à l'aide des formules de BFGS, SR1, etc.

Comme précédemment, nous admettons que le paramètre de troncature des itérations de gradient conjugué ε est assez petit pour que le système réduit *(4.9)* puisse être résolu exactement au voisinage d'une solution de *(4.5)*. La matrice $H = Z^{-\top} M\, Z^{-}$ de ce système fait intervenir B au lieu de L. Dans la démonstration de la proposition suivante, M_k et M_* désignent respectivement

$$M_k = \begin{pmatrix} B_k & 0 \\ 0 & S_k^{-1}\Lambda_{I\,k} \end{pmatrix}, \qquad M_* = \begin{pmatrix} L_* & 0 \\ 0 & S_*^{-1}\Lambda_{I\,*} \end{pmatrix};$$

bien que pouvant prêter à confusion (M_* n'est pas la limite de M_k), cette notation évite d'introduire de nouvelles variables.

Proposition 4.7. *Supposons que (z_k, λ_k) converge vers un point (z_*, λ_*) satisfaisant les conditions suffisantes d'optimalité du second ordre du problème barrière (4.5). Supposons que f, c_E et c_I soient deux fois continûment différentiables. Supposons que (f_k) soit bornée inférieurement, que $(c_{E\,k})$, $(c_{I\,k})$, $(A^{-}_{E\,k})$, $(Z^{-}_{E\,k})$, $(A_{I\,k})$ et les matrices K, P_x, P_s ou P_τ intervenant dans le calcul du pas de restauration soient bornées. Supposons que (B_k) soit bornée, que B_k surestime L_* dans le plan tangent aux contraintes d'égalité, au sens où*

$$\tau_k^{\top}\, {Z^{-}_{E\,k}}^{\top}\, (B_k - L_*)\; Z^{-}_{E\,k}\, \tau_k \geq o\left(|\,\Delta x_k\,|^2\right). \qquad (4.30)$$

Supposons enfin que l'approximation $H_k = {Z^{-}_{E\,k}}^{\top}\big(B_k + {A_{I\,k}}^{\top} S_k^{-1}\,\Lambda_{I\,k}\, A_{I\,k}\big)\, Z^{-}_{E\,k}$ du hessien réduit du lagrangien de (4.5) soit définie positive pour k grand. Alors, si $\omega \in \left]\,0\,;\,0,5\,\right[$ et si (σ_k) est stationnaire à partir d'un certain rang, de valeur $\sigma \geq ||\,\lambda_{\mathrm{NT}\,k}\,||_d + \overline{\sigma}$,

$$\psi_\sigma(z_k + \Delta z_k + e_{z\,k}, \lambda_{I\,k} + \Delta\lambda_{I\,k}) \leq \psi_\sigma(z_k, \lambda_{I\,k}) + \omega\, \psi_\sigma{}'(z_k, \lambda_{I\,k}; \Delta z_k, \Delta\lambda_{I\,k})$$

pour k suffisamment grand.

On notera que l'hypothèse *(4.30)*, et le fait que l'on converge vers un point satisfaisant les conditions suffisantes du second ordre, n'entraînent pas que l'approximation du hessien réduit H_k soit définie positive (car la direction τ_k dans *(4.30)* n'est pas quelconque).

Démonstration. Notre objectif est de vérifier que la quantité

$$\chi_k = \psi_\sigma(z_k + \Delta z_k + e_{z\,k}, \lambda_{I\,k} + \Delta\lambda_{I\,k}) - \psi_\sigma(z_k, \lambda_{I\,k}) - \omega\,\psi_\sigma{}'(z_k, \lambda_{I\,k}; \Delta z_k, \Delta\lambda_{I\,k})$$

est négative pour k suffisamment grand. Comme nous supposons que (z_*, λ_*) vérifie les conditions suffisantes du second ordre du problème *(4.5)*, nous savons que la matrice du système de Newton *(4.6)* est inversible et que son second membre tend vers 0 lorsque k tend vers $+\infty$. Donc sa solution tend aussi vers 0, autrement dit (Δx_k), (Δs_k) et $(\Delta\lambda_{I\,k})$ sont de limite nulle ; ceci sera utilisé à plusieurs reprises.

La première étape a pour but de donner une estimation en $o(|\,\Delta z_k\,|^2)$ et $o(|\,\Delta\lambda_{I\,k}\,|^2)$ des fonctions intervenant dans χ_k. Un développement de Taylor de c donne (Δz_k vérifie la contrainte linéarisée $A(z_k)\,\Delta z_k = -c(z_k)$ et les dérivées secondes de c par rapport à s sont nulles)

$$c(z_k + \Delta z_k) = \frac{1}{2}\,c''(x_k)\,\Delta {x_k}^2 + o\left(|\,\Delta x_k\,|^2\right) = \frac{1}{2}\,c''(x_*)\,\Delta {x_k}^2 + o\left(|\,\Delta x_k\,|^2\right).$$

Nos hypothèses garantissant que (A_k^-) est bornée, nous en déduisons que

$$e_{z\,k} = O(||\,c(z_k + \Delta z_k)\,||) = O\left(|\,\Delta x_k\,|^2\right).$$

Nous avons par ailleurs

$$c(z_k + \Delta z_k + e_{z\,k}) = c(z_k + \Delta z_k) + A(z_k + \Delta z_k)\,e_{z\,k} + o\left(|\,\Delta z_k\,|^2\right).$$

Par définition de $e_{z\,k}$,

$$A(z_k + \Delta z_k)\,e_{z\,k} = A(z_k)\,e_{z\,k} + o\left(|\,\Delta z_k\,|^2\right) = -c(z_k + \Delta z_k) + o\left(|\,\Delta z_k\,|^2\right).$$

Par conséquent,

$$c(z_k + \Delta z_k + e_{z\,k}) = o\left(|\,\Delta z_k\,|^2\right).$$

Nous allons maintenant développer $f_\mu(z_k + \Delta z_k + e_{z\,k})$. On note tout d'abord que

$$-A^-(z_k)^\top \nabla f_\mu(z_k) = \lambda_* - A^-(z_k)^\top \left(\nabla f_\mu(z_k) + A(z_k)^\top \lambda_*\right) = \lambda_* + o(1).$$

Nous avons donc

$$\begin{aligned}
\nabla f_\mu(z_k)^\top e_{z\,k} &= \left(-A^-(z_k)^\top \nabla f_\mu(z_k)\right)^\top c\,(z_k + \Delta z_k) \\
&= \left(\lambda_* + o(1)\right)^\top \left(\frac{1}{2}\,c''(x_*)\,\Delta {x_k}^2 + o\left(|\,\Delta x_k\,|^2\right)\right) \\
&= \frac{1}{2} \sum_{i \in E \cup I} \lambda_{i*}\,\Delta {x_k}^\top \nabla^2 c_i(x_*)\,\Delta x_k + o\left(|\,\Delta x_k\,|^2\right).
\end{aligned}$$

Nous avons également

$$\begin{aligned} f_\mu(z_k + \Delta z_k + e_{z\,k}) &= f_\mu(z_k) + \nabla f_\mu(z_k)^\top (\Delta z_k + e_{z\,k}) \\ &\quad + \frac{1}{2} (\Delta z_k + e_{z\,k})^\top \nabla^2 f_\mu(z_k) (\Delta z_k + e_{z\,k}) + o\left(\left| \Delta z_k + e_{z\,k} \right|^2 \right) \\ &= f_\mu(z_k) + \nabla f_\mu(z_k)^\top (\Delta z_k + e_{z\,k}) \\ &\quad + \frac{1}{2} \Delta z_k{}^\top \nabla^2 f_\mu(z_k) \Delta z_k + o\left(\left| \Delta z_k \right|^2 \right), \end{aligned}$$

en utilisant le fait que $e_{z\,k} = O\left(\left| \Delta z_k \right|^2 \right)$. Le terme de second ordre vérifiant

$$\begin{aligned} \Delta z_k{}^\top \nabla^2 f_\mu(z_k) \Delta z_k &= \Delta z_k{}^\top \nabla^2 f_\mu(z_*) \Delta z_k + o\left(\left| \Delta z_k \right|^2 \right) \\ &= \Delta x_k{}^\top \nabla^2 f(x_*) \Delta x_k + \Delta s_k{}^\top S_*^{-1} \Lambda_{I\,*} \Delta s_k + o\left(\left| \Delta z_k \right|^2 \right) \end{aligned}$$

(car $\mu\, S_*^{-2} = S_*^{-1} \Lambda_{I\,*}$), on en déduit

$$\begin{aligned} f_\mu(z_k + \Delta z_k + e_{z\,k}) &= f_\mu(z_k) + \nabla f_\mu(z_k)^\top \Delta z_k + \frac{1}{2} \Delta x_k{}^\top L_* \Delta x_k \\ &\quad + \frac{1}{2} \Delta s_k{}^\top S_*^{-1} \Lambda_{I*} \Delta s_k + o\left(\left| \Delta z_k \right|^2 \right) \\ &= f_\mu(z_k) + \nabla f_\mu(z_k)^\top \Delta z_k + \frac{1}{2} \Delta z_k{}^\top M_* \Delta z_k + o\left(\left| \Delta z_k \right|^2 \right). \end{aligned}$$

Le dernier terme apparaissant dans $\psi_\sigma(z_k + \Delta z_k + e_{z\,k}, \lambda_{I\,k} + \Delta\lambda_{I\,k})$ est

$$\begin{aligned} v(s_k + \Delta s_k + e_{s\,k}, \lambda_{I\,k} + \Delta\lambda_{I\,k}) &= v(s_k, \lambda_{I\,k}) + v'(s_k, \lambda_{I\,k}; \Delta s_k + e_{s\,k}, \Delta\lambda_{I\,k}) \\ &\quad + \frac{1}{2} v''(s_k, \lambda_{I\,k}) (\Delta s_k + e_{s\,k}, \Delta\lambda_{I\,k})^2 \\ &\quad + o\left(\left| \Delta z_k \right|^2 \right) + o\left(\left| \Delta\lambda_{I\,k} \right|^2 \right). \end{aligned}$$

Comme $(\Delta s_k, \Delta\lambda_{I\,k})$ satisfait la condition de complémentarité perturbée linéarisée,

$$\begin{aligned} v'(s_k, \lambda_{I\,k}; \Delta s_k, \Delta\lambda_{I\,k}) &= \begin{pmatrix} \lambda_{I\,k} - \mu\, S_k^{-1}\, e \\ s_k - \mu\, \Lambda_{I\,k}^{-1}\, e \end{pmatrix}^\top \begin{pmatrix} \Delta s_k \\ \Delta\lambda_{I\,k} \end{pmatrix} \\ &= - \begin{pmatrix} S_k^{-1} \Lambda_{I\,k} \Delta s_k + \Delta\lambda_{I\,k} \\ \Delta s_k + S_k\, \Lambda_{I\,k}^{-1} \Delta\lambda_{I\,k} \end{pmatrix}^\top \begin{pmatrix} \Delta s_k \\ \Delta\lambda_{I\,k} \end{pmatrix} \\ &= -\Delta s_k{}^\top S_k^{-1} \Lambda_{I\,k} \Delta s_k - 2\, \Delta s_k{}^\top \Delta\lambda_{I\,k} - \Delta\lambda_{I\,k}{}^\top S_k\, \Lambda_{I\,k}^{-1}\, \Delta\lambda_{I\,k}. \end{aligned}$$

D'autre part,

$$
\begin{aligned}
v''(s_k,\lambda_{I\,k})\,(\Delta s_k+e_{s\,k},\Delta\lambda_{I\,k})^2 &= \begin{pmatrix}\Delta s_k+e_{s\,k}\\ \Delta\lambda_{I\,k}\end{pmatrix}^{\top}\begin{pmatrix}\mu\,S_k^{-2} & I\\ I & \mu\,\Lambda_{I\,k}^{-2}\end{pmatrix}\begin{pmatrix}\Delta s_k+e_{s\,k}\\ \Delta\lambda_{I\,k}\end{pmatrix}\\
&= \mu\,\Delta {s_k}^{\top} S_k^{-2}\Delta s_k + 2\,\Delta {s_k}^{\top}\Delta\lambda_{I\,k}\\
&\quad + \mu\,\Delta {\lambda_{I\,k}}^{\top}\Lambda_{I\,k}^{-2}\,\Delta\lambda_{I\,k} + o\left(|\,\Delta z_k\,|^2\right)\\
&= \Delta {s_k}^{\top} S_k^{-1}\Lambda_{I\,k}\,\Delta s_k + 2\,\Delta {s_k}^{\top}\Delta\lambda_{I\,k}\\
&\quad + \Delta {\lambda_{I\,k}}^{\top} S_k\,\Lambda_{I\,k}^{-1}\,\Delta\lambda_{I\,k} + o\left(|\,\Delta z_k\,|^2\right) + o\left(|\,\Delta\lambda_{I\,k}\,|^2\right)\\
&= -v'(s_k,\lambda_{I\,k};\Delta s_k,\Delta\lambda_{I\,k}) + o\left(|\,\Delta z_k\,|^2\right) + o\left(|\,\Delta\lambda_{I\,k}\,|^2\right).
\end{aligned}
$$

Enfin,

$$
v'(s_k,\lambda_{I\,k};e_{s\,k},0) = (\lambda_{I\,k}-\mu\,S_k^{-1}\,e)^{\top}e_{s\,k} = o\left(|\,\Delta z_k\,|^2\right).
$$

On obtient, en combinant l'égalité *(4.22)* et les développements de $f_\mu(z_k+\Delta z_k+e_{z\,k})$, $c(z_k+\Delta z_k+e_{z\,k})$ et $v(s_k+\Delta s_k+e_{s\,k},\lambda_{I\,k}+\Delta\lambda_{I\,k})$,

$$
\begin{aligned}
\chi_k = (1-\omega)\;{\psi_\sigma}'(z_k,\lambda_{I\,k};\Delta z_k,\Delta\lambda_{I\,k}) - \frac{\gamma}{2}\,v'(s_k,\lambda_{I\,k};\Delta s_k,\Delta\lambda_{I\,k})\\
+\frac{1}{2}\,\Delta {z_k}^{\top} M_*\,\Delta z_k + o\left(|\,\Delta z_k\,|^2\right) + o\left(|\,\Delta\lambda_{I\,k}\,|^2\right).
\end{aligned}
$$

Nous avons déjà vu (inégalités *(4.23)* et $\sigma_k \geq \|\,\lambda_{\mathrm{NT}\,k}\,\|_d + \overline{\sigma}$) que

$$
{\psi_\sigma}'(z_k,\lambda_{I\,k};\Delta z_k,\Delta\lambda_{I\,k}) \leq -{g_k}^{\top} J_k\,g_k - \overline{\sigma}\,\|\,c_k\,\| + \gamma\,v'(s_k,\lambda_{I\,k};\Delta s_k,\Delta\lambda_{I\,k}).
$$

Il en résulte que

$$
\begin{aligned}
\chi_k &\leq (1-\omega)\,\left(-{g_k}^{\top} J_k\,g_k - \overline{\sigma}\,\|\,c_k\,\|\right) + \left(\frac{1}{2}-\omega\right)\,\gamma\,v'(s_k,\lambda_{I\,k};\Delta s_k,\Delta\lambda_{I\,k})\\
&\quad + \frac{1}{2}\,\Delta {z_k}^{\top} M_*\,\Delta z_k + o\left(|\,\Delta z_k\,|^2\right) + o\left(|\,\Delta\lambda_{I\,k}\,|^2\right)\\
&= \left(\frac{1}{2}-\omega\right)\,\left(-{g_k}^{\top} J_k\,g_k + \gamma\,v'(s_k,\lambda_{I\,k};\Delta s_k,\Delta\lambda_{I\,k})\right) - (1-\omega)\,\overline{\sigma}\,\|\,c_k\,\|\\
&\quad + \frac{1}{2}\,\left(\Delta {z_k}^{\top} M_*\,\Delta z_k - {g_k}^{\top} J_k\,g_k\right) + o\left(|\,\Delta z_k\,|^2\right) + o\left(|\,\Delta\lambda_{I\,k}\,|^2\right).
\end{aligned}
$$

Passons à l'étape suivante. Grâce à la quantité $(1-\omega)\,\overline{\sigma}\,\|\,c_k\,\|$, des estimations en $o\,(\|\,c_k\,\|)$ suffisent à présent. Nous supposons que l'approximation H_k du hessien réduit du lagrangien de *(4.5)* est définie positive et que le système réduit *(4.9)* est résolu de manière exacte, pour k assez grand. Sa solution étant $J_k\,(-g_k+{Z_k^-}^{\top} M_k\,A_k^-\,c_k)$, on a donc

$$
{Z_k^-}^{\top} M_k\,Z_k^-\,J_k\,(-g_k+{Z_k^-}^{\top} M_k\,A_k^-\,c_k) = -g_k+{Z_k^-}^{\top} M_k\,A_k^-\,c_k.
$$

En multipliant cette équation à gauche par $-{g_k}^\top J_k$, et sachant que (g_k) converge vers 0 (car (z_k, λ_k) converge vers un point stationnaire), il vient

$$ {g_k}^\top J_k \, {Z_k^-}^\top M_k \, Z_k^- \, J_k \, g_k = {g_k}^\top J_k \, g_k + o(\| c_k \|) . $$

Nous reportons maintenant l'expression *(4.10)* de Δz_k dans ${\Delta z_k}^\top M_k \, \Delta z_k$, ce qui donne

$$ {\Delta z_k}^\top M_k \, \Delta z_k = {g_k}^\top J_k \, g_k + o(\| c_k \|) . $$

On en déduit

$$ \begin{aligned} {\Delta z_k}^\top M_* \, \Delta z_k - {g_k}^\top J_k \, g_k &= {\Delta z_k}^\top (M_* - M_k) \, \Delta z_k + o(\| c_k \|) \\ &= {\Delta x_k}^\top (L_* - B_k) \, \Delta x_k + o\left(| \Delta s_k |^2 \right) + o(\| c_k \|) . \end{aligned} $$

Il découle du fait que $Z_{E\,k}^- \, \tau_k = \Delta x_k + A_{E\,k}^- \, c_{E\,k}$ tend vers 0 et de l'inégalité *(4.30)* que

$$ {\Delta x_k}^\top (L_* - B_k) \, \Delta x_k = {\tau_k}^\top \, {Z_{E\,k}^-}^\top (L_* - B_k) \, Z_{E\,k}^- \, \tau_k + o(\| c_k \|) \leq o\left(| \Delta x_k |^2 \right) + o(\| c_k \|) . $$

Nous avons prouvé que

$$ {\Delta z_k}^\top M_* \, \Delta z_k - {g_k}^\top J_k \, g_k \leq o\left(| \Delta z_k |^2 \right) + o(\| c_k \|) , $$

et par suite

$$ \begin{aligned} \chi_k \leq \left(\frac{1}{2} - \omega \right) \left(-{g_k}^\top J_k \, g_k + \gamma \, v'(s_k, \lambda_{I\,k}; \Delta s_k, \Delta \lambda_{I\,k}) \right) - (1 - \omega) \, \overline{\sigma} \, \| c_k \| \\ + o\left(| \Delta z_k |^2 \right) + o\left(| \Delta \lambda_{I\,k} |^2 \right) + o(\| c_k \|) . \end{aligned} $$

Ce faisant, nous avons également justifié que

$$ -{g_k}^\top J_k \, g_k = -{\Delta z_k}^\top M_k \, \Delta z_k + o(\| c_k \|) = -{\Delta x_k}^\top B_k \, \Delta x_k - {\Delta s_k}^\top S_k^{-1} \, \Lambda_{I\,k} \, \Delta s_k + o(\| c_k \|) . $$

On rappelle que $\Delta s_k = -(c_{I\,k} + s_k) - A_{I\,k} \, \Delta x_k$, d'où

$$ {\Delta s_k}^\top S_k^{-1} \, \Lambda_{I\,k} \, \Delta s_k = {\Delta x_k}^\top {A_{I\,k}}^\top S_k^{-1} \, \Lambda_{I\,k} \, A_{I\,k} \, \Delta x_k + o(\| c_k \|) . $$

Utilisant l'inégalité *(4.30)* et le fait que ${Z_{E\,*}^-}^\top \left(L_* + {A_{I\,*}}^\top S_*^{-1} \, \Lambda_{I\,*} \, A_{I\,*} \right) Z_{E\,*}^-$ est définie positive, nous avons l'existence d'une constante $C_1 > 0$ telle que

$$ \begin{aligned} -{g_k}^\top J_k \, g_k &= -{\Delta x_k}^\top \left(B_k + {A_{I\,k}}^\top S_k^{-1} \, \Lambda_{I\,k} \, A_{I\,k} \right) \Delta x_k + o(\| c_k \|) \\ &= -{\tau_k}^\top \, {Z_{E\,k}^-}^\top \left(B_k + {A_{I\,k}}^\top S_k^{-1} \, \Lambda_{I\,k} \, A_{I\,k} \right) Z_{E\,k}^- \, \tau_k + o(\| c_k \|) \\ &\leq -{\tau_k}^\top \, {Z_{E\,*}^-}^\top \left(L_* + {A_{I\,*}}^\top S_*^{-1} \, \Lambda_{I\,*} \, A_{I\,*} \right) Z_{E\,*}^- \, \tau_k + o\left(| \tau_k |^2 \right) + o(\| c_k \|) \\ &\leq -C_1 \, | \tau_k |^2 + o\left(| \tau_k |^2 \right) + o(\| c_k \|) . \end{aligned} $$

Il nous reste à majorer de manière adéquate, par des quantités en $|\tau_k|^2$ et $|\Delta\lambda_{Ik}|^2$,

$$\begin{aligned} v'(s_k,\lambda_{Ik};\Delta s_k,\Delta\lambda_{Ik}) &= -\Delta s_k{}^\top S_k^{-1}\Lambda_{Ik}\,\Delta s_k - 2\,\Delta s_k{}^\top\Delta\lambda_{Ik} - \Delta\lambda_{Ik}{}^\top S_k\,\Lambda_{Ik}^{-1}\,\Delta\lambda_{Ik} \\ &= -\left|\left(S_*^{-1}\Lambda_{I*}\right)^{1/2}\Delta s_k\right|^2 - 2\,\Delta s_k{}^\top\Delta\lambda_{Ik} - \left|\left(S_*\,\Lambda_{I*}^{-1}\right)^{1/2}\Delta\lambda_{Ik}\right|^2 \\ &\quad + o\left(|\Delta s_k|^2\right) + o\left(|\Delta\lambda_{Ik}|^2\right). \end{aligned}$$

L'inégalité de Cauchy-Schwarz donne

$$-2\,\Delta s_k{}^\top\Delta\lambda_{Ik} \le 2\left|\left(S_*^{-1}\Lambda_{I*}\right)^{1/2}\Delta s_k\right|\left|\left(S_*\,\Lambda_{I*}^{-1}\right)^{1/2}\Delta\lambda_{Ik}\right|.$$

De l'inégalité

$$2\,|a\,b| = 2\left|\left(\sqrt{\eta}\,a\right)\left(\frac{b}{\sqrt{\eta}}\right)\right| \le \eta\,a^2 + \frac{b^2}{\eta},$$

valable pour tous réels $a, b, \eta > 0$, on déduit que

$$-2\,\Delta s_k{}^\top\Delta\lambda_{Ik} \le \eta\left|\left(S_*^{-1}\Lambda_{I*}\right)^{1/2}\Delta s_k\right|^2 + \frac{1}{\eta}\left|\left(S_*\,\Lambda_{I*}^{-1}\right)^{1/2}\Delta\lambda_{Ik}\right|^2,$$

quel que soit $\eta > 0$. Il existe donc des constantes $C_2, C_3 > 0$ telles que

$$\begin{aligned} v'(s_k,\lambda_{Ik};\Delta s_k,\Delta\lambda_{Ik}) &\le (\eta-1)\,C_2\,|\Delta s_k|^2 + \left(\frac{1}{\eta}-1\right)C_3\,|\Delta\lambda_{Ik}|^2 \\ &\quad + o\left(|\Delta s_k|^2\right) + o\left(|\Delta\lambda_{Ik}|^2\right). \end{aligned}$$

Il existe par ailleurs une constante $C_4 > 0$ telle que $|\Delta s_k|^2 \le C_4\,|\Delta x_k|^2 + o(\|c_k\|)$. Par suite, il existe une autre constante $C_5 > 0$ telle que, pour tout $\eta \ge 1$,

$$\begin{aligned} v'(s_k,\lambda_{Ik};\Delta s_k,\Delta\lambda_{Ik}) &\le (\eta-1)\,C_5\,|\tau_k|^2 + \left(\frac{1}{\eta}-1\right)C_3\,|\Delta\lambda_{Ik}|^2 \\ &\quad + o\left(|\Delta s_k|^2\right) + o\left(|\Delta\lambda_{Ik}|^2\right). \end{aligned}$$

Nous pouvons maintenant conclure cette démonstration. Nous avons établi que

$$\begin{aligned} \chi_k \le \left(\frac{1}{2}-\omega\right)\Bigg(\left(-C_1 + \gamma\,(\eta-1)\,C_5\right)|\tau_k|^2 + \gamma\left(\frac{1}{\eta}-1\right)C_3\,|\Delta\lambda_{Ik}|^2\Bigg) \\ -(1-\omega)\,\overline{\sigma}\,\|c_k\| + o\left(|\tau_k|^2\right) + o\left(|\Delta\lambda_{Ik}|^2\right) + o(\|c_k\|). \end{aligned}$$

Choisissant

$$1 < \eta < 1 + \frac{C_1}{\gamma\,C_5},$$

les coefficients de $|\tau_k|^2$ et $|\Delta\lambda_{Ik}|^2$ dans la somme ci-dessus sont strictement négatifs, ce qui implique que $\chi_k \le 0$ pour k assez grand. □

La convergence quadratique de l'algorithme de Newton tronqué avec correction du second ordre est une conséquence de l'admission asymptotique du pas unité (résultat que nous venons de démontrer), du fait que la correction du second ordre préserve une éventuelle convergence quadratique, et justement de la convergence quadratique de l'algorithme de Newton local (proposition 4.6).

Mais il n'est pas utile de calculer une telle correction du second ordre à chaque itération de l'algorithme. Indiquons comment sont combinées recherches linéaire et curviligne. Nous calculons e_z lorsque le pas unité satisfait la contrainte de positivité sur (s, λ_I) et lorsqu'il ne satisfait pas l'inégalité *(4.24)*. Nous vérifions ensuite que $s + \Delta s + e_s > 0$. Ceci nous assure que les variables d'écart restent strictement positives lorsque l'on se déplace le long de l'arc $\{s + \alpha\,\Delta s + \alpha^2\, e_s : \alpha \in [0,1]\}$, puisque nous avons déjà $s > 0$ et $s + \Delta s > 0$. On effectue finalement une recherche curviligne le long de cet arc si

$$\psi_\sigma(z + \Delta z + e, \lambda_I + \Delta\lambda_I) < \psi_\sigma(z + \Delta z, \lambda_I + \Delta\lambda_I).$$

Dans les autres cas, on se contente d'une recherche linéaire dans la direction $(\Delta z, \Delta\lambda_I)$.

Notre stratégie est donc d'utiliser la correction du second ordre le plus souvent possible à partir du moment où $\alpha_{max} = 1$; nous supposons alors que nous sommes proches de la solution et que corriger Δz apporte un réel bénéfice. Cette approche n'est pas nécessairement optimale en nombre d'appels au simulateur. Une autre possibilité est d'estimer la distance à la variété $c(z) = 0$, en comparant les pas de restauration et pas tangent. On fixe une constante $C > 0$; si $|\, A^- c\,| \leq C\,|\, Z^- \tau\,|$, on est peu éloigné de cette variété et la correction du second ordre est susceptible de favoriser l'admission du pas unité.

4.3. Contrôle du paramètre de perturbation

Pour conclure l'étude de notre méthode de résolution de *(3.1)*, nous nous intéressons aux problèmes liés à la réduction de μ dans le système d'optimalité perturbé *(4.4)*. C'est sur cet aspect des méthodes de points intérieurs non linéaires que ce concentre la recherche actuelle. Mentionnons par exemple les travaux de Nash et Sofer [79] sur le calcul d'un itéré initial pour la résolution du système d'optimalité perturbé après réduction de μ. Les articles de Forsgren, Gill et Shinnerl [43] et Wright [106] traitent de la détérioration du conditionnement de la matrice de Newton, ou du hessien réduit, lorsque μ tend vers 0 et que l'on se rapproche d'une solution de *(3.1)*. Ils montrent que l'on peut, en dépit de ce mauvais conditionnement, résoudre le système *(4.7)* avec précision.

Notre algorithme de Newton tronqué s'inscrit dans le cadre théorique suivant, pour la résolution du problème *(3.1)*. On résout, avec une précision de plus en plus grande, le système *(4.4)* correspondant à un paramètre de perturbation de plus en plus petit.

Proposition 4.8. *Supposons que f, c_E et c_I soient continûment différentiables, que (ν_ℓ) et (μ_ℓ) soient deux suites de réels strictement positifs, décroissantes et de limite 0, et que l'on puisse déterminer, pour chaque $\ell \in \mathbb{N}$, une solution $(x_\ell, s_\ell, \lambda_{E\,\ell}, \lambda_{I\,\ell})$ du système*

$$\left\{\begin{array}{l} \left| \nabla f(x_\ell) + A_E(x_\ell)^\top \lambda_{E\,\ell} + A_I(x_\ell)^\top \lambda_{I\,\ell} \right| \leq \nu_\ell \\ \left| S_\ell\, \lambda_{I\,\ell} - \mu_\ell\, e \right| \leq \nu_\ell \\ \left| c_E(x_\ell) \right| \leq \nu_\ell \\ \left| c_I(x_\ell) + s_\ell \right| \leq \nu_\ell \\ s_\ell,\ \lambda_{I\,\ell} > 0. \end{array}\right. \tag{4.31}$$

Alors tout point d'adhérence x_* de la suite (x_ℓ) est admissible pour le problème (3.1). Si de plus les gradients des contraintes actives en x_* sont linéairement indépendants, il existe des multiplicateurs λ_{E*} et λ_{I*} tels que $(x_*, \lambda_{E*}, \lambda_{I*})$ satisfasse les conditions d'optimalité (4.1) de ce problème.

Démonstration. Considérons une sous-suite $x_{\varphi(\ell)}$ de limite x_* et faisons tendre ℓ vers $+\infty$ dans les inégalités $\left|c_E(x_{\varphi(\ell)})\right| \leq \nu_{\varphi(\ell)}$ et $\left|c_I(x_{\varphi(\ell)}) + s_{\varphi(\ell)}\right| \leq \nu_{\varphi(\ell)}$. On obtient que $c_E(x_*) = 0$, que $(s_{\varphi(\ell)})$ converge, vers une limite $s_* \geq 0$, et que $c_I(x_*) \leq 0$, c'est-à-dire que x_* est un point admissible.

La deuxième inégalité de *(4.31)* montre que $S_{\varphi(\ell)}\, \lambda_{I\,\varphi(\ell)}$ converge vers 0. Si une contrainte d'inégalité, d'indice i, est inactive en x_*, on a nécessairement $s_{i*} > 0$, et par suite $(\lambda_{i\,\varphi(\ell)})$ converge vers $\lambda_{i*} = 0$. Soient K_* l'ensemble des indices $i \in E \cup I$ des contraintes actives en x_*, A_{K_*} la jacobienne de ces contraintes et λ_{K_*} le multiplicateur associé. La première inégalité de *(4.31)* donne

$$\lim\ \nabla f_{\varphi(\ell)} + {A_{K_*\,\varphi(\ell)}}^\top \lambda_{K_*\,\varphi(\ell)} = 0.$$

Sous l'hypothèse d'indépendance linéaire des gradients des contraintes actives en x_*, $A_{K_*\,\varphi(\ell)}$ est surjective, donc $A_{K_*\,\varphi(\ell)}\, {A_{K_*\,\varphi(\ell)}}^\top$ inversible, pour ℓ assez grand. On déduit de l'égalité précédente que $\lambda_{K_*\,\varphi(\ell)}$ converge vers

$$\lim - \left(A_{K_*\,\varphi(\ell)}\, {A_{K_*\,\varphi(\ell)}}^\top\right)^{-1} \nabla f_{\varphi(\ell)} = - \left(A_{K_*}(x_*)\, A_{K_*}(x_*)^\top\right)^{-1} \nabla f(x_*).$$

Par conséquent, si une contrainte, d'indice i, est active en x_*, $\lambda_{i\,\varphi(\ell)}$ a une limite λ_{i*}. Il est clair que $\lambda_{i*} \geq 0$ et $\lambda_{i*}\, c_i(x_*) = 0$, quel que soit $i \in I$. Nous avons vérifié l'existence de $(x_*, \lambda_{E*}, \lambda_{I*})$ solution du système *(4.1)*. □

Signalons que les conditions suffisantes du second ordre peuvent ne pas être satisfaites en ce point limite, même si on résout les équations *(4.31)* de manière exacte (en prenant $\nu_\ell = 0$) et si ces conditions suffisantes sont respectées en chaque solution $(x_\ell, \lambda_{E\,\ell}, \lambda_{I\,\ell})$. Gould et Toint [55] en donnent un contre-exemple. Le schéma précédent reste très général car il ne précise pas comment doivent être choisies les suites (ν_ℓ) et (μ_ℓ). Or ce choix revêt une grande importance du point de vue pratique.

Les algorithmes de suivi de trajectoire en programmation linéaire s'inscrivent dans le cadre précédent. On en distingue plusieurs grandes classes : algorithmes non réalisables (voir Kojima, Meggido et Mizuno [62], Lustig, Marsten et Shanno [68] et Zhang [109]), prédicteurs-correcteurs réalisables (Mizuno, Todd et Ye [75]) ou prédicteurs-correcteurs non réalisables (Mehrotra [74]). Durant l'exécution de ces algorithmes, on s'astreint à rester dans divers voisinages des points centraux, voisinages dont les diamètres tendent vers 0 au fur et à mesure que l'on se rapproche d'une solution du problème d'intérêt. Les itérés successifs sont ainsi naturellement guidés vers cette solution. Un des avantages de ces méthodes est de préciser à quelle distance de la trajectoire centrale il convient de se maintenir – avantage qui revient à fixer ν_ℓ.

Mais ceci est difficile à généraliser à la programmation non linéaire non convexe. Dans nos essais numériques, nous fixons ν (petit, à l'échelle du problème). Nous réduisons μ d'un facteur constant entre chaque résolution du système (*4.31*). Ces règles ne sont pas satisfaisantes et devront être affinées pour obtenir un algorithme plus efficace.

Chapitre 5

Problème du demi-tour en temps minimal

Le système considéré est toujours celui du câble fixé à un navire et tractant une sonde d'exploration sous-marine. On cherche à déterminer une trajectoire du navire et un filage du câble amenant cette sonde d'une position donnée à une autre position donnée, en un temps minimal et en respectant diverses contraintes. On résout ce problème de commande optimale par une approche directe, basée notamment sur la discrétisation de l'équation du câble décrite dans le deuxième chapitre, et utilisant les algorithmes de programmation non linéaire développés dans les troisième et quatrième chapitres.

5.1. Présentation du problème

Dans le modèle du système navire – câble – engin que nous avons proposé dans les deux premiers chapitres, la position relative du câble par rapport au navire, sa vitesse relative et sa tension apparaissent comme des variables d'état. La trajectoire du navire et la longueur du câble au cours du temps sont les commandes agissant sur ce système. On exprime numériquement l'état en fonction de la commande en intégrant en temps l'équation (*2.24*).

Cette modélisation permet de traiter plusieurs problèmes de commande optimale ayant un intérêt pratique. Citons deux problèmes de temps minimal : celui du demi-tour (on dit aussi changement de profil) et celui du suivi de trajectoire. Dans le second cas, on souhaite que l'engin remorqué décrive une trajectoire donnée. Il peut être équipé d'une caméra et devoir suivre une conduite au fond de la mer, pour localiser une éventuelle anomalie. Le problème du demi-tour en temps minimal se pose lors de missions de quadrillage d'une zone, durant lesquelles le poisson est tracté le long de couloirs d'exploration, appelés profils, parallèles les uns aux autres. On veut plus précisément amener l'engin à l'entrée du couloir qui jouxte celui qu'il vient de quitter. C'est ce dernier type de problème que nous allons étudier plus en détails, mais les méthodes mises en œuvre pourraient tout aussi bien s'appliquer au suivi de trajectoire. En mer, la durée de la manœuvre de demi-tour est de plusieurs heures. Il ne fait aucun doute que l'optimiser conduise à des gains en temps significatifs.

Nous avons défini, autour du problème du demi-tour, plusieurs cas-tests ayant en commun un grand nombre de caractéristiques. L'objectif est de minimiser la durée de la trajectoire. La dynamique du système câble – engin, et donc les équations résultant de la discrétisation de (*2.24*), doivent bien sûr être respectées. On considère toujours le même état initial du

système. On impose également toujours le même état final, plus exactement, les mêmes position et vitesse finales. Nous verrons toutefois qu'elles ne sont prescrites qu'avec une certaine précision. Les vitesse et accélération du navire, la vitesse de filage du câble sont bornées. D'autres contraintes, qui rendent compte de la présence d'un relief sous-marin et portent sur la profondeur du poisson, sont également considérées.

Le problème du demi-tour en temps minimal se formule donc comme un problème de commande optimale, caractérisé par des contraintes d'inégalité aussi bien sur l'état que sur la commande. Il nous faut maintenant choisir une méthode appropriée à sa résolution. Les algorithmes pour les problèmes d'optimisation de trajectoire sont de plusieurs types, tous présentés dans le classique Bryson et Ho [21]. L'article de Betts [9] est une synthèse comparative de ces méthodes. Nous allons brièvement les passer en revue.

5.2. Aperçu des méthodes numériques en commande optimale

Commençons par préciser le cadre général des problèmes qui nous intéressent. On étudie des systèmes dont la dynamique est donnée par un système différentiel

$$\begin{cases} \dot{y}(t) = f(y(t), u(t)), & t \in]\,0, t_f\,[\\ y(0) = y_0. \end{cases}$$

On restreint les valeurs admissibles de la commande $u(t)$ à un sous-ensemble $\mathcal{U} \subset \mathbb{R}^m$. L'état est un vecteur $y(t) \in \mathbb{R}^n$ dont la valeur initiale est donnée. Il est possible d'imposer des contraintes algébriques sur l'état et sur la commande, au cours de la trajectoire

$$\phi(y(t), u(t)) = 0, \quad \psi(y(t), u(t)) \leq 0, \qquad t \in]\,0, t_f\,[\,,$$

ou aux instants initial et final. On cherche à optimiser un critère J dans lequel pourront figurer un coût distribué et un coût final,

$$J(y, u, t_f) = \int_0^{t_f} L(y(t), u(t))\, dt + q\,(y(t_f)).$$

Le problème du demi-tour en temps minimal, à longueur de câble constante, entre dans ce formalisme, moyennant de considérer la tension comme une commande, puisque les variables d'état sont toutes des variables différentielles. Cette commande est cependant déterminée de manière unique grâce à la contrainte (algébrique) d'inextensibilité du câble. Bien évidemment, le temps final t_f est une des inconnues du problème.

5.2.1. Programmation dynamique

Une première classe de méthodes est basée sur le principe de programmation dynamique de Bellman [8]. On généralise le problème en faisant varier le temps initial t dans $[\,0, t_f\,]$ et l'état initial x dans l'ensemble des états admissibles. La dynamique du système, les différentes contraintes, le coût distribué sont inchangés – mais ils s'entendent sur $[\,t, t_f\,]$. La solution et la valeur $v(x, t)$ du problème dépendent donc des conditions initiales (x, t).

On démontre que la fonction valeur v est solution (plus exactement, solution de viscosité, voir Bardi et Capuzzo-Dolcetta [6] ou Barles [7]) d'une équation aux dérivées partielles du premier ordre, connue sous le nom d'équation de Hamilton-Jacobi-Bellman,

$$\dot{v}(x,t) + \min_{w \in \mathcal{U}(x)} \Big(L(x,w) + \nabla v(x,t)^\top f(x,w) \Big) = 0. \qquad (5.1.a)$$

Nous avons désigné par $\mathcal{U}(x) \subset \mathcal{U}$ l'ensemble des valeurs admissibles de la commande (satisfaisant les différentes contraintes algébriques). La commande optimale est celle qui, en chaque instant, minimise globalement le hamiltonien $L(x,w) + \nabla v(x,t)^\top f(x,w)$ sur l'ensemble des commandes admissibles ; c'est le principe de Pontryagin [87]. Est également imposée comme condition finale

$$v(x,t_f) = q(x). \qquad (5.1.b)$$

On déduit de l'intégration de l'équation *(5.1)* par différences finies, non seulement une approximation de la valeur $v(y_0,0)$ du problème de départ, mais également de la commande pour laquelle elle est atteinte.

Pour fixer les idées, supposons que lors de cette opération, l'intervalle de temps $[\,0,t_f\,]$ et chaque composante de l'état soient respectivement discrétisés en k_t instants intermédiaires et k_y valeurs discrètes. On va alors devoir calculer et mémoriser en $k_t\,{k_y}^n$ points la valeur de v et la commande optimale à appliquer durant le prochain intervalle de temps. Il est bien clair que cet algorithme devient inapplicable pour un grand nombre n de variables d'état, soit souvent en pratique dès que $n > 3$. Un autre de ses inconvénients est de demander la minimisation globale du hamiltonien, ce qui peut être un problème difficile. Une telle méthode n'est par conséquent pas adaptée au problème du demi-tour, dont la discrétisation en espace à longueur de câble constante (proposée en section 2.2.1) fait apparaître $12\,m$ variables d'état, les position et vitesse des nœuds du câble (m est ici le nombre de nœuds où est calculée la tension).

Néanmoins, lorsqu'elles peuvent être mises en œuvre, les méthodes de programmation dynamique ne manquent pas d'atouts. Elles nous donnent, en chaque état admissible et chaque instant, la commande optimale pour atteindre l'objectif que l'on s'est fixé. On a donc généré une loi de commande optimale *feedback*, fonction de l'état, que l'on pourrait appliquer si, pour quelque raison que ce soit, on s'écartait de la trajectoire optimale. A la différence des méthodes que nous verrons dans la suite, elles tolèrent la présence de variables d'état ou de commande à valeurs discrètes et peuvent assez aisément se dispenser des dérivées des fonctions du problème.

5.2.2. Méthodes de tir

Nous allons maintenant examiner une autre classe de méthodes, fondées sur la résolution des conditions nécessaires d'optimalité du problème. Formulons ce système d'optimalité pour un problème de temps minimal : on cherche à minimiser la durée t_f de la trajectoire d'un système partant d'un état initial $y(0) = y_0$ connu, devant atteindre un état final $y(t_f) = y_f$ imposé, et dont la dynamique est donnée par $\dot{y}(t) = f(y(t),u(t))$. L'ensemble

des valeurs admissibles de la commande est $\mathcal{U}$. Si ce problème admet une solution (y, u, t_f), celle-ci vérifie les équations différentielles

$$\dot{y}(t) = f(y(t), u(t)), \quad \dot{\lambda}(t) = -\lambda(t)^\top \nabla f(y(t), u(t)), \qquad t \in]\,0, t_f\,[\,. \tag{5.2.a}$$

La fonction λ est l'état adjoint, et la deuxième des équations ci-dessus l'équation adjointe. Le principe de Pontryagin se traduit par la contrainte algébrique

$$\lambda(t)^\top f(y(t), u(t)) = \min_{w \in \mathcal{U}} \lambda(t)^\top f(y(t), w), \qquad t \in]\,0, t_f\,[\,. \tag{5.2.b}$$

Enfin, on a comme condition initiale

$$y(0) = y_0, \tag{5.2.c}$$

et comme conditions finales

$$y(t_f) = y_f, \quad 1 + \lambda(t_f)^\top f(y(t_f), u(t_f)) = 0. \tag{5.2.d}$$

La dernière condition est dite de transversalité. Il est d'usage d'appeler problèmes aux deux bouts les systèmes différentiels ou différentiels-algébriques du type *(5.2)*.

On emploie très fréquemment des méthodes de tir pour intégrer le système *(5.2)* ; on pourra consulter Stoer et Bulirsch [97] pour une introduction à ces méthodes. Leur principe est d'estimer $\lambda(0)$ et t_f, ce qui permet de transformer *(5.2)* en un problème avec valeurs initiales et de l'intégrer grâce à un solveur d'équations différentielles-algébriques (c'est la phase de tir). On calcule ensuite l'erreur commise sur les conditions finales *(5.2.d)* et on ajuste $\lambda(0)$ et t_f en conséquence. Ces itérations se poursuivent jusqu'à satisfaire les conditions *(5.2.d)* avec suffisamment de précision. On formalise ce processus en écrivant l'erreur sur les conditions finales comme fonction de $\lambda(0)$ et t_f. Cherchant à annuler cette fonction non linéaire, il vient immédiatement à l'esprit d'appliquer la méthode de Newton. Les méthodes de tir sont locales et demandent donc, dans notre cas, de bonnes approximations initiales de $\lambda(0)$ et t_f si aucune technique de globalisation n'est utilisée.

Inclure dans le problème des contraintes d'égalité $\phi(y(t), u(t)) = 0$, $t \in [\,0, t_f\,]$, ne modifie pas foncièrement la situation. On ajoute dans le système *(5.2)* cette nouvelle contrainte algébrique et on définit un état adjoint supplémentaire. Il s'agit toujours de résoudre un problème aux deux bouts. La prise en compte de contraintes d'inégalité $\psi(y(t), u(t)) \leq 0$ est sensiblement plus délicate. Les difficultés qui se posent sont liées à l'identification des contraintes actives en la solution. On ne connait pas a priori le nombre de phases pendant lesquelles une contrainte donnée est active, ni à quels instants elle devient ou cesse d'être active. Ainsi, ce qui était un problème aux deux bouts est susceptible de se transformer en une suite de problèmes aux deux bouts. Des techniques ont été proposées pour identifier cette structure (Bulirsch, Montrone et Pesch [22], Pesch [82]).

5.2.3. Méthodes de transcription directe

On aura remarqué que les méthodes de tir sont basées sur une discrétisation des conditions d'optimalité du problème de commande optimale. Dans les méthodes directes, à l'inverse, on commence par discrétiser le problème de commande. La variable du problème discret est le vecteur x dont les composantes sont, si le temps final est inconnu, des instants

$$0 = t_1 < t_2 < ... < t_{k_t} = t_f$$

et, dans tous les cas, les valeurs de l'état et de la commande

$$y_i(t_\ell),\ u_j(t_\ell), \quad i = 1, ..., n,\ j = 1, ..., m,\ \ell = 1, ..., k_t$$

en chacun des instants intermédiaires. Puis on résout le système d'optimalité du problème discret grâce à des algorithmes de programmation non linéaire. C'est cette approche que nous utilisons dans la suite.

Un aspect important des méthodes directes est l'intégration numérique de la dynamique $\dot{y}(t) = f(y(t), u(t))$, qui se traduit par un ensemble de contraintes d'égalité $c_E(x) = 0$. Des algorithmes mêlant plusieurs schémas ont été proposés. Par exemple, Bonnans et Launay [13] appliquent la méthode de Runge-Kutta d'ordre 4 dans le calcul des contraintes et de leur jacobienne ; mais ils déterminent une approximation du hessien du lagrangien fondée sur la méthode d'Euler. Dans un autre ordre d'idées, Betts et Huffman [10] résolvent une suite de programmes non linéaires caractérisés par des schémas d'intégration d'ordres de plus en plus élevés, sur des maillages de plus en plus fins. Quant à nous, pour des raisons que nous avons exposées dans la section 2.2.3, nous appliquons un schéma BDF.

Les méthodes directes ont le net avantage de ne pas demander que les différentes phases du problème (les intervalles de temps pendant lesquels les contraintes d'inégalité demeurent actives ou inactives) aient été identifiées. En revanche, les méthodes de tir ont besoin de ces informations et peuvent difficilement être appliquées ex abrupto. Elles sont toutefois réputées pour être plus précises. C'est pourquoi on combine parfois les deux méthodes, transcription directe puis tir (voir von Stryk et Bulirsch [104]).

Compte-tenu des limitations de chacune des précédentes méthodes, l'approche directe, basée en particulier sur l'algorithme de Newton tronqué que nous avons développé dans les chapitres 3 et 4, paraît naturelle pour résoudre le problème du demi-tour en temps minimal. Nous allons tester son efficacité sur différentes variantes du problème, variantes que nous détaillons et écrivons sous la forme *(3.1)* dans la section suivante.

5.3. Définition d'un ensemble de cas-tests

Nous considérons dans tous les cas-tests le repère fixe dont l'origine est la position initiale du navire et la base est orthonormée, directe, unitaire et constituée de deux vecteurs horizontaux et d'un vecteur vertical ascendant. La position du navire à l'instant initial est par conséquent $u^0 = (0, 0, 0)$.

Comme nous l'avons déjà dit, l'objectif de chaque cas-test est de minimiser la durée de la trajectoire, c'est-à-dire $n\,dt$. Le nombre n de pas de temps étant fixé, ceci revient à minimiser le pas de temps dt, qui est une des variables du problème d'optimisation.

5.3.1. Contraintes d'égalité

Jusqu'à présent, les commandes du système étaient connues et nous n'avions pas eu à nous préoccuper de leur discrétisation. Ce sont maintenant des inconnues du problème du demi-tour en temps minimal qui, comme les variables d'état, doivent être discrétisées. Dans ce but, nous récrivons (*2.24*) sous la forme équivalente

$$\begin{cases} \dot{w} = z, \quad \dot{\kappa} = \mu, \quad \dot{\mu} = \nu, \quad \dot{u} = \beta, \quad \dot{\beta} = \gamma \\ M\,\dot{z} = D(w, z, \beta + \mu, \gamma + \nu) - R(w)^\top T \\ C(w) = 0 \\ E(w, \kappa, \ell) = 0. \end{cases} \tag{5.3}$$

Les variables de ce système différentiel-algébrique sont w, z, T, κ, μ, ν, u, β, γ et ℓ. On observe que les positions du navire u et de l'extrémité supérieure (fictive) du câble κ, qui donnent la position absolue du câble $y + u = w + \kappa + u$, jouent des rôles similaires.

On applique le schéma d'Euler implicite à toutes les variables différentielles, c'est-à-dire pour $k = 1, ..., n$,

$$\begin{pmatrix} w^k \\ \kappa^k \\ \mu^k \\ u^k \\ \beta^k \end{pmatrix} = dt \begin{pmatrix} z^k \\ \mu^k \\ \nu^k \\ \beta^k \\ \gamma^k \end{pmatrix} + \begin{pmatrix} w^{k-1} \\ \kappa^{k-1} \\ \mu^{k-1} \\ u^{k-1} \\ \beta^{k-1} \end{pmatrix} \tag{5.4.a}$$

et

$$z^k = dt\, M^{-1} \left(D(w^k, z^k, \beta^k + \mu^k, \gamma^k + \nu^k) - R(w^k)^\top T^k \right) + z^{k-1}. \tag{5.4.b}$$

On a d'autre part les contraintes

$$C(w^k) = 0, \qquad E(w^k, \kappa^k, \ell^k) = 0. \tag{5.4.c}$$

Les composantes verticales de u^k, β^k et γ^k sont implicitement nulles.

Les contraintes d'égalité, qui expriment la dynamique du système câble – engin, sont les équations (*5.4*), dans lesquelles l'état initial est celui du régime stationnaire à vitesse $\beta^0 = (1, 0, 0)$ et longueur de câble $\ell^0 = 4000$ m. Lors d'essais en mer, la vitesse du navire est effectivement d'environ 1 m.s^{-1} durant les phases d'acquisition de données, le long des profils. La vitesse relative initiale v^0 et la vitesse de filage initiale sont nulles. Nous en déduisons $z^0 = \mu^0 = (0, 0, 0)$. On obtient par le calcul w^0 et κ^0.

5.3.2. Contraintes d'inégalité

Certaines contraintes d'inégalité sont communes à tous les cas-tests. Avant d'entamer le demi-tour, le navire progresse à vitesse constante β^0 et le poisson suit un profil 1 (figure 5.1). Nous voulons rendre compte du fait qu'après cette manœuvre, le navire repart à vitesse opposée et le poisson doit suivre un profil 2. Idéalement, au temps final, ce dernier est positionné à l'extrémité du second profil et le câble est dans sa configuration de régime stationnaire à vitesse $\beta^n = (-1, 0, 0)$ et longueur $\ell^n = 4000$ m.

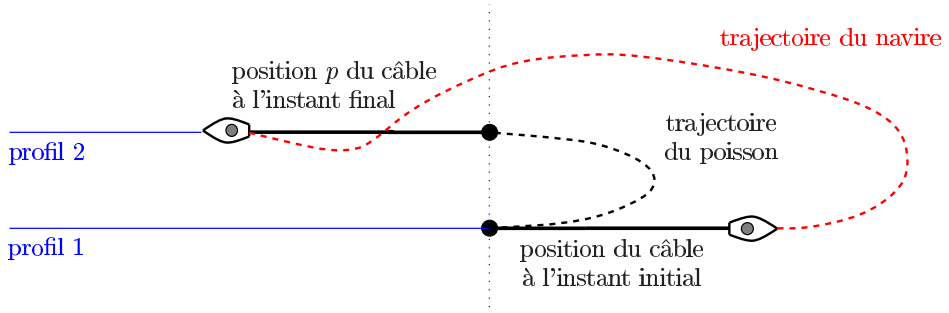


FIG. 5.1 – *changement de profil.*

Notons, pour $k = 0, ..., n$ et $i = 0, ..., 2m - 1$,

$$w^k(i) = \left(w_i^k, w_{i+2m}^k, w_{i+4m}^k\right), \quad z^k(i) = \left(z_i^k, z_{i+2m}^k, z_{i+4m}^k\right)$$

les trois composantes de w^k et z^k dans la base considérée correspondant au $i$ème nœud le long du câble. Soit également $\kappa^k = (\kappa_1^k, \kappa_2^k, \kappa_3^k)$. A l'instant initial, les coordonnées du poisson sont $\left(w_0^0 + \kappa_1^0, 0, w_{4m}^0 + \kappa_3^0\right)$. Si l'on désigne par r la largeur constante des couloirs d'exploration, la position finale du navire devrait être

$$p_{abs,nav} = \left(2\left(w_0^0 + \kappa_1^0\right), r, 0\right)$$

et sa vitesse $-\beta^0$, tandis que la position relative du $i$ème nœud par rapport au navire devrait être

$$p_{rel}(i) = \left(-w_i^0 + \kappa_1^0, 0, w_{i+4m}^0 + \kappa_3^0\right)$$

et sa vitesse relative nulle. Nous avons fixé $r = 500$ m. Puisque rien n'assure que cet état final soit atteignable en un temps fini, nous avons prescrit les position et vitesse finales avec une certaine tolérance, c'est-à-dire

$$\forall\, i = 0, ..., 2m - 1, \qquad \left| w^n(i) + \kappa^n - p_{rel}(i) \right|^2 \leq {\varepsilon_p}^2, \qquad \left| z^n(i) + \mu^n \right|^2 \leq {\varepsilon_v}^2$$

pour chaque nœud le long du câble, et

$$\left| u^n - p_{abs,nav} \right|^2 \leq {\varepsilon_p}^2, \qquad \left| \beta^n + \beta^0 \right|^2 \leq {\varepsilon_v}^2$$

pour le navire. Nous prenons systématiquement $\varepsilon_v = 0,1$ m.s^{-1} mais faisons varier ε_p d'un problème-test à l'autre.

Au cours de la trajectoire de demi-tour, la vitesse et l'accélération du navire sont bornées ; plus précisément,

$$\forall\, k = 1, ..., n, \qquad {\beta_{min}}^2 \leq \left| \beta^k \right|^2 \leq {\beta_{max}}^2, \quad \left| \gamma^k \right|^2 \leq {\gamma_{max}}^2.$$

Sa vitesse de changement de cap est également bornée, ce qui se traduit par

$$\forall\, k = 1, ..., n, \qquad -c_{max} \leq \frac{1}{dt} \arcsin\left(\frac{\left| \beta^{k-1} \wedge \beta^k \right|}{\left| \beta^{k-1} \right| \left| \beta^k \right|} \right) \leq c_{max}.$$

La longueur et la vitesse de filage du câble sont elles aussi bornées,

$$\forall\, k = 1, ..., n, \qquad 0 \leq \ell^k \leq \ell_{max}, \quad \left| \frac{\ell^k - \ell^{k-1}}{dt} \right|^2 \leq {f_{max}}^2.$$

Les bornes sur la vitesse du navire sont $\beta_{min} = 0,5$ m.s^{-1} et $\beta_{max} = 1,5$ m.s^{-1}, celle sur l'accélération $\gamma_{max} = 0,017$ m.s^{-2} (correspondant à une variation de la vitesse de 1 m.s^{-1} par minute). Nous limitons la vitesse de changement de cap à $c_{max} = 0,026$ rad.s^{-1} (soit 90 degrés par minute). La longueur du câble entièrement déroulé est $\ell_{max} = 5000$ m et sa vitesse de filage maximale est $f_{max} = 1$ m.s^{-1}.

On rend compte de la présence d'obstacles sous-marins que le poisson doit éviter grâce à des contraintes

$$\forall\, k = 1, ..., n, \qquad w_z^k(0) + \kappa_z^k \geq \varphi\left(w_x^k(0) + \kappa_x^k + u_x^k\,,\, w_y^k(0) + \kappa_y^k + u_y^k \right). \qquad (5.5.a)$$

La fonction φ varie suivant les cas-tests et le type de relief considéré. Nous y faisons figurer une constante de sécurité, assurant que si *(5.5.a)* est active, le poisson reste cependant à une certaine distance du fond. Comme les obstacles considérés ont tous une partie emergée, nous imposons également

$$\forall\, k = 1, ..., n, \qquad 0 \geq \varphi\left(u_x^k, u_y^k \right) \qquad (5.5.b)$$

pour signifier que le navire doit lui aussi les éviter. Nous n'avons pas jugé utile d'imposer des contraintes similaires sur les nœuds intermédiaires le long du câble.

D'un point de vue pratique, il est important de contrôler la tension du câble, qui doit rester en dessous d'une certaine valeur de rupture, et l'angle entre la tangente au câble au niveau du navire et la vitesse de ce dernier. L'angle en question ne doit pas dépasser un certain seuil, faute de quoi le navire ne peut plus manœuvrer le câble. Ces contraintes n'ont pas été prises en considération.

5.3.3. Définition spécifique des problèmes-tests

On distingue 3 familles de cas-tests, selon leur type de relief sous-marin. Tous utilisent le modèle de câble à longueur variable, à l'exception du cas-test PF. Ce dernier emploie le modèle à longueur constante, fixée à 4000 m. Dans tous les problèmes, le nombre de pas de temps est $n = 15$ et le nombre d'éléments de câble est $2m = 6$.

Les problèmes PF, PVa, PVb et PVc sont sans contrainte de relief sous-marin. Est variable la précision ε_p sur la position finale : $\varepsilon_p = 100$ m pour PVa, $\varepsilon_p = 40$ m pour PF et PVb, et finalement $\varepsilon_p = 5$ m pour PVc. Les instruments de mesure du navire permettent de localiser le poisson avec une erreur de l'ordre de 1% de la longueur de câble immergée. Compte-tenu des 4000 m de longueur de câble en fin de demi-tour, on peut détecter dans les trois premiers cas que la position finale du câble n'est pas atteinte avec la précision souhaitée. Nous n'avons choisi $\varepsilon_p = 5$ m dans PVc que pour rendre le problème plus difficile à résoudre et tester l'algorithme. La résolution de tous ces cas-tests se fait toujours à partir des mêmes trajectoires initiales du navire et du câble, représentées sur la figure 5.2 vues de dessus et vues latéralement.

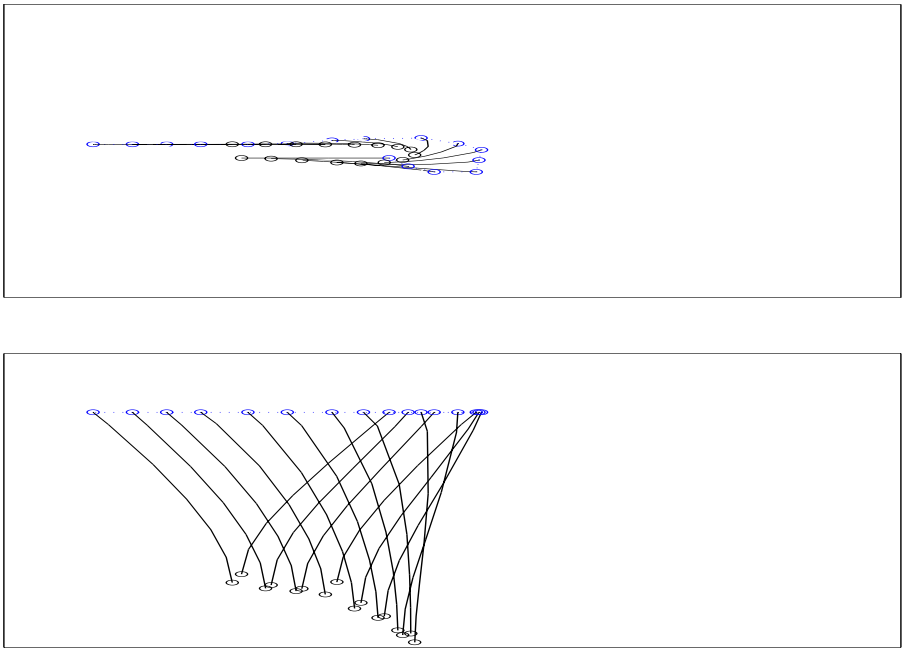


FIG. 5.2 – *trajectoire initiale du navire et du câble des cas-tests sans relief.*

Comme dans tous les cas-tests, ces trajectoires sont obtenues en se donnant une trajectoire du navire et la longueur du câble au cours du temps, c'est-à-dire deux vecteurs $(u^k)_{k=1,\dots,n}$ et $(\ell^k)_{k=1,\dots,n}$. On prend une longueur constante : $\ell^k = 4000$ m. Le pas de temps est calculé de sorte que la vitesse moyenne du navire soit 1 m.s^{-1}. On en déduit la trajectoire du

système câble - engin en intégrant la dynamique *(5.4)*. En d'autres termes, l'itéré initial du problème d'optimisation satisfait toujours ses contraintes d'égalité. En revanche, certaines contraintes d'inégalité ne sont respectées dans aucun problème-test, comme par exemple celles qui portent sur la position et la vitesse finales.

Dans les cas-tests Ia et Ib, nous prenons

$$\varphi(x, y) = \frac{x}{2} - 3600$$

pour modéliser un fond sous-marin plat, dont la profondeur diminue dans la direction de la vitesse initiale du navire (comme si le navire progressait vers une côte). La précision finale est $\varepsilon_p = 40$ m. La différence entre ces deux problèmes-tests tient aux trajectoires initiales du navire et du câble (figures 5.3 et 5.4). Le fond plat représenté ci-après remonte vers la droite. En certains instants, le navire ou le poisson sont sous ce plan – les contraintes *(5.5)* ne sont pas satisfaites. C'est ce qui explique que leurs trajectoires ne soient pas visibles sur toutes leurs longueurs.

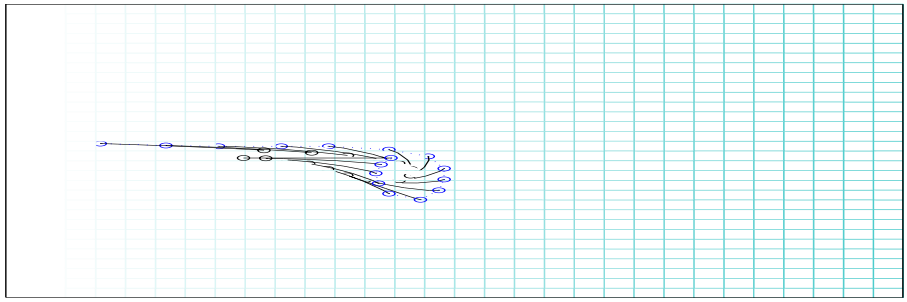


FIG. 5.3 – *trajectoire initiale du cas-test Ia.*

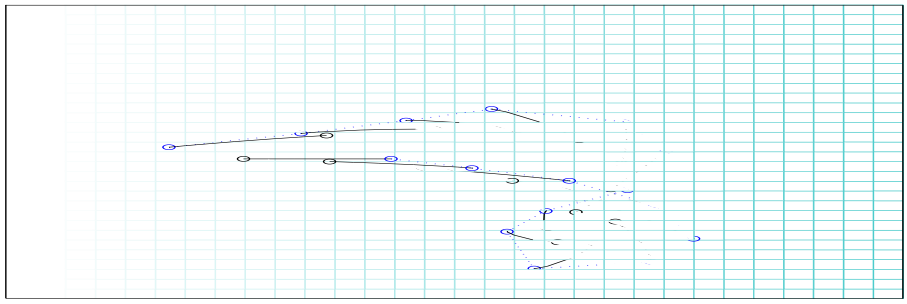


FIG. 5.4 – *trajectoire initiale du cas-test Ib.*

En choisissant une trajectoire initiale pour Ib aussi erratique, et nécessairement très éloignée de la trajectoire optimale que nous cherchons à calculer, nous souhaitons mettre

à l'épreuve la globalisation de l'algorithme par recherche linéaire que nous avons présentée dans les chapitres précédents.

Dans les cas-tests Ca et Cb, nous avons choisi

$$\varphi(x,y) = -\frac{(x-5500)^2+(y+250)^2}{36000} + 1200$$

pour simuler un îlot. A nouveau, les problèmes-tests diffèrent au travers des trajectoires initiales du navire et du câble (figures 5.5 et 5.6). La précision finale est toujours $\varepsilon_p = 40$ m.

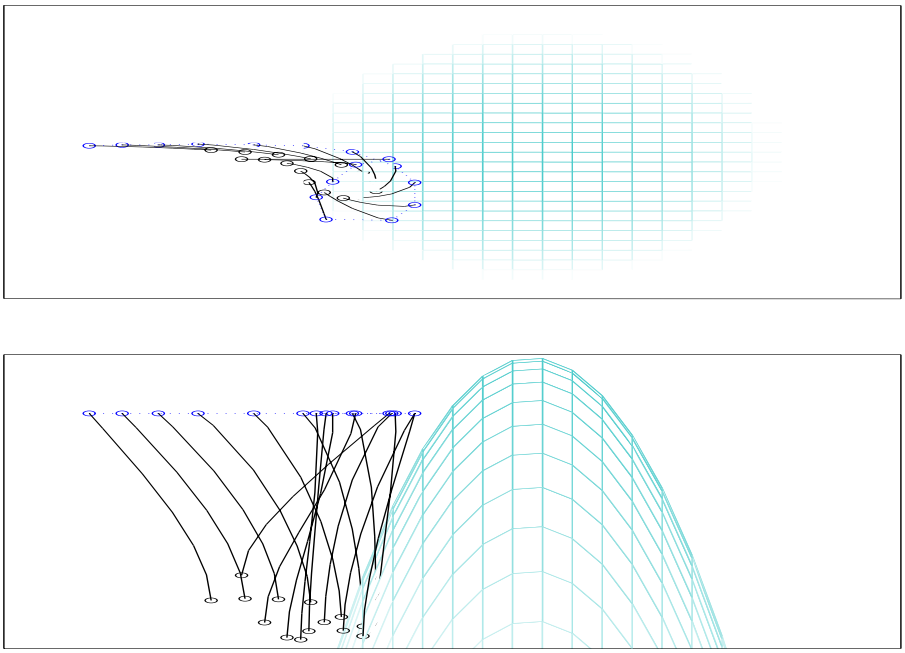


FIG. 5.5 – *trajectoire initiale du cas-test Ca.*

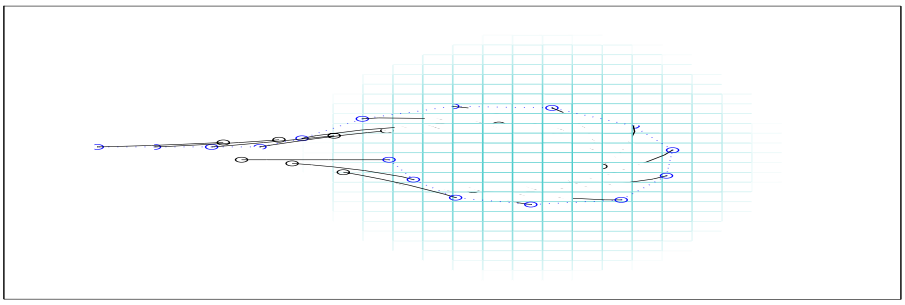


FIG. 5.6 – *trajectoire initiale du cas-test Cb.*

Il semble intuitivement que dans cette situation existent des solution locales. On s'attend à trouver une trajectoire minimisant la durée du demi-tour en contournant l'obstacle, une autre minimisant cette durée sans contourner l'obstacle. C'est ce que nous voulons vérifier en initialisant les trajectoires du navire de part et d'autre de l'îlot.

PF	fond *Plat*, longueur de câble *Fixe*
PVa, PVb, PVc	fond *Plat*, longueur de câble *Variable*, précision $\varepsilon_p = 100$ m, $\varepsilon_p = 40$ m et $\varepsilon_p = 5$ m
Ca, Cb	relief sous forme de *Colline*, trajectoire initiale du navire de part et d'autre de cette colline
Ia, Ib	relief sous forme de plan *Incliné*, trajectoire initiale du navire erratique pour Ib

TAB. 5.1 – *principales caractéristiques des problèmes-tests.*

Nous avons résumé les principales caractéristiques de chacun des problèmes-tests dans le tableau ci-dessus. Ces problèmes sont codés, en Matlab, sous forme (*3.1*). L'optimiseur est également écrit en Matlab.

5.3.4. Initialisation de l'algorithme d'optimisation

Du point de vue de la programmation non linéaire, la dimension de la variable x à optimiser est 631 (problème PF à longueur de câble constante) ou 691. Dans le premier cas, il y 600 contraintes d'égalité $c_E(x) = 0$, et dans les autres cas 645. Le nombre de contraintes d'inégalité $c_I(x) \leq 0$ est 74 pour le problème PF, 134 pour les problèmes PVa, PVb et PVc, et 164 pour le reste des problèmes (dans lesquels figurent des contraintes de profondeur).

Compte-tenu de la manière dont on calcule x_0, itéré initial de l'algorithme d'optimisation, c'est-à-dire en intégrant la dynamique du système câble – engin, les contraintes d'égalité sont initialement satisfaites. Nous avons vu en détaillant les différents cas-tests qu'en revanche, une grande partie des contraintes d'inégalité n'est pas respectée en x_0.

Pour mettre en œuvre notre algorithme, il nous faut initialiser, en plus de x, les variables d'écart $s > 0$, les multiplicateurs λ_I associés aux contraintes d'inégalité et le paramètre de perturbation des conditions d'optimalité μ. En ce qui concerne s, nous avons choisi

$$\forall\, i \in I, \quad s_{i\,0} = \max\,(1, c_i(x_0)).$$

De la sorte, s n'est pas trop petit (chacune de ses composantes est supérieure à 1) et à l'échelle de la contrainte $c_I(x) \leq 0$. Faute de disposer d'une règle pour initialiser μ, nous prenons $\mu = 10^4$, qui s'avère bien convenir aux problèmes-tests (tableau 5.5). Nous en déduisons

$$\lambda_{I\,0} = \mu\, S_0^{-1} e,$$

autrement dit on part d'un point satisfaisant la condition de complémentarité perturbée. Enfin, nous prenons $\lambda_{E\,0} = -A_E^-(x_0)^\top \left(\nabla f(x_0) + A_I(x_0)^\top \lambda_{I\,0}\right)$ dans les tests utilisant le multiplicateur $\lambda_{\text{PQ}\,E}$ et nécessitant donc que λ_E soit initialisé.

5.4. Résultats numériques

Nous nous intéressons dans un premier temps à la résolution du système d'optimalité perturbé *(4.4)* à coefficient μ constant : $\mu = 10^4$. Nous commençons par motiver notre approche en la comparant, dans le cadre des algorithmes de points intérieurs non linéaires, à la méthode de Newton classique, globalisée par recherche linéaire sur la norme du système d'optimalité. Il s'avère que cette dernière méthode ne permet de résoudre aucun de nos problèmes-tests.

5.4.1. Recherche linéaire sur la norme des conditions d'optimalité

Pour la résolution des conditions d'optimalité perturbées *(4.4)*, on pourrait calculer la direction de Newton exacte, solution du système *(4.6)*, puis effectuer une recherche linéaire dans cette direction, en utilisant comme fonction de mérite le carré de la norme euclidienne du système *(4.4)*. La direction de Newton en est effectivement une direction de descente. Le pas initial dans la recherche linéaire respecte bien sûr la contrainte $s, \lambda_I > 0$.

	PF	PVa	PVb	PVc	Ca	Cb	Ia	Ib
it. 0	1.10^7	2.10^6	6.10^6	7.10^6	1.10^6	3.10^6	8.10^5	1.10^6
it. 50	7.10^5	3.10^5	4.10^5	1.10^6	6.10^5	2.10^6	9.10^4	7.10^5
pas	9.10^{-3}	2.10^{-4}	3.10^{-5}	4.10^{-4}	3.10^{-3}	7.10^{-5}	4.10^{-4}	1.10^{-5}

TAB. 5.2 – *norme du système (4.4) aux itérations 0 et 50 et pas à l'itération 50.*

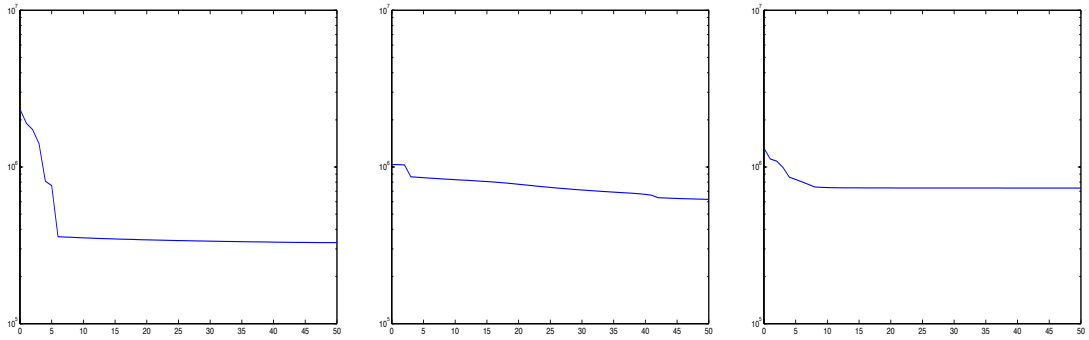


FIG. 5.7 – *représentation à l'échelle logarithmique de la norme du système (4.4) pendant les 50 premières itérations, pour les problèmes-tests PVa, Ca et Ib (de gauche à droite).*

Le tableau 5.2 montre comment évolue la norme des conditions d'optimalité perturbées des différents problèmes-tests entre l'initialisation et la cinquantième itération. D'une manière générale, on a peu progressé, dans le sens où cette norme n'est environ divisée que par 10 en 50 itérations. La figure 5.7 représente cette évolution au cours des itérations, pour les cas-tests PVa, Ca et Ib.

Après une dizaine d'itérations, la décroissance de la norme des conditions d'optimalité perturbées devient extrêmement lente. L'ordre de grandeur du pas obtenu par recherche linéaire se stabilise rapidement (valeurs du tableau 5.2). On distingue deux raisons pour expliquer que ce pas soit si petit. Lors de la résolution des problèmes Cb et Ib, on bute sur les contraintes de positivité de s et λ_I ; le pas ajusté pour satisfaire ces contraintes est ensuite directement admis par la recherche linéaire. Dans les autres cas, c'est la réduction de la norme des conditions d'optimalité perturbées (plus précisément, du gradient du lagrangien) qui impose de prendre le pas petit.

Il est donc clair que l'algorithme calculant la direction de Newton exacte et utilisant le carré de la norme du système d'optimalité *(4.4)* comme fonction de mérite ne permet pas de résoudre nos problèmes-tests en un nombre acceptable d'itérations. Notre méthode de Newton tronquée est foncièrement plus efficace. Nous le vérifions en deux étapes, dans lesquelles l'opérateur de restauration des contraintes mis en œuvre est plus ou moins raffiné.

5.4.2. Algorithmes utilisant l'opérateur de restauration simplifié

Dans le chapitre 3, nous avons présenté deux manières différentes de calculer le pas tangent, basées sur la résolution par itérations de gradient conjugué du système réduit $H\,\tau = -g + Z^{-\top} M\, A^{-} c$ ou bien du système $H\,\tau = -g$. Dans l'analyse de la méthode, ceci revient à considérer deux matrices J (approximations de l'inverse de H) différentes. D'autre part, la convergence globale de l'algorithme a été prouvée lorsque le multiplicateur associé aux contraintes d'égalité (intervenant dans le calcul du hessien du lagrangien L) est le multiplicateur de moindres carrés $\lambda_{\text{MC}\,E}$. Nous nous demandons quels résultats seraient obtenus si on utilisait le multiplicateur du sous-problème quadratique $\lambda_{\text{PQ}\,E}$.

Nous comparons dans le tableau 5.3 les résultats obtenus par les quatre variantes de l'algorithme décrit page 98, variantes combinant les deux calculs du pas tangent et les deux multiplicateurs. Le cas échéant, le calcul de $\lambda_{\text{MC}\,E\,k}$ en début d'itération est remplacé par le calcul de

$$\lambda_{\text{PQ}\,E\,k} = -A_{E\,k}^{-\top} \left(\nabla f_k + L_k\, \Delta x_k + A_{I\,k}^{\top} \left(\lambda_{I\,k} + \Delta \lambda_{I\,k} \right) \right)$$

et la mise à jour $\lambda_{E\,k+1} = \lambda_{E\,k} + \alpha_k \left(\lambda_{\text{PQ}\,E\,k} - \lambda_{E\,k} \right)$ après recherche linéaire. Est employé comme opérateur de restauration celui qui fait intervenir la matrice *(4.11)*, c'est-à-dire l'opérateur sans correction dans l'espace tangent. Nous fixons la valeur de la constante multiplicative du terme $v(s, \lambda_I)$ dans la fonction de mérite *(4.20)* à $\gamma = 0,1$. Ainsi, $f_\mu(z)$ et $\gamma\ v(s, \lambda_I)$ sont de même ordre de grandeur. On observe d'ailleurs que leur ordre de grandeur respectif varie très peu au cours des itérations et d'un cas-test à l'autre. En outre, nous avons vérifié qu'il n'existe pas d'autre valeur de γ qui donne, sur l'ensemble des problèmes-tests, de meilleurs résultats.

Le critère d'arrêt des algorithmes est que la norme des conditions d'optimalité (*4.4*) soit inférieure à 1. Ceci nous assure une précision sur la solution largement suffisante (de l'ordre du dixième de seconde pour le pas de temps, qui est de plusieurs centaines de secondes). Nous avons d'autre part limité le nombre d'itérations. Dans les tableaux suivants, une étoile $\star$ signale les problèmes-tests que l'algorithme n'a pas pu résoudre en moins de 300 itérations.

Dans le cas contraire, nous indiquons le nombre d'itérations avant convergence, le nombre de simulations, la valeur minimale du pas au cours des itérations, le nombre d'itérations où le pas unité a été accepté et le nombre d'itérations où le pas unité a été accepté grâce à une correction du second ordre. L'algorithme ne demande qu'une seule évaluation des dérivées premières et secondes par itérations, sauf en cas de calcul d'une correction du second ordre, où une deuxième évaluation des dérivées premières est nécessaire. Le nombre de simulations désigne donc précisément le nombre d'évaluation des fonctions du problème et non pas de leurs dérivées ; il est équivalent au nombre d'essais de pas dans la recherche linéaire.

Commentons ces résultats, en commençant par quelques remarques générales. Le pas unité est admis par la recherche linéaire dans les dernières itérations de l'algorithme (quand celui-ci converge), et seulement dans les dernières itérations. Le plus souvent, le pas est réduit une première fois pour tenir compte de la contrainte de positivité sur s et λ_I, puis à nouveau pour faire décroître la fonction de mérite (*4.20*). Ceci donne à penser que le domaine de convergence de la méthode de Newton locale est, pour chaque problème-test, relativement petit. Par conséquent, la globalisation par recherche linéaire est, au moins dans la plupart des cas, efficace. On remarque que c'est grâce à la correction du second ordre que le pas unité est asymptotiquement accepté lors de la résolution des problèmes PVc par l'algorithme 1, Ca et Ia par l'algorithme 2.

Les solutions obtenues sont des points stationnaires du problème barrière (*4.5*), avec $\mu = 10^4$. Nous avons vérifié pour chaque cas-test que ce sont bien des minima locaux. Ces points vérifient les conditions suffisantes d'optimalité du second ordre de (*4.5*) : le hessien réduit H y est défini positif. En revanche, H n'est jamais défini positif durant les premières itérations de l'algorithme, lors desquelles on rencontre toujours des directions à courbure négative.

Lorsque les différentes variantes de l'algorithme convergent, elles convergent toujours vers la même solution. Aucune de ces variantes n'est nettement plus efficace que les autres sur l'ensemble des cas-tests. Il est plutôt surprenant que les résultats des versions 1 et 2, calculant L à l'aide du multiplicateur $\lambda_{\text{MC}\,E}$ et de ce fait réputées plus stables, ne soient pas meilleurs que ceux des versions 3 et 4, utilisant $\lambda_{\text{PQ}\,E}$. Il est également difficile de départager les deux modes de calcul du pas tangent (c'est-à-dire les deux matrices J).

Expliquons l'échec de l'algorithme 1 appliqué au cas-test Ib, et plus généralement le nombre important d'itérations pour la résolution des problèmes Ia et Ib, ou PF par l'algorithme 4. Le paramètre de troncature des itérations de gradient conjugué est fixé à $\varepsilon = 10^{-5}$.

	PF	PVa	PVb	PVc	Ca	Cb	Ia	Ib
itérations	45	13	29	107	140	33	216	$\star$
simulations	135	31	82	350	499	91	801	$\star$
pas min.	1.10^{-3}	1.10^{-1}	2.10^{-3}	6.10^{-5}	3.10^{-10}	3.10^{-4}	2.10^{-10}	$\star$
pas unité	5	9	7	8	4	5	7	$\star$
corrections	0	0	0	4	0	0	3	$\star$

alg. 1 : itérations de gradient conjugué sur $H\,\tau = -g + Z^{-\top} M\, A^{-}\, c$, *multiplicateur* $\lambda_{\text{MC}\,E}$

itérations	100	15	57	106	51	24	225	214
simulations	340	36	197	340	158	61	875	753
pas min.	3.10^{-7}	1.10^{-1}	1.10^{-7}	2.10^{-4}	9.10^{-4}	2.10^{-3}	8.10^{-12}	1.10^{-11}
pas unité	5	10	6	4	7	6	8	6
corrections	1	2	1	0	7	0	8	0

alg. 2 : itérations de gradient conjugué sur $H\,\tau = -g$, *multiplicateur* $\lambda_{\text{MC}\,E}$

itérations	85	27	43	89	100	24	168	188
simulations	264	79	130	280	338	65	551	581
pas min.	7.10^{-5}	1.10^{-2}	6.10^{-4}	5.10^{-3}	5.10^{-8}	2.10^{-2}	9.10^{-11}	1.10^{-9}
pas unité	5	6	6	8	4	5	8	7
corrections	1	2	1	0	0	0	0	2

alg. 3 : itérations de gradient conjugué sur $H\,\tau = -g + Z^{-\top} M\, A^{-}\, c$, *multiplicateur* $\lambda_{\text{PQ}\,E}$

itérations	278	29	64	107	44	38	143	152
simulations	1368	83	190	332	130	106	483	433
pas min.	3.10^{-26}	1.10^{-2}	4.10^{-3}	4.10^{-4}	2.10^{-4}	1.10^{-4}	3.10^{-10}	4.10^{-7}
pas unité	5	7	5	11	5	5	8	7
corrections	1	1	0	1	1	0	2	2

alg. 4 : itérations de gradient conjugué sur $H\,\tau = -g$, *multiplicateur* $\lambda_{\text{PQ}\,E}$

TAB. 5.3 – *variantes utilisant l'opérateur (4.11).*

	PF	PVa	PVb	PVc	Ca	Cb	Ia	Ib
itérations	26	14	40	73	38	23	287	⋆
simulations	76	36	119	220	109	61	1075	⋆
pas min.	2.10^{-3}	4.10^{-3}	6.10^{-3}	3.10^{-3}	1.10^{-2}	1.10^{-4}	4.10^{-11}	⋆
pas unité	6	9	6	15	5	4	8	⋆
corrections	0	1	0	0	3	0	3	⋆

alg. 5 : itérations de gradient conjugué sur $H\,\tau = -g + Z^{-\top} M\, A^{-}\, c$, *multiplicateur* $\lambda_{\text{MC}\,E}$

itérations	41	35	39	146	43	27	⋆	⋆
simulations	127	103	135	455	128	70	⋆	⋆
pas min.	6.10^{-4}	5.10^{-3}	6.10^{-5}	2.10^{-3}	1.10^{-4}	2.10^{-5}	⋆	⋆
pas unité	5	7	7	15	4	4	⋆	⋆
corrections	1	2	7	14	2	0	⋆	⋆

alg. 6 : itérations de gradient conjugué sur $H\,\tau = -g$, *multiplicateur* $\lambda_{\text{MC}\,E}$

itérations	67	45	39	87	44	30	35	152
simulations	200	130	112	262	124	83	101	446
pas min.	4.10^{-4}	8.10^{-4}	3.10^{-3}	6.10^{-3}	2.10^{-3}	4.10^{-4}	2.10^{-2}	4.10^{-7}
pas unité	5	7	7	23	5	5	8	10
corrections	1	1	0	15	0	0	3	1

alg. 7 : itérations de gradient conjugué sur $H\,\tau = -g + Z^{-\top} M\, A^{-}\, c$, *multiplicateur* $\lambda_{\text{PQ}\,E}$

itérations	67	47	31	87	63	24	59	⋆
simulations	197	139	89	272	193	63	195	⋆
pas min.	5.10^{-4}	2.10^{-3}	7.10^{-3}	8.10^{-3}	6.10^{-3}	3.10^{-4}	5.10^{-8}	⋆
pas unité	6	7	8	23	5	5	5	⋆
corrections	0	1	2	15	3	0	3	⋆

alg. 8 : itérations de gradient conjugué sur $H\,\tau = -g$, *multiplicateur* $\lambda_{\text{PQ}\,E}$

TAB. 5.4 – *variantes utilisant l'opérateur (4.16), avec* $P_x = diag\,(|\,R_{i\,i}\,|)$ *et* $P_s = \mu\, S^{-2}$.

La valeur propre minimale de H en chacune des solutions des cas-tests est de l'ordre de 10^{-2} (à l'exception de PF, pour lequel $\lambda_{min}(H) \approx 20$). La valeur de ε – à choisir bien entendu a priori, sans connaissance du spectre de H – doit être suffisamment petite si l'on veut résoudre le système réduit de manière exacte au voisinage d'une solution, et par conséquent assurer la convergence quadratique de la méthode.

Dans chaque cas précédent, sont traversés des points $(x, s, \lambda_E, \lambda_I)$ en lesquels H possède des valeurs propres strictement positives proches de 0 (en lesquels donc sa partie définie positive est très mal conditionnée). Pour le problème PF, par exemple, elles s'échelonnent entre 20 et 10^9 en la solution, mais entre 10^{-7} et 10^8 au cours d'une partie des itérations de l'algorithme 4. A cause de ce mauvais conditionnement et d'une valeur certainement trop petite de ε, nous sommes amenés à explorer des directions internes de norme élevée. Ceci se traduit aussi par une matrice J anormalement grande, bien que sa valeur propre maximale reste inférieure à 10^5 et respecte l'inégalité *(3.17)*. Calculant des solutions des systèmes

$$\begin{cases} H\,\tau = -g + Z^{-\top} M\, A^{-} c \\ H\,\tau = -g \end{cases}$$

de normes trop importantes, nous observons que la direction de Newton tronquée Δx, le multiplicateur λ_{NT} et par conséquent le coefficient de pénalisation σ de la fonction de mérite *(4.20)*, sont également trop grands. Le pas maximal le long des directions Δs et $\Delta\lambda_I$ respectant la contrainte de positivité sur s et λ_I est extrêmement petit ; il est encore réduit lors de la recherche linéaire, pour faire décroître *(4.20)*.

Faute d'arrêter les itérations de gradient conjugué avant de rencontrer des directions internes trop grandes, nous calculons un pas tangent lui aussi trop grand. Les performances médiocres des algorithmes sont donc dues au fait que nous ne savons pas quand précisément interrompre la résolution du système réduit. Remarquons, pour étayer cette affirmation, qu'il existe une forte corrélation entre un nombre important d'itérations et un pas minimal petit (signe que le déplacement $(\Delta x, \Delta s, \Delta\lambda_I)$ est trop grand ou mal calculé). On relève par exemple dans le tableau 5.3 que l'algorithme 4, appliqué au cas-test Ca, converge en 44 itérations, au cours desquelles le pas reste supérieur à 2.10^{-4}, alors que l'algorithme 1 converge en 140 itérations et fait décroître le pas jusqu'à 3.10^{-10}.

Nous avions souligné que le choix du paramètre ε est délicat. L'adapter au cours des itérations externes semble indispensable; on gagnerait assurément à le réduire au fur et à mesure que l'on progresse vers une solution de *(4.4)*. En outre, puisque le mauvais conditionnement de H est une difficulté intrinsèque du problème, il serait avantageux de préconditionner le système réduit (Morales et Nocedal [76]), dont il s'agit, pour résumer, de mieux contrôler la résolution.

5.4.3. Algorithmes utilisant un opérateur de restauration corrigé

Nous modifions les algorithmes 1, 2, 3 et 4, en calculant désormais un pas de restauration avec correction dans l'espace tangent (se reporter en section 4.1.4). Il serait fastidieux

de détailler ici les nombreux essais numériques que nous avons effectués. On peut choisir P_s entre μS^{-2} et $S^{-1}\Lambda_I$. On peut aussi prendre P_x parmi les hessiens L et R, les matrices diagonales $\mathrm{diag}(|L_{ii}|)$ et $\mathrm{diag}(|R_{ii}|)$, dont les éléments sont les valeurs absolues des éléments diagonaux de L et R, ou bien encore un multiple de la matrice identité βI (il est dans ce dernier cas difficile de donner une valeur à β) ; rappelons que $R(x, \lambda_{\text{CA}E}, \lambda_{\text{CA}I})$ est le hessien par rapport à x du lagrangien du problème de centre analytique *(4.14)*. Enfin, nous avons essayé $P_\tau = Z_E^{-\top} P_x Z_E^-$ pour chacune des matrices P_x précédentes.

Le tableau 5.4 présente les résultats obtenus en utilisant la matrice *(4.16)*, dans laquelle $P_x = \mathrm{diag}(|R_{ii}|)$ et $P_s = \mu S^{-2}$; ils sont en moyenne meilleurs que pour les autres possibilités d'opérateur de restauration que nous avons énumérées. Dans tous les cas-tests, la version 7 converge en un nombre acceptable d'itérations, à l'exception peut-être de Ib. Mais ce problème a été conçu de telle sorte que sa résolution soit difficile. D'une manière générale, les problèmes PF, PVa, PVb, PVc, Ca et Cb sont plus facilement résolus avec cet opérateur de restauration sophistiqué. Mais paradoxalement, les performances des variantes 1 et 2 appliquées à Ia et Ib se dégradent.

Les mauvais résultats et échecs des algorithmes restent imputables au manque de précision dans la résolution du système réduit. Poussant les itérations de gradient conjugué trop loin, nous calculons un pas tangent et donc une direction de Newton tronquée trop grands. Bien sûr, un changement d'opérateur de restauration n'est pas censé remédier à ceci. Le pas minimal lors de la résolution infructueuse du problème Ib par l'algorithme 5 est de l'ordre de 10^{-25}. Se produit exactement ce que nous avions déjà observé en appliquant la variante 4 à PF.

Cet essai numérique nous semble montrer l'intérêt d'un inverse à droite de type *(4.16)*. Il confirme aussi la nécessité de mieux résoudre le système réduit pour améliorer la méthode de Newton tronquée. En revanche, aucune des versions de l'algorithme n'est clairement désignée comme meilleure que les autres. Si l'algorithme 7 se comporte mieux que les autres sur l'ensemble des cas-tests, l'algorithme 5 est le plus efficace dans la résolution des problèmes PF, PVa, PVc, Ca et Cb.

5.4.4. Réduction du coefficient de perturbation

L'idée centrale des méthodes de points intérieurs est de déterminer un point stationnaire du problème *(3.1)* en résolvant les équations perturbées *(4.4)* pour des valeurs du paramètre μ décroissant vers 0. Une première difficulté est qu'il n'existe aucune règle communément acceptée pour initialiser μ avant ce processus.

Donner une valeur à μ est d'autant plus difficile qu'il n'a pas d'interprétation physique. Nous avons repris, pour chaque $\mu \in \{10^3, 10^4, 10^5\}$, l'itéré initial $(x, s, \lambda_E, \lambda_I)$ détaillé page 126, et noté dans le tableau 5.5 le nombre d'itérations pour résoudre *(4.4)*. Ces itérations de correction sont basées sur l'algorithme 7 de la section précédente.

Des trois valeurs de μ considérées, 10^4 est la seule pour laquelle on puisse toujours résoudre – en moins de 300 itérations – le système perturbé *(4.4)*. Mais ce test illustre essentiellement qu'il y a tout lieu de choisir μ en fonction du problème : $\mu = 10^5$ convient nettement mieux pour PF, Cb et Ib. La question, sans réponse à ce jour, de l'initialisation de μ, est donc de toute importance.

	PF	PVa	PVb	PVc	Ca	Cb	Ia	Ib
$\mu = 10^5$	22	95	135	208	$\star$	25	64	143
$\mu = 10^4$	67	45	39	87	44	30	35	152
$\mu = 10^3$	186	43	242	211	$\star$	50	$\star$	$\star$

TAB. 5.5 – *nombre d'itérations de la première phase de correction.*

Appelons, par analogie avec la programmation linéaire, phase de correction la résolution du système *(4.4)* pour une valeur fixe de μ. Partant de 10^4, nous réduisons maintenant μ d'un facteur constant après chaque étape de correction, en le divisant par 10. Nous avons donc appliqué l'algorithme suivant, pour résoudre le problème avec contraintes d'égalité et d'inégalité *(3.1)*.

Initialisation.
Poser $\ell = 0$. Choisir x_0, $s_0 > 0$, $\lambda_{I\,0} > 0$, $\mu_1 > 0$ et $\nu > 0$. Passer à l'itération 1.

Itération ℓ, $\ell \geq 1$.
a. Calculer $(x_\ell, s_\ell, \lambda_{E\,\ell}, \lambda_{I\,\ell})$, solution à ν près du système d'optimalité *(4.4)* dans lequel $\mu = \mu_\ell$, par la méthode de Newton tronquée, initialisée en $(x_{\ell-1}, s_{\ell-1}, \lambda_{I\,\ell-1})$.
b. Si un test de convergence portant sur la norme des conditions d'optimalité *(4.1)* évaluées en $(x_\ell, \lambda_{E\,\ell}, \lambda_{I\,\ell})$ est satisfait, terminer.
e. Mettre à jour $\mu_{\ell+1} = 0{,}1\,.\,\mu_\ell$ puis passer à l'itération $\ell + 1$.

La solution $(x, s, \lambda_E, \lambda_I)$ d'une phase de correction devient l'itéré initial de la phase de correction suivante. La méthode de Newton tronquée mise en œuvre est l'algorithme 7. Figurent dans le tableau 5.6 le nombre d'itérations de correction aux valeurs succesives de μ (dites itérations internes) et le nombre de ces itérations où le pas unité a été accepté. Les résultats sont les mêmes pour Ia et Ib (sauf dans la première étape de correction, à $\mu = 10^4$) et ne sont donc indiqués que pour Ia.

Ayant constaté que x n'évolue pratiquement plus, nous arrêtons cette procédure à $\mu = 1$. Les conditions d'optimalité *(4.1)* – non pertubées – sont alors résolues avec une précision tout à fait satisfaisante. A l'exception du problème-test Cb, le nombre d'itérations internes décroît avec μ, et dans de nombreux cas, plus de la moitié de ces itérations est consacrée

à la résolution du premier système *(4.4)*. On observe que le pas unité est fréquemment admis dans les itérations à $\mu \leq 10^3$. Il est donc bien plus aisé de suivre la trajectoire centrale que de s'en approcher en partant d'un point éloigné.

		PF	PVa	PVb	PVc	Ca	Cb	Ia
$\mu = 10^4$	itérations	67	45	39	87	44	30	35
	pas unité	5	7	7	23	5	5	8
$\mu = 10^3$	itérations	7	9	11	22	37	8	13
	pas unité	7	6	6	15	16	5	5
$\mu = 10^2$	itérations	4	5	5	11	4	23	6
	pas unité	4	5	3	8	4	4	4
$\mu = 10$	itérations	2	3	4	3	3	31	3
	pas unité	2	3	4	3	2	6	3
$\mu = 1$	itérations	1	1	2	2	1	4	2
	pas unité	1	1	2	2	1	4	2
nombre total d'itérations de Newton		81	63	61	125	89	96	61
temps de calcul (minutes CPU sur une station DEC Alpha 500)		24	27	26	54	40	43	27

TAB. 5.6 – *itérations de Newton tronquées pour des valeurs de μ décroissantes.*

Globalement, le nombre d'itérations internes est peu élevé et la résolution des problèmes rapide (ce qui est également dû à une discrétisation assez grossière). Nous avons vérifié que les solutions obtenues à $\mu = 1$ (que nous savons être très proches des solutions à $\mu = 0$) sont bien des minima locaux du problème *(3.1)* correspondant à chaque problème-test. Nous avons donc bien déterminé des trajectoires du système navire – câble – engin en temps localement minimal.

5.4.5. Solutions des problèmes-tests

Les figures 5.8 et 5.9 représentent, dans les cas-tests Ia et Ib, les trajectoires solutions du système d'optimalité *(4.4)*, respectivement pour $\mu = 10^4$ et $\mu = 0$. La durée de la première est de 2 h 20 min et celle de la seconde 1 h 31 min. La figure 5.10 montre que la trajectoire initiale ne satisfaisait aucune des contraintes sur la position du câble au temps final, sur la vitesse du navire ou encore sur la profondeur de l'engin remorqué.

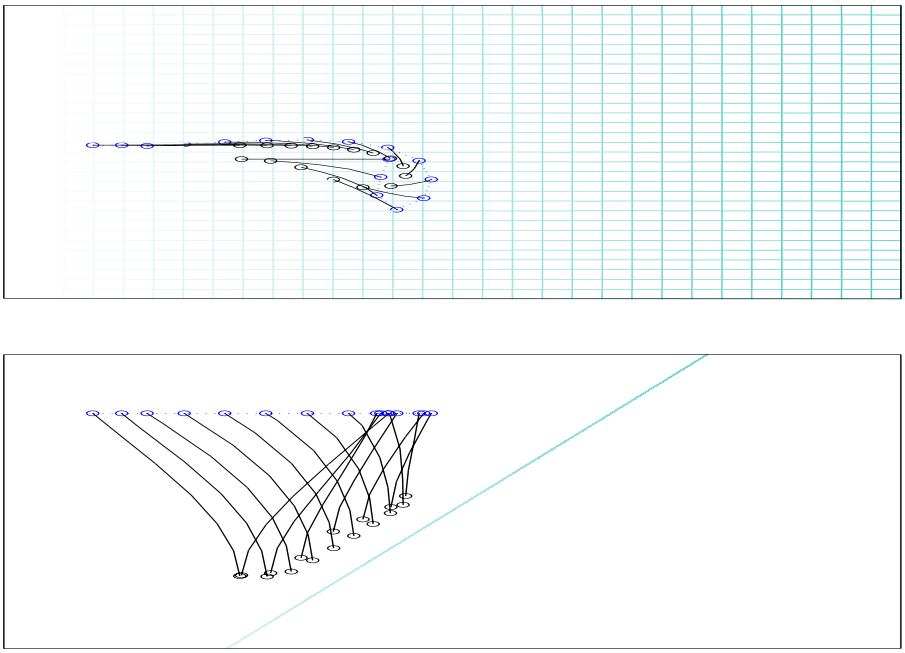


FIG. 5.8 – *trajectoire solution du système (4.4) avec* $\mu = 10^4$*, pour les cas-tests Ia et Ib.*

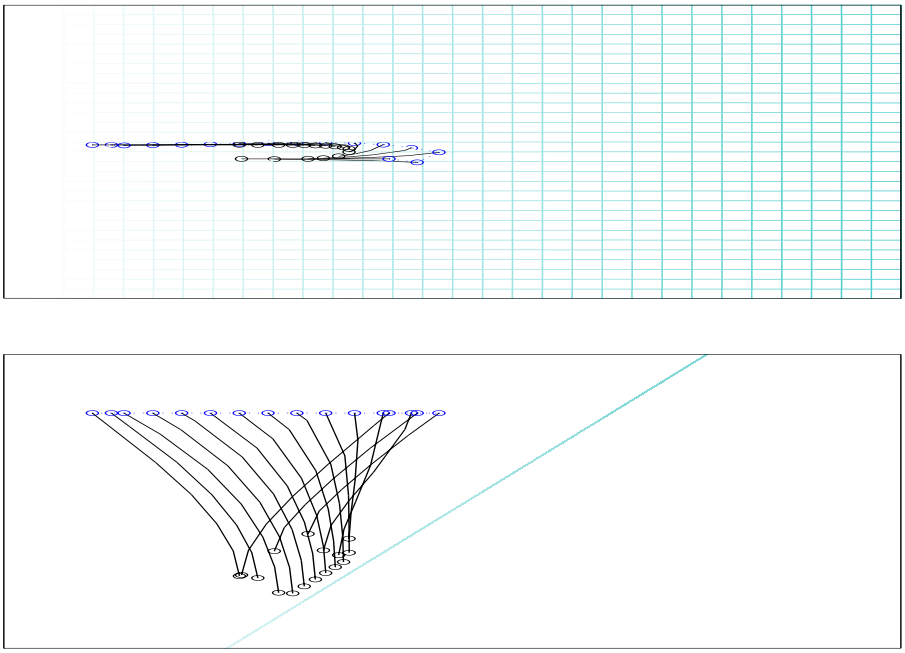
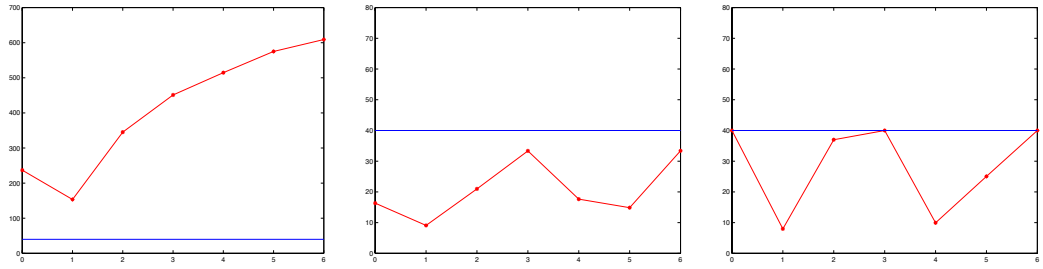
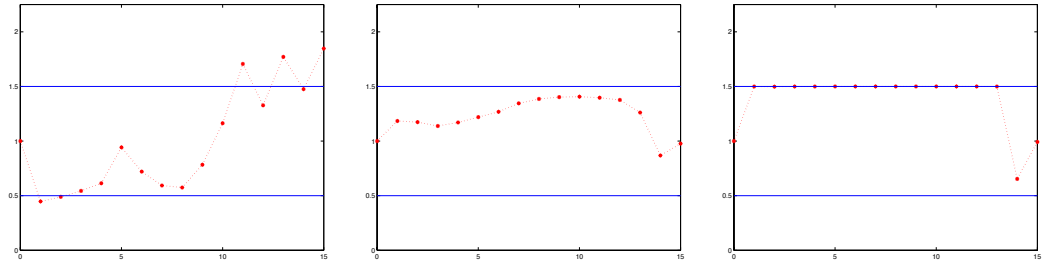


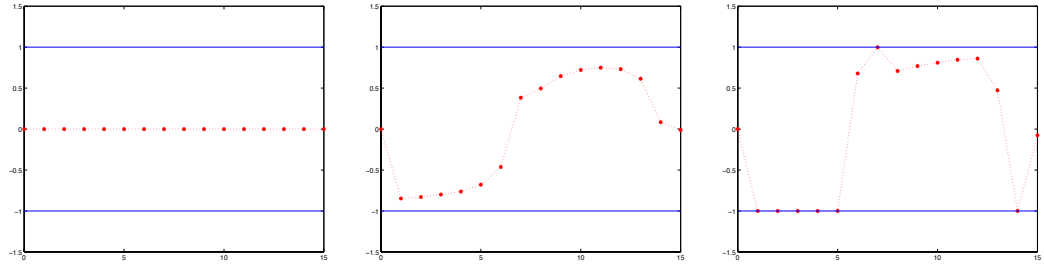
FIG. 5.9 – *trajectoire optimale des cas-tests Ia et Ib.*



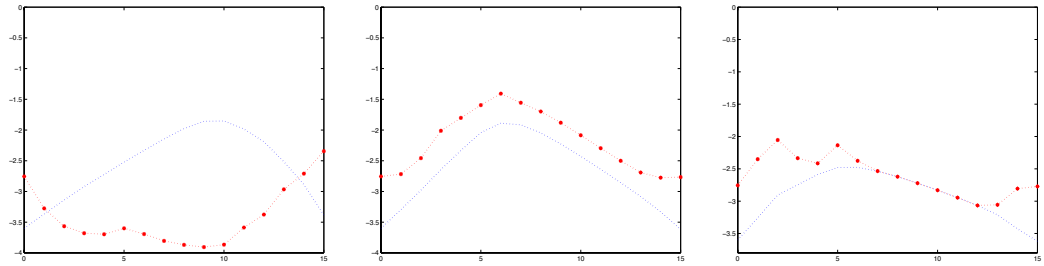
contrainte sur la position finale du câble



contrainte sur la vitesse du navire



contrainte sur la vitesse de filage du câble.



contrainte sur la profondeur du poisson

FIG. 5.10 – *évolution de certaines contraintes d'inégalité au cours de l'algorithme (de gauche à droite, trajectoires initiale, solutions à $\mu = 10^4$ et $\mu = 0$), pour le cas-test Ia.*

Le fait que la solution à $\mu = 10^4$ soit un point strictement intérieur, en lequel les contraintes précédentes sont inactives, se traduit par une trajectoire du poisson nettement au-dessus du fond sous-marin. La trajectoire du navire est assez surprenante. On l'explique de la manière suivante : elle permet de remonter le poisson vers la surface suffisamment tôt pour éviter le plan incliné sans toutefois enrouler le câble trop rapidement.

A $\mu = 0$, les contraintes sont saturées à certains instants. On fait fonctionner le treuil à plein régime. La contrainte sur la profondeur du poisson est active ; s'il ne heurte pas le fond, c'est grâce à la distance de sécurité que nous incluons dans (*5.5.a)*. C'est maintenant en enroulant le câble à vitesse maximale que l'on satisfait cette contrainte. Sa vitesse de filage était d'ailleurs peu prévisible : on déroule le câble après avoir évité l'obstacle, pour l'enrouler à nouveau en fin de trajectoire. Nous conjecturons que cette manœuvre a un effet stabilisant.

Nous indiquons dans le tableau 5.7 les durées des différentes trajectoires optimales. Pour les problèmes sans relief sous-marin et à longueur de câble variable, ces durées sont d'autant plus importantes que l'écart ε_p toléré sur la position finale du câble est faible (100 m pour PVa, 40 m pour PVb et 5 m pour PVc).

PV	PVa	PVb	PVc	Ca	Cb	Ia & Ib
1 h 54 min	1 h 16 min	1 h 27 min	1 h 40 min	1 h 28 min	3 h 35 min	1 h 31 min

TAB. 5.7 – *durées optimales calculées dans les différents problèmes-tests.*

On pouvait également prévoir que la durée optimale serait plus grande dans le cas PF que dans le cas PVb. Tous deux considèrent $\varepsilon_p = 40$ m, mais PF est basé sur le modèle de câble à longueur constante. Les trajectoires admissibles pour PF sont donc admissibles pour PVb. Le fait de pouvoir enrouler et dérouler le câble réduit la durée de la manœuvre de 25 min, soit un gain en temps de 22% relativement important. Un essai numérique détaillé dans [27] démontre d'ailleurs que cette possibilité est d'autant plus intéressante que ε_p est petit (lorsque $\varepsilon_p = 1$ m, on réduit de 40% la durée du demi-tour en bobinant le câble).

Les durées optimales sont très proches dans les problèmes Ia, Ca et PVb. On constate sur les figures 5.9, 5.11 et 5.13 que les trajectoires du navire sont similaires. Le fond sous-marin n'influence essentiellement que le filage du câble et par conséquent la profondeur de l'engin. Comme nous le pressentions, la trajectoire optimale dans le cas-test Cb contourne l'îlot et n'est pas globalement optimale ; elle a ceci de particulier que le câble est totalement remonté durant une partie du demi-tour, ce qui autorise le navire à passer au plus près de l'îlot (figure 5.12).

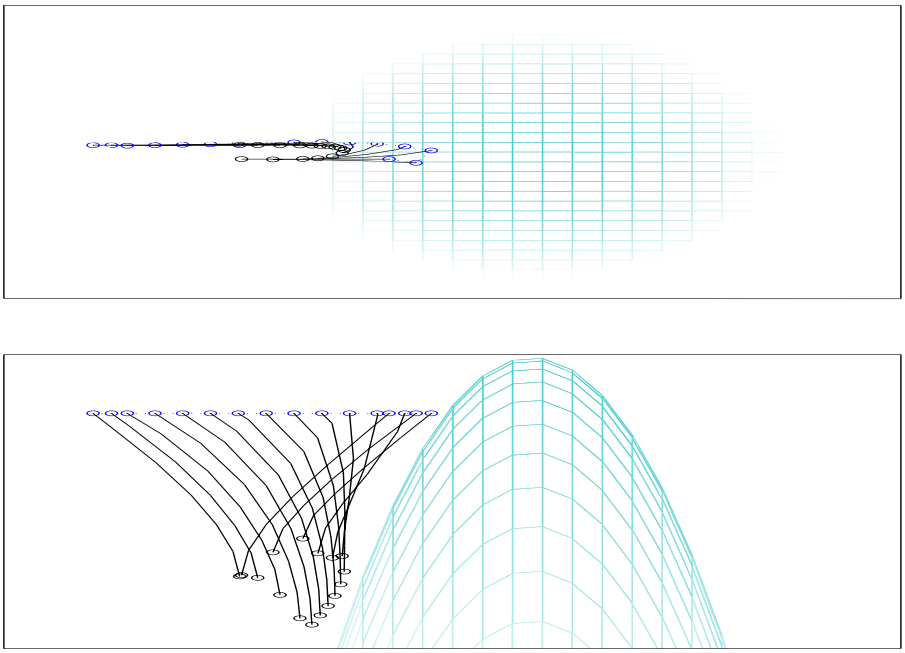


FIG. 5.11 – *trajectoire optimale du cas-test Ca.*

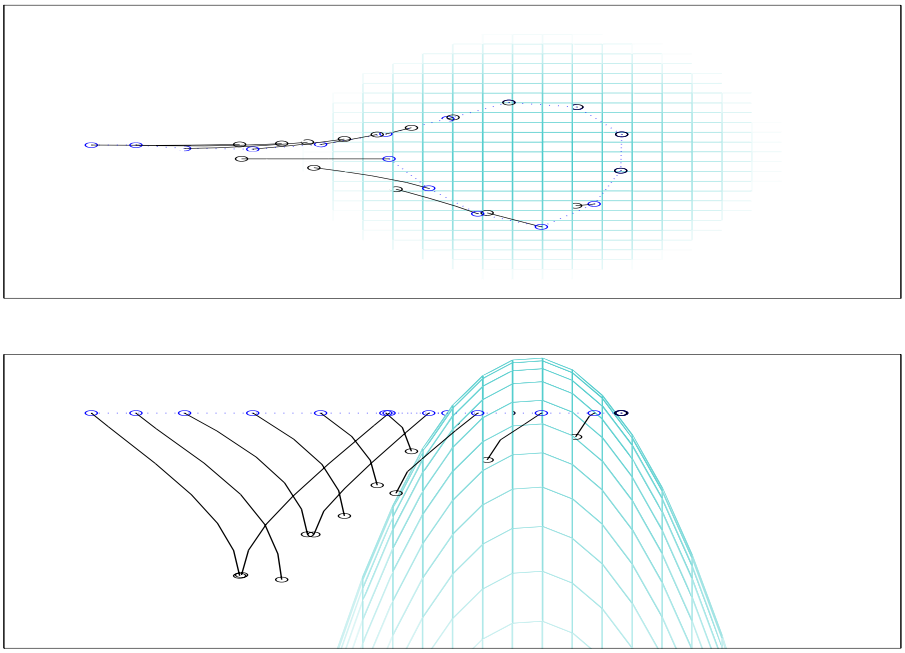


FIG. 5.12 – *trajectoire optimale du cas-test Cb.*

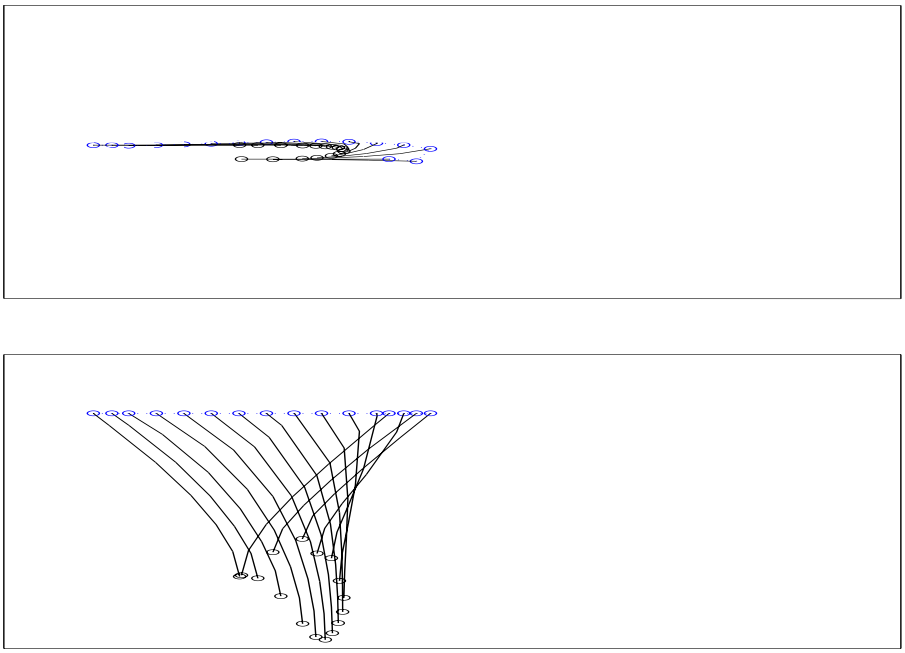


FIG. 5.13 – *trajectoire optimale du cas-test PVb.*

Le temps de calcul des solutions que nous venons de présenter (ou, de façon équivalente, le nombre d'itérations de Newton) est tout à fait raisonnable. L'algorithme mis en œuvre a été conçu dans le but de résoudre des problèmes de commande optimale avec des contraintes d'inégalité sur l'état (telles que les contraintes sur la position finale du câble ou sur la profondeur du poisson, dans nos cas-tests). Il nous semble avoir montré son potentiel ; revenons sur quelques-uns de ses points faibles.

5.4.6. Perspectives de développement de l'algorithme

Au sujet des problèmes avec contraintes d'égalité seulement, nous avons déjà insisté sur l'obligation de mieux résoudre le système réduit (en mettant au point une règle plus précise d'arrêt des itérations de gradient conjugué, en le préconditionnant). Ceci est impératif pour s'attaquer à la résolution de problèmes plus finement discrétisés, où les difficultés déjà rencontrées se font sentir de manière encore plus aigüe. Un autre axe de recherche important est la gestion des directions à courbure négative détectées lors de la résolution du système réduit.

En ce qui concerne les problèmes avec contraintes d'égalité et d'inégalité, la réflexion sur le choix d'un opérateur de restauration des contraintes mérite d'être poursuivie. Une règle de décroissance du paramètre de perturbation μ devra être mise au point (en s'inspirant des algorithmes prédicteurs-correcteurs de la programmation linéaire, par exemple). La recherche linéaire pourrait être modifiée : lorsque le pas devait être réduit,

nous utilisions dans nos tests une formule d'interpolation basée sur un modèle quadratique de la fonction de mérite, ce qui est classique. Or un des termes composant cette fonction est un logarithme. On peut dès lors s'interroger sur l'adéquation d'un tel modèle.

En vue de ces améliorations, les résultats figurant dans cette thèse ne marquent donc qu'une étape, certes prometteuse, dans le développement d'une version industrielle de notre algorithme d'optimisation.

Références

[1] K. M. Anstreicher et J. P. Vial. *On the Convergence of an Infeasible Primal-Dual Interior-Point Method for Convex Programming.* Optimization Methods and Software, 1994, vol. 3, pp. 273–283.

[2] P. Armand, J. C. Gilbert et S. Jan-Jégou. *A Feasible BFGS Interior Point Algorithm for Solving Strongly Convex Minimization Problems.* Rapport technique 3500, INRIA, 1998. A paraître dans SIAM Journal on Optimization.

[3] L. Armijo. *Minimization of Functions having Lipschitz Continuous First Partial Derivatives.* Pacific Journal of Mathematics, 1966, vol. 16, pp. 1–3.

[4] U. Ascher et L. Petzold. *Computer Methods for Ordinary Differential Equations and Differential Algebraic Equations.* SIAM, 1998.

[5] I. Babuška. *The Finite Element Method with Lagrange Multipliers.* Numerische Mathematik, 1973, vol. 20, pp. 179–192.

[6] M. Bardi et I. Capuzzo-Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations.* Birkhäuser, 1997.

[7] G. Barles. *Solutions de viscosité des équations de Hamilton-Jacobi.* Mathématiques et applications, Springer, 1994.

[8] R. Bellman. *Dynamic Programming.* Princeton University Press, 1957.

[9] J. T. Betts. *Survey of Numerical Methods for Trajectory Optimization.* Journal of Guidance, Control and Dynamics, 1998, vol. 21, pp. 193–207.

[10] J. T. Betts et W. P. Huffman. *Mesh Refinement in Direct Transcription Methods for Optimal Control.* Optimal Control Applications & Methods, 1998, vol. 19, pp. 1–21.

[11] P. Boggs et J. Tolle. *Sequential Quadratic Programming.* Acta Numerica, 1995, vol. 4, pp. 1–51.

[12] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal et C. Sagastizábal. *Optimisation numérique.* Mathématiques et applications, Springer, 1997.

[13] J. F. Bonnans et G. Launay. *Large-Scale Direct Optimal Control Applied to a Re-Entry Problem.* Journal of Guidance, Control, and Dynamics, 1998, vol. 21, pp. 996–1000.

[14] J. F. Bonnans et A. Shapiro. *Optimization Problems with Perturbations: a Guided Tour.* SIAM Review, 1998, vol. 40, pp. 228–264.

[15] J. F. Bonnans et A. Shapiro. *Perturbation Analysis of Optimization Problems.* Springer-Verlag, 1999.

[16] K. E. Brenan, S. L. Campbell et L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations.* North-Holland, 1989.

[17] K. E. Brenan et B. E. Engquist. *Backward Differentiation Approximations of Nonlinear Differential Algebraic Equations.* Mathematics of Computation, 1988, vol. 51, pp. 659–676.

[18] H. Brezis. *Analyse fonctionnelle. Théorie et applications.* Mathématiques appliquées pour la maîtrise, Masson, 1983.

[19] F. Brezzi. *On the Existence, Uniqueness and Approximation of Saddle Point Problems Arising from Lagrangian Multipliers.* Revue française d'Automatique, Informatique et Recherche Opérationnelle : analyse numérique, 1974, vol. 8, pp. 129–151.

[20] F. Brezzi et M. Fortin. *Mixed and Hybrid Finite Element Methods.* Springer Series in Computational Mathematics, Springer-Verlag, 1991.

[21] A. E. Bryson et Y. C. Ho. *Applied Optimal Control.* Hemisphere Publishing Company, 1975.

[22] R. Bulirsch, F. Montrone et H. J. Pesch. *Abort Landing in the Presence of a Windshear as a Minimax Optimal Control Problem. Part I: Necessary Conditions. Part II: Multiple Shooting and Homotopy.* Journal of Optimization Theory and Applications, 1991, vol. 70, pp. 1–23 et 223–254.

[23] J. J. Burgess. *Bending Stiffness in a Simulation of Undersea Cable Deployment.* International Journal of Offshore and Polar Engineering, 1993, vol. 3, pp. 197–204.

[24] R. H. Byrd, J. C. Gilbert et J. Nocedal. *A Trust Region Method Based on Interior-Point Techniques for Nonlinear Programming.* Rapport technique 2896, INRIA, 1996.

[25] R. H. Byrd, M. E. Hribar et J. Nocedal. *An Interior Point Algorithm for Large Scale Nonlinear Programming.* SIAM Journal on Optimization, 1999, vol. 9, pp. 877–900.

[26] R. M. Chamberlain, C. Lemaréchal, H. C. Pedersen et M. J. D. Powell. *The Watchdog Technique for Forcing Convergence in Algorithms for Constrained Optimization.* Mathematical Programming Study, 1982, vol. 16, pp. 1–17.

[27] L. Chauvier, G. Damy, J. C. Gilbert et N. Pichon. *Optimal Control of a Deep-Towed Vehicle by Optimization Techniques.* Actes de la conférence Oceans'98, Nice, 1998.

[28] S. H. Cheng et N. J. Higham. *A Modified Cholesky Algorithm Based on a Symmetric Indefinite Factorization.* Rapport technique 289, Department of Mathematics, University of Manchester, 1996.

[29] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland, 1978.

[30] A. R. Conn, N. I. M. Gould et P. Toint. *A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds.* SIAM Journal of Numerical Analysis, 1991, vol. 28, pp. 545–572.

[31] A. R. Conn, N. I. M. Gould et P. Toint. *A Note on Using Alternative Second-Order Models for the Subproblems Arising in Barrier Function Methods for Minimization.* Numerische Mathematik, 1994, vol. 68, pp. 17–33.

[32] R. Courant et D. Hilbert. *Methods of Mathematical Physics.* Interscience, 1962.

[33] M. Crouzeix et A. L. Mignot. *Analyse numérique des équations différentielles.* Mathématiques appliquées pour la maîtrise, Masson, 1984.

[34] C. Cuvelier, A. Segal et A. A. van Steenhoven. *Finite Element Methods and Navier Stokes Equations.* Mathematics and Its Applications, D. Reidel Publishing Company, 1986.

[35] G. Damy, M. Joannides, F. Le Gland, M. Prévosto et R. Rakotozafy. *Integrated Short Term Navigation of a Towed Underwater Body.* Actes de la conférence Oceans'94, Brest, 1994, pp. 577–582.

[36] R. Dautray et J. L. Lions. *Analyse mathématique et calcul numérique pour les sciences et les techniques.* Collection Commissariat à l'Energie Atomique, Masson, 1984.

[37] R. S. Dembo et T. Steihaug. *Truncated-Newton Algorithms for Large-Scale Unconstrained Optimization.* Mathematical Programming, 1983, vol. 26, pp. 190–212.

[38] D. den Hertog. *Interior-Point Approach to Linear, Quadratic, and Convex Programming.* Kluwer Academic Publisher, 1994.

[39] J. E. Dennis, M. El-Alem et M. C. Maciel. *A Global Convergence Theory for General Trust-Region Based Algorithms for Equality Constrained Optimization.* SIAM Journal on Optimization, 1997, vol. 7, pp. 177–207.

[40] A. S. El-Bakry, R. A. Tapia, T. Tsuchiya et Y. Zhang. *On the Formulation and Theory of the Newton Interior-Point Method for Nonlinear Programming.* Journal of Optimization Theory and Applications, 1996, vol. 89, pp. 507–541.

[41] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, 1987.

[42] A. Forsgren et P. E. Gill. *Primal-Dual Interior Methods for Nonconvex Nonlinear Programming.* SIAM Journal on Optimization, 1998, vol. 8, pp. 1132–1152.

[43] A. Forsgren, P. E. Gill et J. R. Shinnerl. *Stability of Symmetric Ill-Conditioned Systems Arising in Interior Methods for Constrained Optimization.* SIAM Journal on Matrix Analysis and Applications, 1996, vol. 17, pp. 187–211.

[44] C. Führer et B. J. Leimkuhler. *Numerical Solution of Differential Algebraic Equations for Constrained Mechanical Motion.* Numerische Mathematik, 1991, vol. 59, pp. 55–69.

[45] D. Gabay. *Reduced Quasi-Newton Methods with Feasibility Improvement for Nonlinearly Constrained Optimization.* Mathematical Programming Study, 1982, vol. 16, pp. 18–44.

[46] W. Gautschi. *Numerical Analysis.* Birkhäuser, 1997.

[47] D. M. Gay, M. L. Overton et M. H. Wright. *A Primal-Dual Interior Method for Nonconvex Nonlinear Programming.* In *Advances in Nonlinear Programming* (Y. Yuan). Kluwer Academic Publisher, 1998, pp. 31–56.

[48] C. W. Gear. *The Simultaneous Numerical Solution of Differential-Algebraic Equations.* IEEE Transactions on Circuit Theory, 1971, pp. 89–95.

[49] C. W. Gear, B. Leimkuhler et G. K. Gupta. *Automatic Integration of Euler-Lagrange Equations with Constraints.* Journal of Computational and Applied Mathematics, 1985, vol. 12-13.

[50] P. Germain. *Mécanique (tome I).* Ellipses, 1986.

[51] P. E. Gill, W. Murray, M. A. Saunders et M. H. Wright. *Some Theoretical Properties of an Augmented Lagrangian Merit Function.* In *Advances in Optimization and Parallel Computing* (P. M. Pardalos). North-Holland, 1992, pp. 101–128.

[52] P. E. Gill, W. Murray et M. H. Wright. *Practical Optimization.* Academic Press, 1981.

[53] V. Girault et P. A. Raviart. *Finite Elements Methods for Navier-Stokes Equations. Theory and Algorithms.* Springer-Verlag, 1986.

[54] R. Glowinski, T. W. Pan et J. Périaux. *A Fictitious Domain Method for Dirichlet Problem and Applications.* Computer Methods in Applied Mechanics and Engineering, 1994, vol. 111, pp. 283–303.

[55] N. I. M. Gould et P. Toint. *A Note on the Second-Order Convergence of Optimization Algorithms Using Barrier Functions.* Rapport technique 97/15, Rutherford Appleton Laboratory, 1997.

[56] E. Hairer et G. Wanner. *Solving Ordinary Differential Equations. Tome II: Stiff and Differential Algebraic Problems.* Springer-Verlag, 1996.

[57] S. P. Han. *A Globally Convergent Method for Nonlinear Programming.* Journal of Optimization Theory and Applications, 1977, vol. 32, pp. 297–309.

[58] S. P. Han et O. L. Mangasarian. *Exact Penalty Functions in Nonlinear Programming.* Mathematical Programming, 1979, vol. 17, pp. 251–269.

[59] U. Hetmaniuk et O. Paget. *Etude d'un câble inextensible.* Rapport de stage, Ecole Polytechnique, 1996.

[60] F. S. Hover. *Methods for Positioning Deeply-Towed Underwater Cables.* Thèse de doctorat, Woods Hole Oceanographic Institution, Massachussets Institute of Technology, 1993.

[61] N. Karmarkar. *A New Polynomial Time Algorithm for Linear Programming.* Combinatorica, 1984, vol. 4, pp. 373–395.

[62] M. Kojima, N. Megiddo et S. Mizuno. *A Primal-Dual Infeasible-Interior-Point Algorithm for Linear Programming.* Mathematical Programming, 1993, vol. 61, pp. 261–280.

[63] M. Lalee, J. Nocedal et T. Plantenga. *On the Implementation of an Algorithm for Large-Scale Equality Constrained Optimization.* SIAM Journal on Optimization, 1998, vol. 8, pp. 682–706.

[64] J. L. Lions et E. Magenes. *Problèmes aux limites non homogènes.* Dunod, 1968.

[65] P. Lötstedt et L. Petzold. *Numerical Solution of Nonlinear Differential Equations with Algebraic Constraints. Part I: Convergence Results for Backward Differentiation Formulas.* Mathematics of Computation, 1986, vol. 46, pp. 491–516.

[66] C. Lubich, U. Nowak, U. Pöhle et C. Engstler. *MEXX - Numerical Software for the Integration of Constrained Mechanical Multibody Systems.* Rapport technique SC 92-12, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 1992.

[67] D. G. Luenberger. *Introduction to Linear and Nonlinear Programming.* Addison-Wesley, 1973.

[68] I. Lustig, R. E. Marsten et D. F. Shanno. *Computational Experience with a Primal-Dual Interior Point Method for Linear Programming.* Linear Algebra and Its Applications, 1991, vol. 152, pp. 191–222.

[69] O. L. Mangasarian et S. Fromovitz. *The Fritz John Necessary Optimality Conditions in the Presence of Equality and Inequality Constraints.* Journal of Mathematical Analysis and Applications, 1967, vol. 17, pp. 37–47.

[70] N. Maratos. *Exact Penalty Function Algorithms for Finite Dimensional and Control Optimization Problems.* Thèse de doctorat, Imperial College, Londres, 1978.

[71] D. Marichal. *Contribution à l'étude statique et dynamique des câbles sous-marins.* Thèse de doctorat, Université de Nantes, 1979.

[72] D. Marichal, B. Dassonville et O. Lefort. *Etude dynamique des systèmes sous-marins remorqués à l'aide de câbles de longueur variable.* Actes de l'Association Technique Maritime et Aéronautique, Paris, Session 1987.

[73] D. Q. Mayne et E. Polak. *A Superlinearly Convergent Algorithm for Constrained Optimization Problems.* Mathematical Programming Study, 1982, vol. 16, pp. 45–61.

[74] S. Mehrotra. *On the Implementation of a Primal-Dual Interior-Point Method.* SIAM Journal on Optimization, 1992, vol. 2, pp. 575–601.

[75] S. Mizuno, M. J. Todd et Y. Ye. *On Adaptative-Step Primal-Dual Interior-Point Algorithms for Linear Programming.* Mathematics of Operations Research, 1993, vol. 18, pp. 964–981.

[76] J. L. Morales et J. Nocedal. *Automatic Preconditioning by Limited Memory Quasi-Newton Updating.* Soumis pour publication dans SIAM Journal on Optimization, 1997.

[77] J. D. Mudie et K. A. Kastens. *Computer Aided Piloting of a Deeply Towed Vehicle.* Ocean Engineering, 1980, vol. 7, pp. 743–754.

[78] R. M. Murray. *Trajectory Generation for a Towed Cable System using Differential Flatness.* Actes du congrès IFAC, San-Francisco, 1996, pp. 395–400.

[79] S. G. Nash et A. Sofer. *Why Extrapolation Helps Barrier Methods.* Rapport technique, Operations Research and Engineering Department, George Mason University, Fairfax, 1998.

[80] E. O. Omojokun. *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints.* Thèse de doctorat, University of Colorado at Boulder, 1991.

[81] J. P. Penot. *On Regularity Conditions in Mathematical Programming.* Mathematical Programming Study, 1982, vol. 19, pp. 167–199.

[82] H. J. Pesch. *A Practical Guide to the Solution of Real-Life Optimal Control Problems.* Control Cybernetics, 1994, vol. 23, pp. 7–60.

[83] N. Pichon. *Modélisation, simulation et commande optimale pour le remorquage d'engins sous-marins profonds.* Thèse de doctorat, Université Paris IX-Dauphine, 1997.

[84] W. T. Pinto. *On the Dynamics of Low Tension Marine Cables.* Thèse de doctorat, Department of Mechanical Engineering, University College, Londres, 1995.

[85] O. Pironneau. *Finite Element Methods for Fluids.* John Wiley & Sons – Masson, 1989.

[86] L. Pode. *Tables for Computing the Equilibrium Configuration of a Flexible Cable in a Uniform Stream.* Rapport technique 687, David Taylor Model Basin, 1951.

[87] L. S. Pontryagin. *The Mathematical Theory of Optimal Processes.* Wiley-Intersciences, 1962.

[88] M. J. D. Powell. *A Fast Algorithm for Nonlinearly Constrained Optimization Calculations.* In *Numerical Analysis* (G. A. Watson). Springer-Verlag, 1978, pp. 144–157.

[89] B. Pshenichnyi et Y. Danilin. *Numerical Methods for Extremal Problems.* MIR, 1978.

[90] P. A. Raviart et J. M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles.* Mathématiques appliquées pour la maîtrise, Masson, 1983.

[91] M. Reeken. *The equation of motion of a chain.* Mathematische Zeitschrift, 1977, vol. 155, pp. 219–237.

[92] M. Reeken. *Classical solutions of the chain equation, part I and II.* Mathematische Zeitschrift, 1979, vol. 165 et 166, pp. 143–169 et 67–82.

[93] A. Robin. *Modélisation quasi-statique 3D d'un câble et identification de paramètres.* Rapport technique DITI SOM 92-373, IFREMER, 1992.

[94] S. M. Robinson. *Stability Theorems for Systems of Inequalities. Part II: Differentiable Nonlinear Systems.* SIAM Journal on Numerical Analysis, 1976, vol. 13, pp. 497–513.

[95] J. V. Sanders. *A Three-Dimensional Dynamic Analysis of a Towed System.* Ocean Engineering, 1982, vol. 9, pp. 483–499.

[96] R. L. Sani, P. M. Gresho, R. L. Lee et D. F. Griffiths. *The Cause and Cure of the Spurious Pressures Generated by Certain FEM Solutions of the Incompressible Navier-Stokes Equations. Parts I and II.* International Journal for Numerical Methods in Fluids, 1981, vol. 1, pp. 17–43 et 171–204.

[97] J. Stoer et R. Bulirsch. *Introduction to Numerical Analysis.* Springer-Verlag, 1993.

[98] Y. Sun et J. W. Leonard. *Dynamics of Ocean Cables with Local Low-Tension Regions.* Ocean Engineering, 1998, vol. 25, pp. 443–463.

[99] R. Temam. *Navier-Stokes Equations.* North-Holland, 1979.

[100] T. Terlaky, J. P. Vial et K. Roos. *Theory and Algorithms for Linear Programming : an Interior Point Approach.* Wiley Intersciences, 1997.

[101] R. J. Vanderbei et D. F. Shanno. *An Interior-Point Algorithm for Nonconvex Nonlinear Programming.* Rapport technique SOR-97-21, Statistics and Operation Research Department, Princeton University, 1997.

[102] J. P. Vial. *Computational Experience with a Primal-Dual Interior-Point Method for Smooth Convex Programming.* Rapport technique 1992.7, Logilab - HEC, Université de Genève, 1992.

[103] M. C. Villalobos, R. A. Tapia et Y. Zhang. *The Sphere of Convergence of Newton's Method on Two Equivalent Systems from Nonlinear Programming.* Rapport technique TR99-15, Department of Computational and Applied Mathematics, Rice University, Houston, 1999.

[104] O. von Stryk et R. Bulirsch. *Direct and Indirect Methods for Trajectory Optimization.* Annals of Operations Research, 1992, vol. 37, pp. 357–373.

[105] M. H. Wright. *Why a Pure Primal Newton Barrier Step May Be Infeasible.* SIAM Journal on Optimization, 1995, vol. 5, pp. 1–12.

[106] M. H. Wright. *Ill-Conditioning and Computational Error in Primal-Dual Interior Methods for Nonlinear Programming.* SIAM Journal on Optimization, 1998, vol. 9, pp. 84–111.

[107] S. J. Wright. *Primal-Dual Interior-Point Methods.* SIAM, 1997.

[108] E. E. Zajac. *Dynamics and Kynematics of the Laying and Recovery of Submarine Cables.* Rapport technique, Bell Laboratories, 1957.

[109] Y. Zhang. *On the Convergence of a Class of Infeasible-Interior-Point Methods for the Horizontal Linear Complementarity Problem.* SIAM Journal on Optimization, 1994, vol. 4, pp. 208–227.

Résumé

On s'intéresse à un câble immergé, tractant un engin fixé à son extrémité inférieure et remorqué par un navire en surface. Un treuil permet d'enrouler ou dérouler le câble. On souhaite calculer une trajectoire du navire et un filage du câble emmenant l'engin d'une position à une autre position données en temps minimal : problème du demi-tour en temps minimal.

On considère un modèle de câble inextensible et parfaitement flexible. Son équation d'évolution, en position et tension, est du type ondes et couplée à une contrainte algébrique traduisant l'inextensibilité. La tension est le multiplicateur associé à cette contrainte. On discrétise en espace l'équation à longueur de câble constante par éléments finis mixtes. En résulte un système différentiel-algébrique que l'on intègre en temps en appliquant un schéma de différences rétrogrades. Puis on généralise aux câbles de longueur variable grâce à une méthode de domaine fictif.

Dans un second temps, on développe des algorithmes d'optimisation non linéaire pour la commande optimale. Une méthode de Newton tronquée est proposée pour les problèmes avec contraintes d'égalité seulement. Le calcul de la direction permet de prendre en compte leur non-convexité, par troncature des itérations de gradient conjugué dans l'espace tangent aux contraintes. La globalisation est effectuée par recherche linéaire sur une fonction de pénalisation exacte. On étend cet algorithme aux problèmes avec contraintes d'inégalité grâce à une méthode de points intérieurs non linéaire.

Enfin, dans la formulation du problème du demi-tour, on considère, en plus des contraintes d'égalité provenant de la discrétisation des équations d'état, des contraintes d'inégalité sur la vitesse du navire, la profondeur de l'engin, la position finale du câble, etc. Le problème est résolu à l'aide de l'algorithme présenté ci-dessus.

Mots-clés : câble immergé, commande optimale, optimisation non linéaire, méthode de Newton tronquée, méthode de points intérieurs.

Optimal control of towed submarine vehicles with constraints

Abstract: we consider an immersed vehicle deployed through a cable several thousand meters long. The cable may be wound or unwound from the surface ship. Our objective is to determine the ship trajectory and winding velocity that bring the vehicle from one given position to another, in minimum time.

We assume the cable to be inextensible and infinitely flexible. Its evolution equation, formulated in position and tension, is similar to the wave equation, coupled with an algebraic constraint to express inextensibility. The tension is the multiplier associated with this constraint. We first discretize the cable equation in space when its length is constant, using mixed finite elements. The resulting differential-algebraic equation is integrated in time using a BDF scheme. We then generalize to variable length cables thanks to a fictitious domain method.

In a second part, we develop nonlinear programming algorithms for optimal control. A truncated Newton's method is proposed for equality constrained problems. The direction computation takes into account their nonconvexity, via truncation of the conjugate gradient iterations in the constraint tangent space. A linesearch procedure, on an exact penalty function, globalizes this algorithm. We next extend the algorithm to inequality constrained problems by means of a nonlinear interior point approach.

Finally, to address the minimum time problem, we consider inequality constraints on ship velocity, vehicle depth, the final cable position, etc, in addition to the equality constraints arising from the discretized state equation. The preceding algorithm is applied to solve this problem.

Keywords: immersed cable, optimal control, nonlinear programming, truncated Newton's method, interior point method.